On Defining, Computing and Guaranteeing Statistical Quality-of-Service in High-Speed Networks [1] [2]

Ramesh Nagarajan[3], James F. Kurose[4] and Don Towsley[5]

COINS Technical Report TR 90-123
Computer and Information Science Department
University of Massachusetts Amherst
MA 01003

Abstract

Future high-speed networks (HSNs) are expected to support a wide variety of services such as voice and video and provide a guaranteed quality-of-service (QOS). In this report, we examine in detail the issues of computing and guaranteeing QOS. Traditionally, the computation of user-oriented performance criteria such as the average delay has been carried out via steady state analysis of queueing theoretic models of communication networks. In this paper, we show that the steady state analysis is not entirely sufficient for QOS purposes in future high-speed networks as it yields long-run performance measures that are not appropriate for envisaged applications in HSNs. New QOS criteria for such applications are proposed and their computation detailed for some simple queueing models. One *key result* of this paper is that for *Poisson traffic the carried load in the network may be maintained at fairly high levels*, in comparison to those predicted by standard steady state analysis, while providing fairly good QOS (as defined in this report) to applications. On the other hand, it appears that for *correlated traffic (such as for ADPCM packet voice) large values of the carried load may not be supportable while providing an acceptable QOS.*

[1]This work was supported in part by the Office of Naval Research under contract N00014-90-J-1293 and the Defense Advanced Research Projects Agency under contract NAG2-578/

[2]Presented in part at IEEE INFOCOM'92, Florence, Italy, 1992

[3]Dept. of Electrical and Computer Engineering, UMASS, Amherst

[4]Department of Computer and Information Science, UMASS, Amherst

[5]Department of Computer and Information Science, UMASS, Amherst

KEYWORDS: Quality-of-service, High-speed networks, Statistical guarantees, Transient analysis

1 Introduction

Unlike traditional data networks, future broadband-ISDN (BISDN) wide-area networks will be required not only to carry a broad range of traffic classes but to do so while providing a *guaranteed quality-of-service* to some of these traffic classes. The actual QOS performance metrics of interest are likely to vary from one application to another, but are projected to include such measures as cell loss, delay, and delay jitter guarantees. The need to provide such end-to-end QOS guarantees (essentially a circuit-like performance requirement) while still taking advantage of the resource gains offered by a statistically-multiplexed transport mechanism remains an important, yet largely unsolved problem facing BISDN architects. Indeed, the recent CCITT Study Group XVIII recommendations on BISDN [I.390] and QOS in BISDN [I.388] leave not only the mechanisms for achieving QOS guarantees, but also the very definition of QOS itself, as topics for "further study."

QOS criteria for connections in communication networks are typically guaranteed by performance analysis of appropriate queueing models. The performance analysis of queueing models is usually based on assumptions of stationary traffic and steady state conditions [Rei82] which yields long-run (infinite horizon) performance criteria. However, this approach may be insufficient for the purposes of QOS guarantees for a number of reasons. First, under realistic network conditions connection durations are finite and hence the value of long-run averages in providing QOS guarantees is questionable. Secondly, long-run criteria appear inappropriate for envisaged voice and video services in BISDN [Ram88, RW90, Bra90, Sha90, KV89, Bie91, PZ93]; short-term metrics such as the packet delays in talkspurts [RW90, Bra90] and the packet loss within blocks of packets for packet video [Sha90, KV89, Bie91, PZ93] are deemed more appropriate. We will hence consider the issue of appropriate QOS criteria and its guaranteeing in this report. We will be particularly interested in providing *QOS guarantees* i.e., a measure of "confidence" that the provisioned QOS criteria are indeed being met. Next, we elaborate on some of these issues in the context of specific queueing models.

Consider, for illustrative purposes, a voice multiplexer and connections at this queue each of five minute duration and having a packet loss QOS requirement of 6.66×10^{-03} defined over the connection duration. Standard steady state analysis of this queueing model leads to the requirement that the offered load to the queue be restricted to below 125 sources in order to meet the QOS requirement. This result, however, holds only if the connection durations are infinite. For the finite connection durations in this example, steady state analysis does not provide any guarantees. We will be able to show, however, with the aid of transient analysis that at an offered load of 120 (125) sources approximately 3% (43%) of the five-minute voice calls violate the QOS requirement.

In the above discussion, the QOS criteria was taken to be implicitly defined over the connection duration. In this report, we additionally examine the suitability of defining the QOS metric over the entire connection duration. As pointed out earlier, several researchers [Ram88, RW90, Bra90, Sha90, KV89, Bie91, PZ93] argue that long-run averages are inappropriate for voice and video services and instead espouse finite horizon QOS metrics such as the packet voice delay in talkpurts. In this context, we define new QOS criteria called the *interval QOS* and *block QOS* wherein the QOS criteria is defined over intervals of time and finite groups of connection packets respectively rather than connection durations or infinite horizons. As discussed above, this new QOS criteria appears to be more appropriate for projected services in BISDN. Our interest will be in the QOS criteria defined over these intervals (groups) and in *guaranteeing* that the fraction of intervals (groups) that violate the QOS criteria is below a specified value.

