ABSTRACT

The peakedness characterization of a teletraffic stream has long been recognized as an extremely useful one for the analysis of loss systems. In this paper, we present a preliminary investigation into this concept as well. Restricting attention, for simplicity, to single server queueing systems $G/G/1$ with FCFS service, we derive an explicit approximation to the waiting time distribution, parameterized by quantities which can be easily computed from the peakedness functional of the arrival stream and the first two moments of the service time. The peakedness functional characterization allows arbitrary stationary arrival streams to be treated, and avoids focusing attention on aspects of arrival streams which in many circumstances may be misleading, such as interarrival time statistics. The accuracy of this approximation is assessed by several examples, and rules of thumb are given predicting its degree of accuracy, or lack thereof, in potential applications.

INTRODUCTION AND SUMMARY

1. INTRODUCTION AND SUMMARY

1.1 Introduction

The performance analysis of modern telecommunication systems invariably involves predicting queuing delays at various points in the system in terms of either measured or assumed traffic and system parameters. Examples are packet switching systems for voice and/or data where transport and switching delays can significantly affect the viability of service. More often than not, system complexity causes the arrival streams at queues to have complex structures which can be expected to have non-negligible effects on the magnitudes of queuing delays. For example, both data and packetized voice traffic sources often are "bursty," and this burstiness tends to increase delays; on the other hand, data links often "smooth" a packet process, and this smoothing tends to decrease delays at subsequent switches.

While recent advances in the formulation and analytic solution, as well as effective numerical solutions, of quite elaborate queueing models allow many complex systems to be almost exactly modeled and analyzed, the application of such models is often inappropriate due to the time and effort that is required. Moreover, the dimension of the parameter space associated with such a model is often so large that it is difficult, if not impossible, to estimate appropriate values of model parameters from traffic measurements.

Thus, there is the need for simple modeling and approximation approaches that capture the most important effects of traffic characteristics on queuing delays. Such simple methods would prove extremely useful in deriving "quick and dirty" approximations for the early performance assessment of alternative system designs, and in understanding performance sensitivities to various traffic characteristics so that appropriate traffic measurements can be planned. Another important benefit derived from simple modeling and approximation methods is new insight into decomposition approaches for the analysis of large systems.

In this paper we consider only cases where service requirements of individual arrivals at a service system are mutually independent, and also independent of the arrival stream itself. Thus it is meaningful to focus attention on characterizing the arrival stream, which can be mathematically modeled as a point process. Two basic approaches can be taken to describe such a process. In one approach, attention is focused on the sequence of interarrival times, $T_i, i \geq 1$, separating successive arrivals to the system (if arrivals can occur in batches, then some of the $T_i$'s may equal zero). In the second approach, attention is focused on the counting process, $N(t), t \geq 0$, associated with the point process; $N(t)$ is a piece-wise constant random process which increases at arrival epochs by integer quantities of magnitude equal to the number of arrivals at those epochs.

These are clearly two equivalent descriptions of an arrival stream when a complete characterization is available. However, the differing focuses of these two descriptions give rise to different approximation philosophies when a complete characterization of an arrival stream is not available. For example, when attention is focused on interarrival times, it is inviting to assume that successive interarrival times are independent, i.e., that the stream is a renewal process, since this assumption tremendously reduces the complexity of the arrival stream characterization and the analysis of its interactions with service systems. However, in many cases this independence of interarrival times is not present, and to assume its presence will result in significant characterization inaccuracies, regardless of how accurately the interarrival time distribution may have been determined. The difference between these two approaches has been pointed out in [1], where partial characterization methods based on $\{T_i\}$ are called stationary-interval methods, while those based on $\{N(t)\}$ are called asymptotic methods.

1.2 Peakedness as an arrival stream characterization

Peakedness is a traffic characteristic which was originally introduced to aid in the approximate analysis of complex loss systems. More recently, it has been shown in [2] that a generalized concept of peakedness is equivalent to a complete second order point process characterization of an arrival stream. No specific assumptions need be made about the structure of the stream (e.g., Poisson, renewal, orderly, etc.) except that it is stationary. The arrival stream is characterized in terms of a peakedness functional, $Z[\cdot]$, which takes complementary holding time distributions as arguments, and maps them into peakedness values. For a given complementary holding time distribution $G$, the peakedness value $Z[G]$ is defined as

