Single Server Queueing Models for Communication Systems

Moshe Sidi and Asad Khamisy
Department of Electrical Engineering
Technion — Israel Institute of Technology
Haifa 32000, Israel

Abstract

One of the most important performance measures in the analysis of data networks is the average delay required to deliver a packet from origin to destination. The packet delay will affect the choice and performance of several network algorithms such as routing and flow control. Queueing theory is the primary methodological framework for analyzing network delay. It provides a basis for adequate delay approximations, as well as valuable qualitative results and worthwhile insights.

The literature on queueing models for communication systems is abundant. In this paper we concentrate on single server queueing models for communication systems. We describe some of the more recent models that were developed for modeling the behavior of nodes of modern communication networks. These models are fundamental and provide the foundation for more elaborate models with multiple servers.

We focus on the queueing and transmission delays of packets. First, we describe some classical queueing models, namely, the $M/M/1$, $M/G/1$ and $G/M/1$ queueing models. These models assume that all interarrival times and service times are independent. This assumption does not hold for communication systems and we describe extensions of the classical queueing models to allow for such dependency. It is demonstrated that with this dependency the delays in the buffers preceding communication links are smaller. Next, we describe completely different approaches for the characterization of arrival processes. These characterizations are motivated by the need to model real-time applications in high-speed communication networks where it is important to capture the burstiness of the arrival processes. Finally, we describe another performance measure, namely, the message delay. The motivation for this measure arises from the process of segmentation and reassembly of messages into packets which is natural for packet-switched networks.

1 Introduction

Queueing models and analytical techniques for evaluating the performance of communication systems evolved during the recent decades hand by hand with the progress of these systems. In its most general formulation, the problem addressed by a performance analyst is to characterize the service provided by a system when it is loaded with some given load. In this context the system is known by a full description of the behavior of all its components (in a communication system these are the nodes and the links of the system), the load is given as some known stochastic processes (in a communication system the load corresponds to the traffic generated by the users of the system), and the service is to be characterized by the distribution of its parameters (such as the traffic in its links, the delays it causes, etc.).

The particular importance of the problem of performance analysis of communication systems stems from the variety of settings in which it is encountered. For instance, when a new communication network is designed, which is supposed to guarantee some predetermined parameters of service, one wishes to calculate the amount and size of resources (e.g., buffer sizes at the nodes, capacity of the links, etc.) which are needed to fulfill the requirements. The difficulty of the problem arises from the competitive and heterogeneous demands for such resources from the users of the communication system. In fact, this problem is a special case of the resource allocation problem which is the root of most of the problems in the field of information processing and beyond this field. The capacity assignment problem [40] is another variant of this problem in which one wishes to minimize the average delay of a packet in the network for a given topology, packet arrival processes and the total capacity of the links of the network. Another example is when a new user wishes to join an already operational network. Since modern real-time applications require some minimal performance guarantees (such as very rare packet losses, or very short delays), the question here is whether the new user can be admitted, and accept its needed service, while the performance degradation other existing users sense will not violate their needs. One of the most difficult aspects of this problem is that of treating all users fairly when accepting a new user to the system. The problem of admission control in packet-switched networks was studied in the past, see e.g. [4] for a discussion of this problem and a definition of fairness in such networks. Recently, many papers have studied this problem in the context of high speed packet-switched networks in which a fast-to-implement policy is needed, see e.g. [26, 30, 31].

A certain special case of analysis of a communication system, which has been addressed and studied very intensively, is the case of single node networks. Queueing models (see, for example, [39]) that consist of a single service station with a single server that is fed by a single stream of customers has been used. The input process of new customers is assumed to be some known stochastic point-process, and the service is described by the distribution of the time to process each customer. The customers in these models correspond to packets in a communication system. Many other cases of single node networks have also been studied, see [1, 28, 36, 41, 48, 56] and many others. These works consider a variety of arrival processes, service distributions and service disciplines, and address cases where the system has several input links or output links. The fact of the matter is that it is impossible to discuss all the various models that appeared in the literature in a single paper. Even a complete book will be too short to contain all of them. In this paper we will try to describe some of the more recent models that were developed for modeling the behavior of nodes of modern communication networks.

The case in which the communication network consists of more than a single node, has proved to be very complex, even in simple settings. This complexity is essentially due to the complicated way in which different traffic streams interact with each other within the network, and to the dependencies which are imposed by these interactions. One can distinguish between networks that are served by a single server and networks with a server in each of the nodes. The former usually corresponds to a communication system with users that share a common resource, such as a single cell wireless system, an Ethernet or a Token-Ring local area network, etc. Various queueing models are used to analyze the performance of such systems, such as polling models, priority models, etc. Networks with servers in each node usually correspond to wide area computer communication networks, such as SNA, DECNET, ARPANET and TYMNET, see [4, 59] for a brief description of these networks. Closed-form solutions for the number of packets in the nodes of the network in this case is known only for a limited class of networks, known as product-form networks. A good example for this class is the Jackson network (see [37, 39]), where the queue lengths of all the nodes behave as if they were independent.

It is apparent that queueing theory is the primary methodological framework for analyzing the performance of communication systems such as the delay of a packet in the system. Its use often requires substantial simplifying assumptions since, unfortunately, the theory is still limited and realistic assumptions make meaningful analysis extremely difficult. For this reason, it is sometimes impossible to obtain accurate quantitative performance measure predictions on the basis of queueing models. Nevertheless, these models often provide a basis for adequate delay approximations, as well as valuable qualitative results and worthwhile insights, as we will attempt to describe in this paper.

The literature on queueing models for communication systems is abundant. It is probably impossible to cover all subjects in a single paper. Therefore, in this paper we decided to concentrate on single server queueing models for communication systems. These models are fundamental and provide the foundation for more elaborate models with multiple servers.

In the following sections we describe several studies that analyze the packet delays in communication systems. In Section 3 we describe several classical queueing systems, namely, the $M/M/1$, $M/G/1$ and $G/M/1$ systems. We describe several results which hold under very general assumptions and summarize the main results for each model. We proceed with more realistic models for communication systems in Section 4, where we describe models which relax the independence assumption of the interarrival and service times used in the classical models. In Section 5 we describe completely different approaches for the characterization of arrival processes. These novel characterizations are motivated by the need to model real-time applications in high-speed communication networks where it is important to capture the *burstiness* of the arrival processes. In Section 6 we describe another performance measure, namely, the message delay. We review some of the works which analyze the behavior of the message delay process in some classical queueing systems. Numerical examples are provided throughout the paper to clarify the differences between the various models.

2 General model

We consider a communication link with given transmission capacity of C bits per second. Several traffic streams (e.g., sessions or virtual connections) are multiplexed on the link. The manner that the capacity is allocated between these traffic streams affects the performance characteristics (e.g., delay and loss) of each traffic stream. The most common scheme used in packet-switched networks is called *statistical multiplexing.* In this scheme, the packets of all traffic streams are merged into a single queue at the front of the communication link and transmitted on a First-Come First-Served (FCFS) basis. The variability of the arrival processes from the different traffic streams can cause the arrival rate of packets scheduled for transmission on the communication link to momentarily exceed the transmission rate, and hence a queue of packets can build up at the front of the link. Typical arrival rates range from one arrival per many minutes to one million arrivals per second. Simple models for the variability of the arrivals include Poisson arrivals, deterministic arrivals and uniformly distributed arrivals. Typical packet lengths vary roughly from a few bits to 10^8 bits, with file transfer applications at the high end and interactive sessions at the low end. Simple models for packet length distribution include an exponentially decaying probability distribution, fixed length packets and uniform probability distribution.

One of the most important performance measures of a data network is the average delay required to deliver a packet from origin to destination. The packet delay will affect the choice and performance of several network algorithms such as routing and flow control. In what follows, we will focus on packet delay in a single server queueing system representing the communication link. This delay consists of four components: (1) The processing delay which corresponds to the delay between the time the packet is received at the input link and the time it is assigned to an output link queue for transmission. This delay can be very large in today's communication networks, where an average of a few thousands instructions are performed for each packet received in the node; but is negligible in high-speed networks, such as ATM [60], where the packet header is processed by a very fast dedicated hardware. This delay is usually independent of the amount of traffic handled by the node. (2) The queueing delay which corresponds to the delay between the time the packet is assigned to a queue for transmission and the time it starts being transmitted. During this time the packet waits in the queue while other packets are being transmitted. (3) The transmission delay which corresponds to the delay between the times that the first and the last bits of the packet are transmitted. (4) The propagation delay which corresponds to the delay from the time the last bit of the packet is transmitted at one end of the link until the time it is received at the other end of the link. This delay depends on the physical characteristics of the link and is independent of the traffic carried on the link.

In what follows we will focus on the queueing and transmission delays of packets. First, we describe some classical queueing models, namely, the $M/M/1$, $M/G/1$ and $G/M/1$ queueing systems. The first letter indicates the nature of the arrival process, where M stands for Poisson process and G stands for general distribution of interarrival times. The second letter indicates the nature of the probability distribution of the service (transmission) time, where M and G stand for exponential and general distributions, respectively. The third letter indicates the number of servers, where in our case there is one server (transmission line). The server do not idle when there are packets waiting in the queue (work conserving server). Moreover, these models assume that all interarrival times and service times are independent. This assumption is the key difference between a communication system and its corresponding queueing system model and it is related to the well known *independence assumption*, see Kleinrock [40]. The independence assumption for Poisson arrival processes and exponentially distributed service times states that each time a packet is received at a node within the system, a new service time is chosen independently from an exponential distribution. This is clearly inadequate for a communication system since packets maintain their lengths as they pass through the network. However, it was suggested by Kleinrock that merging several packet streams on a transmission line has an effect akin to restoring the independence of interarrival times and packet service times.