In summary, we will show that steady state queueing analysis of queueing models is not entirely sufficient for the purposes of guaranteeing QOS in future HSNs. We propose appropriate QOS criteria and provide mechanisms for computing QOS criteria and most importantly *guaranteeing* that the QOS criteria will be met.

The rest of this report is organized as follows. Section 2 defines in more detail our performance metrics and QOS criteria. Section 3 details the computational aspects of these performance metrics and QOS criteria. Section 4 considers specific queueing models for illustrating the computation of proposed performance metrics and QOS criteria. Section 5 considers the issue of quality-of-service guarantees. Finally, we summarize our work in Section 6.

2 Performance Metrics and QOS criteria

In this section, we first define two queue performance metrics of interest. These performance metrics are then used to define the QOS criteria. In order to define the performance metrics we introduce some notation. The notation used corresponds to a queueing theoretic model of a communication system. The focus is on a single queue (switch port or multiplexer). In addition to the notation, we also detail some of the assumptions essential to the ensuing analysis.

A_i : Arrival instant of customer i to the queue. It is assumed that the arrival process is stationary. We will assume that the arrival instants and the service times are independent of each other.

$X_i = A_i - A_{i-1}$: Interarrival-time random variable (RV). The interarrival-times are taken to be independent, identically-distributed (iid) RVs with $X_i \sim F(\cdot)$, i.e., the arrival process is a renewal process and $E[X_i] < \infty$. Also, we denote by $F_e(\cdot)$ the equilibrium or residual-lifetime distribution [Wol89] of X_i.

$\{N_t, t \geq 0\}$: A counting process for the arrivals to the queue. We assume that the sample paths of N_t are right continuous.

λ : Arrival rate to queue, $0 < \lambda = 1/E[X_1] < \infty$.

S_i : Service-time of customer i, $\{S_i,\ i > 0\}$ is assumed to be a sequence of i.i.d. RVs.

Service Discipline: The service mechanism is assumed to be FCFS.

K: Buffer capacity of the queueing system.

$\{X(t), t \geq 0\}$: A stochastic process that represents the number in the queueing system (including the one in service, if any). We assume that the sample paths of $X(t)$ are left continuous, thus ensuring that arrivals do not see themselves.

$E_m[Z]$: Conditional expectation of random variable Z given $X(0) = m$.

$\mathrm{Var}_m(Z)$: Variance of random variable Z given $X(0) = m$.

$P_m(Z \leq x)$: Conditional distribution of the random variable Z given $X(0) = m$.

$P_{mn}(t)$: $P(X(t) = n \mid X(0) = m)$.

$E_m^c[Z]$: Conditional expectation of random variable Z given $X(A_1) = m$.

$\mathrm{Var}_m^c(Z)$: Variance of random variable Z given $X(A_1) = m$.

$P_m^c(Z \leq x)$: Conditional distribution of the random variable Z given $X(A_1) = m$.

$P_{mn}^c(i)$: $P(X(A_i) = n \mid X(A_1) = m)$.

ρ : Offered load i.e., ratio of average arrival rate to service rate $(1/E[S_i] > 0)$

$f(.)$: A bounded, measurable, real-valued function defined on the state space of $X(t)$. Unless otherwise specified we assume no particular properties of $f(.)$.

Given our interest in the performance of queues over finite horizons, we need to define an appropriate performance metric over this horizon. Before proceeding to define performance metrics, it is instructive to briefly review some of the work on the transient analysis of queues which focusses on the performance of queues over finite horizons. The work on transient analysis can be broadly classified into two different categories. Most of the work focusses on *time-dependent point statistics* of system random variables such as the mean number in queue at some time t [Mor58, Moo75, Kot78, OR83, As86, Gra77, L$^+$90]. More recent work, however, focusses on *cumulative measures* rather than on point measures.

The work of [Fil88, Gra87, RT88, S$^+$88] details computation of the so-called time averages which are cumulative performance measures over finite time intervals. Grassman [Gra87] considers the computation of the means and variances of time averages in Markovian environments. Smith *et al.* [S$^+$88] propose cumulative measures for studying computer system performance metrics such as "accumulated reward" and "useful work accomplished". Trivedi and Reibman [RT89] evaluate two different techniques for computing time-averaged measures such as "interval availability" and accumulated reward. Our focus in this section will be on such cumulative measures as they directly reflect the user-oriented QOS measures in which we are interested.