$$\lim_{\mu \to 0} \frac{\Var[K_{\mu}]}{\E[K_{\mu}]}$$

where $K_{\mu}$ is the number of busy servers in a Poissonian infinite server group, with holding time distribution $1-G$, to which the arrival stream is hypothetically offered. The intuitive concept is that, if a given complementary holding time distribution characterizes the "reaction time" of the arrival stream with a system, then the resulting peakedness value is a potentially useful measure of stream variability with respect to that system. It is often convenient to restrict the domain of $Z[\cdot]$ to exponential holding time distributions, and this defines an exponential peakedness function, $Z_{exponential}[\cdot]$, mapping the holding time distribution to a peakedness value, $Z_{exponential}[\cdot]$, i.e., the reciprocal of the mean holding time, to a peakedness value.

The reader is referred to [2] for a summary of the origins of the peakedness concept and for details on representations and properties.
of the peakedness functional used in this paper. Particularly
noteworthy is the fact that the pair \( \alpha, Z[G] \) where 
\( \lambda = \lim E[N_t]/t \) is the intensity of the arrival stream, provides 
a complete second order characterization of the counting process \( N_t \). 
Equivalent second order characterizations are the pairs \( \alpha, V(\alpha) \),
\( \alpha, \kappa(\alpha) \), and \( \alpha, U(\alpha) \), where \( V(\alpha) \) is the variance-time curve, \( \kappa(\alpha) \) is the 
covariance density, and \( U(\alpha) \) is the expectation function of the 
counting process.

The characterization in terms of the expectation function is par-
ticularly relevant in many situations; this function is the generaliza-
tion of the renewal function associated with a renewal process, and is 
defined as:

\[
U(x) = \text{the expected number of arrivals following, an arbitrarily chosen arrival, for } x \geq 0,
\]

(At first glance it may appear that \( U(\alpha) \) provides only a first order 
characterization of the stream; however, it is due to the fact that 
time in (1.1) is being measured from an arbitrarily chosen arrival 
that \( U(\alpha) \) provides second order information.) Defining \( \hat{U}(\alpha) \) to be 
the Laplace-Stieltjes transform of \( U(\alpha) \):

\[
\hat{U}(s) = \int e^{-sx} dU(x)
\]

it can be shown that the value of the exponential peakedness func-
tion at the holding time rate \( \beta \) is given by

\[
\exp(\beta) = 1 + \hat{U}(-\alpha) - \lambda/\beta
\]

(1.3)

Finally, a parameter that is often relevant in describing an 
arrival stream is the quantity \( V/M \), defined as the variance-to-mean 
ratio for the number of arrivals occurring in a very long interval of 
time, i.e.,

\[
V/M = \lim \frac{\text{Var}[N_t]}{E[N_t]}
\]

It can be shown that

\[
V/M = 2 \lim_{\beta \to 0} \exp(\beta) = 1 + 2 \lim_{\beta \to 0} \hat{U}(-\alpha) - \lambda/\beta
\]

(1.4)

(1.5)

(Equation (1.5) is an illustration of second order information con-
tained in the expectation function.)

In many cases, it is of interest to estimate the value of \( \exp(\beta) \), 
for an actual arrival stream, for a very small value of \( \beta \); for such 
cases, (1.4) provides a convenient method. For larger values of \( \beta \),
(1.3) can be used, based on the following procedure for estimating 
\( \hat{U}(\alpha) \):

i) Set \( i = 1 \).

ii) Generate a pseudo-random integer-valued variable \( K_i \) with a
'suitably large' mean value.

iii) Generate an exponentially distributed pseudo-random variable 
\( E_i \) with mean \( 1/\beta \).

iv) Allow \( K_i \) arrivals to occur, and select the next arrival. Set a
timer to the value \( E_i \), and record the number of subsequent 
arrivals that occur before the timer expires; denote this as 
\( N(E_i) \).

v) Increment i, and loop back to step ii.

It can be easily seen that the sample mean of the quantities \( N(E_i) \),
generated via the above procedure will be an unbiased estimate of 
\( \hat{U}(\alpha) \), as long as the mean of the quantities \( K_i \) is sufficiently large 
(the \( K_i \)'s simply provide a convenient method for 'selecting an arbi-
trary arrival.').