To further clarify this point, we consider two transmission lines in tandem where the interarrival times and service times of packets at the first transmission line are independent. The interarrival time between two successive packets on the second transmission line can certainly be no less than the service time for the second of these packets on the first transmission line. Since the service time for this packet on the second transmission line is directly related to its previous service time (and therefore highly correlated with the interarrival time between the two packets on the second transmission line), we see that the arrival process of packets to a node due to the internal traffic in the network is not independent of the service time these packets receive at that node. In what follows we describe extensions of the classical queueing models to allow for this dependency between the interarrival and service times. We further describe additional motivations for such models and several results related to the packet delay distribution.

Another crucial point that we address in this paper is related to the burstiness characterization of arrival processes. We describe several recent approaches for such a characterization that are motivated by the need to model real-time applications in high-speed communication networks. The common feature of these approaches is the bounding (either deterministic or stochastic) of the burstiness of the arrival processes that yields computable bounds on many performance measures in the network.

In many systems the message delay (where a message consists of a block of consecutive packets), 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 [60], TCP/IP [17] and TDMA based systems [53] 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 packets 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. We will describe some of the recent papers which analyze the behavior of the message delay for some classical queueing systems.

3 Classical queueing models

Here, we describe the $M/M/1$, $M/G/1$ and $G/M/1$ queueing models. The analysis of these models can be found in almost every book related to queueing theory. We refer to Kleinrock [39] for the details of the analysis of each of these models.

3.1 General queueing results

Before proceeding with the analysis of the packet delays in classical queueing models, we mention three general results which hold for a $G/G/1$ queueing system. The first general result is known as Little's theorem [46]. It states that the average number of customers (packets) in a queueing system is equal to the average arrival rate of customers to that system, times the average time spent in that system. The derivation of this theorem is simple and it doesn't depend on any specific assumptions regarding the arrival distribution or the service time distribution; nor does it depend on the number of servers in the system or upon the particular queueing discipline within the system. Note that, the system in this case can correspond to the queue and the server, the queue only or the server only, in which case relations are obtained between the corresponding entities.

The second general result is related to the stability of the $G/G/1$ queueing system. Define ρ as the average arrival rate of customers, times the average service time of a customer. Then, $\rho < 1$ is a sufficient condition for the stability of the $G/G/1$ queueing system, given that the arrival and the service processes are ergodic [2]. Stability here refers to the fact that limiting distributions of all random variables of interest (such as the delay in the system) exist, and that all customers are eventually served. We will assume that this condition holds for all queueing systems we consider, unless otherwise specified.

Another result which holds under very general assumptions is that in steady-state, the system appears statistically identical to an arriving and departing customer. That is, in steady-state, the number of customers in the system just before an arrival equals in distribution to the number of customers just after a departure. The only requirement for this result to hold is that the system reaches a steady-state with positive steady-state probabilities to have any n customers in the system, and that the number in the system changes by unit increments. These assumptions hold for all queueing systems we consider.

3.2 The $M/M/1$ system

The arrival process is Poisson with rate (inverse of average interarrival time) λ. The service time is exponentially distributed with rate (inverse of average service time) μ. The analysis of the $M/M/1$ system is based on the theory of Markov chains [55], where the states of the Markov chain correspond to the number of customers in the system. In particular, it is based on a special case of a Markov process named the birth-death process in which two successive states can only differ by a unity. From the state transition rate diagram of the birth-death process the so called detailed balance equations are obtained from which the steady-state probability p_n of having n customers in the system is obtained. All quantities of interest can then be obtained from this probability distribution. Below we summarize the results for the $M/M/1$ system.

- Utilization factor (proportion of time the server is busy) $\rho = \frac{\lambda}{\mu}$.
- Probability of n customers in the system $p_n = \rho^n(1-\rho) \quad n = 0, 1, 2, \ldots$.
- Average number of customers in the system $N = \frac{\rho}{1-\rho}$.
- Average customer time in the system $T = \frac{1}{\mu-\lambda}$.

- Average number of customers in queue $N_Q = \frac{\rho^2}{1-\rho}$.
- Average waiting time in queue of a customer $W = \frac{\rho}{\mu-\lambda}$.
- Probability density function (pdf) of the system time $t(y) = (\mu-\lambda)\exp^{-(\mu-\lambda)y}$, $y \geq 0$. That is, the system time is exponentially distributed with parameter $\mu - \lambda$.
- Probability density function of the waiting time $w(y) = (1-\rho)u_0(y)+(\mu-\lambda)\exp^{-(\mu-\lambda)y}$, $y \geq 0$, where $u_0(y)$ is the unit impulse function.
- Laplace Stieltjes Transform (LST) of the pdf of the system time $\mathcal{T}(s) = \frac{\mu-\lambda}{s+\mu-\lambda}$.

From the expressions for N and T, note that, increasing the arrival rate and the service rate by a factor of K, $K > 1$, does not change N but decreases T by a factor of K. In other words, a transmission line K times as fast will accommodate K times as many packets per second at K times smaller average delay per packet.

Finally, we will describe two additional results related to this system. The first is known as the PASTA (Poisson Arrivals See Time Averages) property [54]. It states that when the arrival process is Poisson, an arriving customer finds the system in a "typical" state. That is, the probability distribution of the number of customers in the system just before an arrival equals to the steady-state probability distribution. This holds for queueing systems with Poisson arrivals regardless of the distribution of the service time. The second is known as Burke's theorem [7]. It states that the departure process of an $M/M/1$ system is itself Poisson with parameter λ and is independent of the other processes in the system.

3.3 The $M/G/1$ system

The arrival process is Poisson with rate λ. The service time has arbitrary distribution with LST denoted by $\mathcal{B}(s)$. Let μ and x^2 be the average and the second moments of the service time. The analysis here is based on the method of the embedded Markov chain at the departure instants from service. Below we summarize the results for the $M/G/1$ system.

- Utilization factor (proportion of time the server is busy) $\rho = \frac{\lambda}{\mu}$.

- The generating function $\mathcal{Q}(z) \triangleq \sum_{i=0}^{\infty} p_i z^i$ of the probability distribution $p_i,\ i \geq 0$ of the number of customers in the system in steady-state (also at departure and arrival instants) $\mathcal{Q}(z) = \mathcal{B}(\lambda - \lambda z)\frac{(1-\rho)(1-z)}{\mathcal{B}(\lambda-\lambda z)-z}$.

- Average number of customers in the system $N = \rho + \frac{\lambda^2 x^2}{2(1-\rho)}$.

- Average customer time in the system $T = \frac{1}{\mu} + \frac{\lambda x^2}{2(1-\rho)}$.

- Average number of customers in queue $N_Q = \frac{\lambda^2 x^2}{2(1-\rho)}$.

- Average waiting time in queue of a customer $W = \frac{\lambda x^2}{2(1-\rho)}$.

- LST for the system time $\mathcal{T}(s) = \mathcal{B}(s)\frac{s(1-\rho)}{s-\lambda+\lambda\mathcal{B}(s)}$.

- LST for the waiting time $\mathcal{W}(s) = \frac{s(1-\rho)}{s-\lambda+\lambda\mathcal{B}(s)}$.

Since the $M/D/1$ system yields the minimum possible value for x^2 for a given μ, it follows that the values of W, T, N_Q and N for the $M/D/1$ system are lower bounds to the corresponding quantities for an $M/G/1$ system of the same λ and μ.

It turns out that the average number of customers in the system is the same for any order of servicing customers (and not only for the FCFS discipline assumed throughout the paper) as long as the order is determined independently of the required service times.

3.4 The $G/M/1$ system

This is in fact the "dual" of the $M/G/1$ system. The interarrival times have a general LST $\mathcal{A}(s)$ with a mean time between arrivals equals $1/\lambda$. The service times of customers are exponentially distributed with mean $1/\mu$. The analysis is based on the method of the embedded Markov chain at the arrival instants to the system. All the results are expressed in terms of a root σ that is the unique root in the range $0 < \sigma < 1$ of the functional equation $\sigma = \mathcal{A}(\mu - \mu\sigma)$. Once σ is evaluated, the following results are immediately available.

- Utilization factor (proportion of time the server is busy) $\rho = \frac{\lambda}{\mu}$.

- Probability of n customers found in the system by an arrival $r_n = (1 - \sigma)\sigma^n$, $n = 0, 1, 2, \ldots$.
- Probability distribution function of the waiting time $W(y) = 1 - \sigma \exp^{-\mu(1-\sigma)y}$, $y \geq 0$.
- Average waiting time in queue of a customer $W = \frac{\sigma}{\mu(1-\sigma)}$.