We propose two basic performance metrics of interest. We will refer to the two as the *time average* and the *customer average* respectively. Both metrics are appropriate averages over finite horizons. The customer average is the average service (delay, loss etc.,) received by defined groups of customers or by those that arrive over a defined interval while the time average is the average queue performance over the entire time period. The customer average *directly* defines our QOS criteria over connection durations and with some additions also defines a new proposed QOS criteria over intervals and groups of packets within a connection. The time average, in contrast, seems to have limited application as a QOS criteria. However, it will serve to highlight certain aspects of the behaviour of cumulative measures and for comparison with the customer averages.

Customer Average

In order to provide a measure of the customer's quality-of-service we define the customer average as

$$A_C(t_1, t_2) = \frac{\sum_{i=N_{t_1}+1}^{N_{t_2}} f(X(A_i))}{N(t_1, t_2)}, \quad N(t_1, t_2) > 0. \tag{1}$$

where $f(X(A_i))$ is a function of the system state as seen by the ith arrival and $N(t_1, t_2) = N_{t_2} - N_{t_1}$; the customer average is then simply the queue performance averaged over arrivals in the time interval (t_1, t_2). The customer average is undefined for $N(t_1, t_2) = 0$. We also propose an alternative metric in terms of arrivals

$$A_p(n_1, n_2) = \frac{\sum_{i=n_1}^{n_2} f(X(A_i))}{(n_2 - n_1 + 1)}, \quad n_2 > n_1 \tag{2}$$

where (n_1, n_2) is the block of arrivals over which the average is computed.

A QOS metric of considerable interest is the fractional loss [N+91]. In order to evaluate this metric define

$$f(x) = \begin{cases} 1 & x = K, \\ 0 & \text{Otherwise.} \end{cases} \tag{3}$$

The customer average with $f(.)$ as defined above yields the fractional loss over finite intervals or a finite number of arrivals. The above fractional loss metric will be of primary interest in this report.

To illustrate the computation of alternate QOS metrics we consider also the average number metric. Here, we define

$$f(x) = x \tag{4}$$

We will primarily use this metric to contrast the time and customer averages rather than as a QOS criteria. If the interest is in delay-based metrics such as the average delay then alternate system state descriptors such as the virtual work will have to be considered. We will not consider such a formulation here but note that for the case that the service-time distribution is deterministic such as in ATM-based networks [I.188] the above definition suffices for delay-based metrics as well.

The *connection-duration based QOS criteria* is now simply defined as $A_C(0, T)$ where T is the connection duration and the connection is assumed to commence at $t = 0$. Note, however, that $A_C(0, T)$ is the performance averaged over the aggregate stream of arrivals. However, our interest is in the QOS for individual connections that compose this aggregate stream rather than the composite stream. The customer average over intervals (arrivals) is not necessarily identical, in general, for the composite stream and the individual streams composing it even if the streams are identical. We will, however, return to this issue in Section 5 and indicate as to how the value of the customer average evaluated for the aggregate stream may be employed for the individual streams.

Next, we define an *interval QOS criteria* for connections by considering intervals of time rather than the connection duration. The definition of this QOS criteria utilizes the basic definition of the customer average over intervals. Consider a connection of duration T and intervals of size Δ. Define,

$$I_{(t_1,t_2)} = \begin{cases} 1 & A_C(t_1, t_2) \text{ violates the QOS criteria} \\ 0 & \text{Otherwise} \end{cases} \tag{5}$$

That is, the indicator function takes the value 1 if the QOS criteria is *not met* during the interval (t_1, t_2). We then define the interval QOS as,

$$A_I(T, \Delta) = \frac{\sum_{i=1}^{L} I_{((i-1)\Delta, i\Delta)}}{L} \tag{6}$$

where $L = \lfloor \frac{T}{\Delta} \rfloor$, assuming the connection starts at $t = 0$. Hence, $A_I(T, \Delta)$ is the fraction of intervals that violate the QOS criteria. Alternatively, we could consider groups of arrivals (packets) based on the customer average over arrivals. Consider a connection that generates P packets. Define,

$$B_{(k_1,k_2)} = \begin{cases} 1 & A_p(k_1, k_2) \text{ violates the QOS criteria} \\ 0 & \text{Otherwise.} \end{cases} \tag{7}$$

That is, the indicator function takes the value 1 if the QOS criteria is *not met* over the block of packets (k_1, k_2). We then define the block QOS as,

$$A_B(P, n) = \frac{\sum_{k=1}^{M} B_{((k-1)n+1,kn)}}{M} \tag{8}$$

where $M = \lfloor \frac{P}{n} \rfloor$.

