Analysis of a Correlated Queue in a Communication System

Israel Cidon, *Senior Member, IEEE*, Roch Guérin, *Senior Member, IEEE*, Asad Khamisy, and Moshe Sidi, *Senior Member, IEEE*

Abstract—**A family of queues where the service time B_n of customer n depends on the interarrival time I_n between customers $n-1$ and n is studied. In particular, the focus is on dependencies that arise naturally in the context of communication systems, where the finite speed of the communication links constrains the amount of data that can be received in a given time interval. Specifically, queues are studied where the random variables I_n and B_n exhibit some form of *proportionality* relation. Such dependencies can have significant impact on system performance and it is, therefore, critical to develop tractable models that account for them. The paper starts with the simple case of a deterministic proportionality relation between the service time of a customer and its preceding interarrival time. This is then extended to allow for the addition of an independent, generally distributed *overhead* to the service time of each customer. Next, several models that capture the ON–OFF behavior of communication links in packet networks are considered. In all cases, expressions for the delay experienced by a packet in the system are provided. Numerical examples that illustrate the impact of dependencies through comparison with less accurate models are also supplied. While the paper is clearly motivated by problems that originated in the field of communications and in particular packet switching networks, its results should be of relevance to other environments as well.**

Index Terms—**Correlated queues, dependent queues, proportional dependency, ON–OFF processes.**

I. Introduction

THE FOCUS of this paper is on a family of queues where service and interarrival times exhibit some form of dependency. The initial motivation for this study was the modeling of a communication link in a packet-switched network carrying variable size packets. It should, however, be pointed out that the results obtained are also of general interest as they provide simple and new tools to analyze the impact of certain types of dependencies in queues. The general issue of dependencies in queueing systems is clearly an important one, and has been extensively studied in the literature. The reader is referred to [1] for a review on the various types of dependencies that exist in packet queues, and a study of their impact on different system performance measures.

In this paper, we focus on a particular type of dependencies, where the service time associated with a packet, e.g., its transmission time on the link, is correlated with its interarrival time. Such correlations arise, for example, in the context of a packet-switched network where variable length packets are forwarded from one node to another. The finite speed of network links then results in large packets having correspondingly large interarrival times, i.e., for a link of speed S the amount of work received in a time interval τ cannot exceed $S \times \tau$. This strong positive correlation between interarrival and service times, can greatly improve the delay characteristics in the buffers preceding communication links. It is, therefore, important to provide models that account for this effect while remaining tractable.

A number of earlier works have considered the issue of correlation between service and interarrival times. In particular, the impact of correlated interarrival and service times has been investigated in the context of two important queueing systems, i.e., tandem queues and fluid-flow models. Although these studies and the techniques they rely on are not directly applicable to our model, they address similar issues and provide further motivations for the models we develop in the paper. After reviewing the more relevant aspects and results from these works, we focus on a series of papers that analyzed queueing systems closely related to those considered in this paper. We highlight their key features, and then identify how the results of this paper extend them to provide more accurate and yet simple models that are applicable to many communication systems.

One of the earliest work to systematically investigate the issue of correlation between service and interarrival times is [2]. In [2], Kleinrock studied the impact of correlated message lengths and interarrival times in the context of a queueing network model for communication networks. The intractability of the general problem led him to formulate the well-known and useful *independence assumption*, which amounts to ignoring correlations. This approach is reasonably accurate in the presence of sufficient traffic *mixing* in the network, but can significantly overestimate delays in systems where there is a strong positive correlation between service and interarrival times as in tandem queues (see [3, 4, 5]), where little or no traffic mixing is present.

The dependency between interarrival times and the amount of work that can be brought into a system, has also been

Manuscript received December 26, 1991; revised July 6, 1992. This work was done while A. Khamisy and M. Sidi were visiting the IBM T. J. Watson Research Center, Yorktown Heights, NY.

I. Cidon is with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel and the IBM T.J. Watson Research Center, P.O. Box 704, Yorktown Heights, NY 10598.

R. Guérin is with the IBM T.J. Watson Research Center, P.O. Box 704, Yorktown Heights, NY 10598.

A. Khamisy and M. Sidi are with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel.

IEEE Log Number 9203979.

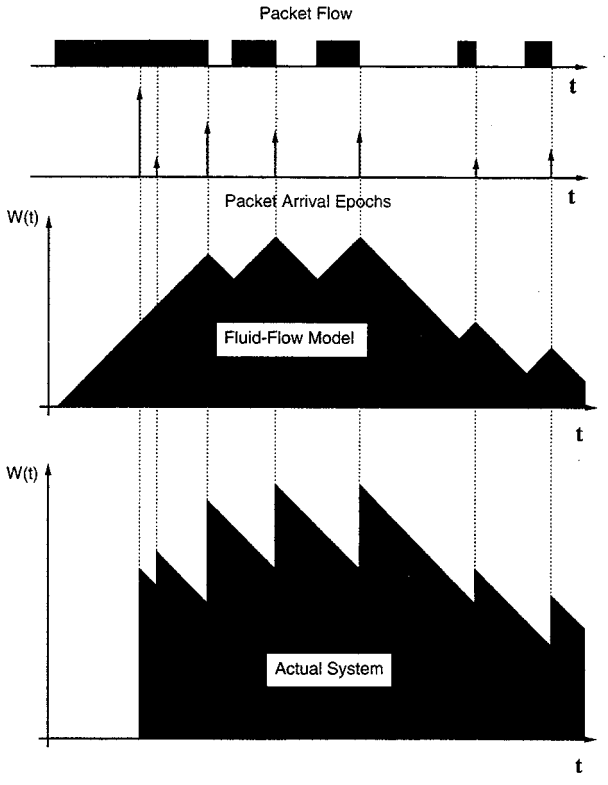

Fig. 1. Comparison of ON–OFF fluid-flow and discrete arrivals models.

studied in the context of fluid-flow models [6]–[12] which assume that work arrives into and is removed from a system at continuous and possibly varying rates. A particularly popular and simple example is that of an ON–OFF source feeding a buffer, which is emptied at a constant rate. The finite input and output rates account for the dependency between the amount of data received and the elapsed time t, i.e., the amount of data received is proportional to both the input rate and t.

While fluid-flow models capture some of the dependencies that exist between arrivals and service times in communication systems, they do not account for the *granularity* of arrivals and services. Rather, they assume that both arrivals and departures are progressive, with the work in the system being a continuous function of time. This may not always be an adequate assumption for communication systems (especially not in store-and-forward networks), i.e., packets must typically be fully received before they can be forwarded. As illustrated in Fig. 1, this can result in significant inaccuracies when estimating system performance (see also [13]). It is one of the purposes of this paper to propose and to analyze models, that not only account for the type of dependencies captured by fluid-flow models, but also preserve the discrete nature of arrivals and services that is characteristic of many communication systems. Before proceeding with the description of these models, we complete our review of earlier works by discussing several papers [14]–[21] that are directly relevant to this study.

A. A Correlated Queue

One of the early works to consider a queueing system with explicit correlations between interarrival and service times is [14]. It analyzes a system with Poisson arrivals at rate λ, where the service time B_n of the nth customer is proportional to the interarrival time I_n between the $(n-1)$st and the nth customers. In other words, the service time is a deterministic function of the interarrival time, with $B_n = \alpha I_n$ ($\alpha < 1$ for stability). This system can be used to model a buffer connected to a unit speed communication link, that receives an uninterrupted string of packets with exponentially distributed length from an upstream link of speed α. An explicit expression for the delay distribution is obtained in [14], while the initial busy period, the system state, and the output process are studied in [16]. Numerical comparisons between this system and related M/M/1, M/D/1 and D/M/1 queues were carried out in [15].

