Distribution of the Loss Period for Some Queues in Continuous and Discrete Time

Henning Schulzrinne
Dept. of Electrical and Computer Engineering
University of Massachusetts
Amherst, MA 01003

James F. Kurose
Dept. of Computer and Information Science
University of Massachusetts
Amherst, MA 01003

Abstract

For soft real-time communication systems, packet loss due to excessive delay rather than average delay becomes the critical performance issue. While most previous studies of real-time systems measure loss as a time-average fraction of excessively delayed packets, this paper characterizes the stochastic properties of time-out loss periods for infinite queues, that is, uninterrupted intervals during which the virtual wait is at or above some fixed threshold. We present analytic expressions and numerical techniques for computing both "time-based" measures such as the distribution of periods during which all arriving packets are lost due to excessive delay as well as "packet-based" measures such as the distribution of the number of consecutively lost packets and the number of successful packets between such periods of loss. Both continuous and discrete-time systems are examined. It is shown that the assumption of random packet loss severely underestimates the number of consecutively lost packets. Also, the loss period is found to be independent of the waiting time threshold for the G/M/1 queue and $D^{[\mathrm{Geo}]}/D/1$ queue, with very little influence for other queueing models.

1 Introduction

For soft real-time communication systems, packet loss rates rather than average delay becomes the critical performance measure. Interactive voice, video and distributed measurement and control are examples of such systems that impose delay constraints, but also show a certain tolerance for lost packets. Most previous studies of real-time systems only measure loss as a time-average fraction of missing packets (e.g., [1, 2]). In order to judge the "quality" of loss-prone communication, however, it is important to know not just how *many* packets are being lost, but also whether losses occur in clusters or randomly. Papers on packet voice reconstruction typically assume random, uncorrelated occurrence of packet loss [3, 4]; as we will show, this assumption might be overly optimistic. The work by Shacham and Kenney [5] shows that loss correlation in a finite-buffer system could completely eliminate the advantage of three orders of magnitude predicted for forward-error correction under the assumption of independent (Bernoulli) losses.

In this paper, we consider a single FCFS queue with infinite buffer in isolation in which an arriving customer[1] which reaches the server after waiting h or longer is considered "lost" (even though it still receives service). We characterize the stochastic properties of the *time-out loss period*, the uninterrupted interval of time during which the virtual work in the queue exceeds the given threshold h. Given this time-based quantity, we also arrive at measures of customer-based quantities such as the distribution of the number of consecutive customers which each spends more than an allotted time waiting for service. We show that the assumption that each customer is lost independently from the previous one leads to a significant underestimation of the duration of such loss periods. Using elementary methods of probability, we also prove that for certain classes of queues the duration of the loss period is independent of the value of the threshold, while we show through numerical examples that for other important classes the influence of the threshold is minimal for interesting probabilities of loss. Our numerical results also indicate that the expected number of consecutively lost customers (for the same loss probability) varies by as much as 50% depending on the batch arrival distribution used. We also derive measures of the time between loss periods. Throughout the paper, emphasis is placed on providing results that can be readily numerically evaluated.

A number of authors have investigated related aspects of the problem. Kamitake and Suda [6] consider a discrete-time queue in which traffic is generated by a changing number of active callers, with each active caller generating a packet in a slot according to a Bernoulli process. They compute the steady state loss rate for a given number of active callers and then consider the variation in the number of active callers in computing the amount of uninterrupted time from when this loss rate first exceeds a value ζ until it drops below another value η, with $\eta < \zeta$. Our work differs from [6] in that we directly characterize those periods of time in which arriving customers are lost, rather than characterizing loss as being "quasi-stationary" during periods of times during which the number of active sources remains constant.

Leland [7] mentions, but does not elaborate on measuring consecutive losses per connection in an ATM simulation experiment. Woodruff and Kositpaiboon [8] mention the desirability of specifying the probability and duration of periods of high cell loss rates. Ferrandiz and Lazar [9] investigate the distribution of gaps, that is, consecutive losses, due to blocking and clipping (see below) in a multiclass $G/G/m/B$ queueing system; we discuss similarities and differences between our work and [9] in the following sections. Van Doorn [10] and Meier-Hellstern [11] characterize the overflow process from a finite Markovian queueing system. [5] investigates the effect of correlated errors on forward-error correction only through simulation. A large

This work is supported in part by the Office of Naval Research under contract N00014-90-J-1293, the Defense Advanced Research Projects Agency under contract NAG2-578 and an NSF equipment grant, CERDCR 8500332.

[1] The terms "customer" and "packet" will be used interchangeably.

body of literature analyzes the overflow process of blocked-calls-cleared systems, but the results do not seem directly applicable to our problem.

The paper is organized as follows. After defining more precisely the systems and measures of interest in the section below, continuous-time queues are investigated in Section 2. In Section 4 we then apply similar methods to derive the corresponding measures for a discrete-time queue of interest in packet switching. We conclude by summarizing the work presented and pointing out some issues to be investigated. Additional details and examples can be found in a companion technical report [12].

2 Performance Measures

This paper focuses on a single-server queue, where customers are processed in the order of arrival. Customers that spend more than a deterministic, fixed amount of time h waiting for service are tagged as lost on leaving the queue, but are still served. (Ferrandiz and Lazar [9] refer to this as clipping loss.) This definition of loss is motivated by considerations of traffic with soft real-time constraints, where packets that are excessively delayed are worthless to the receiver. The loss as defined here differs from that studied in our earlier work [1], where excessively delayed customers depart before occupying the server.

A *loss period* (LP) is an uninterrupted interval during which all arriving customers would experience a waiting time exceeding h. For infinite queues with FCFS service, the loss period equals the interval during which the virtual work in the queue is greater than the threshold h. Loss periods and busy periods are related in that a busy period is a special case of a loss period, with threshold value $h = 0$. Also, both busy periods and loss periods start with the arrival of a customer. They differ, however, in that a busy period ends with a departure of a customer, while the end of a loss period is not connected with a customer departure. A *noloss period* is the interval between two loss periods. For $h = 0$, noloss periods become idle periods of the queue.

While the loss period is of independent interest, we are particularly concerned with measuring the number of consecutively lost customers, called *loss run* for brevity. Note that the number of customers arriving in a loss period is not identical to the number of consecutively lost customers. In particular, the first customer triggering a loss period, i.e., the customer that pushes the virtual work above h, is itself not lost. Thus, loss periods consisting of a single arrival do not contribute to customer loss. Note that there may also be several noloss periods interspersed between two customer losses if each of the loss periods separating the noloss periods consists of only the arrival triggering a loss period. Similar to loss runs, *success runs* denote the number of consecutive customer without any loss.