Note that, the number found in the system by an arrival is geometrically distributed with parameter σ, independent of the form of the interarrival time distribution (except insofar as it affects the value for σ). We comment that, the steady-state probabilities p_n of n customers in the system at an arbitrary instant differs from r_n in that $p_0 = 1 - \rho$ whereas $r_0 = 1 - \sigma$ and $p_n = \rho(1 - \sigma)\sigma^{n-1} = \rho r_{n-1}$ for $n = 1, 2, \ldots$ (see [16]). In the $M/G/1$ system we saw that $p_n = r_n$. In the $M/M/1$ system $\sigma = \rho$ and the equations are the same as for the $M/M/1$ system. Note also, that the waiting times are exponentially distributed (with an accumulation point at the origin), independent of the form of the interarrival time distribution.

4 Correlated queue models

The focus here is on a family of queues where service and interarrival times exhibit some form of dependency. The initial motivation for such models was the modeling of a communication link in a packet-switched network carrying variable size packets. 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 [27] 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 what follows we describe two types of dependencies which arise in communication systems. In the first type, the service time B_n of packet n depends on the interarrival time I_n between packets $n - 1$ and n. The discussion is based mainly on [11]. In the second type, the interarrival time I_{n+1} between packets n and $n + 1$ depends on the service time B_n of packet n. The discussion is based mainly on [13].

4.1 Queues with service times proportional to interarrival times

In this part 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 C the amount of work received in a time interval τ cannot exceed $C \times \tau$. This strong positive correlation between interarrival and service times, can greatly improve the delay characteristics in the buffers preceding communication links as will be demonstrated later. It is, therefore, important to provide models that account for this effect while remaining tractable.

One of the earliest work to systematically investigate the issue of correlation between service and interarrival times is [38]. In [38], Kleinrock studied the impact of correlated packet 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* (see discussion in Section 2), 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 [6, 8, 9]), 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 studied in the context of fluid-flow models [1, 25, 35, 36, 41, 48, 58] 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

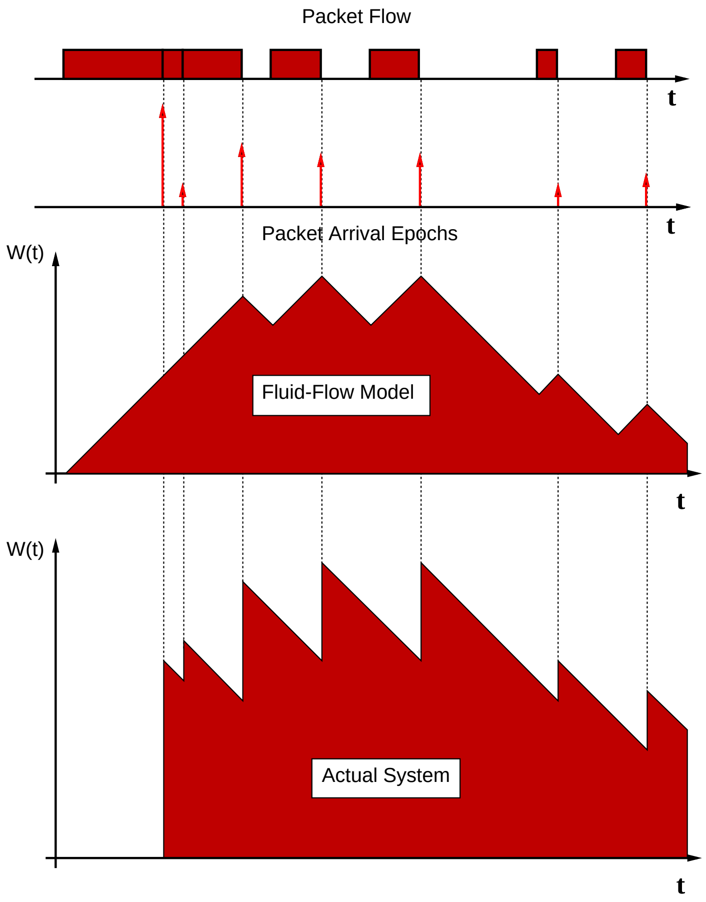

Figure 1: Comparison of ON-OFF Fluid-Flow and Discrete Arrivals Models

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 Figure 1 (where $W(t)$ stands for the amount of unfinished work in the system at time t), this can result in significant inaccuracies when estimating system performance (see also [52]). In [11], the authors propose and 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 [18, 19, 20, 21, 32, 33, 44, 45] that are directly relevant to these models.

One of the early works to consider a queueing system with explicit correlations between interarrival and service times is [18]. It analyzes a system with Poisson arrivals at rate

λ, where the service time B_n of the n-th customer is proportional to the interarrival time I_n between the $(n-1)$-st and the n-th 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 lengths from an upstream link of speed α. An explicit expression for the delay distribution is obtained in [18], while the initial busy period, the system state, and the output process are studied in [20]. Below we summarize the results for the correlated $M/M/1$ system.

- Utilization factor (proportion of time the server is busy) $\rho = \alpha$.
- The evolution equation of the system delay of a packet $t_{n+1} = (t_n - I_{n+1})^+ + \alpha I_{n+1}$, where $(x)^+ \stackrel{\Delta}{=} max(x, 0)$.
- The LST of the system delay of a packet $\mathcal{T}(s) = \prod_{i=1}^{\infty} \frac{1-\alpha^i}{1-(1-s/c)\alpha^i}$, where $c \stackrel{\Delta}{=} \lambda/(1-\alpha)$.
- The average of the system delay of a packet $T = \frac{1-\alpha}{\lambda} \sum_{i=1}^{\infty} \frac{\alpha^i}{1-\alpha^i}$.

Again, we see that a transmission line K times as fast will accommodate K times as many packets per second at K times smaller average delay per packet.

α	$E[t]$	$E[t^c]$	$var[t]$	$var[t^c]$
0.2	0.250	0.241	0.063	0.041
0.5	1.000	0.803	1.000	0.248
0.9	9.000	2.709	81.00	1.164
0.95	19.00	3.470	361.0	1.365

Table 1: Averages and variances of delay time of $M/M/1$ (t) and correlated $M/M/1$ (t^c) systems.

Numerical example: Table 1 compares between the average and the variance of the delay time of a packet in a $M/M/1$ and a correlated $M/M/1$ systems for average interarrival time $\lambda = 1$ and for different loads $\alpha = 0.2$, 0.5 , 0.9 , 0.95. Table 1 shows that the correlated system drastically reduces the average delay time as compared with the $M/M/1$ system.

More general correlations were considered in [21] using a bivariate exponential distribution to characterize the correlation between interarrival and service times. This work was

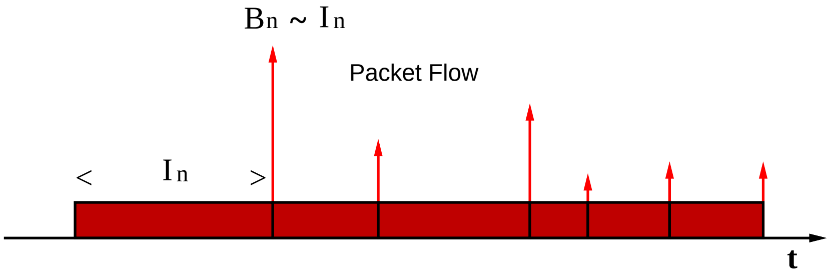

Figure 2: System with Proportional Interarrival and Service Times

subsequently extended in several papers. The delay density was shown to have a hyperexponential distribution in [32], while [33] studied the sensitivity of this distribution to the value of the correlation coefficient. The busy period was investigated in [45], while a system with infinitely many servers was considered in [44]. Borst *et al.,* [5], analyzed a variant of the $M/G/1$ queue in which the service times of arriving customers depend on the length of the interval between their arrival and the previous arrival. They obtained the LST of the delay time of a customer and proved that the average delay time of a customer is smaller or equal than the average delay time of a customer in the corresponding $M/G/1$ system without dependency.

Cidon *et al.*, [11], expanded the model of [18] in several new directions that makes it more applicable to the modeling of actual communication systems. A system similar to that of [18] (see Figure 2), was introduced and a simple derivation for the LST of the delay in the system was presented. This simple derivation was 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 [18, 20, 21, 32, 44, 45]. The LST of the delay was then obtained by applying results from the theory of linear functional equations [42] 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 to tackle more involved systems as illustrated in [11] and in the following.

The first extension considered in [11] consists of the addition of an independent, generally distributed, non-negative random variable to the service time. Using the notations introduced above, the service time of the n-th customer is now of the form $B_n = \alpha I_n + J_n$, where J_n is an independent, non-negative 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. It was observed in [11] that the average delay time of a packet depends on the entire probability distribution of the service time and not only on its first and second moments, as in the $M/G/1$ system.

The simple model of [18] 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 corresponds 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 [11] to 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 $g_2 = 1 - g_1$ (with $\alpha_1 g_1 < 1$ for a stable system). 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 Figure 3. It is also possible to add an independent and generally distributed "overhead" to each packet. An explicit expressions for the LST and the average of the delay time of a packet were obtained in [11].

Numerical example: Figure 4 shows 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. Figure 4 shows also the average delay of a packet in an equivalent $M/G/1$ system, in which the service time of a customer has the same distribution as in the correlated system, but is sampled independently of any other event in the system. Figure 4 shows that the average delay of the equivalent $M/G/1$ system is always larger than the average delay of the system with 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 the correlated system is 6.55 (13.82) while in the equivalent $M/G/1$ system it is 38.46 (109.9).