More general correlations were considered in [17] using a bivariate exponential distribution to characterize the correlation between interarrival and service times. This work was subsequently extended in several papers. The delay density was shown to have a hyperexponential distribution in [18], while [19] studied the sensitivity of this distribution to the value of the correlation coefficient. A system with infinitely many servers was considered in [20]. Recently, a variant of the M/G/1 queue, in which service time and interarrival time are positively correlated, was studied in [21].

B. Scope of the Paper

The work presented in this paper begins with a system similar to that of [14], and expands it in several new directions that makes it more applicable to the modeling of actual communication systems. As mentioned earlier, our goals are to account for both the dependencies between packet interarrival and transmission times due to the finite speed of communication links, and the discrete nature of these events. In this paper, we propose new models that not only accurately capture these effects, but, more importantly remain tractable. The models and results are sufficiently general so that they are also of interest and potentially useful to model dependencies in systems outside the field of communications. We now sketch out the different results obtained, and point to their significance while outlining the structure of the paper.

In Section II, we introduce a system similar to that of [14] (see Fig. 2), and present a simple derivation for the Laplace–Stieltjes Transform (LST) of the delay in the system. This simple derivation is obtained by directly focusing on the steady state equations, rather than on the transient evolution equations as was done in all earlier works on similar correlated queues [14], [16]–[18], [20], [21]. The LST of the delay is then obtained by applying results from the theory of linear functional equations [22] and the analytic properties of the LST. This approach not only provides a formal framework for such problems, but it also results in a solution method that is applicable to a more general class of problems. In particular, it allows us to tackle more involved systems as illustrated in the rest of the paper.

The first extension we consider consists of the addition of an independent, generally distributed, non-negative random variable to the service time. Using the notations introduced

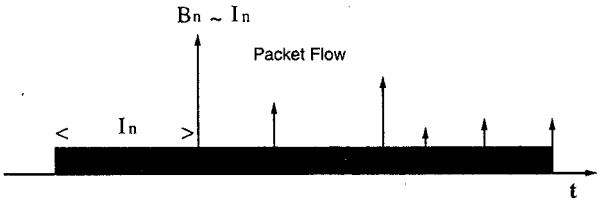

Fig. 2. System with proportional interarrival and service times.

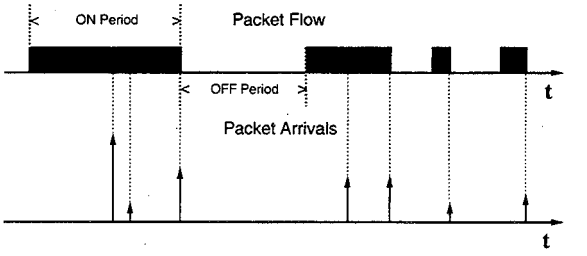

Fig. 3. System with ON–OFF source and multiple discrete arrivals.

above, the service time of the nth customer is now of the form $B_n = \alpha I_n + J_n$, where J_n is an independent, nonnegative random variable with a general distribution. This extension is useful to model systems where each packet needs additional service in excess of its raw transmission time. The additional service may be due to some overhead such as a header appended to the original data, or correspond to some processing that needs to be performed for each packet.

The simple model of Section II is useful to capture the impact of dependencies between packet interarrival and service times. However, from a modeling point-of-view, it imposes a number of limiting constraints. In particular, it requires that the input correspond to a "saturated link" with a transmission rate lower than that of the output link ($\alpha < 1$). In order to overcome this limitation, the model is further extended in Section III, where we allow the input process to alternate between active and idle periods. This is achieved by allowing the proportionality constant α to be itself a random variable, that takes value $\alpha_1 > 0$ with probability g_1 and $\alpha_2 = 0$ with probability $g_2 = 1 - g_1$. This results in an ON–OFF input process with exponentially distributed ON and OFF periods, a geometric number of packets in each ON period, and exponentially distributed packet sizes. Specifically, after an exponentially distributed time interval of duration I_n and mean $1/\lambda$, a packet of size $\alpha_i I_n$ is generated with probability g_i, $i = 1, 2$. This creates exponentially distributed active and idle periods on the link, with means $1/\lambda(1 - g_1)$ and $1/\lambda g_1$, respectively. The resulting arrival process is illustrated in Fig. 3. Note that, as with the model of Section III, it is also possible to add an independent and generally distributed "overhead" to each packet.

This model, although reminiscent of a fluid-flow model for a two-state Markovian ON–OFF source, exhibits a number of key differences. First, data arrival does not take place gradually over the duration of an ON period. Rather, work accumulates for some interval of time, and it is only upon its completion that a packet is generated to the system. This provides a more accurate representation of the discrete nature of packet arrivals. Second, the model allows for the partition of a single ON period into multiple packets. This is in contrast to a fluid-flow model, where data arrival is interrupted over the duration of the entire ON period, and the transmission of bits rather than packets is considered. Despite its increased flexibility, the model of Section 3 still has a number of limitations. In particular, it requires that the average length of the active and the idle periods on the link be proportional, i.e., within a factor $g_1/(1 - g_1)$. This implies that for a given link utilization, the average duration of incoming bursts is fixed. Burst duration is, however, a key performance factor [13], [23], and it is of interest to develop models that allow burst duration and utilization to vary independently. This is the topic of Section IV.

In Section IV, we consider a model where the arrival process corresponds to an *extended* ON–OFF process. As in Section III, we allow multiple packets with exponentially distributed lengths to be generated during a single ON period, but this is now achieved without imposing any constraint on the duration of OFF periods and hence on the utilization. Specifically, the link is assumed to remain active for an exponentially distributed time I_n with mean $1/\lambda$, at the end of which a packet of size αI_n is generated. The link then starts a new ON period with probability $1 - p$ or enters an OFF period with probability p. The duration of an OFF period is exponentially distributed with mean $1/\mu$, and the link returns to the ON state at the end of an OFF period. This allows us to construct ON periods, where the number (geometrically distributed) of consecutive packets that are generated is independent of the length of OFF periods.

Note that the arrival process of Section III can be viewed as a special case of this extended ON–OFF process with $p = g_2$ and $\mu = \lambda g_1$, whose analysis is much simpler. Similarly, the more traditional ON–OFF process where each ON period corresponds to a single packet and is always followed by an OFF period, corresponds to the special case $p = 1$. The arrival process introduced in Section IV is therefore, quite general and provides us with the necessary flexibility to investigate the influence of different parameters on system performance. In addition, it is again possible to further enhance the model by allowing the addition of an independent and generally distributed overhead to each packet.

Numerical examples that illustrate the results obtained for the models described in the paper are provided. These examples help identify the impact and significance of the successive refinements allowed by the three models. Some comparisons with models that assume either independent interarrival and service times or rely on fluid-flow approximations are also provided. A brief conclusion summarizes the findings of the paper.

II. Deterministic Proportional Dependency

Here, we start with the model of [14]. We consider a system with Poisson arrivals at rate λ, where the service time B_n of the nth packet is proportional to the interarrival time I_n between the $(n-1)$st and nth packets, i.e., $B_n = \alpha I_n$ ($\alpha < 1$ for stability). Let W_n denote the amount of unfinished work just after the nth packet arrival. Assuming service is according

to the FIFO discipline, W_n is also the delay of the nth packet and it evolves according to

$$W_{n+1} = (W_n - I_{n+1})^+ + \alpha I_{n+1}, \qquad n \geq 0, \qquad (1)$$

where $X^+ = \max(X, 0)$.