Fig. 1 depicts a virtual work sample path and the measures defined above. Time proceeds along the abscissa from left to right, while the ordinate shows the virtual work at a given instant. In the figure, arrivals that will be tagged as lost on departure are marked with $\otimes$, the others with $\odot$. The extent of loss periods (LP) and busy periods (BP) are indicated by horizontal bars. Bold arrows represent *initial jumps*, the amount of virtual work above h at the beginning of a loss period. The height of the vertical jumps indicates the amount of work brought to the system by an arriving customer. The unfinished work decreases at a unit rate as the server processes jobs.

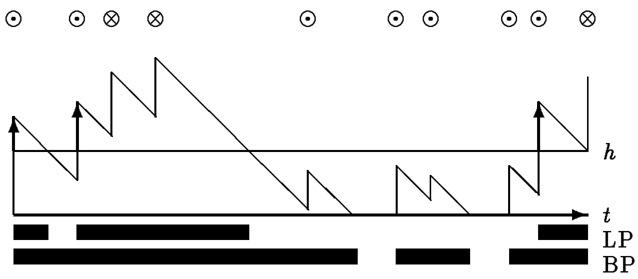

Figure 1: Virtual work sample path

3 Continuous Time Systems: $G/M/1$

As a first model, we investigate a system with general arrival process of rate λ and exponential service with rate μ. The buffer is infinite, so that only loss due to excessive waiting time occurs. All packets, regardless of whether they exceed the waiting time threshold or not, are served. (A system where late packets are dropped will be treated later.) The special case of Poisson arrivals, i.e., the $M/M/1$ system, will be treated in detail since it yields closed-form expressions for the measures of interest. The $M/M/1/\infty$ system was also investigated by Ferrandiz and Lazar [9] as a special case of their $G/G/m/B$ analysis. Their analysis seems considerably more involved, does not readily yield numerical results and does not make use of the simple connection to the busy period pointed out here. The model is applicable, if only in approximation, to systems with variable packet sizes, for example the PARIS network [13] or a variation of the Knockout switch [14].

Throughout this section, we will verify and illustrate our calculations through a running example consisting of an $M/M/1$ queue with arrival rate $\lambda = 0.8$, service rate $\mu = 1$ (and thus a load of $\rho = \lambda/\mu = 0.8$) and a clipping threshold of $h = 3$ so that $\alpha = 1 - \rho e^{(\lambda-\mu)h} = 43.905\%$ of all arrivals experience a delay of more than 3 (are "lost"). The methods presented below apply equally well at lower loss probabilities; we have chosen this (impractically) high loss to speed simulation verification of our results through simulation.

3.1 Distribution of the $G/M/1$ Initial Jump and Loss Period

As pointed out above, a loss period commences when an arrival causes the virtual work to cross the threshold h from below. In order to establish the distribution of the loss period, the following lemma describes the distribution that governs the initial jump, i.e., the virtual work immediately after the arrival of the first customer in a loss period.

Lemma 1 *For a $G/M/1$ queue, the initial jump has the same distribution as the service time and does not depend on the threshold h.*

PROOF Let the random variable J represent the height of the initial jump and $f_J(j)$ its density. The density $f_J(j)$ can be expressed through conditioning, where we define V as the virtual work just prior to the arrival of the customer whose new work pushes the virtual work above h.

$$f_J(j) = \int_0^h P[V = v | V \le h] \cdot$$
$$P[\text{new work} = h + j - v | \text{new work} \ge h - v] \, dv \quad (1)$$

The second conditional probability can be rewritten as

$$\frac{\mu e^{-\mu(h+j-v)}}{e^{-\mu(h-v)}} = \mu e^{-\mu j},$$

which follows immediately from the memorylessness property of the exponential distribution.

Now we can rewrite the jump density as

$$\begin{aligned} f_J(j) &= \mu e^{-\mu j} \int_0^h \frac{P[v = y \cap y \le h]}{P[v \le h]} \, dy \\ &= \mu e^{-\mu j} \frac{1}{P[v \le h]} \int_0^h P[v = y] \, dy \\ &= \mu e^{-\mu j} \frac{1}{P[v \le h]} P[V \le h] \\ &= \mu e^{-\mu j} \qquad \square \end{aligned}$$

Given this property of the initial jump, the following theorem follows immediately:

Theorem 1 *In a $G/M/1$ queueing system, a loss period is stochastically identical to a busy period. The loss period distribution is independent of the threshold h. For $G/G/1$ queues, a loss period is stochastically identical to a busy period with special first service.*

Busy periods with special (or exceptional) first service are covered by Wolff [15, p. 392-394].

The independence of the loss behavior from the threshold recalls a similar observation made by Li [16] regarding the buffer overflow process in a packet voice system. There, the time spent in the overload state was found to be independent of the buffer size.

3.2 Consecutive Customers Lost

While the duration of a loss period is distributed like the duration of a busy period, we recall from Section 2 that the number of consecutive customers lost does not have the same distribution as the number of customers in a busy period. Defining C_C and C_B as the number of consecutively lost customers and the number of customers in a busy period, respectively, we have

$$P[C_C = n] = \frac{P[C_B = n+1]}{P[C_B > 1]}, \; n > 0$$

where $P[C_B > 1]$ is the probability that the busy period contains more than one customer.

Let us apply these results to the $M/M/1$ queue. With $P[C_B = n]$ given by [17, Eq. (5.157)]

$$P[C_B = n] = \frac{1}{n} \binom{2n-2}{n-1} \rho^{n-1} (1+\rho)^{1-2n}, \; n > 0$$

we compute

$$P[C_B > 1] = 1 - P[C_B = 1] = 1 - \frac{1}{1+\rho} = \frac{\rho}{1+\rho}.$$

Thus,

$$P[C_C = n] = \frac{1}{n+1} \binom{2n}{n} \frac{\rho^{n-1}}{(1+\rho)^{2n}}, \; n > 0.$$

Note that this result differs markedly from the geometric distribution postulated by Ferrandiz [9, Corollary 5.3].