The interval QOS has been proposed by organisations such as CCITT and CCIR for data and voice services [GL83, YW89, G.880] in the context of error performance of digital transmission systems. More recently, a proposal by the Internet Research Task Force (IRTF) [Par92] recommends an interval-based loss QOS metric. Woodruff *et al.* [WK90] suggest the evaluation of performance metrics such as the duration and occurence of high loss periods. Ramaswamy and Willinger [RW90] and Bradlow [Bra90] propose that the delay in talkspurts for the ADPCM voice source be studied rather than the typical long-run averages. Such metrics are proposed as more appropriate for the engineering of playout buffers at the receiver. The interval and block-based QOS criteria will also be useful for the design and analysis of forward error correction schemes such as the Block encoding scheme proposed in Karlsson *et al.* [KV89], Shacham [Sha90] and Biersack [Bie91]. Finally, Pancha and Zarki [PZ93] propose the fractional packet loss in a video frame as a measure of QOS for evaluating certain bandwidth reservation mechanisms for MPEG video sources. The nature of the applications in HSNs will thus require the network service provider to guarantee transmission performance over shorter time scales of burst and block lengths. Hence, it seems natural and more meaningful to define the QOS over these time scales rather than over the entire connection.

Time Average

The time average is defined as a random variable,

$$A_T(t) = \frac{\int_0^t f(X(s))ds}{t} \tag{9}$$

Several researchers [RT89, S$^+$88, Gra87] have investigated in detail the efficient computation of the time averages for Markovian queueing systems and hence our treatment here is brief. We, however, do consider it briefly and provide some numerical examples.

To illustrate the usefulness of the time average consider the following definition of $f(.)$,

$$f(x) = \begin{cases} 1.0 & x > 0 \\ 0.0 & \text{Otherwise} \end{cases} \tag{10}$$

Hence $f(.)$ is the instantaneous server (link) utilization of the queueing system. Substituting in equation (9) yields the server utilization over the appropriate time period (the CPU (server)

utilization over time intervals is a commonly used metric in the evaluation of computer system performance, for e.g., see [H+88] as well as communication system performance, for e.g., see [Zha90, BZ90]). We shall consider this performance metric briefly in our analysis of the $M/M/1/K$ queue.

3 Computing performance metrics and QOS criteria

In this section, we consider various issues in the computation of the time and customer averages defined in the previous section. We present a general approach towards this computation and in later sections will illustrate the technique for specific queueing models. Since the time and customer averages are random variables, we will be concerned primarily with computing some distributional information for the customer averages. Moments such as the mean and variance of the customer average are computed in Appendices A and B. We will also consider the complexity of our analysis for the customer average. We make only brief comments on the time averages. Finally, we outline the computation of the interval and connection-based QOS criteria.

3.1 Performance metrics

First, we consider time averages. The computation of the mean and variance of the time averages for Markovian queueing systems has been considered previously [S+88, Gra87, RT89]. Hence, we will not consider this performance metric in detail. However, in the case of the $M/M/1/K$ queue, closed-form expressions for the time averages are available and we will use these to outline the behaviour of the time averages and compare it with that of the corresponding customer averages.

We next consider the customer average. The analysis, in the following, considers primarily the $M/M/1/K$ queue. Extensions to the $M/G/1/K$ and $GI/M/1/K$ queues are discussed at the end of this section and in Section 4. Our interest will be in computing performance metrics over an arbitrary finite time-interval $(0,t)$ with $X(0)=m,\ m=0,1,2,\cdots,K$.

Note that $\{X(A_i),\ i\geq 1\}$ for the $M/M/1/K$ queue is a discrete-time Markov chain (DTMC) on a finite state space. We assume that it is aperiodic, irreducible and hence positive-recurrent. Let $\lim_{i\to\infty} P(X(A_i)=k)=p_k,\quad k=0,1,\cdots,K$. The transition probability matrix $\boldsymbol{P}$ of the DTMC that governs its evolution can be computed as in [GH85]. Also, $X(t)$ is regenerative with finite mean cycle length and let $\lim_{t\to\infty} P(X(t)=k)=\pi_k,\ \ k=0,1,\cdots,K$.

Distribution of the customer average

We consider the distribution of the customer average and in particular the computation of the mass at zero i.e., the probability of the customer average being zero (this happens for example when there is no loss in an interval in the case of the fractional loss metric). The utility of this performance metric will be clear when we consider QOS guarantees in Section 5. We assume in the following that $f(x)\geq 0$.