1.3 The issue of appropriate holding time distributions for peaked-
ness determination

We have stated that \( \alpha, Z[G] \) provides a complete second order 
characterization of the arrival stream. However, except for those 
cases where the structure of the arrival stream is extremely well 
understood, it may be impossible to exactly determine the peaked-
ness functional, i.e., the mapping \( Z: G \rightarrow \text{peakedness value} = Z[G] \),
one must focus instead on determining one or more peakedness 
values directly. This is especially the case when measurements must 
be taken on an actual arrival stream. Moreover, if we are interested 
in using peakedness to approximately characterize the interaction of 
the arrival stream with a given service system, then it is clear that 
for many G's the resulting \( Z[G] \) will bear little relation to the prob-
lem at hand. It is in part the determination of an appropriate G 
that makes the use of the peakedness concept somewhat of an art.

For loss systems there is an obvious G that makes intuitive sense; 
simply using the complementary distribution of the servers' holding 
times results in a peakedness value that may be used in such computa-
tional procedures as the equivalent random method [3] and 
Hayward's approximation [4]. However, for delay systems, 
the appropriate choice of G is not so obvious, due principally to the fact 
that G should appropriately characterize the 'reaction time' of the 
arrival stream with the delay system, and it is essentially this reac-
tion time that we are attempting to determine by analyzing the 
delay system. Specific questions that need answers are:

• When there is access to only a single peakedness value, at what 
  G should \( Z[G] \) be evaluated to provide maximum information 
  relevant to quantifying delays?

• How can the entire peakedness functional \( Z[·] \), or the values 
  \( Z[G] \) at several \( G \)'s, best be utilized in quantifying delays?

• How much information relevant to quantifying delays is con-
tained in the peakedness functional, and how well can delays be 
  approximated in terms of a single \( Z[G] \) as compared with other 
  simple delay approximations?

Addressing these questions for general delay systems, with arbi-
trary numbers of servers and arbitrary service disciplines, is well 
beyond the scope of this study. The objective of this paper is to 
address these questions for the simplest class of queues with general 
stationary input: GI/G/1 queues with FCFS service discipline (we 
do not restrict attention to queues with renewal input, i.e., GI/G/1). 
It is hoped that insights gained from this simpler study will eventu-
ally impact on the issues of more general delay systems.

The essence of the peakedness-based delay approximation that 
will be derived below is that in many cases an appropriate choice for 
\( G \) is

\[
G(x) = P(W > x) \quad \text{for } x > 0
\]

(1.6)

where \( W \) is the waiting time in the queue. Thus, answers to the 
first and second questions above are that if only a single peakedness 
value is available, it should be with respect to the best a priori esti-
mate of the conditional delay distribution (e.g., a design objective 
delay distribution may suffice), while if several peakedness values 
are available, they should be with respect to a set of complementary 
distributions that span likely delay distributions. The answer to the 
third question will be provided by some of the examples later in this 
paper.

1.4 Outline of the remainder of the paper

The remainder of this paper is organized as follows. In Section 2 
we briefly consider the GI/M/1 queue, and show that: i) knowledge 
of \( \exp(\beta) \) uniquely determines the waiting time distribution; and ii) knowledge 
of \( \exp(\beta) \) provides a better approximation to the waiting time distri-
bution than does knowledge of the interarrival time variance, when-
ever \( \beta^{-1} \geq E[W] \). In Section 3 we derive an approximation to the 
waiting time distribution in terms of the function \( \exp(\beta) \) and 
the first two moments of service time. This approximation is then 
tested on several examples in Section 4, where it is shown that in 
many cases the approximation yields better results than would be 
obtained from complete knowledge of the interarrival time distri-
bution and an assumption of renewal process arrivals. Finally, Section 5 
contains some concluding remarks.

2. OBSERVATIONS ON THE GI/MI WAITING TIME

A convenient starting point in the investigation of the effects of 
peakedness on delays is the solution to the GI/M/1 queue (renewal 
input, exponential service time distribution), where the existence of 
an exponential delay distribution is well known.
where $\mu$ is the service rate and, with $\phi()$ denoting the Laplace-Stieltjes transform of the interarrival time distribution, $\omega$ is the unique real number in the interval $(0, 1)$ which satisfies

$$\omega = \phi(1-\omega)$$  \hspace{1cm} (2.2)

Since for a renewal process the expectation function is just the renewal function, for which the Laplace-Stieltjes transform is known to be $\phi(1-\phi(x))$, it is seen from (1.3) that

$$\phi(x) = \frac{\lambda + x \epsilon \exp(x) - 1}{\lambda + x \epsilon \exp(x)}$$  \hspace{1cm} (2.3)

It follows that the waiting time distribution is just

$$P[W \leq x] = 1 - (1-\beta/\mu) e^{-\beta x}, \quad x \geq 0$$  \hspace{1cm} (2.4)

where $\beta$ satisfies

$$\beta = \frac{\mu - \lambda}{\epsilon \exp(\beta)}$$  \hspace{1cm} (2.5)

Thus, knowledge of $\epsilon \exp(\beta)$ allows exact determination of the waiting time distribution in $GI/M/1$!