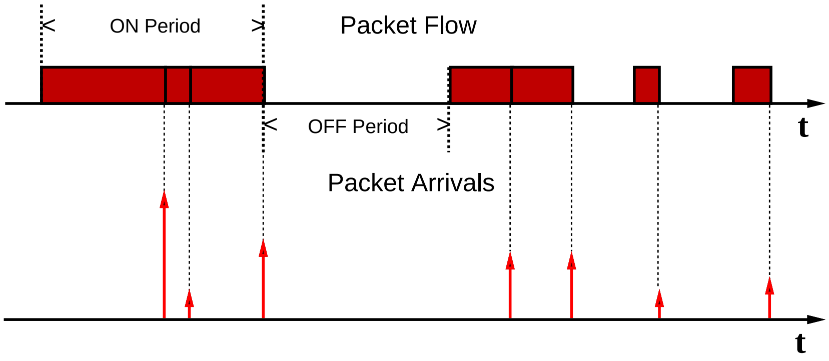

Figure 3: System with ON-OFF Source and Multiple Discrete Arrivals

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 uninterrupted over the duration of the entire ON period, and the transmission of bits rather than packets is considered. Despite its increased flexibility, this model 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 [30, 52], and it is of interest to develop models that allow burst duration and utilization to vary independently.

In order to overcome this limitation, a model where the arrival process corresponds to an *extended* ON-OFF process was presented and analyzed in [11]. This model allows 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

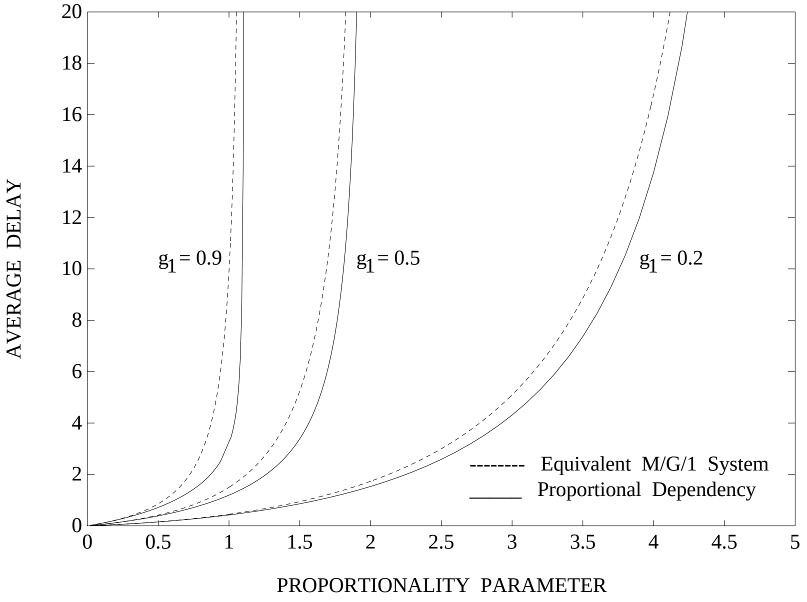

Figure 4: Average delay versus proportionality parameter α

with mean $1/\mu$, and the link returns to the ON state at the end of an OFF period. The stability condition for this system is $\alpha < 1 + p\lambda/\mu$. This allows 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 the first ON-OFF model 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 of the extended ON-OFF model is, therefore, quite general and provide 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 randomly distributed overhead to each packet.

Numerical example: An interesting numerical example which computes the average delay for the case $p = 1$, i.e., a single packet is generated at the end of the ON state, and compares it to the values obtained assuming equivalent $G/M/1$ and fluid-flow models was provided in [11]. The equivalent $G/M/1$ system has independent interarrival times with a

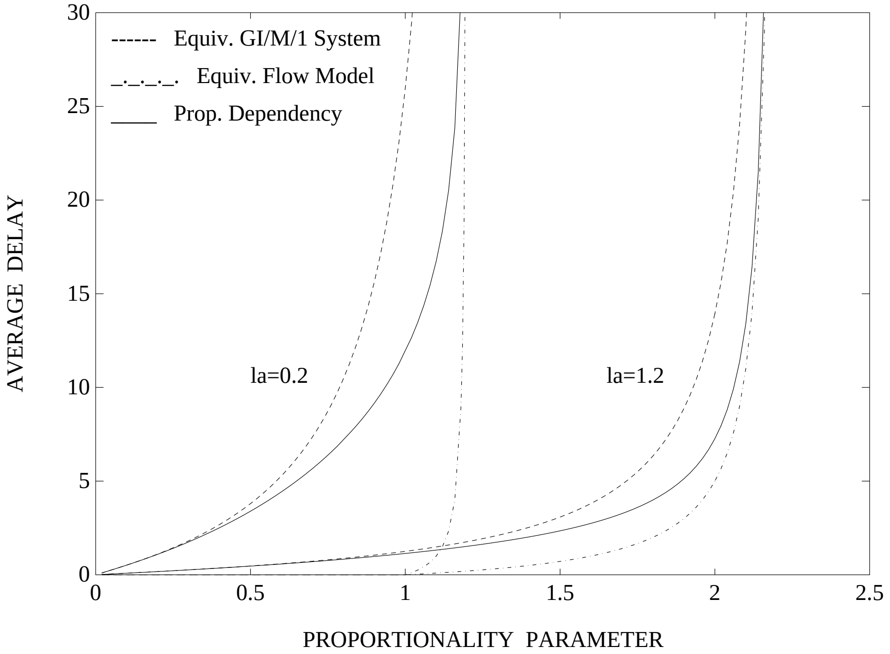

Figure 5: Average delay versus proportionality parameter α; $\mu = 1$.

probability distribution whose LST is $\mathcal{A}(s) = \frac{\mu\lambda}{(s+\mu)(s+\lambda)}$. 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 $\frac{\alpha-1}{\lambda-\mu(\alpha-1)}$, if $\alpha > 1$ and zero otherwise.

The average delays for all three models are plotted in Figure 5 as a function of the proportionality parameter α. Note that $\alpha < 1 + \lambda/\mu$ is the stability condition for all three models. Two cases were considered; 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 [11] in a sense "bridge the gaps" left by previous approaches. Specifically, while traditional point-process models such as the $G/M/1$ account for the granular nature of customer arrivals and departures, they typically ignore dependencies between interarrival and service times. As demonstrated in Figure 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 Figure 1 and illustrated in Figure 5, this can in turn yield an overly optimistic view of system behavior, in particular at light and medium loads. The models developed in [11], because they are able to retain both aspects, provide more accurate estimates of actual system performance for all load values.

4.2 Queues with interarrival times proportional to service times

In this part we consider queueing systems in which the interarrival time between two packets depends on the service time of the first packet. We focus on *proportional* dependency which is very natural in packet switching networks. Here the interarrival time between two consecutive packets arriving over a communication link is proportional to the size (in bits) of the first packet and consequently to the time it will take to forward this packet over the next link. The motivation for studying these queueing systems originated in the context of packet-switched networks, and especially in high-speed networks. The issue of interest was the effect of input policing functions that have been proposed to control the flow of packets in such networks. The basic goal of these policing functions is to ensure adequate network performance by regulating the amount of data that can arrive to a link within any given time interval. These controls result in significant dependencies between the amount of work brought in by packets and the time between the arrival of successive packets. Such dependencies have a significant impact on system performance as described in [13].

While numerous previous papers have studied the effect of many different dependencies in queueing systems, very little work seems to have been done on the type of dependencies described in this part. In particular, the case where the service time of a packet depends on the time since the *previous* arrival has been thoroughly studied [18, 19, 33, 11, 20, 21, 32, 44, 45]. The first paper that addresses the dual problem where the time to the *next* arrival depends on the service time of the arriving packet was [13]. Previous work on the dual problem has been essentially limited to the study of general conditions for either stability [47] or finite moments of the busy period [29].