$P_m(A_C(0,t)=0)$ can be computed as

$$P_m(A_C(0,t)=0|N_t>0) \;=\; \sum_{n=1}^{\infty} P_m(\frac{\sum_{i=1}^{n} f(X(A_i))}{n}=0|N_t=n)P(N_t=n|N_t>0)$$

$$\approx \sum_{n=1}^{\infty} P_m(\sum_{i=1}^{n} f(X(A_i)) = 0) P(N_t = n | N_t > 0)$$
$$= \sum_{n=1}^{\infty} P_m(\bigcap_{i=1}^{n} f(X(A_i)) = 0) P(N_t = n | N_t > 0). \quad (11)$$

The above equation is only an approximation (referred to as Approx1) as we have ignored the dependence of the system state distribution observed by a particular arrival on the number of arrivals in t (note for example that with a large number of arrivals in t it is more likely that an arrival will see a busy server). Now Equation (11) is just a weighted sum of probabilities. This can be evaluated to a desired accuracy by appropriate truncation of the infinite series [RT88] (see also Appendix E).

We also consider a second approximation (referred to as Approx2) for the customer averaged performance metric. For an arbitrary time instant t, define

$$N = \lfloor E[N_t] \rfloor = \lfloor \lambda t \rfloor \quad (12)$$

The second approximation then is:

$$P_m(A_C(0,t) = 0 | N_t > 0) = \sum_{n=1}^{\infty} P_m(\frac{\sum_{i=1}^{n} f(X(A_i))}{n} = 0 | N_t = n) P(N_t = n | N_t > 0)$$
$$\approx P_m(f(X(A_1)) = 0, \cdots, f(X(A_N)) = 0) \quad (13)$$

The joint probability term in the above equation is easily computed by appropriately grouping states and considering corresponding submatrices of the transition probability matrix $\boldsymbol{P}$. The analysis is detailed in Appendix C. The mass at zero converges to zero as $t \to \infty$, as is expected; the proof is presented in Appendix C.

The computation of the customer average over arrivals follows in the same manner as for the customer average over intervals and hence is omitted. Note that our second approximation is essentially the customer average over arrivals if we interpret N to be the number of arrivals (rather than the expected number of arrivals in a given time interval) over which the averaging is being carried out. The computation is *exact* in this case.

Finally, we comment on the analysis for the more general $GI/M/1/K$ and $M/G/1/K$ queues. The analysis for the $GI/M/1/K$ queue is a rather straightforward extension of the above analysis for the $M/M/1/K$ queue modified to account for the first arrival in the interval. The time to the first arrival should then be drawn from the residual inter-arrival time distribution[6], $F_e(\cdot)$, rather than from the regular inter-arrival time distribution. The remainder of the analysis remains identical.

For the $M/G/1/K$ queue further approximations are necessary since $X(A_i)$ is not Markovian. The nature of the approximation depends on the particular service-time distribution. For example, in the $M/E_r/1/K$ queue the number of phases of service in the system, $\{X_p(A_i), i > 0\}$, evolves as a DTMC and can be used to evaluate statistics of $f(X(A_i))$ at the arrival instants (see Appendix H).

[6]The time to the first arrival may actually depend on the initial system state $X(0)$. However, it is known that for the $GI/M/1$ queue, a lack of bias assumption [MW90] holds, i.e., the conditional intensity of the arrival process at time t given $X(t)$ and $X(t)$ are uncorrelated. This suggests that perhaps the time to the first arrival after time t given $X(t)$ is independent of $X(t)$. If this is not the case, the above may be viewed as an additional approximation.

For the $M/D/1/K$, we assume that the residual service-time of the customer being serviced at arrival instants (if any) is drawn from an uniform distribution at each arrival instant (see Section 4.2 for further details).

Computational Complexity

Finally, we comment on the computational complexity of the analysis presented above. The truncation of the infinite series in equation (11) yields $O(\lambda t)$ terms, for a fixed accuracy requirement (proved in Appendix F), while computing the customer average for time interval t. The computation of the customer-based statistics, such as $P_m(\cdots)$, involve $O(\lambda t)$ matrix multiplications of $(K+1)\times(K+1)$ matrices. Hence, the computation time and number of arithmetic operations are $O(K^3\lambda T)$ (where T is the maximum time value for which one wishes to compute the customer average). Although this value does increase relatively quickly with buffer size, the computation yields the performance for all initial occupancies (values of m).

3.2 QOS criteria

Here we consider the computation of the interval-based (packet-group based) and connection-based QOS criteria. The QOS criteria over connection durations is the customer average whose computation has been outlined in detail in the previous section. We consider first the QOS criteria computed over packet groups and then the interval-based measure.