Let $\mathcal{W}_n(s) = E\left[e^{-sW_n}\right] (\mathrm{Re}(s) \geq 0)$ be the LST of W_n. Then

$$\mathcal{W}_{n+1}(s) = E\left[e^{-s\left[(W_n - I_{n+1})^+ + \alpha I_{n+1}\right]}\right], \qquad n \geq 0.$$

Therefore, letting $f_{W_n}(w)$ denote the probability density function (pdf) of W_n, we obtain

$$\begin{aligned}\mathcal{W}_{n+1}(s) &= \int_0^\infty f_{W_n}(w)\, dw \left[\int_0^w \lambda e^{-\lambda x} \cdot e^{-s(w-x+\alpha x)}\, dx + \int_w^\infty \lambda e^{-\lambda x} \cdot e^{-s\alpha x}\, dx\right] \\ &= \frac{\lambda}{\lambda - (1-\alpha)s}\left[\mathcal{W}_n(s) - \mathcal{W}_n(\lambda + \alpha s)\right] + \frac{\lambda}{\lambda + \alpha s}\mathcal{W}_n(\lambda + \alpha s).\end{aligned}$$

Letting $n \to \infty$, we obtain the following functional relation for the LST of the delay $\mathcal{W}(s)$ in steady state:

$$\mathcal{W}(s) = \frac{\lambda}{(1-\alpha)(\lambda + \alpha s)}\, \mathcal{W}(\lambda + \alpha s) \qquad (2)$$

Such relations are called homogeneous linear functional equations (see [22]).

Define $s_i = \lambda(1-\alpha^i)/(1-\alpha) + \alpha^i s$, $i \geq 0$, $c = \lambda/(1-\alpha)$ and $b(s) = c/(\lambda + \alpha s)$ to obtain for any $k \geq 0$,

$$\mathcal{W}(s) = \mathcal{W}(s_{k+1}) \prod_{i=0}^{k} b(s_i).$$

Letting $k \to \infty$, we have (recall that $\alpha < 1$),

$$\mathcal{W}(s) = \mathcal{W}(c) \prod_{i=0}^{\infty} b(s_i).$$

Using the fact that $\mathcal{W}(0) = 1$, we obtain

$$\mathcal{W}(c) = \prod_{i=1}^{\infty} (1 - \alpha^i),$$

and therefore,

$$\mathcal{W}(s) = \prod_{i=1}^{\infty} \frac{1-\alpha^i}{1 - (1 - s/c)\alpha^i}.$$

Calculating the first moment $E[W] = -\frac{d\mathcal{W}(s)}{ds}\Big|_{s=0}$ we obtain

$$E[W] = \frac{1-\alpha}{\lambda} \sum_{j=1}^{\infty} \frac{\alpha^j}{1-\alpha^j}.$$

A. Additional Random Overhead

The first extension we consider consists of the addition of an independent, generally distributed, nonnegative random variable to the service time. Using the notation introduced above, the service time of the nth customer is now of the form $B_n = \alpha I_n + J_n$, where J_n is an independent, nonnegative random variable with a general distribution. As mentioned earlier, this extension is useful to model systems where each packet service time includes an additional overhead in excess of its raw transmission time. This overhead may be in the form of a header appended to the original data, or corresponds to some processing that needs to be performed for each packet.

The delay W_n of the nth customer evolves according to,

$$W_{n+1} = (W_n - I_{n+1})^+ + \alpha I_{n+1} + J_{n+1}, \qquad n \geq 0,$$

where J_n is a nonnegative random variable, independent of n, and independent of I_k for any k. The pdf of J_n is $f_J(y)$ and its LST is $\mathcal{J}(s)$. By definition,

$$\mathcal{W}_{n+1}(s) = E\left[e^{-s\left[(W_n - I_{n+1})^+ + \alpha I_{n+1} + J_{n+1}\right]}\right], \qquad n \geq 0.$$

Therefore,

$$\begin{aligned}\mathcal{W}_{n+1}(s) &= \int_0^\infty f_{W_n}(w)\, dw \cdot \left[\int_0^w \lambda e^{-\lambda x} \cdot e^{-s(w-x+\alpha x)}\, dx \cdot \int_0^\infty f_J(y) e^{-sy}\, dy + \int_w^\infty \lambda e^{-\lambda x} \cdot e^{-s\alpha x}\, dx \cdot \int_0^\infty f_J(y) \cdot e^{-sy}\, dy\right] \\ &= \frac{\lambda \mathcal{J}(s)}{\lambda - (1-\alpha)s}\mathcal{W}_n(s) + \frac{\lambda \mathcal{J}(s) s}{(\lambda + \alpha s)[(1-\alpha)s - \lambda]} \cdot \mathcal{W}_n(\lambda + \alpha s).\end{aligned}$$

Letting $n \to \infty$ we obtain the following functional equation for LST $\mathcal{W}(s)$ of the delay in steady state:

$$\mathcal{W}(s) = b(s)\mathcal{W}(\lambda + \alpha s), \qquad (3)$$

where

$$b(s) = \frac{\lambda s \mathcal{J}(s)}{[(1-\alpha)s - \lambda(1 - \mathcal{J}(s))](\lambda + \alpha s)},$$

which is consistent with (2), when $\mathcal{J}(s) = 1$. Note that if $\alpha = 0$ the results reduce to the well-known M/G/1 Pollaczek–Khinchin waiting time relation (see [24]). Also note that the condition for steady state in this case is $\alpha < 1 - \lambda/\delta$ where $1/\delta = E[J_n]$.

As before, we obtain from (3) that

$$\mathcal{W}(s) = \mathcal{W}(c) \prod_{i=0}^{\infty} b(s_i),$$

where we recall that $s_i = \lambda(1-\alpha^i)/(1-\alpha) + a^i s$, $i \geq 0$ and $c = \lambda/(1-\alpha)$. To determine the constant $\mathcal{W}(c)$ we use

TABLE I
AVERAGE DELAY VERSUS THE PROPORTIONALITY PARAMETER α

α	0.1	0.2	0.3	0.4	0.45	0.49	0.495
Exponential	1.2882	1.6905	2.3420	3.8888	6.5745	26.7529	51.7778
Discrete	1.2847	1.6835	2.3311	3.8722	6.5537	26.7258	51.7497

the normalization condition, i.e., $\mathcal{W}(0) = 1$. With the aid of L'Hospital's law we obtain

$$\mathcal{W}(c) = \left(1 - \frac{\lambda/\delta}{1-\alpha}\right) \prod_{i=1}^{\infty} \frac{\mathcal{J}(\xi_i) - \alpha^i}{\mathcal{J}(\xi_i)}, \tag{4}$$

where $\xi_i = \lambda(1-\alpha^i)/(1-\alpha)$. It should be noted that when the stability condition $\alpha < 1 - \lambda/\delta$ holds, $\mathcal{W}(c)$ is strictly positive. This follows from the fact that $\mathcal{J}(\xi_i) \geq 1 - \xi_i/\delta > \alpha^i$.

Differentiating $\mathcal{W}(s)$ at $s = 0$, we get the expression for the average delay, (see equation at bottom of page) where $\mathcal{J}'(\xi_i) = d\mathcal{J}(s)/ds|_{s=\xi_i}$ and $\mathcal{J}''(0) = d^2\mathcal{J}(s)/ds^2|_{s=0}$.

Numerical Examples: From (5), we observe that the average delay depends on the entire probability distribution and not only on the first and the second moments of the random variable J_n. In Table I, we compute the average delays for two distributions for J_n that have the same first and second moments. Specifically, one distribution is exponential with parameter $\delta = 2$, and the other distribution is discrete with $\text{Prob}(J_n = 0) = \text{Prob}(J_n = 1) = 0.5$. Note that the differences between the averages are small.