Since the average number of customers per busy period is $1/(1-\rho)$, we have that for the $M/M/1$ queue the average number of consecutive customers lost is

$$E[C_C] = \frac{1}{P[C_B > 1]} \left(\frac{1}{1-\rho} - 1 \right) = \frac{1+\rho}{1-\rho}. \qquad (2)$$

This result differs markedly from that obtained under the assumption that losses occur independently as Bernoulli events with the time-average loss probability α. In that case, the conditional probability mass function (pmf), given one loss, for the number of consecutive losses would be distributed geometrically as

$$P[\hat{C_C} = n] = \alpha^{n-1}(1-\alpha)$$

with an average value of $E[\hat{C_C}] = 1/(1-\alpha)$. For our running example, the independence assumption leads one to conclude that a loss period consists of 1.78 customers on average, while our analysis above shows that the actual number for this system is 9. Thus, customer losses are far more clustered than the assumption of independent losses would suggest.

An additional characterization of loss periods is provided by the *conditional probability of packet loss* given that the previous packet was lost, denoted here by r. It is directly related to the average loss run length, $E[C_C]$, through [9, eq. (5.1)]

$$\begin{aligned} E[C_C] &= \frac{1}{1-r} \approx 1 + r + r^2 + \ldots \\ r &= 1 - \frac{1}{E[C_C]}. \end{aligned}$$

For the $M/M/1$ case,

$$r = \frac{2\rho}{1+\rho}.$$

The clustering of losses in a queueing system is naturally also reflected in this measure. For our $M/M/1$ example, the conditional loss probability equals 0.89, while the assumption of independent losses would result in a conditional loss probability r equal to the loss probability α, which evaluates to 0.44 for our running example.

3.3 Distribution of Noloss Period

The distribution of the time between loss periods is more difficult to determine. This interloss time comprises the interval between the time the virtual work W drops below h from above up to the first time instance it rises above this mark again, i.e., the event $\min\{t : W(t) = h | W(0) = h\}$. The sample path with respect to time t of this stochastic process is continuous in time and right-continuous in state, with drift of rate t and jumps of exponential height at Poisson intervals. The difficulty appears since the process is "sticky" at the zero line, with dwell time corresponding to the interarrival (or queue idle time) distribution. We are interested in the distribution of the time to absorption of this process at the $W = h$ barrier.

To make the problem tractable, a Markovian arrival process has to be assumed; otherwise the duration of the noloss period would depend on the time of the last arrival during the preceding loss period. Thus, the computation of this section will be limited to the $M/G/1$ model.

We select a model involving an approximation that is based on the so-called Moran dam model [18]. In this model, the water content of a dam or reservoir is represented by a continuous or discrete-state, discrete-time homogeneous Markov process. For reasons of computability, we choose a discrete-state representation, yielding a discrete-time Markov chain (DTMC). Time is discretized in quantities of τ, a fraction of the service time, and the Markov chain tracks the virtual work W_n in the queue at epochs of integer multiples of τ. Thus, the Markov chain has $k = h/\tau$ states. At the end of each discretization interval, at $(n\tau)^-$, the virtual work, if positive, decreases by one unit, reflecting the fact that the virtual work decreases at unit rate. Arrivals bring in an amount of work X_n, again in multiples of τ, and occur just after the beginning of the discretization interval, at $(n\tau)^+$. The noloss period ends as soon as the virtual work reaches h or state $k-1$. We model this by making state $k-1$ an absorbing state and compute the duration of the noloss period as the time to absorption into $k-1$. Given this description of the DTMC, we can write the state evolution recursively as

$$W_{n+1} = \min(k, W_n + X_n) - \min(1, W_n + X_n).$$

Let us define a_k as the probability that k units of work of size τ arrive. Also, denote the complementary cumulative distribution function g_j as

$$g_j = \sum_{i=j}^{\infty} a_j = 1 - \sum_{i=0}^{i-1} a_j.$$

The state transition matrix follows readily:

$$\mathbf{P} = \left[\begin{array}{c|ccccc} & 0 & 1 & \dots & k-2 & k-1 \\ \hline 0 & a_0 + a_1 & a_2 & & a_{k-1} & g_k \\ 1 & a_0 & a_1 & & a_{k-2} & g_{k-1} \\ \dots & & & & & \\ k-2 & 0 & 0 & & a_1 & g_2 \\ k-1 & 0 & 0 & & 0 & 1 \end{array}\right] \quad (3)$$

The last row stems from the fact that state $k-1$ is absorbing. The state transition probabilities are computed as

$$a_j = P[\tau j \leq X \leq \tau(j+1)], \quad (4)$$

where X is the amount of work arriving in a slot. We know that the accumulated work from n arrivals in a $G/M/c$ system is Erlang-n distributed with density

$$f(x) = \frac{\mu n (\mu n x)^{n-1}}{\Gamma(n)} e^{-\mu n x}.$$

and cumulative distribution function $F(x) = P(n, \mu x)$, where $P(a, x)$ is the normalized incomplete gamma function

$$P(a, x) \triangleq \frac{\gamma(a, x)}{\Gamma(a)} = \frac{1}{\Gamma(a)} \int_0^x e^{-t} t^{a-1}\, dt.$$

The distribution of arriving work needed in evaluating Eq. (4) is hence given by

$$\begin{aligned} P[X < x] &= \sum_{n=0}^{\infty} P[X < x | n \text{ arrivals}] P[n \text{ arrivals}] \\ &= e^{-\lambda\tau} + \sum_{n=1}^{\infty} P(n, \mu x) e^{-\lambda\tau} \frac{(\lambda\tau)^n}{n!}. \end{aligned}$$

The distribution of the time to absorption can be computed in two basic ways. First, since the probability of having been absorbed by the nth transition is simply the probability that the DTMC is in state $k-1$ after n steps, we can use the basic state probability equation in its recursive or matrix-power form,

$$\pi^{(n)} = \pi^{(n-1)} \mathbf{P} = \pi^{(0)} \mathbf{P}^n$$

where $\pi^{(0)} = (0, 0 \dots 0, 1, 0)$, i.e., $k-2$ is the initial state. The matrix power form can be evaluated as a special case of the general relationship for any functional f of a matrix, given by $f(\mathbf{P}) = \mathbf{V} f(\Lambda) \mathbf{V}^{-1}$, where $\mathbf{V}$ is the matrix of eigenvectors of $\mathbf{P}$ and the function f is applied element-by-element to the diagonal matrix of eigenvalues Λ [19, p. 8]. The eigenvalue approach may be more accurate for large values of n.