Cidon *et al.,* [13], consider systems, where the interarrival time I_{n+1} between packets n and $n+1$ depends on the service time B_n of packet n. Specifically, they consider cases where the dependency between I_{n+1} and B_n is a *proportionality* relation, and B_n is an

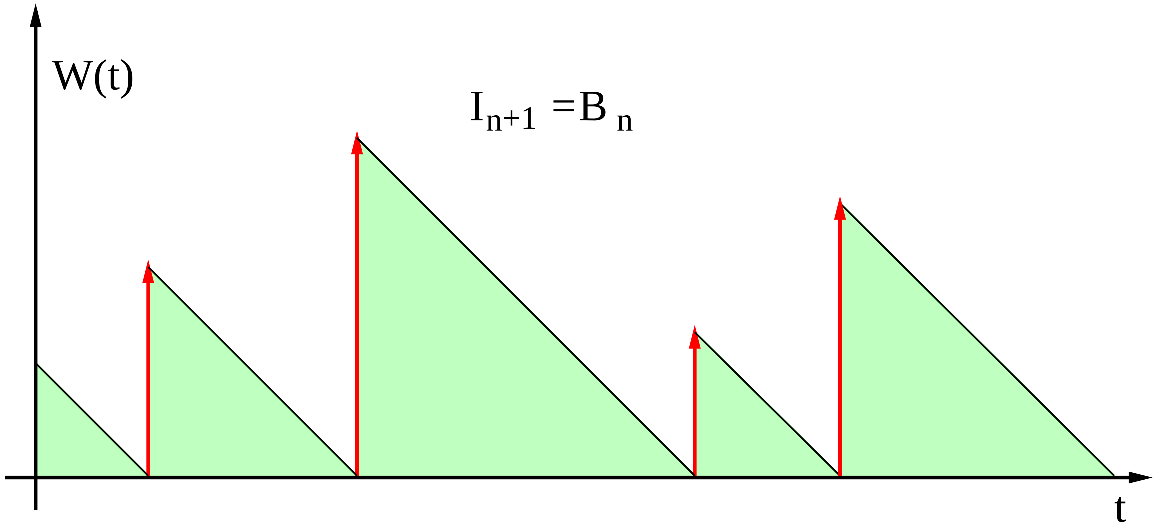

Figure 6: Workload in Queue with Interarrival Time Proportional to Service Time

exponentially distributed random variable. The proportional dependency was motivated by the modeling of some policing functions used in packet switched networks. For instance, let us illustrate how a simple *spacer* controller that is used to limit the peak rate at which a source can generate data into a network (see [3, 10, 26]), introduces such dependencies. The enforcement of a maximum rate is achieved by requiring that after sending a packet of size B, a *space* of duration B/R be inserted before the next packet can be sent. The rate R is then the maximum allowable rate for the source. (Note that the existence of a maximum network packet size is assumed here). This rate R is typically equal to the source peak rate, but can be set to a lower value when low speed links are present in the connection's path [3, 10] or if the network traffic has to be smoothed [31].

Assuming that the above spacer is saturated by a source of rate R bits per second whose traffic is fed to a link of speed C bits per second, interarrival and service times (in seconds) at the link are then proportional with $I_{n+1} = \alpha B_n$ and $\alpha = C/R$. As shown in Fig. 6, which plots the evolution of the workload at the network link for the extreme case $\alpha = 1$, i.e., equal source and link rates (stability requires $\alpha \geq 1$), the analysis of this simple case is of little interest. However, there exists a number of extensions to this basic model, that make it non-trivial although still tractable, and more important, useful in modeling actual communication systems. These extensions were presented and analyzed in [13] in which the LST, $\mathcal{W}(s)$, of the waiting time in steady-state was obtained using the spectral analysis method typical to G/G/1 queues [40]. In what follows we describe some of these extensions and their importance in the modeling of communication systems.

The simplest extension consists of adding an independent random variable to the interarrival time. The presence of such a random component in the interarrival time, allows to model more accurately how the traffic generated by a spacer controller arrives at an internal network link. First, such a model can capture the effect of interactions between packets from a given source and other traffic streams inside the network. In particular, the gaps that the spacer initially imposes between packets, are modified according to the different delays that consecutive packets observe through the network. The arrival process at a link can then be modeled as consisting of a deterministic component (the spacing imposed by the spacer at the network access), to which a random network jitter has been added. Second, the addition of a random component also allows to relax the assumption of a saturated spacer queue since it can be used to model the time between packets that arrive to the spacer. Finally, another useful application is when the spacer itself randomizes the gaps between successive packets. This randomization in the spacer may be useful to avoid correlation between traffic streams of distinct sources. In particular, it helps to prevent (malicious) sources from harming network performance by cooperating to generate a large burst of data into the network. Here we summarize the results of this model for a system with positive jitter J_n, exponentially distributed with parameter δ.

- Utilization factor (proportion of time the server is busy) $\rho = \frac{\delta}{\alpha\delta+\mu}$, where B_n is exponentially distributed with parameter μ.
- The evolution equation of the waiting time of a packet $w_{n+1} = (w_n + (1-\alpha)B_n - J_n)^+$.
- The LST of the waiting time of a packet $\mathcal{W}(s) = 1 - \gamma\delta + \gamma\delta\frac{\frac{1-\gamma\delta}{\gamma}}{s+\frac{1-\gamma\delta}{\gamma}}$, where $\gamma \triangleq (1-\alpha)/\mu$.

The waiting time in the system in steady-state at the arrival instants thus behaves as the waiting time in $M/M/1$ queue with arrival rate δ and service rate $1/\gamma = \frac{\mu}{1-\alpha}$. However, the arrival rate to the system is actually $\lambda = \frac{\mu\delta}{\mu+\alpha\delta}$ and the service rate is clearly μ. Comparing this system with a $M/M/1$ system with parameters λ,μ, observe that both systems have the same stability condition. However, in the $M/M/1$ system the term that governs the exponent (with a minus sign) of the pdf is $\mu\left(1 - \frac{\delta}{\mu+\alpha\delta}\right)$. In this system the term is

$\mu\left(\frac{1}{1-\alpha} - \frac{\delta}{\mu}\right)$ which is larger when the stability condition ($\rho < 1$) holds. This implies that the tail probability decreases much faster in this system when compared to the corresponding $M/M/1$ system. An extension to this system where J_n can also take negative values (but the interarrival period is of-course kept positive) can be found in [13].

Another extension of the basic model is to allow the proportionality constant to be itself a random variable. Specifically, cases where the proportionality factor is randomly chosen from a finite set of values were considered in [13]. This allows the modeling of a generalized spacer, where the factor used to compute the enforced spacing is allowed to vary. For example, the arrival of a high priority packet that is sensitive to access delay could be handled by allowing earlier transmission. The interarrival times at the network link would then depend on both packet priorities and the size of the previous packet. The analysis of this system is more involved than the previous one, see [13].

Numerical example: The following numerical example was provided in [13]. Consider a system with $I_{n+1} = \Omega_n B_n$, where Ω_n is a random variable with a finite support, independent of any other random variable in the system. Specifically, consider the case where $\Omega_n = \alpha_i$ with probability a_i for $1 \leq i \leq N + M$ for some integers $N, M \geq 1$. Clearly, $\sum_{i=1}^{N+M} a_i = 1$. Consider the case $1 < \alpha_1 < \alpha_2 < \cdots < \alpha_N$ and $\alpha_{N+1} < \alpha_{N+2} < \cdots < \alpha_{N+M} \leq 1$. The stability condition for the system is $\sum_{i=1}^{N+M} a_i \alpha_i > 1$.

Figures 7-8 show the average and the variance of the waiting time of a system with random proportional dependency with $N = 3$, $M = 2$ and $a_i = 0.2$, $1 \leq i \leq 5$. It also shows the same quantities in an equivalent $G/M/1$ system in which the service time is exponentially distributed with parameter μ, and the interarrival times are independent and sampled from a probability distribution whose LST is $\mathcal{A}(s) = \sum_{i=1}^{N+M} a_i \frac{\mu/\alpha_i}{\mu/\alpha_i + s}$. Namely, with probability a_i, the interarrival time is exponentially distributed with parameter μ/α_i. The sum $\sum_{i=1}^{N+M} a_i \alpha_i$ is kept constant. In particular, $\alpha_1 = 1.2$, $\alpha_2 = 1.3$, $\alpha_4 = 0.1$ and $\alpha_3 + \alpha_5$ is kept constant. The average and the variance of the waiting time are depicted as a function of the largest proportional parameter α_3.

Figures 7-8 show that the equivalent $G/M/1$ system exhibits much larger averages and variances of the waiting time. This implies that correlations between service times and interarrival times have a smoothing effect on the system. This has also been observed in [11, 18, 19, 20, 21] for different types of correlations. It is interesting to note that although

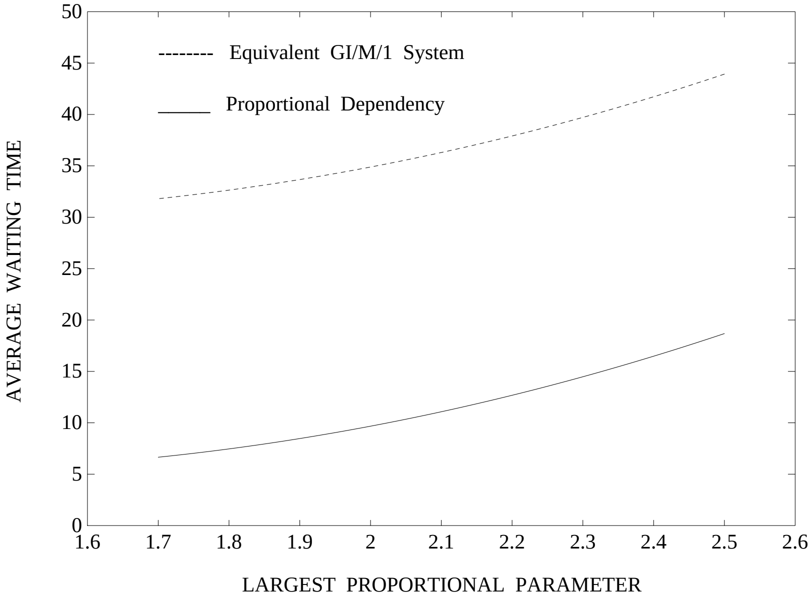

Figure 7: Average waiting time versus largest proportional parameter: $a_i = 0.2$, $1 \leq i \leq 5$; $\alpha_1 = 1.2$, $\alpha_2 = 1.3$, $\alpha_4 = 0.1$ $\alpha_3 + \alpha_5 = 2.6$.