III. RANDOM PROPORTIONAL DEPENDENCY

We begin this section by considering the basic model of the previous section, but now assuming that the proportionality parameter between the service time and the interarrival time is itself a random variable. Specifically, we assume that the delay $\mathcal{W}_n$ of the nth customer evolves according to

$$W_{n+1} = (W_n - I_{n+1})^+ + \Omega_{n+1} I_{n+1}, \qquad n \geq 0,$$

where Ω_n is a random variable that takes the value α_l with probability g_l, $1 \leq l \leq L$, for some integer $L\left(\sum_{l=1}^{L} g_l = 1\right)$. The stability condition of this system is $E[\Omega_n] < 1$, i.e., $\sum_{l=1}^{L} \alpha_l g_l < 1$. As in Section II, it is assumed that I_n is exponentially distributed with parameter λ.

As before, the LST $\mathcal{W}_{n+1}(s)$ can be expressed as

$$\mathcal{W}_{n+1}(s) = E\left[e^{-s\left[(W_n - I_{n+1})^+ + \Omega_{n+1} I_{n+1}\right]}\right], \qquad n \geq 0.$$

Therefore,

$$\begin{aligned}\mathcal{W}_{n+1}(s) &= \sum_{l=1}^{L} g_l \left[\int_0^{\infty} f_{W_n}(w)\, dw \left(\int_0^{w} \lambda e^{-\lambda x} \cdot e^{-s(w-x+\alpha_l x)}\, dx + \int_w^{\infty} \lambda e^{-\lambda x} \cdot e^{-s\alpha_l x}\, dx\right)\right] \\ &= \sum_{l=1}^{L} g_l \left[\frac{\lambda}{\lambda - (1-\alpha_l)s} \left[\mathcal{W}_n(s) - \mathcal{W}_n(\lambda + \alpha_l s)\right] + \frac{\lambda}{\lambda + \alpha_l s} \mathcal{W}_n(\lambda + \alpha_l s)\right]. \end{aligned} \tag{6}$$

Letting $n \to \infty$ we obtain the following functional equation for the LST $\mathcal{W}(s)$ of the delay in steady state:

$$\left(1 - \sum_{l=1}^{L} \frac{g_l \lambda}{\lambda - (1-\alpha_l)s}\right) \mathcal{W}(s) = \sum_{l=1}^{L} \frac{g_l \lambda s}{[(1-\alpha_l)s - \lambda](\lambda + \alpha_l s)} \mathcal{W}(\lambda + \alpha_l s). \tag{7}$$

Such a relation is a nonhomogeneous linear functional equation, and its solution is an open problem as indicated in [22].

In the following, we will restrict ourselves to a particular case of (7) that captures some of the aspects of ON–OFF behavior of arrival processes in networks. Specifically, we assume that $L = 2$, $\alpha_1 = \alpha$ and $\alpha_2 = 0$. With these assumptions we essentially have an ON–OFF input process with exponentially distributed ON and OFF periods, a geometric number of packets in each ON period, and exponentially distributed packet sizes. Specifically, after an exponentially distributed time interval of duration I_n and mean $1/\lambda$, a packet of size $\alpha_i I_n$ is generated with probability g_i, $i = 1, 2$. This creates exponentially distributed active and idle periods on the link, with mean $1/\lambda(1-g_1)$ and $1/\lambda g_1$, respectively. The resulting arrival process is illustrated in Fig. 3. Note that some customers arrive at the system but do not require any service from it. The delays computed in the following analysis include the delay of these customers. If one is interested only in the delays of customers that require service, the expressions we derive should be slightly modified and the derivation is straightforward.

$$E[W] = \frac{\alpha}{\lambda} + \frac{1}{\delta} + \frac{1}{2} \frac{\mathcal{J}''(0)}{1 - \alpha - \lambda/\delta} + \sum_{i=1}^{\infty} \frac{\xi_i^2(1-\alpha)(\alpha\mathcal{J}(\xi_i) - \mathcal{J}'(\xi_i)) + \lambda\mathcal{J}(\xi_i)(1-\mathcal{J}(\xi_i))(1-\alpha\xi_i) + \lambda\xi_i\mathcal{J}'(\xi_i)}{\xi_i\mathcal{J}(\xi_i)[(1-\alpha)\xi_i - \lambda(1-\mathcal{J}(\xi_i))](\lambda + \alpha\xi_i)} \alpha^i \tag{5}$$

A. Analysis of the Case $\alpha_1 = \alpha$, $\alpha_2 = 0$

When $\alpha_1 = \alpha$, $\alpha_2 = 0$, we obtain from (7) that

$$a(s)\mathcal{W}(s) = b(s)\mathcal{W}(\lambda + \alpha s) + c(s), \tag{8}$$

where

$$a(s) = s(1-\alpha) - \lambda(1 - \alpha g_1),$$
$$b(s) = \lambda g_1 (s - \lambda)/(\lambda + \alpha s),$$
$$c(s) = g_2 \mathcal{W}(\lambda)[s(1-\alpha) - \lambda].$$

Note that the constant $\mathcal{W}(\lambda)$ is yet unknown, but using the normalization condition, i.e., $\mathcal{W}(0) = 1$, we obtain that

$$\mathcal{W}(\lambda) = 1 - \alpha g_1$$

and we assume that the stability condition $\alpha g_1 < 1$ holds.

By successive substitutions, we obtain from (8) that

$$\mathcal{W}(s) = \mathcal{W}(s_{n+1}) \prod_{j=0}^{n} \frac{b(s_j)}{a(s_j)} + \sum_{i=0}^{n} \frac{c(s_i)}{a(s_i)} \prod_{j=0}^{i-1} \frac{b(s_j)}{a(s_j)},$$

where we recall that $s_i = \lambda(1-\alpha^i)/(1-\alpha) + \alpha^i s$ for $i \geq 0$. When $n \to \infty$, we have

$$\mathcal{W}(s) = \mathcal{W}(s_\infty) \prod_{j=0}^{\infty} \frac{b(s_j)}{a(s_j)} + \sum_{i=0}^{\infty} \frac{c(s_i)}{a(s_i)} \prod_{j=0}^{i-1} \frac{b(s_j)}{a(s_j)}, \tag{9}$$

where an empty product equals 1 and (9) holds only when all the quantities involved are finite.

We now distinguish between the two cases of $\alpha < 1$ and $\alpha > 1$. Note that for the former case $\mathcal{W}(s_\infty) = \mathcal{W}(c)$, where we recall that $c = \lambda(1-\alpha)$, while in the latter case $\mathcal{W}(s_\infty) = \mathcal{W}(\infty)$. The case $\alpha = 1$ is handled similarly to the case $\alpha > 1$, but it involves some technical difficulties for some values of g_1. For purposes of clarity and simplification we focus on cases where $\alpha \neq 1$.