The other alternative defines $f_{il}^{(n)}$ as the probability that the system *first* enters state l after n steps, given the initial state is i. To use the first-passage formulation, the matrix P has to be returned to its recurrent form by replacing the last row with $(0, \dots, 0, a_0, g_1)$. It is readily seen that this first-passage pmf is recursively defined for all transition matrices as

$$\begin{aligned} f_{il}^{(1)} &= P_{il} \text{ for } i = 0, 1, \dots & (5) \\ f_{il}^{(n)} &= \sum_{j, j \neq l} P_{ij} f_{jl}^{(n-1)} & (6) \end{aligned}$$

We obtain a lower bound on the cumulative probability by using $k-2$ as the initial state, an upper bound by using $k-1$.

Sample calculations showed that state equations, matrix computations and the approach using f_{il} yield the same numerical result to within four significant figures, indicating that roundoff errors are not a serious problem here. Also, the computational effort is about the same.

The discretization error incurred by using a particular value of τ can be estimated by computing the expected duration of the noloss period. Since the fraction of packets lost, α, is related to the expected loss period $E[L]$ and the expected noloss period $E[N]$ by[2]

$$\alpha = \frac{E[L]}{E[L] + E[N]} \quad (7)$$

the expected noloss period can be computed as

$$E[N] = E[L] \left(\frac{1}{\alpha} - 1 \right).$$

[2] Replacing α by the load, ρ, noloss and loss periods by idle and busy periods yields the well-known relation for busy cycles, again underlining the strong connection between loss and busy periods.

Given the DTMC approximating the virtual work process, the expected noloss period (equivalent to the time to absorption) can be computed as

$$E[N] = \left(\frac{1}{\pi'_{k-1}} - 1\right)\tau \tag{8}$$

where π'_{k-1} is the *steady-state* probability that the return process corresponding to the DTMC is in state $k-1$. The transition probability matrix of the return process is derived from P by replacing the last row with all zeros, except for a one in column $k-2$. This relationship is derived in [20, p. 112, Problem 3] for the case of two absorbing states, but the result generalizes readily to any number of absorbing states (see also [21, p. 103]).

For our example, the exact value of $E[N]$ is 6.388. For the discretization with $\tau = 0.1$, Eq. (8) yields a value of 6.109, which improves to 6.237 and 6.327 for $\tau = 0.05$ and $\tau = 0.02$, respectively.

3.4 Customers per Noloss Period

It appears difficult to derive an expression for the distribution of the number of customers in a noloss period. The expected value, however, is readily available since the average number of customer arrivals during loss periods, $E[C_L]$, and noloss periods, $E[C_N]$, are related in a similar fashion as the respective duration measures, Eq. (7), yielding

$$\begin{aligned} \alpha &= \frac{E[C_L]}{E[C_L]+E[C_N]} \\ E[C_N] &= E[C_L]\left(\frac{1}{\alpha}-1\right). \end{aligned}$$

The difficulty in determining the distribution arises due to the fact that noloss periods do not "see" the same Poisson arrival process with rate λ as a random observer, just as the the arrival rate measured during busy periods is higher than the arrival rate measured over all customers. Thus, the conditional probability of the number of arrivals given a noloss period duration cannot be readily computed.

All results were confirmed by simulation experiments. Even for the relatively coarse quantization of $\tau = 0.1$, the noloss duration agrees quite closely with the approximate analysis. Details can be found in [12].

4 Clip Loss in Discrete-time Systems

We now turn our attention to a queueing model that is commonly used for packet switches and ATM-type networks [22, 23]. In this model, time is discretized, with deterministic service (of duration $\tau = 1$) and batch arrivals, which, in many instances, allow somewhat simpler solutions than their continuous-time counterparts. Batches arrive at the beginning of a slot of unit width, while at most one customer departs at the end of a slot. (Hunter [24, p. 193] refers to this as an early arrival system.) We allow the batch size, A, to have a general distribution, but require the batch sizes to be independent from slot to slot and independent of the state of the queue itself.

We will say that batch sizes are geometrically distributed if their probability mass function (pmf) is $a_n = pq^n$, with $q = 1-p$ and an average batch size of $\rho = q/p$. We will also cover the Poisson distribution with mean ρ,

$$a_n = \frac{e^{-\rho}\rho^n}{n!} \text{ for } n \in [0, \infty),$$

and the binomial distribution with mean $\rho = \nu p$,

$$a_n = \begin{cases} \binom{\nu}{n} p^n q^{\nu-n} & \text{for } n \in [0,\nu] \\ 0 & \text{otherwise} \end{cases}.$$

While numeric solutions are possible for general batch size distributions, a queue with geometric batch sizes, i.e., the system $D^{[\mathrm{Geo}]}/D/1$, will be shown to share the identity between busy period and loss period described in Section 2 for the continuous-time $GI/M/1$ system. Also, restricting batch sizes to be geometrically distributed will yield closed-form expressions for many distributions of interest. Due to the close relationship between busy and loss periods, we will investigate busy periods in some detail in Section 4.1, followed by derivations of the properties of loss and noloss periods in sections 4.2 and 4.3, respectively.

4.1 The Busy and Idle Period

In discrete-time systems, we define that a *busy period* begins when the first customer in a batch experiences no waiting, i.e., finds the server free. Note that, unlike in continuous time, a server may be continuously occupied for more than one busy period. This occurs if the last customer in a busy period, departing in (n^-, n), is immediately followed by one or more new arrivals in (n, n^+). Thus, the first customer in that batch enters service immediately, starting a new busy period, while the server experiences no idle slot. Later, we will discuss the composite busy period which encompasses time intervals without server idling, consisting of one or more busy periods.

Let us return now to the busy period and compute its distribution. Because each customer occupies the server for one slot, duration and number of customers served are equal in the discrete-time case. For geometric batches[3], we can compute the number of customers in a busy period by making use of Takács combinatorial arguments [25, p. 102f], [17, p. 225f]. Let B be the number served in a busy period and $\tilde{A}_n$ the number of arrivals during the service times of customers 1 through n, where no assumption is made as to whether these n customers belong to the same busy period(s).