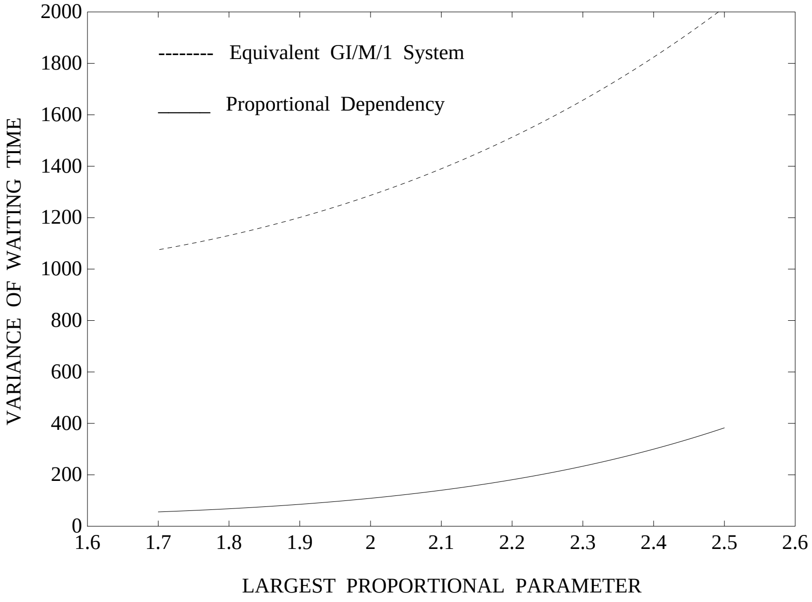

Figure 8: Variance of waiting time versus largest proportional parameter: $a_i = 0.2$, $1 \leq i \leq 5$; $\alpha_1 = 1.2$, $\alpha_2 = 1.3$, $\alpha_4 = 0.1$ $\alpha_3 + \alpha_5 = 2.6$.

$\alpha_3 + \alpha_5$ is kept constant, both the average and the variance increase with increasing α_3 (and decreasing α_5). This implies that decreasing α_5 has a more pronounced effect on the performance of the queueing system. The reason is that increasing α_3 and decreasing α_5 while keeping their sum constant, increases the variability of the arrival process, and hence increases the average and the variance of the waiting time as shown in Figures 7-8.

5 Burstiness characterization Models

In this section we will describe completely different approaches for characterization of arrival processes. These novel characterizations are motivated by the need to model real-time applications in high-speed communication networks where it is important to capture the *burstiness* of the arrival processes.

The first approach we describe is deterministic and has been introduced by Cruz ([22, 23, 24]) by characterization of the concept of *a burstiness constraint.* According to Cruz a traffic stream with rate $R(t)$ at time t is said to satisfy such a constraint if there exist some constants ρ and σ such that

$$\int_s^t R(u)du \leq \rho(t-s) + \sigma$$

holds for all $0 \leq s < t$. In that case $R(t)$ is said to be a (σ, ρ) process and the notation used is $R(t) \sim (\sigma, \rho)$. Note that this characterization ignores any stochastic nature of the traffic stream, and must hold for any sample path of it. The analysis Cruz presents achieves two major results regarding the performance of a single-server system. The first says that if all the input traffics to a single-server system satisfy burstiness constraints, then so do the output traffics from that system (not necessarily with the same parameters, though). In other words, there exist constants ρ' and σ' such that the output process $R^{out}(t)$ satisfies $R^{out}(t) \sim (\sigma', \rho')$. This claim is proved by Cruz for a variety of systems, with various service disciplines. For instance, for a work-conserving multiplexer with a general service discipline that is fed by two streams, $R_1(t) \sim (\sigma_1, \rho_1)$ and $R_2(t) \sim (\sigma_2, \rho_2)$, the output stream satisfies

$$R^{out}(t) \sim (\frac{C(\sigma_1 + \sigma_2)}{C - \rho_1 - \rho_2}, \rho_1 + \rho_2)$$

where C ($C > \rho_1 + \rho_2$) is the service rate of the multiplexer. The second result says that in the above case the delay suffered by each bit that enters the system is upper bounded by a constant D, which depends, of course, on the parameters of the entering traffic streams and the nature of the examined system. For instance, for the above multiplexer the constant D is give by

$$D = \frac{\sigma_1 + \sigma_2}{C - \rho_1 - \rho_2}$$

It is clear that the main advantages of Cruz characterization are that the output process of various systems has the same characterization as the input processes (with different parameters), and that it allows to derive deterministic bounds on queue lengths and delays in the system. Further progress within this characterization has been presented by Parekh and Gallager in [49, 50, 51], for a special service discipline — the Packet-based Generalized Processing Sharing (PGPS).

The second approach we describe is stochastic in nature and has been introduced by Kurose ([43]). It attempts to overcome the deterministic nature of Cruz characterization. Suppose that for every $\tau > 0$ there exists a random variable $X(\tau)$ such that for all $0 \leq s < t$ with $t - s = \tau$, the arrival stream $R(t)$ satisfies

$$\Pr\left\{\int_s^t R(u)du \geq x\right\} \leq \Pr\{X(\tau) \geq x\}$$

for all x. This means that $X(\tau)$ is a stochastic bound on $\int_s^t R(u)du$. We define $|X(\tau)| = sup\{x : \ \Pr\{X(\tau) \geq x\} > 0\}$. It is not difficult to see that $|X(\tau)|$ is a generalization of the term $\rho(t-s) + \sigma$ in Cruz characterization. Kurose introduced the characterization of a traffic stream in discrete-time by stochastically bounding the amount of data it might carry in any fixed length interval of time. In other words, a traffic stream with rate $R(t)$ $(t = 0, 1, 2, \ldots)$ is characterized by a series $\{(R_k, k);\ k \in \mathbb{N}\}$ of random variables if

$$\Pr\left\{\int_s^t R(u)du \geq x\right\} \leq \Pr\{R_{t-s} \geq x\}$$

holds for all $0 \leq s < t$ (s and t are non-negative integers) and all $x \geq 0$. This characterization is denoted by $R(t) \sim \{(R_k, k); k \in \mathbb{N}\}$. Kurose was also able to show that the output process of various systems has the same characterization as the input processes (with different parameters). For instance, for a switch with two input stream $R_1(t) \sim \{(R^1_{n_1}, 1), (R^1_{n_1+1}, 2), \ldots\}$ and $R_2(t) \sim \{(R^2_{n_2}, 1), (R^2_{n_2+1}, 2), \ldots\}$, the output process of stream i $(i = 1, 2)$ satisfies $R_i^{out}(t) \sim \left\{(R^i_{n_i+l}, 1), (R^i_{n_i+l+1}, 2), \ldots\right\}$ where l is the least integer that satisfies $|R^1_{n_1+l}| + |R^2_{n_2+l}| \leq l + 1$. In addition, the delays suffered by each packet in the switch are upper bounded by some constant D. Note that although this approach is stochastic in nature, the bound on l is a deterministic bound. In that sense, both approaches described so far require the intervals over which the sum of the input peak rates to a system exceeds its output capacity to be bounded. This is not the case for most commonly used for modeling input streams to a communication system, such as the simple Bernoulli process or the well known Poisson process (which might have bursts of any length).

The third approach we describe is also stochastic and has been introduced by Yaron and Sidi [61]. It attempts to overcome the limitations of the first two approaches. With this approach, rather than assuming that a traffic stream has a bounded burstiness, it is assumed that the distribution of its burst length has an exponential decay. Processes with such a characterization are said to have *Exponentially Bounded Burstiness (E.B.B.).* If $R(t)$ is the traffic rate of a stream, then the stream has Exponentially Bounded Burstiness (E.B.B.) if there exist constants ρ, A, α such that

$$\Pr\left\{\int_s^t R(u)du \geq \rho(t-s)+\sigma\right\} \leq Ae^{-\alpha\sigma}$$

for all $\sigma \geq 0$ and all $0 \leq s < t$. In that case we use the notation $R(t) \sim (\rho, A, \alpha)$. The basic advantage of this bound over the ones mentioned earlier is that it holds for natural processes such as Bernoulli, Poisson and other processes. It is evident that it holds for most of the processes one might encounter in modeling communication networks. Two additional advantages make it applicable in many settings were the aforementioned characterizations cannot be used. First, it imposes no deterministic requirements on the bounded processes, and hence allows the analysis of systems were the peak rates of the input streams might exceed the system capacity over arbitrarily long intervals of time. Second, it allows the analysis of compound systems with correlated input processes — no independence assumption is needed when applying the analysis.