Our interest is in

$$\begin{aligned} E_m^c[A_B(P,n)] &= \frac{\sum_{k=1}^{M} E_m^c[B_{((k-1)n+1,kn)}]}{M} \\ &= \frac{\sum_{k=1}^{M}\sum_{l=0}^{K} P_l^c(B_{(1,n)} \text{ violates QOS criteria}) P_{ml}^c((k-1)n+1)}{M}. \end{aligned} \tag{14}$$

Since $\lim_{k\to\infty} P_{ml}^c((k-1)n+1) = p_l$, we have

$$\lim_{P\to\infty} E_m^c[A_B(P,n)] = \sum_{l=0}^{K} P_l^c(B_{(1,n)} \text{ violates QOS criteria}) p_l. \tag{15}$$

Similarly, we have for the interval-based metric that

$$\lim_{T\to\infty} E_m[A_I(\Delta,T)] = \sum_{l=0}^{K} P_l(A_C(0,\Delta) \text{ violates QOS criteria})\pi_l. \tag{16}$$

4 Analysis of Queueing Models

In this section, we illustrate the computation of the time and customer averages for some simple queueing models. First, we consider the $M/M/1/K$ queue. Our purpose in examining the $M/M/1/K$ queue is two-fold. Closed-form expressions for the time-averaged behaviour in the

$M/M/1/K$ queue will be used to demonstrate qualitative convergence characteristics of the cumulative performance measures. Then we will contrast the behaviour of the customer and time averages in this queue.

Since the target transport mechanism for BISDN is the ATM scheme [I.188] which is characterized by fixed size packets, we consider the $M/D/1/K$ queue next. This queue better models switch ports and multiplexers operating under the ATM scheme than the $M/M/1/K$ queue.

4.1 The $M/M/1/K$ queue

In the $M/M/1/K$ queue the interarrival and service times form a sequence of i.i.d exponential RVs which are mutually independent.

First, we consider time averages. Although the time-averages have limited usefulness as a QOS criteria, we will use the time-averages to demonstrate several interesting properties of the cumulative measures. First, we will use the time-averaged utilization to highlight some convergence characteristics of cumulative measures to steady state values. Second, we will contrast the time and customer averages. Finally, we evaluate the accuracy of the analysis for the customer averaged loss in this queue.

Our starting point will be the closed-form solution for the transient state (number in system) probabilities of the $M/M/1/K$ queue. This is derived in [Mor58] and also appears in [RT88]. From [Mor58], we have for the time-dependent state probabilities,

$$\begin{aligned} P_{mn}(t) &= \pi_n + \frac{2\rho^{\frac{n-m}{2}}}{K+1}\sum_{i=1}^{K}\frac{1}{x_i}((\sin(\frac{im\pi}{K+1}) - \sqrt{\rho}\sin(\frac{i(m+1)\pi}{K+1})) \\ &\quad \times((\sin(\frac{in\pi}{K+1}) - \sqrt{\rho}\sin(\frac{i(n+1)\pi}{K+1}))e^{-\gamma_i t} \end{aligned} \tag{17}$$

Here, γ_i are the eigenvalues of the system (infinitesimal generator of $X(t)$) given by [Mor58],

$$\begin{aligned} \gamma_i &= \lambda + \mu - 2\sqrt{\lambda\mu}\cos(\frac{i\pi}{K+1}) \\ x_i &= \frac{\gamma_i}{\mu} \quad 1 \le i \le K \end{aligned} \tag{18}$$

Now define $f(.)$ as in equation (10). Then the mean time averaged utilization is obtained as

$$\begin{aligned} E_m[A_T(t)] &= E_m[\frac{\int_0^t f(X(t))dt}{t}] \\ &= \frac{\int_0^t E_m[f(X(t))]dt}{t} \end{aligned} \tag{19}$$

Using the definition of $f(.)$ we have,

$$\begin{aligned} E_m[f(X(t))] &= P(X(t) > 0 \mid X(0) = m) \\ &= \sum_{n=1}^{K} P_{mn}(t) \end{aligned} \tag{20}$$

Table 1: Distribution of customer averaged loss in a M/M/1/10 queue

Time t (service units)	$\rho = 0.8, m = 10$	
	$P_m(A_C(t) = 0)$	Simulation (90 % C.I.)
10.000	0.2939	0.2919 (8.725e-03)
20.000	0.2499	0.2339 (2.203e-02)
30.000	0.2308	0.2312 (6.936e-03)
40.000	0.2174	0.2181 (6.793e-03)
50.000	0.2059	0.2043 (6.633e-03)
100.0	0.1580	0.1588 (7.185e-03)

Hence the mean time-averaged utilization is easily evaluated as

$$E_m[A_T(t)] = \sum_{n=1}^{K} \pi_n + \sum_{n=1}^{K} \frac{2\rho^{\frac{n-m}{2}}}{K+1} \sum_{i=1}^{K} \frac{1}{x_i}((\sin(\frac{im\pi}{K+1}) - \sqrt{\rho}\sin(\frac{i(m+1)\pi}{K+1})) \times((\sin(\frac{in\pi}{K+1}) - \sqrt{\rho}\sin(\frac{i(n+1)\pi}{K+1}))\frac{1-e^{-\gamma_i t}}{\gamma_i t} \tag{21}$$