Case 1) $\alpha > 1$: In this case, we use the fact that the infinite product in (9) vanishes for any $\mathrm{Re}(s) > 0$ since the degree of the denominator is higher than the degree of the numerator for each term in the product and since s_j is monotonically increasing with j. Therefore, we obtain

$$\mathcal{W}(s) = \sum_{i=0}^{\infty} \frac{c(s_i)}{a(s_i)} \prod_{j=0}^{i-1} \frac{b(s_j)}{a(s_j)}.$$

Note that in this case $a(s_i)$ does not vanish for any $\mathrm{Re}(s) > 0$ and any i (since the stability condition $\alpha g_1 < 1$ holds), so as required the function $\mathcal{W}(s)$ is analytic in the region $\mathrm{Re}(s) > 0$. By taking the derivative of $\mathcal{W}(s)$ with respect to s at $s = 0$, we obtain the average delay $E[W]$,

$$\begin{aligned}\lambda E[W] = {} & g_1(1+\alpha) + g_1(\alpha - 1) \\ & \cdot \left[\frac{\alpha}{1-\alpha g_1} + \frac{1}{1-g_1}\right] + \frac{g_1^2 \alpha}{(1+\alpha)(\alpha - g_1)} \\ & + g_1^2(\alpha-1)^2 \sum_{i=3}^{\infty} \\ & \cdot \frac{1}{(1-g_1\alpha^{1-i})(\alpha^{i-1}-1)(\alpha^i - 1)} \\ & \cdot \prod_{j=2}^{i-1} \frac{g_1}{g_1 - \alpha^{j-1}}.\end{aligned}$$

Case 2) $\alpha < 1$: In this case, we observe from (9) that we need to compute the constant $\mathcal{W}(s_\infty) = \mathcal{W}(c)$. We do that by exploiting the analytic properties of $\mathcal{W}(s)$ for $\mathrm{Re}(s) > 0$. Rewrite (9) as

$$\mathcal{W}(s) = \frac{\left[\mathcal{W}(s_\infty) + \sum_{i=0}^{\infty} \frac{c(s_i)}{b(s_i)} \prod_{j=i+1}^{\infty} \frac{a(s_j)}{b(s_j)}\right] \prod_{j=0}^{\infty} b(s_j)}{\prod_{j=0}^{\infty} a(s_j)}. \tag{10}$$

Note that the denominator of $\mathcal{W}(s)$ vanishes only for values of s for which $a(s_j)$ vanishes for some j. Consider the root σ_0 for which $a(s) = a(s_0) = 0$, i.e.,

$$\sigma_0 = \frac{\lambda(1 - \alpha g_1)}{(1-\alpha)}.$$

Since $\sigma_0 > 0$, the numerator of $\mathcal{W}(s)$ must also vanish at $s = \sigma_0$. Therefore,

$$\mathcal{W}(s_\infty) = -\sum_{i=0}^{\infty} \frac{c(s_i|_{s=\sigma_0})}{b(s_i|_{s=\sigma_0})} \prod_{j=i+1}^{\infty} \frac{a(s_j|_{s=\sigma_0})}{b(s_j|_{s=\sigma_0})}. \tag{11}$$

Simple computation yields

$$a(s_i|_{s=\sigma_0}) = \lambda g_1 \alpha (1 - \alpha^i),$$
$$b(s_i|_{s=\sigma_0}) = \frac{\lambda g_1 \alpha (1 - g_1 \alpha^i)}{1 - g_1 \alpha^{i+2}},$$
$$c(s_i|_{s=\sigma_0}) = -\lambda g_1 (1-g_1)(1-\alpha g_1)\alpha^{i+1},$$

and therefore, after some algebra, we obtain

$$\mathcal{W}(s_\infty) = \sum_{i=0}^{\infty} \frac{(1-g_1)(1-\alpha g_1)\alpha^i}{(1-g_1\alpha^i)(1-g_1\alpha^{i+1})} \prod_{j=i+1}^{\infty} (1-\alpha^j). \tag{12}$$

It is important to note that the denominator of $\mathcal{W}(s)$ vanishes also at other values of s with $\mathrm{Re}(s) > 0$. This occurs when $a(s_j)$ vanishes for some $j \geq 1$. However, we will show that these values yield the same equation for $\mathcal{W}(s_\infty)$. For instance, assume that for some $k \geq 1$, $a(s_k) = 0$ and let σ_k denote the single solution of that equation, i.e., $\sigma_k = \lambda(1 - g_1\alpha^{1-k})/(1-\alpha)$. Note that we consider only values of k for which $\sigma_k > 0$. Then from (10), we obtain

$$\begin{aligned}\mathcal{W}(s_\infty) &= -\sum_{i=0}^{\infty} \frac{c(s_i|_{s=\sigma_k})}{b(s_i|_{s=\sigma_k})} \prod_{j=i+1}^{\infty} \frac{a(s_j|_{s=\sigma_k})}{b(s_j|_{s=\sigma_k})} \\ &= -\sum_{i=k}^{\infty} \frac{c(s_i|_{s=\sigma_k})}{b(s_i|_{s=\sigma_k})} \prod_{j=i+1}^{\infty} \frac{a(s_j|_{s=\sigma_k})}{b(s_j|_{s=\sigma_k})},\end{aligned} \tag{13}$$

where we used the fact that $a(s_k|_{s=\sigma_k}) = 0$. Now using the identity $s_i|_{s=\sigma_k} = s_{i-k}|_{s=\sigma_0}$ for $i \geq k$ we have that (13) is equivalent to (11).

Once the quantity $\mathcal{W}(s_\infty)$ is determined (see (12)), the LST of the delay is completely determined. By taking the derivative of $\mathcal{W}(s)$ with respect to s at $s = 0$, we obtain the average delay

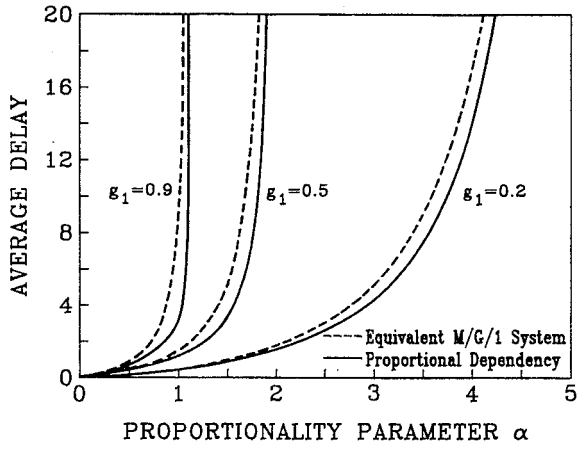

Fig. 4. Average delay versus proportionality parameter α.

in this case. Assuming that $g_1 \neq \alpha^i$ for any $i \geq 1$, we obtain

$$\lambda E[W] = g_1(1+\alpha)+g_1(\alpha-1)\left[\frac{\alpha}{1-\alpha g_1}+\frac{1}{1-g_1}\right]+\frac{g_1^2\alpha}{(1+\alpha)(\alpha-g_1)} +g_1^2(\alpha-1)^2\sum_{i=3}^{\infty}\frac{1}{(1-g_1\alpha^{1-i})(\alpha^{i-1}-1)(\alpha^i-1)} \cdot\prod_{j=2}^{i-1}\frac{g_1}{g_1-\alpha^{j-1}}+\mathcal{W}(s_\infty)\frac{g_1^2(1-\alpha)^2}{(1-\alpha g_1)(1-g_1)}\prod_{j=2}^{\infty}\frac{g_1}{g_1-\alpha^{j-1}},$$

where $\mathcal{W}(s_\infty)$ is given in (12). When $g_1 = \alpha^i$ for some $i \geq 1$ the expression obtained is slightly more involved and, therefore, omitted.

Numerical Examples: In Fig. 4, we depict the average delay as a function of the proportionality parameter α for various values of g_1 assuming that $\lambda = 1$. As expected, the average delay grows monotonically with α and with g_1. For comparison purposes we also plot the average delay of a customer in an equivalent M/G/1 system, in which the service time of a customer has the same distribution as in our system, i.e., the service time of the nth customer is distributed as $\Omega_n I_n$, but it is sampled independently of any other event in the system. We observe that the average delay of the equivalent M/G/1 system is always larger than the average delay of our system with the random proportional dependency. When $g_1 = 0.9$, the difference gets very large when the system is heavily loaded. For instance, for $\alpha = 1.08$ (1.1), the average delay in our system is 6.55 (13.82) while in the equivalent M/G/1 system it is 38.46 (109.9). Similar behaviors have also been observed in previous studies of correlated queues [14].