The probability mass function of the number served in a busy period is given by [17, Eq. (5.166)]

$$P[B=n] = \frac{1}{n} P[\tilde{A}_n = n-1].$$

For the case of deterministic service and batch size distribution a_n, the probability on the right-hand side can be readily derived:

$$P[\tilde{A}_n = n-1] = P[n-1 \text{ arrivals in } n \text{ slots}] = a_{n-1}^{n*}.$$

Here, a_n^{n*} denotes the n-fold convolution of a_n with itself. For the case of geometrically distributed batches, the convolution becomes the negative binomial or Pascal distribution with probability mass function

$$a_n^{r*} = \binom{r+n-1}{n} p^r q^n.$$

[3] The ensuing development requires that the work arriving during the service of each customer be i.i.d.. For batch arrivals, the arrivals that occur while the first customer is being serviced consist of the batch starting the busy period minus one. The distribution of this "shortened" batch has the same distribution as a regular batch only if batches are geometrically distributed. For other batch distributions, the following calculations can serve as an approximation.

Thus, the busy period for geometric batches is distributed according to

$$P[B=n] = \frac{1}{qn}\binom{2n-2}{n-1}(qp)^n \qquad (9)$$

$$= \frac{1}{n}\binom{2n-2}{n-1}\rho^{n-1}(1+\rho)^{1-2n}. \qquad (10)$$

The last transformation makes use of the fact that the system load ρ is related to the batch distribution parameter p through $p = 1/(1+\rho)$. We recognize the last expression as the distribution of the number of customers served in an $M/M/1$ busy period [17, p. 218].

The z-transform of the number served in a busy period, $B(z)$, follows from the $M/M/1$-derivation [17, p. 218]:

$$B(z) = \frac{1+\rho}{2\rho}\left[1 - \sqrt{1 - \frac{4\rho z}{(1+\rho)^2}}\right] = \frac{1-\sqrt{1-4pqz}}{2q} \qquad (11)$$

The expected number served (and arriving) in a busy period can be computed by evaluating an infinite series using Eq. (11) or directly copying the $M/M/1$ result:

$$E[B] = \sum_{n=1}^{\infty} nP[B=n] = \frac{p}{\sqrt{1-4pq}} = \frac{1}{1-\rho}.$$

The *idle period* I for general batch sizes is geometrically distributed with density $P[I=n] = a_0^n(1-a_0), n \geq 0$. Recall that a_0 is the probability of a batch having zero members. Thus, the average idle period is given by $a_0/(1-a_0)$. For geometric batch sizes, $a_0 = p$ and thus an idle period last an average of $1/\rho$ slots.

In contrast to the continuous-time case, an idle period may have zero duration. This occurs if a new batch arrives immediately after the last customer of a busy period departs. We will call a period of continuous server occupancy spanning several busy periods a *composite busy period* and identify random variables associated with it by a tilde. A composite busy period consists of ϕ busy periods with the geometric probability $P[\phi = k] = (1-a_0)^{k-1}a_0$. In the z-transform domain, the number of customers in the composite busy period, $\tilde{B}$, is determined as $\tilde{B}(z) = \phi(B(z))$, where

$$\phi(z) = \frac{pz}{1-qz}$$

is the probability generating function of the number of busy period constituting a composite busy period.

For geometric batches, $\tilde{B}(z)$ can be expanded using Eq. (11):

$$\begin{aligned}\tilde{B}(z) &= \frac{p}{q}\frac{1-\sqrt{1-4pqz}}{1+\sqrt{1-4pqz}} \\ &= \frac{p}{q}\frac{1-2\sqrt{1-4pqz}+1-4pqz}{4pqz} \\ &= \frac{1}{zq}\left[\frac{1-\sqrt{1-4pqz}}{2q}\right] - \frac{p}{q}\end{aligned}$$

The bracketed fraction is recognized as $B(z)$ and thus $\tilde{B}(z)$ can be inverted easily, yielding

$$\begin{aligned}P[\tilde{B}=n] &= \frac{1}{q}\frac{1}{n+1}\binom{2n}{n}\frac{\rho^n}{(1+\rho)^{2n+1}} \\ &= \frac{1}{n+1}\frac{(2n)!}{(n!)^2}\frac{\rho^{n-1}}{(1+\rho)^{2n}} \qquad (n \geq 1).\end{aligned}$$

The expected number of customers in a composite busy period (and its expected duration in slots) is seen to be

$$E[\tilde{B}] = E[B]\cdot E[\phi] = \frac{E[B]}{a_0},$$

in general or

$$E[\tilde{B}] = \frac{1+\rho}{1-\rho}$$

for geometric batches.

4.2 The Loss Period

A *loss period* begins when one or more customers arriving in a batch see h or more customers already in the system (in other words, if their wait is equal to or greater than h.) Thus, a busy period is (again) a special case of a loss period with h having the value zero. A loss period ends when there are h or fewer customers left after a departure. Just as discussed above for the case of busy periods, a loss period may be followed immediately by another loss period. This occurs if the number of customers reaches h at some point n, i.e., the loss period ends, and a batch arrives in (n, n^+), starting a new loss period. An uninterrupted interval where the number of customers in the system just prior to an arrival never drops below h (or, equivalently, where the occupancy after the arrival instants remains above h) will be referred to as a *composite loss period.* Clearly, it consists of an integral number of loss periods.[4] The random variables L and $\tilde{L}$ represent the loss period and composite loss period, respectively, while the random variable V represents the occupancy on the slot boundary, equivalent to the work in the queue seen by an arriving batch (virtual work). Because of the deterministic service time and the slotted arrivals, duration and customer count properties are the same, i.e., a loss period of l slots leads to l consecutive losses.

Fig. 2 depicts an example of a sample path showing a composite loss period made up of two loss periods (LPs) for a threshold of $h = 3$.

Just like for the continuous-time case, we are interested in determining the distribution of the initial jump J, that is, the work load or, equivalently, the number of customers that begins a loss period. Fig. 2, for example, shows two initial jumps, one occurring at the arrival of batch $\{c \ldots f\}$, with a height of two, and the second at the arrival of $\{g\}$, with a height of one.

Lemma 2 *The initial jump J into a loss period has the distribution*

$$P[J=j] = \frac{\sum_{a=j}^{\min(h+j,\nu)} P[V_h = h+j-a]P[A=a]}{\sum_{a=1}^{\min(h,\nu)} P[V_h > h-a]P[A=a] + P[A>h]},$$

where the distribution of the batch size random variable A is zero outside the range $[0,\nu]$ and the conditional system occupancy distribution seen by an arriving batch is defined as

$$P[V_h = v] \triangleq \frac{P[V=v]}{P[V \leq h]}.$$

Proof We sum the probabilities of all events leading to a jump of size j (given that a jump occurred), noting that

[4] However, unlike in the $M/M/1$-case, we do not have to factor out the single-customer busy periods.