Now, if only a single peakedness value, $z_0 = \epsilon \exp(\beta_0)$, were available, how well could the delay distribution be determined? It can be shown, e.g., using explicit results in [5], that the function $z_\epsilon(\beta)$ can be bounded sharply (i.e., the bounds can be attained) in terms of $\lambda, \beta_0, z_0$ and $z_\epsilon$ as follows:

$$z_\epsilon(\beta; \lambda, \beta_0, z_0) \leq z_\epsilon(\beta_0; \lambda, \beta_0, z_0), \quad \text{for } 0 \leq \beta \leq \beta_0$$  \hspace{1cm} (2.6a)

$$z_\epsilon(\beta; \lambda, \beta_0, z_0) \leq z_\epsilon(z_0; \lambda, \beta_0, z_0), \quad \text{for } \beta \geq \beta_0$$  \hspace{1cm} (2.6b)

where, with $x_0$ being the unique root in $(0, \infty)$ to

$$x_0 = \frac{\lambda + \beta_0 z_0}{\lambda - \beta_0} (1 - e^{-\beta x_0})$$  \hspace{1cm} (2.7)

the functions $z_\epsilon(\beta)$ and $z_\epsilon(z_0)$ are

$$z_\epsilon(\beta; \lambda, \beta_0, z_0) = \frac{\lambda x_0}{\lambda - \beta_0} - \frac{\lambda}{\beta_0}$$  \hspace{1cm} (2.8)

$$z_\epsilon(z_0; \lambda, \beta_0, z_0) = \left[1 - \frac{(\lambda + \beta_0 z_0 - 1)}{\lambda + \beta_0 z_0}ight]^{-1} - \frac{\lambda}{\beta_0}$$  \hspace{1cm} (2.9)

These sharp bounds on $z_\epsilon(\beta)$ will now yield sharp bounds on $\beta$, via (2.5), in terms of $\lambda, \beta_0$ and $z_0$. Clearly, a better approximation will result from knowledge of $z_\epsilon(\beta_0)$ the closer $\beta_0$ is to the quantity $\mu(1-\omega)$, the reciprocal of the conditional mean $E[W|W>0]$.

A set of sharp bounds analogous to (2.6) can be derived for $z_\epsilon(\beta)$ in terms of $\lambda$ and $\epsilon$, the variance of the interarrival times. It is shown in [5] that these bounds are not as tight as those in (2.6), in particular for the range $\beta \geq \beta_0$. We thus conclude that in the $GI/M/1$ queue, a single peakedness value provides more information relevant to quantifying delays than does the variance of interarrival times, as long as the peakedness value results from a $\beta_0$ no larger than $1/E[W|W>0]$.

3. A PEAKEDNESS-BASED DELAY APPROXIMATION FOR $GI/G/1$ QUEUES WITH FCFS SERVICE

In this section we derive an approximation for the delay distribution for $GI/G/1$ queues with FCFS service in terms of the peakedness functional of the arrival stream. This approximation results from a series of simplifying assumptions, the first of which is that the complementary delay distribution is approximately exponential, i.e.

$$P[W > x] = \alpha e^{-\beta x}, \quad x \geq 0$$  \hspace{1cm} (3.1)

This approximation is at least partly justified by the facts that i) with few exceptions (3.1) will be asymptotically true as $x \rightarrow \infty$, and ii) the range of $x$-values over which (3.1) is nearly correct increases as the server utilization, $\rho$, increases. We next introduce probabilistic arguments and further simplifying approximations to determine the parameters $\alpha$ and $\beta$.