As with the two previous characterizations, Yaron and Sidi were able to show that for various systems that are fed with E.B.B. processes, the output streams are also bounded similarly. Furthermore, the delays these systems cause, and the length of the queues which are built up within them, all have exponentially decaying distributions. For instance, consider a work-conserving multiplexer that is fed by two streams $R_1(t) \sim (\rho_1, A_1, \alpha_1)$ and $R_2(t) \sim (\rho_2, A_2, \alpha_2)$. If the service rate of the multiplexer is C and $\rho_1 + \rho_2 < C$, Yaron and Sidi show that the output process $R(t)$ has exponentially bounded burstiness. In particular, $R(t) \sim (\rho', A', \alpha')$ with $\rho' = \rho_1 + \rho_2$, $A' = (A_1 + A_2)/(1 - e^{-\alpha'(C-\rho_1-\rho_2)})$ and $\alpha' = 1/(\alpha_1^{-1} + \alpha_2^{-1})$. Furthermore, the backlog $W(t)$ at the multiplexer satisfies $\Pr\{W(t) \geq \sigma\} \leq A'e^{-\alpha'\sigma}$ for all σ.

Numerical example: Consider a two input multiplexer with output capacity $C = 1$, and with bounded input capacities $C_1 = C_2 = 1$, and assume it is fed with two Bernoulli input processes, each with parameter $P = \frac{1}{8}$. An appropriate E.B.B. characterization of the input processes can be derived, which allows a tradeoff between the upper rate ρ and the decay factor α as illustrated in Table 2. Notice that the larger the gap one allows between the true mean rate P and the upper rate ρ of the characterization, the better the decay factor one gets.

ρ	A	α
0.15	1	0.41
0.2	1	1.06
0.3	1	2.06

Table 2: E.B.B. characterizations of a $P = \frac{1}{8}$ Bernoulli source.

Using these bounds for the input processes, one can compute exponential bounds for the queue length in the multiplexer. Figure 9 presents these bounds for the case of independent input processes, with some actual simulation results distributions. The dashed lines show

Figure 9: Queue length distributions and its E.B. bounds: independent sources.

the computed bounds for two different characterizations of the input processes (with $\rho = 0.2$ and with $\rho = 0.3$). The solid lines show the actual distributions at five different time stops ($t = 10, 50, 100, 500, 1000$) found by simulating the behavior of a multiplexer which has an empty queue at $t = 0$ for $100,000$ times. Considering these actual distributions, it is apparent that the proposed analysis captures the transient, as well as the steady-state, behavior of the multiplexer queue. Similar analysis can be carried out for dependent sources (see [61]).

6 Message delay processes

As mentioned in Section 2, in many systems the message delay, and not the packet delay, is the measure of interest for the network designer. Here, we describe some results for the message queueing delays in a node of a communication system. A message consists of a block of consecutive packets 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 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 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. Due to the finite speed of the communication links and the multiplexing of packets from different sessions, the dispersed generation model is more adequate for communication networks.

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 packets. 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., [4]). 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.

Traditionally, the analysis of the message delay was associated with batch arrival processes (see, e.g., [39, 34, 53]), 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. Kleinrock [39] analyzed Bulk (Batch) arrival queueing system, where a single server queueing system with bulk Poisson arrivals (i.e., the arrival instants are Poisson and at each arrival instant a bulk of arbitrary size sampled from an independent integer random variable is brought to the system and it corresponds to a message) and exponentially distributed service times was considered. Halfin [34] analyzed a discrete-time queueing system with batch arrivals, where batch sizes have geometric distribution and the queue discipline is indifferent to batch sizes and service times. He proved that the packet delay distribution is the same as the batch (message) delay distribution, were delay is defined to be the delay of the last served packet in the batch. The proof was based on a discrete-time analog of the PASTA theorem. The message delay distribution for TDMA systems with a generalized arrival process was presented in [53]. The analysis was based on a generating functions approach.

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 [57] for Poisson Cluster Processes (PCP). Here, messages arrive to the system according to 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 [57], the average delay of packets was approximated for a $PCP/D/1$ system.

In [12] a new technique to analyze the message delay in systems with dispersed arrivals was introduced. It has been shown that the correlation between the delays of packets which belong to the same message has a strong effect on the performance of the system. For example, it was shown that evaluating the timeout for message retransmissions under the assumption that the packet delays are independent is quite pessimistic. The model used in [12] 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. Every n consecutive packets are grouped into a message. An exact analysis of the message delay was presented. In particular, an efficient recursive procedure to obtain the LST of the message delay for different arrival models and different number of sessions was presented. It was shown that, the assumption that the delays of packets are independent from packet to packet can lead to wrong conclusions. This fact was demonstrated by comparing the exact variance of the message delay with the variance of the message delay as obtained from the above independence assumption. Numerical examples were provided to show that the variance of the message delay may be over-estimated by the above independence assumption for a wide range of message sizes.

It was observed in [12] 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. The second is the time delay of an arbitrary packet (stands for the last packet of that message) in the system. These two components are of course dependent random variables. However, the average message delay can be obtained directly from the sum of the averages of these two components. Cidon *et al.,* [12] suggests a simple way to approximate the message delay by assuming that these two components are independent random variables. The numerical results in [12] demonstrated the relative error of such an approximation and the so called "negative feedback" effect that governs the message delay process. If the message's packet arrival 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 above independence assumption of the message arrival time and the last packet delay time.

Numerical example: The following numerical example was provided in [12]. Consider a $M/M/1$ system with arrival rate λ, service rate μ and messages of fixed size n. The average message delay equals $\frac{n-1}{\lambda} + \frac{1}{\mu-\lambda}$. The variance of the message delay using the above approximation equals $\frac{n-1}{\lambda^2} + \frac{1}{(\mu-\lambda)^2}$. The relative variance error of the message delay, defined as $100 * [(approximated\ variance)/(exact\ variance) - 1]$ is plotted in Figure 10 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.

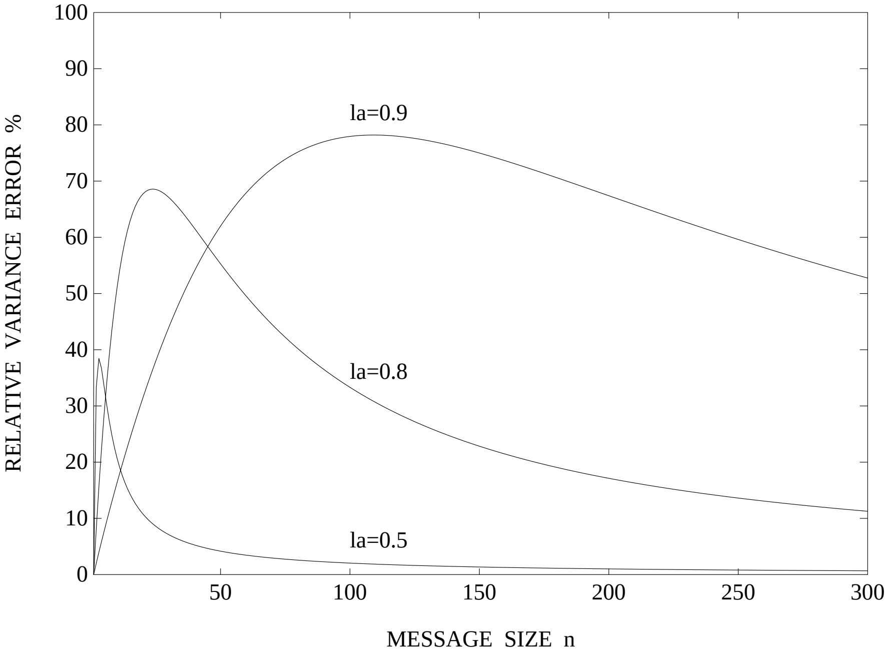

Figure 10: The relative variance error of the message delay versus the message size n.

Another two important quantities, namely, the maximum delay of a packet in a message and the number of packets in a message whose delays exceed a pre-specified time threshold were analyzed in [14] for the dispersed generation model. These quantities are important for the proper design of playback algorithms [15] and time-out mechanisms for retransmissions. In [14], a new analytical approach that yields efficient recursions for the computation of the probability distribution of each quantity was presented. It has been shown that the correlation between packet delays of the same message has a strong effect on each of these quantities.

References

[1] D. Anick, D. Mitra and M. M. Sondhi, "Stochastic Theory of a Data-Handling System with Multiple Sources", *Bell Sys. Tech. J. (B.S.T.J.)*, 61(8):1871–1894, October 1982.

[2] F. Baccelli and P. Bremaud, *Elements of Queueing Theory*, Springer Verlag, 1994.

[3] K. Bala, I. Cidon and K. Sohraby, "Congestion Control for High-Speed Packet Switch", *In Proc. INFOCOM'90*, San Francisco, 1990.

[4] D. Bertsekas and R. Gallager, *Data Networks,* Prentice-Hall International Editions, 1987.

[5] M.B. Combe, S. C. Borst and O. J. Boxma, "Collection of Customers: A Correlated M/G/1 Queue", *Performance Evaluation Rev.*, Vol. 20, pp. 47-59, 1991.

[6] O. J. Boxma, "On a Tandem Queueing Model with Identical Service Times at Both Counters, Parts I and II", *Adv. Appl. Prob.*, 11:616–659, 1979.

[7] P. J. Burk, "The Output of a Queueing System", *Journal of Operations Research,* 4, 699-704, 1966.

[8] S. B. Calo, "Delay Properties of Message Channels", *ICC,* pp. 43.5.1–43.5.4, 1979.