IV. An ON–OFF Source

In this section, the source that generates packets is modeled as an ON–OFF source which alternates between ON and OFF *states*. While in the ON state packets are generated at the end of each ON *period* (there may be multiple ON periods while in the ON state) with a size proportional to the duration of the ON period. ON and OFF periods are exponentially distributed with parameters λ and μ, respectively. At the end of an ON period the source either starts another ON period with probability $q = 1-p$ or an OFF period with probability p. At the end of an OFF period the source always begins an ON period. This results in a transition rate from ON state to OFF state equals to $p\lambda$, while transitions from OFF state to ON state occur at rate μ. We let n, $n = 0, 1, \cdots$, be the packet arrival index, with the nth packet arriving into the system at the end of the nth ON period. The random variable I_n^{ON} is used to denote the duration of the nth ON period. Similarly, in the case where the nth ON period is immediately preceded by an OFF period, this OFF period will then be termed the nth OFF period and the random variable I_n^{OFF} is used to denote its length. We further denote by I_n the interarrival time between the $(n-1)$st and the nth packets and by B_n the service time of the nth packet. We then have $B_n = \alpha I_n^{\mathrm{ON}}$, where α is the (nonnegative) proportionality constant relating the ON period duration and the service time of the nth packet, and $\alpha < 1 + p\lambda/\mu$ is the stability condition of the system. The evolution of the delay W_n of the nth packet is then given by,

$$W_{n+1} = (W_n - I_{n+1})^+ + \alpha I_{n+1}^{\mathrm{ON}}, \qquad n \geq 0. \tag{14}$$

Let δ_n be the state of the source just after the nth packet arrives. It is clear that δ_n = OFF with probability p and δ_n = ON with probability q. Using the evolution equation (14) and the definition of the LST, we have

$$\begin{aligned}\mathcal{W}_{n+1}(s) &= \Pr(\delta_n = \mathrm{OFF}) \cdot E\left[e^{-s\left[(W_n - I_{n+1}^{\mathrm{OFF}} - I_{n+1}^{\mathrm{ON}})^+ + \alpha I_{n+1}^{\mathrm{ON}}\right]}\right] \\ &\quad + \Pr(\delta_n = \mathrm{ON}) E\left[e^{-s\left[(W_n - I_{n+1}^{\mathrm{ON}})^+ + \alpha I_{n+1}^{\mathrm{ON}}\right]}\right] \\ &= p\int_0^\infty f_{W_n}(w)\left[\int_0^w \lambda e^{-\lambda x} \cdot\left(\int_0^{w-x}\mu e^{-\mu y} \cdot e^{-s(w-(1-\alpha)x-y)}\,dy + \int_{w-x}^{\infty}\mu e^{-\mu y}e^{-s\alpha x}\,dy\right)dx + \int_w^\infty \lambda e^{-\lambda x}e^{-s\alpha x}\,dx\right]dw \\ &\quad + q\int_0^\infty f_{W_n}(w)\left[\int_0^w \lambda e^{-\lambda x}e^{-s(w-(1-\alpha)x)}\,dx + \int_w^\infty \lambda e^{-\lambda x}e^{-s\alpha x}\,dx\right]dw.\end{aligned} \tag{15}$$

Assuming $\alpha < 1 + p\lambda/\mu$ and letting $n \to \infty$, we get

$$a(s)\mathcal{W}(s) = b(s)\mathcal{W}(\lambda + \alpha s) + c(s), \tag{16}$$

where

$$\begin{aligned} a(s) &= (\lambda+\alpha s)(\lambda-\mu+\alpha s)(\lambda p + (\alpha-1)(s-\mu)), \\ b(s) &= \lambda(s-\mu)(\mu - q(\lambda+\alpha s)) \\ c(s) &= \lambda p(\lambda+\alpha s)(\lambda - s + \alpha s)\mathcal{W}(\mu). \end{aligned} \tag{17}$$

By successive substitutions, we obtain from (16) that

$$\mathcal{W}(s) = W(s_\infty)\prod_{j=0}^{\infty}\frac{b(s_j)}{a(s_j)} + \sum_{i=0}^{\infty}\frac{c(s_i)}{a(s_i)}\prod_{j=0}^{i-1}\frac{b(s_j)}{a(s_j)}, \quad (18)$$

where an empty product equals 1 and we recall that $s_i = \lambda(1-\alpha^i)/(1-\alpha) + \alpha^i s$, $i \geq 0$. As in Section III, the above expression is only meaningful when all the quantities involved are defined, i.e., the infinite product is bounded. This can be shown to hold when $\alpha \neq 1$, but as before technical difficulties exist in the case $\alpha = 1$. For purposes of clarity and simplification, we focus on cases where $\alpha \neq 1$.

In order to uniquely determine $\mathcal{W}(s)$ we need to identify the two unknowns $\mathcal{W}(\mu)$ and $\mathcal{W}(s_\infty)$. We shortly describe how they can be obtained using the normalization condition and the analytic properties of $\mathcal{W}(s)$ in the region $\mathrm{Re}(s) > 0$. Let

$$\gamma_i = \prod_{j=0}^{i}\frac{b(s_j)}{a(s_j)}\bigg|_{s\to 0} = \frac{\mu(1-\alpha)}{\beta_{i+1}}\prod_{j=0}^{i}\frac{\beta_{j+1}}{(1-\alpha^{j+1})(\beta_j+\lambda p)} \quad (19)$$

and

$$d_1 = \frac{1}{\mathcal{W}(\mu)}\sum_{i=0}^{\infty}\frac{c(s_i)}{a(s_i)}\prod_{j=0}^{i-1}\frac{b(s_j)}{a(s_j)}\bigg|_{s\to 0} = (1-\alpha)\lambda^2 p\sum_{i=0}^{\infty}\frac{\alpha^i(\alpha^{i+1}-1)\gamma_i}{\beta_i\eta_{i+1}}, \quad (20)$$

where $\beta_i \triangleq \mu(1-\alpha) - \lambda(1-\alpha^i)$, $\eta_i \triangleq \mu(1-\alpha) - \lambda q(1-\alpha^i)$, $i \geq 0$. In the rest of the analysis, we assume that $\beta_i, \eta_i \neq 0$. If $\beta_k = 0$ for some $k \geq 1$, a different equation for d_1 can be obtained using the fact that $a(s_k)\sum_{i=k}^{\infty}\frac{c(s_i)}{a(s_i)}\prod_{j=0}^{i-1}\frac{b(s_j)}{a(s_j)}\Big|_{s\to 0} = 0$. Then, d_1 can be obtained using L'Hospital's law at $s = 0$. Similarly, if $\eta_k = 0$ for some $k \geq 1$, we have $d_1 = \frac{1}{\mathcal{W}(\mu)}\sum_{i=0}^{k}\frac{c(s_i)}{a(s_i)}\prod_{j=0}^{i-1}\frac{b(s_j)}{a(s_j)}\Big|_{s\to 0}$. We omit the derivations for these cases as they are straightforward and do not add to the understanding of the solution technique. We now proceed with the determination of the two unknowns $\mathcal{W}(\mu)$ and $\mathcal{W}(s_\infty)$ for which we distinguish between the cases $\alpha < 1$ and $\alpha > 1$.