First note that from (3.1)

$$\alpha/\beta = E[W]$$  \hspace{1cm} (3.2a)

$$= \tau_r E[S_n] + \tau_1 E[Q_d]$$  \hspace{1cm} (3.2b)

$$= \tau_r P[W>0] + \tau_1 E[Q_d]$$  \hspace{1cm} (3.2c)

where $S_n$ and $Q_d$ denote the number of jobs in service in service (0) and in queue, respectively, just prior to an arbitrary arrival, $Q_d$ denotes the number of jobs left in queue by an arbitrary job departing the queue (including those that do not wait in the queue), $\tau_1$ is the mean service time of an arbitrary job, and $\tau_r$ is the mean remaining service time for the job in service at an arbitrary arrival if $S_n > 0$. Equation (3.2c) follows from (3.2b) by the equivalence of the queueing distributions at arrivals to, and departures from, the queue; see [3].

We now note that $Q_d$ equals the number of arrivals to the queue during the waiting time of an arbitrary arrival, and introduce the approximation that

$$E[Q_d|W] = U(W)$$  \hspace{1cm} (3.3)

where $U(\cdot)$ is the expectation function of the arrival process. For renewal arrivals this approximation is exact, and for many arrival processes and queueing systems this approximation will be quite good; however, there are situations (one will be illustrated in an example) where this estimate of $E[Q_d|W]$ is systematically biased, and thus will yield a poor approximation. For example, for streams characterized by a slowly varying instantaneous arrival rate, some values of which may produce server utilizations close to or greater than one, for large values of $W$ it will typically be that $E[Q_d|W] > U(W)$; also, for certain arrival streams with finite source characteristics, it may be that $E[Q_d|W] < U(W)$ for large values of $W$.) Accepting (3.3) as a reasonable approximation for most cases, we now have

$$E[Q_d] = E(U(W)) = \alpha \hat{U}(\beta)$$  \hspace{1cm} (3.4)

from (3.1) and an integration by parts.

Combining (3.2c) and (3.4), and noting that $P[W>0] = \alpha$, we find that $\beta$ can be determined via

$$\beta = \frac{1}{\tau_r + \tau_1 \hat{U}(\beta)}$$  \hspace{1cm} (3.5)

or, by making use of (1.3),

$$\beta = \frac{\mu - \lambda}{\tau_1 + z_\epsilon(\beta_0) - 1}$$  \hspace{1cm} (3.6)

where $\mu/\tau_1$ is the service rate. Finally, we need an approximate expression for $\tau_r$, and a simple approximation for which the accuracy tends to increase with increasing server utilization is just

$$\tau_r = \frac{\tau_2}{2 \tau_1}$$  \hspace{1cm} (3.7)

where $\tau_2$ is the second moment of the service times; (3.7) is exact when the arrival stream is Poisson.

To determine $\alpha$, we use the fact that

$$P[W>0] = 1 - \frac{1}{E[\# jobs served in a busy period]}$$  \hspace{1cm} (3.8)

Similarly to (3.4), we approximate

$$E[\# jobs served in a busy period|B] = 1 + U(B)$$  \hspace{1cm} (3.9)

where $B$ denotes the duration of a busy period. If we were to assume that the busy periods are approximately exponentially distributed, with parameter $\gamma$, then

$$E[\# jobs served in a busy period] = 1 + \hat{U}(\gamma)$$  \hspace{1cm} (3.10)

Observing that for both the $GI/M/1$ and the $M/G/1$ queues the ratio of the mean duration of a busy period to the mean conditional waiting time is just $\tau_1/\tau_r$, where $\tau_r$ is given in (3.7), we are now motivated to use $\beta, \tau_1$ for $\gamma$ in (3.10). This leads to the following approximation for $\alpha$:  

\[ 3.1A-4-3 \]
Thus, the final delay approximation is (3.1) with \( \alpha \) and \( \beta \) computed via (3.6), (3.7), and (3.12). This approximation is exact for all GI/M/1 queues, since all of the approximating assumptions above are exact in these cases. Moreover, it can be seen that, while (3.1) is not exact for M/G/1 queues in general, it does result in exact values for \( P[W > 0] \) and \( E[W] \) (recall that \( \epsilon_{\text{approx}}(\cdot) = 1 \) for Poisson arrivals.

**4. EXAMPLES OF THE PEAKEDNESS-BASED DELAY APPROXIMATION**

In this section we summarize conclusions drawn from several examples that have been used to test the accuracy of the delay approximation derived in Section 3. These examples by no means give a complete evaluation of the accuracy of the approximation, but do serve to illustrate cases where it can be expected to be good, and other cases where it cannot. The examples also serve as an illustration of how easily the exponential peakedness function, \( \epsilon_{\text{approx}}(\cdot) \), can be determined. To conserve space, numerical results are not given here; rather we summarize the conclusions we have drawn from examination of many numerical examples.