[9] S. B. Calo, "Message Delays in Repeated-Service Tandem Connections", *IEEE Trans. Commun.*, COM–29(5):670–678, May 1981.

[10] I. Cidon, I. Gopal and R. Guérin, "Bandwidth Management and Congestion Control in PlaNET", *IEEE Commun. Mag.*, 29(10):54–63, October 1991.

[11] I. Cidon, R. Guérin, A. Khamisy and M. Sidi, "Analysis of a Correlated Queue in Communication Systems", *IEEE Transactions on Information Theory,* Vol. 39, No. 2, pp. 456–465, March 1993.

[12] I. Cidon, A. Khamisy and M. Sidi, "On Queueing Delays of Dispersed Messages", To appear in *Queueing Systems - Theory and Applications,* 1994. (see also *In Proc. of INFOCOM'93,* pp. 843–849, April 1993).

[13] I. Cidon, R. Guerin, A. Khamisy and M. Sidi, "On Queues with Inter-Arrival Times Proportional to Service Times", *INFOCOM'93,* pp. 237–245, April 1993.

[14] I. Cidon, A. Khamisy and M. Sidi, "Dispersed Messages in Discrete-time Queues: Delay, Jitter and Threshold Crossing", To appear in *In Proc. of INFOCOM'94,* June 1994. Also in Technion EE PUB No. 894, August 1993.

[15] J-Y Cochennec et al., "Asynchronous Time-division Networks: Terminal Tynchronization for Video and Sound Signals", *In Proceedings of GLOBECOM'85*, pp. 791–794, December 1985.

[16] J. W. Cohen, *The Single Server Queue*, North-Holland, 1969.

[17] D. Comer, *Internetworking with TCP/IP, Principles, Protocols, and Architectures*, Prentice-Hall, Englewood Cliffs, 1988.

[18] B. W. Conolly, "The Waiting Time Process for a Certain Correlated Queue", *Operations Research*, 16:1006–1015, 1968.

[19] B. W. Conolly and N. Hadidi, "A Comparison of the Operational Features of Conventional Queues with a Self-Regulating System", *Appl. Stat.*, 18:41–53, 1969.

[20] B. W. Conolly and N. Hadidi, "A Correlated Queue", *J. Appl. Prob.*, 6:122–136, 1969.

[21] 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.*, 37(2):263–275, October 1979.

[22] R. L. Cruz, "A Calculus for Network Delay and a note on Topologies of Interconnection Networks", Ph.D. thesis, University of Illinois, Urbana-Champaign, 1987.

[23] R. L. Cruz, "A Calculus for Network Delay, Part I: Network Elements in Isolation", *IEEE Transactions on Information Theory*, Vol. 37, No. 1, pp. 114–131, January 1991.

[24] R. L. Cruz, "A Calculus for Network Delay, Part II: Network Analysis", *IEEE Transactions on Information Theory*, Vol. 37, No. 1, pp. 132–141, January 1991.

[25] A. I. Elwalid, D. Mitra and T. E. Stern, "Statistical Multiplexing of Markov Modulated Sources: Theory and Computational Algorithms", *In Proceedings of the 13th International Teletraffic Congress (ITC-13)*, pp. 495–500, Copenhagen, 1991.

[26] A. I. Elwalid and D. Mitra, "Analysis and Design of Rate-Based Congestion Control of High-Speed Networks, I: Stochastic Fluid Models, Access Regulation", *QUESTA*, 9:29–64, September 1991.

[27] K. W. Fendick, V. R. Saksena and W. Whitt, "Dependence in Packet Queues", *IEEE Trans. Commun.*, COM-37(11):1173–1183, November 1989.

[28] D. P. Gaver Jr., R. G. Miller Jr., "Limiting Distributions for some Storage Problems," *Studies in Applied Probability and Management Science*, Editors K. J. Arrow, S. Karlin and H. Scarf, Stanford University Press, 1962, pp. 110–126.

[29] S. Ghahramani and R. W. Wolff, "A New Proof of Finite Moment Conditions for GI/G/1 Busy Periods", *Queueing Systems Theory Appl. (QUESTA)*, 4(2):171–178, June 1989.

[30] R. Guérin, H. Ahmadi and M. Naghshineh, "Equivalent Capacity and its Application to Bandwidth Allocation in High-Speed Networks", *IEEE J. Select. Areas Commun.*, SAC-9(7):968–981, September 1991.

[31] R. Guérin and L. Gün, "A Unified Approach to Bandwidth Allocation and Access Control in Fast Packet-Switched Networks", *In Proc. INFOCOM'92*, pp. 1–12, Florence, Italy, 1992.

[32] N. Hadidi, "Queues with Partial Correlation", *SIAM J. Appl. Math.*, 40(3):467–475, June 1981.

[33] N. Hadidi, "Further Results on Queues with Partial Correlation", *Operations Research*, 33:203–209, 1985.

[34] S. Halfin, "Batch Delays Versus Customer Delays", *The Bell System Technical Journal,* Vol. 62, No. 7, September 1983.

[35] O. Hashida and M. Fujiki, "Queueing Models for Buffer Memory in Store-and-Forward Systems", *In Proceedings of the 7th International Teletraffic Congress (ITC 7)*, pages 323/1–323/7, Stockholm, 1973.

[36] H. Kaspi and M. Rubinovitch, "The Stochastic Behavior of a Buffer with Non-Identical Input Lines", *Stochastic Processes and their Applications*, 3:73–88, 1975.

[37] F. P. Kelly, *Reversibility and Stochastic Networks*, John Wiley & Sons, 1979.

[38] L. Kleinrock, *Communication Nets*, McGraw-Hill, New York, 1964 (Also Dover, 1972).

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

[40] L. Kleinrock, *Queueing Systems, Volume 2: Computer Applications*, John Wiley & Sons, New York, 1976.

[41] L. Kosten, "Liquid Models for a Type of Information Storage Problems", *Delft Progress Report: Mathematical Engineering, Mathematics and Information Engineering*, 11:71–86, 1986.

[42] M. Kuczma, B. Choczewski and R. Ger, *Iterative Functional Equations*, Encyclopedia of Mathematics and its Applications, Cambridge University Press, Cambridge, 1990.

[43] J. Kurose, "On Computing Per-Session Performance Bounds in High-Speed Multi-Hop Computer Networks", *Performance '92*, Newport, June 1992.

[44] C. Langaris, "A Correlated Queue with Infinitely Many Servers", *J. Appl. Prob.*, 23:155–165, 1986.

[45] C. Langaris, "Busy-Period Analysis of a Correlated Queue with Exponential Demand and Service", *J. Appl. Prob.*, 24:476–485, 1987.

[46] J. Little, "A Proof of the Queueing Formula $L = W$", *Journal of Operations Research,* 18:172-174, 1961.

[47] R. M. Loynes, "The Stability of a Queue with Non-Independent Inter-Arrival and Service Times", *Proc. Cambridge Philos. Soc.*, 58(3):497–520, 1962.

[48] D. Mitra, "Stochastic Theory of a Fluid Model of Producers and Consumers Coupled by a Buffer", *Adv. Appl. Prob.*, 20:646–676, September 1988.

[49] A. K. Parekh, "A Generalized Processor Sharing approach to flow control in Integrated Services Networks", Ph. D. thesis, Department of Electrical Engineering and Computer Science, MIT, 1992.

[50] A. K. Parekh, R. G. Gallager, "A Generalized Processor Sharing approach to flow control — The Single Node Case", *IEEE/ACM Transactions on Networking,* Vol. 1, pp. 344–357, June 1993.

[51] A. K. Parekh, R. G. Gallager, "A Generalized Processor Sharing approach to flow control — The Multiple Node Case", *MIT Laboratory for Information and Decision Systems Technical Report 2076*, 1991. Also in*Proc. IEEE INFOCOM'93*, 1993, pp. 521–530.

[52] J. W. Roberts, "Variable-Bit-Rate Traffic Control in B-ISDN", *IEEE Commun. Mag.*, 29(9):50–56, September 1991.

[53] R. Rom and M. Sidi, *Multiple Access Protocols; Performance and Analysis*, Springer-Verlag, 1990.

[54] Ronald W. Wolff, "Poisson Arrivals See Time Averages", *Operations Research*, Vol. 30, No. 2, March-April 1982.

[55] S. M. Ross, *Introduction to Probability Models*, new York, 1980.

[56] M. Rubinovitch, "The Output Of a Buffered Data Communication System," *Stochastic Processes and their Applications*, 1 1973, pp. 375–382.

[57] K. Sohraby, "Delay Analysis of a Single Server Queue with Poisson Cluster Arrival Process Arising in ATM Networks", *GLOBECOM'89*, pp. 611–616, 1989.

[58] T. E. Stern and A. I. Elwalid, "Analysis of Separable Markov-Modulated Models for Information-Handling Systems", *Adv. Appl. Prob.*, pp. 105–139, March 1991.

[59] A. S. Tanenbaum, *Computer Networks,* Prentice-Hall International Editions, 1989.

[60] UNI Specification Draft 2.4, Technical Report, ATM Forum: Technical Committee, August 1993.

[61] O. Yaron and M. Sidi, "Calculating Performance Bounds in Communication Networks", *IEEE/ACM Transactions on Networking,* Vol. 1, pp. 372–385, June 1993.