Case 1) $\alpha > 1$*:* In this case, $s_\infty = \infty$ and it can be shown that all the quantities involved are finite (in fact, the infinite product $\prod_{j=0}^{\infty} b(s_j)/a(s_j)$ vanishes as the degree of $a(s)$ is larger than that of $b(s)$). Therefore we have

$$\mathcal{W}(s) = \sum_{i=0}^{\infty}\frac{c(s_i)}{a(s_i)}\prod_{j=0}^{i-1}\frac{b(s_j)}{a(s_j)}. \quad (21)$$

Note that (assuming $\beta_i, \eta_i \neq 0, i \geq 1$) $a(s_i)$ does not vanish for any $\mathrm{Re}(s) > 0$ and any i as long as the stability condition holds. The function $\mathcal{W}(s)$ is, therefore, as required analytic in the region $\mathrm{Re}(s) > 0$. The second unknown $\mathcal{W}(\mu)$ is then obtained by applying the normalization condition $(\mathcal{W}(0) = 1)$ in equation (21), which gives

$$\mathcal{W}(\mu) = d_1^{-1} = \left((1-\alpha)\lambda^2 p\sum_{i=0}^{\infty}\frac{\alpha^i(\alpha^{i+1}-1)\gamma_i}{\beta_i\eta_{i+1}}\right)^{-1}.$$

Once $\mathcal{W}(s)$ has been determined, we can proceed to obtain an expression for the average delay in the system, $E[W]$. Let

$$\theta_i = \frac{d}{ds}\left(\prod_{j=0}^{i}\frac{b(s_j)}{a(s_j)}\right)\bigg|_{s\to 0} = \sum_{j=0}^{i}\alpha^j\left(\frac{\alpha}{\beta_{j+1}} + \frac{1}{\beta_j+\lambda p} - \frac{1}{\beta_j} - \frac{\alpha q}{\eta_{j+1}} - \frac{\alpha}{\lambda(1-\alpha^{j+1})}\right) \quad (22)$$

and

$$d_2 = \frac{1}{\mathcal{W}(\mu)}\frac{d}{ds}\left(\sum_{i=0}^{\infty}\frac{c(s_i)}{a(s_i)}\prod_{j=0}^{i-1}\frac{b(s_j)}{a(s_j)}\right)_{s\to 0} = (1-\alpha)^2\lambda^2 p\sum_{i=0}^{\infty}\frac{\alpha^i(\alpha^{i+1}-1)}{\beta_i\eta_{i+1}} \cdot\left(\frac{2\alpha^{i+1}-1}{\lambda(1-\alpha^{i+1})} + \frac{q\alpha^{i+1}}{\eta_{i+1}} + \frac{\alpha^i}{\beta_i} + \theta_i\right). \quad (23)$$

Then, we find that

$$E[W] = -\frac{d\mathcal{W}(s)}{ds}\bigg|_{s\to 0} = -\mathcal{W}(\mu)d_2.$$

Case 2) $\alpha < 1$*:* In this case all the quantities involved can again be shown to be finite and we have $\mathcal{W}(s_\infty) = \mathcal{W}\left(\frac{\lambda}{1-\alpha}\right)$. We, therefore, need to determine the two unknowns $\mathcal{W}(\mu)$ and $\mathcal{W}\left(\frac{\lambda}{1-\alpha}\right)$. The normalization condition gives us one equation,

$$d_1\mathcal{W}(\mu) + \gamma_\infty\mathcal{W}\left(\frac{\lambda}{1-\alpha}\right) = 1, \quad (24)$$

where $\gamma_\infty \triangleq \lim_{i\to\infty}\gamma_i$ and γ_i is given in (19).

Another equation is obtained by exploiting the analyticity of $\mathcal{W}(s)$ in the region $\mathrm{Re}(s) > 0$. For this purpose, we rewrite (18) as

$$\mathcal{W}(s) = \frac{\left[\mathcal{W}(s_\infty) + \sum_{i=0}^{\infty}\frac{c(s_i)}{b(s_i)}\prod_{j=i+1}^{\infty}\frac{a(s_j)}{b(s_j)}\right]\prod_{j=0}^{\infty}b(s_j)}{\prod_{j=0}^{\infty}a(s_j)} \quad (25)$$

Note that the denominator of $\mathcal{W}(s)$ vanishes only for values of s for which $a(s_j) = 0$ for some j. Consider the root σ_0 for which $a(s) = a(s_0) = 0$, i.e.,

$$\sigma_0 = \mu + \frac{\lambda p}{1-\alpha}.$$

Since $\sigma_0 > 0$, the numerator of $\mathcal{W}(s)$ must also vanish at $s = \sigma_0$. Therefore,

$$\mathcal{W}\left(\frac{\lambda}{1-\alpha}\right) = d_3\mathcal{W}(\mu), \quad (26)$$

where

$$d_3 = -\frac{1}{\mathcal{W}(\mu)} \sum_{i=0}^{\infty} \frac{c(s_j)}{b(s_j)} \prod_{j=i+1}^{\infty} \frac{a(s_i)}{b(s_i)} \Bigg|_{s=\sigma_0}$$
$$= (1-\alpha)p(\lambda - \mu(1-\alpha)) \cdot \sum_{i=0}^{\infty} \frac{\alpha^i(\xi_{i+1} + \mu(1-\alpha))}{(1-q\alpha^{i+1})\xi_i\xi_{i+1}} \cdot \prod_{j=i+1}^{\infty} \frac{(1-\alpha^j)(\xi_{j+1} + \mu(1-\alpha))}{\lambda(1-q\alpha^{j+1})}$$

and $\xi_i \triangleq \lambda(1-q\alpha^i) - \mu(1-\alpha)(1-\alpha^i)$, $i \geq 0$. For simplicity, we assume that $\xi_i \neq 0$, $i \geq 1$. If $\xi_i = 0$ for some $i \geq 1$, d_3 can again be obtained by applying L'Hospital's rule at $s = \sigma_0$.

It is important to note that the denominator of $\mathcal{W}(s)$ can also vanish at the other values of s with $\mathrm{Re}(s) > 0$. This occurs when $a(s_i)$ vanishes for some $i \geq 1$. However, as in Section III, it can be shown again that all these values yield the same equation (26). Now, from (24) and (26), we have

$$\mathcal{W}(\mu) = \frac{1}{d_1 - \gamma_\infty d_3}, \qquad \mathcal{W}\left(\frac{\lambda}{1-\alpha}\right) = \frac{d_3}{d_1 - \gamma_\infty d_3}. \quad (27)$$

The average delay $E[W]$ can then be found to be

$$E[W] = -\left(\mathcal{W}(\mu)d_2 + \mathcal{W}\left(\frac{\lambda}{1-\alpha}\right)\gamma_\infty\theta_\infty\right), \quad (28)$$

where $\theta_\infty \triangleq \lim_{i\to\infty} \theta_i$ and θ_i is given in (22).

Numerical Examples: We now provide some numerical examples that illustrate the results developed in this section. In particular, we compute the average delay for the case $p = 1$, i.e., a single packet is generated at the end of the ON state, and compare it to the values obtained assuming equivalent GI/M/1 and fluid-flow models. The equivalent GI/M/1 system has independent interarrival times with a probability distribution whose LST is $\mathcal{A}(s) = \mu\lambda/(s+\mu)(s+\lambda)$. The service times are independent of the interarrival times and exponentially distributed with parameter λ/α. The equivalent fluid-flow model is such that the output rate is 1 and the input rate in the ON state is $\alpha > 1$ (for $\alpha < 1$ the unfinished work in the system is always zero). For this model, the packet delay is defined from the time the last bit of the packet is received until the time it completely departs the system. Therefore, the average delay for the equivalent fluid-flow model is $(\alpha-1)/\lambda - \mu(\alpha-1)$, if $\alpha > 1$ and zero, otherwise.