**4.1 The MR(2)/M/1 queue**

A fairly general class of stationary arrival processes is the class of Markov-renewal processes governed by the evolution of an \( n \)-state Markov chain, which we denote as the class of MR(\( n \)) processes. A typical such process is completely described by an \( n \times n \) matrix of (usually defective) distribution functions:

\[
A(x) = [A_{j,k}(x)]
\]

where, if \( t_i \) denotes the \( i \)th arrival epoch, \( X_i \) denotes the Markov chain state at \( t_i \), and \( T_{i-1} = t_{i-1} - t_{i-2} \) denotes the \( i \)th interarrival time, then

\[
A_{j,k}(x) = P[X_{m} = k, T_{i+1} < x | X_{m} = j; X_{m}, m \leq i-1; T_{m}, m \leq i](4.2)
\]

The ordinary renewal process is a special case where \( n = 1 \) and appropriate \( A_{j,k} \)'s, various types of dependences between interarrival times may be modeled.

A quantity of basic interest is the matrix of Markov chain transition probabilities at arrival epochs:

\[
P = [P_{j,k}], \quad P_{j,k} = \lim_{x \to \infty} A_{j,k}(x) \quad (4.3)
\]

The invariant vector, \( \pi \), of \( P \), defined as

\[
\pi = \pi P, \quad \sum_{j} \pi_{j} = 1 \quad (4.4)
\]

is the vector of Markov chain state probabilities at an arbitrary arrival. It is convenient to define the matrix of Laplace-Stieltjes transforms of the components of \( A \):

\[
\Phi(s) = [\phi_{j,k}(s)], \quad \phi_{j,k}(s) = \int_{0}^{\infty} e^{-sx}dA_{j,k}(x) \quad (4.5)
\]

and the Laplace-Stieltjes transform of the Markov-renewal kernel

\[
H(s) = [H_{j,k}(s)] = \Phi(s)[I - \Phi(s)]^{-1} \quad (4.6)
\]

Although an MR(\( n \)) is in general not a renewal process, one can partially characterize it by its interarrival time distribution, which is easily seen to have the Laplace-Stieltjes transform

\[
\phi(s) = \sum_{j,k} \pi_{j} P_{j,k} \phi_{j,k}(s) \quad (4.7)
\]

Also, the expectation function of the process can be shown to have the Laplace-Stieltjes transform

\[
\tilde{U}(s) = \sum_{j,k} \pi_{j} H_{j,k}(s) \quad (4.8)
\]

and the exponential peakedness function can be obtained via (1.3). It is of interest to compare the quality of the peakedness-based delay approximation with an approximation that uses \( \phi(s) \) in a GI/M/1 approximation, i.e., via (2.1) and (2.2) (the latter would be a typical approximation approach if one were concentrating on interarrival time statistics).

The solution of the MR(\( n \))/M/1 queue has been known for some time, (see, e.g., [6]), and in these test examples we have focused on the special case where \( n = 2 \), where it can be shown that the complementary delay distribution is of the form

\[
P[W > 1] = a_{0} e^{-\beta_{1}} + a_{1} e^{-\beta_{2}} \cdot x \geq 0 \quad (4.9)
\]

Moreover, there is a well defined procedure for obtaining the quantities \( \alpha \) and \( \beta \).

A considerable number of MR(\( 2 \))/M/1 examples have been considered numerically, and the above mentioned two approximations have been compared with the exact solutions, with the following conclusions derived:

- Uniformly over all examples considered, the peakedness-based approximation yields more accurate estimates of the mean waiting time and the asymptotic decay rate of the delay distribution that does the renewal-process-based approximation.
- For most examples considered, the peakedness-based approximation also yields a more accurate estimate of \( P[W > 0] \).
- The accuracy of the peakedness-based approximation for mean delay increases with the server utilization, and is typically within 10% of the true mean for server utilizations \( \rho > 0.50 \), and much better than this for \( \rho > 0.75 \).
- On the other hand, the amount of error in the renewal-process-based approximation may be arbitrarily large, depending on the characteristics of the MR(\( 2 \))/M/1 queue being approximated.