Consider now the convergence of the time-averaged utilization to the steady-state value with time. The time-dependent term in the equation above is of the form $\frac{1-e^{-\gamma t}}{\gamma t}$. The exponential term dominates at small time while at large time scales the behaviour is $O(1/t)$. Contrast this to the general observation in [OR83] that the time-dependent values approach the steady state values exponentially fast for large time. *Clearly, the behaviour of cumulative measures are far different from the time-dependent behaviour of point values explored in previous papers.*

We next consider the mean time-averaged number in queue and compare it to the mean customer-averaged number. The mean number at time t in the $M/M/1/K$ queue is obtained as,

$$E_m[X(t)] = \sum_{j=1}^{K} j P_{mj}(t) \tag{22}$$

with $P_{mj}(t)$ as defined earlier. The mean time-averaged number is then obtained as

$$E_m[A_T(t)] = \frac{\int_0^t E_m[X(s)]ds}{t} \tag{23}$$

The mean customer-averaged number is easily evaluated as in Appendix A. Figure 1 compares the mean time and customer-averaged number in queue. We know from PASTA that the customer and time averages are identical over an infinite horizon for Poisson arrivals. However, it is clear from the figure that they are not identical except for large values of t. *Hence, for finite time intervals the customer and time averages will have to be computed separately as in this report.*

Finally, we numerically illustrate the computation of the mass at zero for the customer averaged loss which was detailed in Section 3.1. The probability of no loss for the $M/M/1/10$ and $M/M/1/20$ queues are shown in Tables 1, 2 and 3. It is seen that the analytical values are fairly accurate.

Table 2: Distribution of customer averaged loss in a $M/M/1/20$ queue

Time t (service units)	$\rho = 0.625, m = 20$	
	$P_m(A_C(t) = 0)$	Simulation (90 % C.I.)
10.0	0.4269	0.4084 (1.237e-02)
20.0	0.3948	0.3827 (1.066e-02)
30.0	0.3844	0.3845 (8.653e-03)
40.0	0.3795	0.3871 (8.013e-03)
50.0	0.3776	0.3819 (7.993e-03)
100.0	0.3752	0.3685 (7.936e-03)

Table 3: Distribution of customer averaged loss in a $M/M/1/20$ queue

Time t (service units)	$\rho = 0.625, m = 10$	
	$P_m(A_C(t) = 0)$	Simulation (90 % C.I.)
50.000	0.99618	0.9961 (1.116e-03)
60.000	0.99552	0.9955 (1.101e-03)
80.000	0.99476	0.9938 (1.291e-03)
100.00	0.99443	0.9934 (1.332e-03)

4.2 The $M/D/1/K$ queue

The $M/D/1/K$ queue is characterized by Poisson arrivals and deterministic service times i.e., the A_i are iid exponential and the service times are iid deterministic. We consider the customer averages only here.

Because service-times are deterministic, $X(A_i)$ in the $M/D/1/K$ queue is not described by a Markovian process. This considerably complicates the analysis without any approximations. Consider the number in the system at arrival instants. The one-step transition probability matrix for $X(A_i)$, i.e., the distribution of $X(A_{i+1})$ given $X(A_i)$, is exactly computed since an arbitrary arrival observes a residual service-time that is uniformly distributed on $[0, \tau]$, where τ is the deterministic service-time. However, this is insufficient to calculate the occupancy distribution, $X(A_{i+n})$ $n > 1$, seen by future arrivals since it requires knowledge of the particular residual service time seen by this random arrival i and those seen by future arrivals. We ignore this complication and take the distribution seen by the nth arrival to be the n-fold product of the one-step transition matrix. A similar approximation is employed by Yuan and Silvester [YS89] in a different context. We conjecture that the approximation is pessimistic in general since it effectively elongates the service time (by choosing the residual service time uniformly for each arrival) and hence overestimates the occupancy distribution. However, we conjecture that this effect will be insignificant at lower traffic intensities since it is unlikely that consecutive arrivals will see the same customer in service. The computation of the one-step transition probability matrix is detailed in Appendix I. The remainder of the analysis is identical to that in Section 3.1.

The approximation for the probability of no loss for the $M/D/1/10$ and the $M/D/1/40$ queue with different initial occupancies is shown in Tables 4, 5, 6 and 7 respectively. We see that our approximation for the probability of no loss, $P_m(A_C(t) = 0)$, agrees well with simulation.