The average delays for all three models are plotted in Fig. 5 as a function of the proportionality parameter α. Note that $\alpha < 1 + \lambda/\mu$ is the stability condition for all three models. We consider two cases with $\mu = 1$, $\lambda = 0.2$ and $\mu = 1$, $\lambda = 1.2$. The results for both cases illustrate the fact that the models developed in this paper in a sense "bridge the gaps" left by previous approaches.

Specifically, while traditional point-process models such as the GI/M/1 account for the granular nature of customer arrivals and departures, they typically ignore dependencies between interarrival and service times. As demonstrated in Fig. 5 and many previous studies, this often results in overly pessimistic estimates of system performance, especially at high loads. Conversely, fluid-flow models successfully capture the dependencies that exist between interarrival and service times, but they fail to preserve the discrete nature of these events. As alluded to in Fig. 1 and illustrated in Fig. 5, this can in turn yield an overly optimistic view of system behavior, in particular at light and medium loads. The models developed in this paper, because they are able to retain both aspects, provide more accurate estimates of actual system performance for all load values.

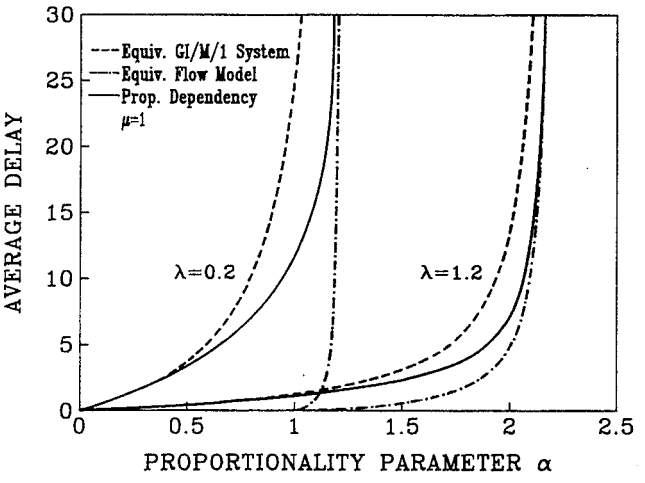

Fig. 5. Average delay versus proportionality parameter α.

V. Summary

In this paper, we have developed several models and solution methods, that allow accurate and yet tractable analysis of a number of key aspects in packet communication systems. In particular, the models allow us to capture the strong positive correlation that finite transmission speeds introduce between packet sizes and interarrival times. The dual scenario, where the interarrival time exhibits a proportional dependency on the service time of the previous packet was considered in a companion paper [25]. As illustrated through a number of numerical results, these dependencies can have a significant impact on system performance, which simpler and more traditional models failed to account for.

The scenarios considered in the paper and the results that were derived, extend the work started in [14] in a number of directions. The approach taken, based on functional equations, provides a general framework for the study of such systems and allows us to greatly extend the type of problems that can be studied. In particular, this enables us to define and analyze new models which accurately describe communication links in packet networks. This gives a greater insight into the actual behavior of such systems.

Although packet networks originally motivated the investigation carried out in this paper, the models and techniques that were developed should be of general interest and applicable to other environments. As a case in point, the initial study of correlated queues carried out in [14] did not even arise in a queueing context. Rather, it was intended as a model for a chase involving a hunter and its quarry.

REFERENCES

[1] K. W. Fendick, V. R. Saksena, and W. Whitt, "Dependence in packet queues." *IEEE Trans. Commun.*, vol. COM-37, pp. 1173–1183, Nov. 1989.

[2] L. Kleinrock, *Communication Nets.* New York: McGraw-Hill, 1964.

[3] O. J. Boxma, "On a tandem queueing model with identical service times at both counters, parts I and II," *Adv. Appl. Probab.*, vol. 11, pp. 616–659, 1979.

[4] S. B. Calo, "Delay properties of message channels," *ICC*, pp. 43.5.1–43.5.4, 1979.

[5] ——, "Message delays in repeated-service tandem connections," *IEEE Trans. Commun.*, vol. COM-29, pp. 670–678, May 1981.

[6] O. Hashida and M. Fujiki, "Queueing models for buffer memory in store-and-forward systems," in *Proc. 7th Int. Teletraffic Congress (ITC 7)*, Stockholm, 1973, Paper 323, pp. 323/1-323/7.

[7] H. Kaspi and M. Rubinovitch, "The stochastic behavior of a buffer with non-identical input lines," *Stochastic Processes Applicat.*, vol. 3, pp. 73–88, 1975.

[8] D. Anick, D. Mitra, and M. M. Sondhi, "Stochastic theory of a data-handling system with multiple sources," *Bell Syst. Tech. J.*, vol. 61, pp. 1871–1894, Oct. 1982.

[9] L. Kosten, "Liquid models for a type of information storage problems," *Delft Progress Report: Mathematical Engineering, Mathematics and Information Engineering*, vol. 11, pp. 71–86, 1986.

[10] D. Mitra, "Stochastic theory of a fluid model of producers and consumers coupled by a buffer," *Adv. Appl. Probab.*, vol. 20, pp. 646–676, Sept. 1988.

[11] A. I. Elwalid, D. Mitra, and T. E. Stern, "Statistical multiplexing of Markov modulated sources: Theory and computational algorithms," in *Proc. 13th Int. Teletraffic Cong.*, Copenhagen, 1991, pp. 495–500.

[12] T. E. Stern and A. I. Elwalid, "Analysis of separable Markov-modulated models for information-handling systems," *Adv. Appl. Probab.*, vol. 23, pp. 105–139, Mar. 1991.

[13] J. W. Roberts, "Variable-bit-rate traffic control in B-ISDN," *IEEE Commun. Mag.*, vol. 29, pp. 50–56, Sept. 1991.

[14] B. W. Conolly, "The waiting time process for a certain correlated queue," *Operations Res.*, vol. 16, pp. 1006–1015, 1968.

[15] B. W. Conolly and N. Hadidi, "A comparison of the operational features of conventional queues with a self-regulating system," *Appl. Statist.*, vol. 18, pp. 41–53, 1969.

[16] ——, "A correlated queue," *J. Appl. Probab.*, vol. 6, pp. 122–136, 1969.

[17] B. W. Conolly and Q. H. Choo, "The waiting time process for a generalized correlated queue with exponential demand and service," *SIAM J. Appl. Math.*, vol. 37, no. 2, pp. 263–275, Oct. 1979.

[18] N. Hadidi, "Queues with partial correlation," *SIAM J. Appl. Math.*, vol. 40, no. 3, pp. 467–475, June 1981.

[19] ——, "Further results on queues with partial correlation," *Operations Res.*, vol. 33, pp. 203–209, 1985.

[20] C. Langaris, "A correlated queue with infinitely many servers," *J. Appl. Probab.*, vol. 23, pp. 155–165, 1986.

[21] M. B. Combe, S. C. Borst, and O. J. Boxma, "Collection of customers: A correlated M/G/1 queue," *Performance Eval. Rev.*, vol. 20, pp. 47–59, 1991.

[22] M. Kuczma, B. Choczewski, and R. Ger, *Iterative Functional Equations* (Encyclopedia of Mathematics and Its Applications). Cambridge, MA: Cambridge Univ. Press, 1990.

[23] R. Guérin, H. Ahmadi, and M. Naghshineh, "Equivalent capacity and its application to bandwidth allocation in high-speed networks," *IEEE J. Select. Areas Commun.*, vol. 9, no. 7, pp. 968–981, Sept. 1991.

[24] L. Kleinrock, *Queueing Systems, Volume 1: Theory.* New York: John Wiley, 1975.

[25] I. Cidon, R. Guérin, A. Khamisy, and M. Sidi, "On queues with inter-arrival times proportional to service times," to appear in *INFOCOM'93*, Apr. 1993.