**4.2 The PCPG(B)/M/1 queue**

To investigate the quality of the approximation when the service times are not exponentially distributed, a PCP/G(B)/1 queue was considered. In this system, the arrivals constitute a Poisson Cluster Process (PCP) with clusters of arrivals being spawned by an underlying Poisson process, and each cluster consisting of a random number of equally spaced arrivals. Individual clusters will typically overlap. The PCP is characterized by \( \lambda \), the rate at which clusters originate; \( B \), the deterministic spacing between arrivals in a cluster; \( g(\cdot) \), the probability generating function of the number of arrivals in a cluster; and \( n_{1} \) and \( n_{2} \), the first and second moments of the number of arrivals in a cluster. The service times in the queueing system are characterized by the first two moments \( \tau_{1} \) and \( \tau_{2} \), and in addition it is assumed that the service times are bounded from below by the spacing of arrivals in a single cluster (thus, the "B" in the notation G(B) denotes "bounded service"):

\[
P[\text{service time } \geq \delta] = 1 \quad (4.10)
\]

From a generalization of an observation made by Farell ([7]), it can be shown that the mean waiting time in this model is given by

\[
E[W] = \frac{\rho s_{2} \tau_{1} + [\tau_{1} + (\rho - 1)\lambda s_{2} s_{1} - 2\tau_{2}]}{1 - \rho} \quad (4.11)
\]

where \( \rho = \lambda n_{1} \tau_{1} \) is the server utilization.

For the PCP it can be shown that the expectation function is just

\[
U(x) = \lambda n_{1} x + \sum_{i=1}^{\infty} u(x - i\delta) P[K \geq i] \quad (4.12)
\]

where the random variable \( K \) is defined as the number of subsequent arrivals in the same cluster as an arbitrarily chosen arrival, and \( u(\cdot) \) is the unit step function:

\[
u(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases} \quad (4.13)
\]

It then follows from a length-biasing argument that...

3.1A-4-4
\[ \hat{U}(s) = \frac{\lambda n_1}{s} + e^{-s}\left[ \frac{1-e^{-\lambda t_k}(e^{-s\lambda}-1)n_1}{(1-e^{-s\lambda})^2}\right] \]  
(4.14)

and \( z_{exp}(\cdot) \) can be obtained via (1.3).

The peakedness-based approximation was tested on this example for a wide range of model parameter values, with the same basic conclusions as were derived in Section 4.1, i.e., that the approximation is quite good, especially as the server utilization increases.

4.3 A packetized voice example

The final example that we discuss is the application of the peakedness-based delay approximation to the prediction of delays that would be experienced at a single trunk in a packet switching network over which the packets associated with many packetized voice virtual circuits are being transmitted. This situation may be modeled as a PCP/D/1 queueing system, but the picture is somewhat different from that in Section 4.2 in that now the (deterministic) service times are very much less (typically .02 as large) than the inter-packet spacing within a packet cluster.

The peakedness-based delay approximation was calculated and compared with results obtained via a simulation. For server utilization up to \( \rho = .75 \) the agreement was very good; however, for higher utilizations, the approximation rapidly lost accuracy.

There is an explanation for this behavior of the approximation: at high utilizations the validity of (3.3) in the approximation derivation becomes more questionable. In this queueing system, large delays typically occur when a large number of clusters happen to be overlapping; thus, if an arrival experiences a large delay, the mean number of arrivals during its waiting time is larger than \( U(W) \).

That is, there is positive correlation between an arrival's waiting time and the number of subsequent arrivals during its waiting time which is not captured entirely by the functional relationship \( U(W) \).

5. CONCLUSIONS

In this paper we have demonstrated that a simple delay approximation based on generalized peakedness provides an effective way to approximately predict delays. However, the problem to which the approximation is being applied should be scrutinized for phenomena such as was illustrated in Section 4.3, which can reduce the accuracy of the approximation. The accuracy of the delay approximation seems to depend critically on the validity of the approximation (3.3); if this is approximately true, then the delay approximation can be expected to yield accurate results, especially for relatively large server utilizations. Equation (3.3) is a poor approximation when the arrival stream infrequently enters periods of sustained instantaneous arrival rates which push the server utilization close to or greater than one.

Finally, it has also been shown that the value of \( z_{exp}(1/E[W|W>0]) \) is the most relevant peakedness information, from a single peakedness value, for predicting delays.

6. REFERENCES