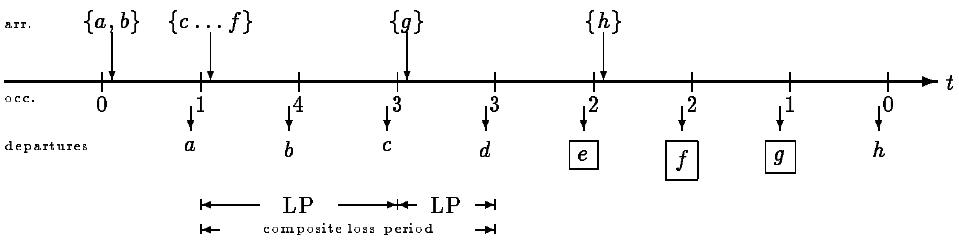

Figure 2: Loss periods in discrete time ($h = 3$)

the random variables A and V are independent. Thus, we have for jumps into loss periods,

$$\begin{aligned} & P[J=j] \\ &= \sum_{v+a=h+j} P[V_h = v \cap A = a | A + V_h > h] \\ &= \frac{\sum_{v+a=h+w} P[V_h = v \cap A = a \cap A + V_h > h]}{P[A + V_h > h]} \\ &= \frac{\sum_{v+a=h+w} P[V_h = v] P[A = a]}{\sum_{v+a>h} P[V_h = v] P[A = a]} \\ &= \frac{\sum_{a=1}^{\min(h+j,\nu)} P[V_h = h + j - a] P[A = a]}{\sum_{a=1}^{\min(h,\nu)} P[V_h > h - a] P[A = a] + P[A > h]} \end{aligned}$$

Among loss periods, we have to distinguish those that form the first loss period in a composite loss period. While an arriving batch that starts a loss period may see between 0 and h in the system, a batch that initiates a composite loss period, i.e., the initial jump conditioned on the fact that the loss period is the first in a composite loss period, can see at most $h-1$ customers. Thus, the following lemma provides a separate expression for the jump into a composite loss period, $\tilde{J}$.

Lemma 3 *The initial jump into a composite loss period $\tilde{J}$ has the distribution*

$$P[\tilde{J} = j] = \frac{\sum_{a=j+1}^{\min(h+j,\nu)} P[V_{h-1} = h + j - a] P[A = a]}{\sum_{a=2}^{h} P[V_{h-1} > h - a] P[A = a] + P[A > h]}.$$

The derivation of $P[\tilde{J}]$ proceeds as in the proof of the previous lemma.

Finally, initial jumps of all but the first loss period within a composite loss period always start at h. Since the queue occupancy seen by an arriving batch and the batch size itself are independent, the jump into these "interior" loss periods is distributed like a regular non-zero batch.

In close parallel to the continuous-time case, the memorylessness property of the geometric distribution makes the distribution of both types of initial jump a shifted version of the batch distribution, independent of h. We formulate more precisely in a lemma:

Lemma 4 *For the $D^{[\mathrm{Geo}]}/D/1$ queue, the initial jump J and the initial jump into composite loss periods $\tilde{J}$ are distributed like non-zero batches, with pmf $P[J = j] = P[\tilde{J} = j] = pq^{j-1} = \rho^{j-1}/(\rho+1)^j$.*

The proof can be found in [12].

Given the characteristics of the initial jump, the (composite) loss period are seen to be stochastically identical to (composite) busy periods with a special first service given by the initial jump distribution. By the previous lemma, loss periods in a system with geometric batch arrivals have the same distribution as regular busy periods of the same system. Since the members of a batch that initiate a loss period also experience delays of at least h, the number of customers lost in a loss period and the number served in a busy period with the above-mentioned special first service are stochastically identical as well.

For general batch-size distributions, the probabilities for loss periods of length one can easily be written down exactly:

$$\begin{aligned} P[L=1] &= P[J=1] \\ P[\tilde{L}=1] &= P[J=1] a_0 \end{aligned}$$

Closed-form expressions for measures of L and $\tilde{L}$ for longer durations seem difficult to obtain, however. We therefore model the queue state during loss periods as a discrete-time Markov chain with an absorbing state. Since the composite loss period is of greater practical significance, indicating the number of consecutively lost customers, we will focus on this random variable for the remainder of this section. The states of the chain indicate the amount of unfinished work above h just after a batch has arrived. For computational reasons, we truncate the transition matrix to $K+2$ states and write

$$\mathbf{P} = \left[\begin{array}{c|cccccc} & 0 & 1 & 2 & \dots & K & K+1 \\ \hline 0 & 1 & 0 & 0 & \dots & 0 & 0 \\ 1 & a_0 & a_1 & a_2 & \dots & a_K & 1 - \sum_{j=0}^{K} P_{1j} \\ 2 & 0 & a_0 & a_1 & \dots & a_{K-1} & 1 - \sum_{j=0}^{K} P_{2j} \\ \vdots & & & & & & \\ K+1 & 0 & 0 & 0 & \dots & a_0 & 1 - a_0 \end{array} \right]$$

Note that the zero state is absorbing since the loss period ends when the amount of unfinished work above h reaches zero. We therefore obtain the duration of a composite loss period by evaluating the chain's absorption probabilities over time, given the initial state probabilities determined by the distribution of the initial jump, also truncated to $K+2$ values.[5]

[5] For the tail of the loss-period, this random walk with drift and one absorbing barrier may be represented by the corresponding Brownian motion [26, p. 437].

Since the distribution of the composite loss probability typically has very long tails, computing its expected value through summation of the weighted probabilities was found to be numerically inaccurate and computationally expensive. However, an alternative approach exists [27, p. 425]. Let d_j be the expected time to absorption into state 0, starting at state j. By the law of total probability, we can write a system of linear equations

$$d_j = \sum_{k=1}^{K+1} \mathbf{P}_{jk} d_k + 1.$$

In matrix from, the linear system consists of the $\mathbf{P}$ matrix with its first row and column removed.

4.3 The Noloss Period

The state evolution during a noloss period can also be modeled by a discrete-time transient Markov chain with initial state h. Unlike in the continuous-time case, this model is exact. Since the number of possible states is limited to $h+1$ (including the absorbing state $h+1$ representing a new loss period), no truncation error is incurred.

We track the number of customers in the system just after arrivals, at n^+. Since we cannot easily accommodate zero first-passage times, we compute the conditional pmf of noloss periods lasting at least one slot. The pmf of the unconditional loss period is then simply the conditional pmf scaled by a_0. We denote the respective random variables by N' and N.

The transition probability matrix is similar to the one used to approximate the noloss period for the continuous-time case, Eq. (3), but since there are no departures when the system is empty, the first and second row are identical.

$$P = \begin{bmatrix} & 0 & 1 & \dots & h-1 & h & h+1 \\ \hline 0 & a_0 & a_1 & & a_{h-1} & a_h & g_{h+1} \\ 1 & a_0 & a_1 & & a_{h-1} & a_h & g_{h+1} \\ \dots & & & & & & \\ h-1 & 0 & 0 & & a_1 & a_2 & g_3 \\ h & 0 & 0 & & a_0 & a_1 & g_2 \\ h+1 & 0 & 0 & & 0 & 0 & 1 \end{bmatrix}$$

where

$$g_k = \sum_{j=k}^{\infty} a_j = 1 - \sum_{j=0}^{k-1} a_j.$$

The expected length of the noloss period can be obtained as in the continuous-time case by evaluating the steady-state probabilities of the return process. The state transition matrix of the return process is derived from the matrix $\mathbf{P}$ by replacing the last row $(h+1)$ with zeros except for a one in column h (see Eq. (8)).

The expected number of consecutive successful customers can be computed as discussed in Section 3.4.

4.4 Numerical Examples

For a first impression of the behavior of the loss period, we plot the mean value of the composite loss period as a function of the system load ρ in Fig. 3. As expected, the curves follow the characteristic pattern of loss probability and delay curves for queueing systems, with a gradual rise up to a "knee" point, followed by a region of high sensitivity to ρ for high loads. Exhibiting a pattern that will be seen in the following curves as well, the geometric case is clearly separated from the other distributions, which track each other closely in mean composite loss period.

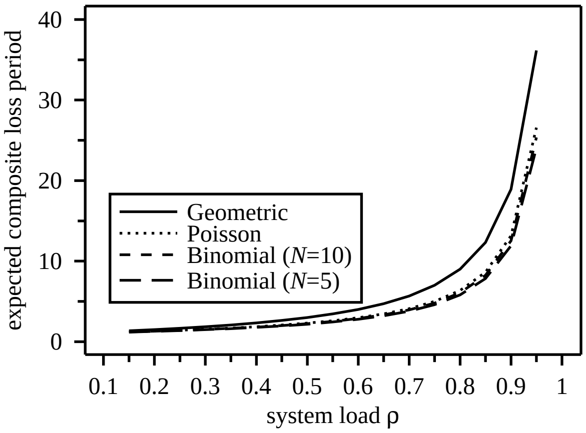

Figure 3: Expected composite loss period as a function of system load for $h = 5$

We stated as Lemma 4 that the distribution of the initial jump and (composite) loss period of the $D^{[\mathrm{Geo}]}/D/1$ queue are independent of the threshold h. It seems therefore natural to investigate to what extent these quantities depend on h for other batch distributions commonly used for modeling data communication systems. As an example, consider the Poisson distribution and binomial distribution[6] with an average batch size of $\rho = 0.8$. For values of h ranging from 3 on up (corresponding to losses of about 30% and less), Table 1 shows that h plays no significant role in $E[\tilde{J}]$ and consequently $E[\tilde{L}]$. (The same observation also holds for the distribution, not shown here.) It other words, for loss probabilities of practical interest, the loss period is basically independent of the threshold.

	Poisson			Binomial ($\nu = 2$)		
h	α	$E[\tilde{J}]$	$E[\tilde{L}]$	α	$E[\tilde{J}]$	$E[\tilde{L}]$
0	1.0000	1.453	7.264	1.0000	1.25	6.25
1	0.6936	1.304	6.520	0.5556	1.00	5.00
2	0.4569	1.275	6.376	0.2469	1.00	5.00
3	0.2974	1.272	6.358	0.1097	1.00	5.00
4	0.1933	1.272	6.358	0.0488	1.00	5.00
5	0.1256	1.272	6.358	0.0217	1.00	5.00
6	0.0816	1.272	6.358	0.0096	1.00	5.00
7	0.0531	1.272	6.358	0.0043	1.00	5.00
8	0.0345	1.272	6.358	0.0019	1.00	5.00
10	0.0146	1.272	6.358	0.0004	1.00	5.00

Table 1: Probability of loss, expected composite loss period and jump for Poisson batches as a function of h for $\rho = 0.8$

The distribution of the composite loss period is shown in Fig. 4. It is seen here that the distributions differ little

[6]The value of $\nu = 2$ used here is the smallest non-trivial value. As ν increases, the behavior should approach that of the Poisson distribution.

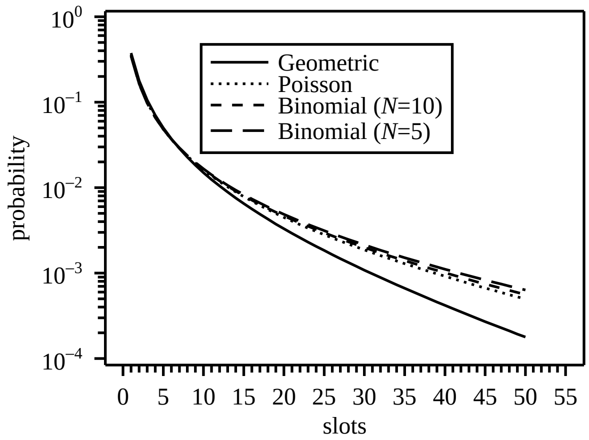

Figure 4: Probability mass function of the composite loss period for $\alpha = 0.1$ and $h = 5$

for small composite loss periods, with most of the difference in expectation caused by the divergence in the tail of the distribution. The loss period for geometric batches tails off significantly faster than those for the either the Poisson or the binomial distribution.

5 Summary and Future Work

In the preceding sections, we have developed probabilistic measures for the behavior of losses in single-server queues commonly used in analyzing high-speed networks and fast packet switches. These measures should be useful in the design and test of packet recovery systems, which face a much harder task than predicted by the optimistic assumption of independent losses.

We found that for certain important queues, the distribution of the loss period is independent of the threshold value used, while for all discrete-time batch distributions investigated the threshold value has very little influence on the loss period. It remains to be seen whether there exists a certain value of h above which the initial jump (and, hence, the loss period) changes no further with increases in h.

It seems natural to extend these results to finite queues in order to analyze periods of buffer overflow, as was done in [16] for a packet voice arrival process.

For $G/M/1$ and $D^{[\mathrm{Geo}]}/D/1$ queues, a busy period can be regarded as a special case of a loss period. Thus, computation of the distribution of busy periods is of particular interest in studying loss phenomena. Our computation of the busy period using combinatorial arguments applied strictly only to geometric batches, but it might provide a readily computable approximation for other distributions.

It is anticipated that the study of correlated arrival processes and non-FIFO service disciplines will require significantly more sophisticated techniques than those found sufficient here.

References

[1] H. Schulzrinne, J. F. Kurose, and D. Towsley, "Congestion control for real-time traffic in high-speed networks," in *Proceedings of the Conference on Computer Communications (IEEE Infocom)*, (San Francisco, CA), pp. 543–550, June 1990.

[2] L. J. Forys, "Performance analysis of a new overload strategy," in *Proceedings of the Tenth International Teletraffic Congress (ITC-10)*, (Montreal), p. 5.24, IAC, North Holland, June 1983.

[3] N. S. Jayant, "Effects of packet losses in waveform-coded speech," in *International Conference on Computers and Communications*, (Atlanta, Georgia), pp. 275–280, IEEE, Oct. 1980.

[4] L. A. DaSilva, D. W. Petr, and V. S. Frost, "A class-oriented replacement technique for lost speech packets," in *Proceedings of the Conference on Computer Communications (IEEE Infocom)*, (Ottawa, Canada), pp. 1098–1105, IEEE, Apr. 1989.

[5] N. Shacham and P. McKenney, "Packet recovery in high-speed networks using coding and buffer management," in *Proceedings of the Conference on Computer Communications (IEEE Infocom)*, (San Francisco, CA), pp. 124–131, IEEE, June 1990.

[6] T. Kamitake and T. Suda, "Evaluation of an admission control scheme for an ATM network considering fluctuations in cell loss rate," in *Proceedings of the Conference on Global Communications (GLOBECOM)*, (Dallas, Texas), pp. 1774–1780, IEEE, Nov. 1989.

[7] W. E. Leland, "Window-based congestion management in broadband ATM networks: the performance of three access-control policies," in *Proceedings of the Conference on Global Communications (GLOBECOM)*, (Dallas, Texas), pp. 1794–1800, IEEE, Nov. 1989.

[8] G. M. Woodruff and R. Kositpaiboon, "Multimedia traffic management principles for guaranteed ATM network performance," *IEEE Journal on Selected Areas in Communications*, vol. 8, pp. 437–446, Apr. 1990.

[9] J. M. Ferrandiz and A. A. Lazar, "Consecutive packet loss in real-time packet traffic," in *Proceedings of the Fourth International Conference on Data Communication Systems*, (Barcelona), pp. 306–324, IFIP TC6, June 1990.

[10] E. A. van Doorn, "On the overflow process from a finite Markovian queue," *Performance Evaluation*, vol. 4, pp. 233–240, Nov. 1984.

[11] K. S. Meier-Hellstern, "The analysis of a queue arising in overflow models," *IEEE Transactions on Communications*, vol. 37, pp. 367–372, Apr. 1989.

[12] H. Schulzrinne, J. F. Kurose, and D. Towsley, "The correlation of losses in finite and infinite queues," Tech. Rep. TR91-3, Department of Computer and Information Science, University of Massachusetts, Amherst, MA, 1991.

[13] I. Cidon and I. S. Gopal, "PARIS: An approach to integrated high-speed private networks," *International Journal of Digital and Analog Cabled Networks*, vol. 1, pp. 77–85, April–June 1988.

[14] K. Y. Eng, M. G. Hluchyj, and Y. S. Yeh, "A knockout switch for variable-length packets," in *Conference Record of the International Conference on Communications (ICC)*, (Seattle, WA), pp. 794–799, IEEE, June 1987.

[15] R. W. Wolff, *Stochastic Modeling and the Theory of Queues*. Englewood Cliffs, NJ: Prentice Hall, 1989.

[16] S. Li, "Study of information loss in packet voice systems," *IEEE Transactions on Communications*, vol. 37, pp. 1192–1202, Nov. 1989.

[17] L. Kleinrock, *Queueing Systems — Theory*, vol. 1. New York, NY: Wiley-Interscience, 1975.

[18] P. A. P. Moran, *The Theory of Storage*. Methuen's Monographs on Applied Probability and Statistics, London/New York, NY: Methuen/Wiley, 1959.

[19] L. Takács, *Stochastic Processes — Problems and Solutions*. London/New York, NY: Methuen/Wiley, 1960.

[20] S. Karlin and H. M. Taylor, *A First Course in Stochastic Processes*. San Diego, CA: Academic Press, 2nd ed., 1975.

[21] U. N. Bhat, *Elements of Applied Stochastic Processes*. New York, NY: John Wiley, 2nd ed., 1984.

[22] M. J. Karol, M. G. Hluchyj, and S. P. Morgan, "Input versus output queueing on a space-division packet switch," *IEEE Transactions on Communications*, vol. COM-35, pp. 1347–1356, Dec. 1987.

[23] Y.-S. Yeh, M. G. Hluchyj, and A. S. Acampora, "The knockout switch: A simple, modular architecture for high-performance packet switching," *IEEE Journal on Selected Areas in Communications*, vol. SAC-5, pp. 1274–1282, Oct. 1987.

[24] J. J. Hunter, *Mathematical Techniques of Applied Probability – Discrete Time Models: Techniques and Applications*, vol. 2. New York, NY: Academic Press, 1983.

[25] L. Takács, *Combinatorial Methods in the Theory of Stochastic Processes*. New York, NY: John Wiley, 1967.

[26] P. A. P. Moran, *An Introduction to Probability Theory*. Oxford, Great Britain: Clarendon Press, 1968.

[27] W. Feller, *An Introduction to Probability Theory and Its Applications*, vol. 1. New York, NY: John Wiley and Sons, third ed., 1968.