Table 4: Distribution of customer averaged loss in a M/D/1/10 queue

Time t (service units)	$\rho = 0.8, m = 10$	
	$P_m(A_C(t) = 0)$	Simulation (90 % C.I.)
10.0	0.25996	0.2445 (7.071e-03)
20.0	0.22883	0.2203 (6.818e-03)
30.0	0.21687	0.2102 (6.704e-03)
40.0	0.21034	0.2031 (6.618e-03)
50.0	0.20575	0.2071 (6.666e-03)
100.0	0.18849	0.1901 (6.455e-03)

Table 5: Distribution of customer averaged loss in a M/D/1/10 queue

Time (service units)	$\rho = 0.8, m = 4$	
	$P_m(A_C(t) = 0)$	Sim. (90% C.I.)
10.0	0.99631	0.9867 (1.886e-03)
20.0	0.97629	0.9639 (3.065e-03)
30.0	0.95491	0.9554 (3.396e-03)
40.0	0.93647	0.9427 (3.823e-03)
50.0	0.91978	0.9269 (4.279e-03)
100.0	0.84466	0.8734 (5.47e-03)

Table 6: Distribution of customer averaged loss in a M/D/1/10 queue

Time t (service units)	$\rho = 0.6, m = 10$	
	$P_m(A_C(t) = 0)$	Simulation (90 % C.I.)
10.0	0.42360	0.4144 (8.34e-03)
20.0	0.40646	0.4009 (8.062e-03)
30.0	0.40226	0.3955 (8.044e-03)
40.0	0.40085	0.4021 (8.066e-03)
50.0	0.40023	0.4031 (8.069e-03)
100.0	0.39882	0.3984 (8.053e-03)

Table 7: Distribution of customer averaged loss in a M/D/1/40 queue

Init. occ. (m)	$\rho = 0.8, t = 50$	
	$P_m(A_C(t) = 0)$	Simulation (90 % C.I.)
0	1.0	1.0 (0.0)
10	1.0	1.0 (0.0)
20	1.0	1.0 (0.0)
25	0.99956	0.9990 (1.645e-03)
30	0.99076	0.9869 (5.89e-03)
35	0.89621	0.9060 (1.519e-02)
40	0.20742	0.2189 (2.152e-02)

5 Quality-of-Service Guarantees

In this section, we revisit the computation of the customer-based and interval-based QOS criteria outlined in Section 3.2. Our interest here will be on mechanisms for ensuring that the specified QOS criteria are indeed being met i.e., in *guaranteeing QOS criteria.* We consider first the issue of guaranteeing the connection-based QOS criteria and subsequently the interval-based QOS criteria.

In order to investigate the issue of guaranteeing the connection-based QOS for finite-duration connections, we propose and analyze the following simple connection model. We assume a *fixed* number of connections that have *identical* fixed connection durations. We also assume that the connection traffic characteristics are stochastically identical. The constant value of the number of connections is maintained by immediately replacing a departing connection with an identical connection.

Given the above scenario, our interest will be in *bounding* the fraction of connections that violate the QOS criteria defined over the finite connection duration. We will refer to the fraction of connections that meet the specified QOS criteria as the *guarantee level* and the QOS criteria as the *QOS requirement.* The fraction of connections that *violate* the QOS requirement will be referred to as the *violation level.* Hence, for a given QOS requirement the larger the value of the guarantee level, or alternatively the smaller the violation level, the better the "quality" of the service. Before detailing the computation of the guarantee levels, we introduce additional notation

T : Connection duration in packet transmission (service) time units and, where appropriate, in real-time units.

Q^c : The real-valued QOS requirement over the connection duration. For example, Q^c may be a loss probability of 10^{-03} defined over arrivals from a connection of 5 minute duration. We will adopt the notation of a two-tuple for the QOS requirement when there two sets of results corresponding to two different values of the QOS requirement. This will be the case, for example, when we compare the Poisson and voice multiplexers in Section 5.3.

Q^I : The real-valued QOS requirement over a specified interval duration.

$G^c(T, Q^c)$: Guarantee level i.e., the fraction of connections that meet the QOS requirement.

$V^c(T, Q^c)$: Violation level i.e., the fraction of connections that violate the QOS requirement.

$G^I(\Delta, Q^I)$: Guarantee level i.e., the fraction of intervals of size Δ that meet the QOS requirement Q^I.

$V^I(\Delta, Q^I)$: Violation level i.e., the fraction of intervals that violate the QOS requirement Q^I.

L^s : Steady state loss probability at a given offered load.

The guarantee level is then computed as

$$G^c(T, Q^c) \;=\; 1 - \sum_{i=0}^{K} P_i(A_C(0,T) > Q^c)\pi_i \tag{24}$$

and the violation level is simply $V^c(T, Q^c) = 1 - G^c(T, Q^c)$. Our interest in this section will be in the packet loss metric and in particular in guaranteeing small values of the packet loss metric. The

acceptable values of the packet loss ratio depend to a great extent on the coding schemes, nature of the error detection and correction techniques and on the nature of the service itself and hence is beyond the scope of this report. However, packet loss QOS requirements of 10^{-04} to 10^{-09} may be expected [YW89]. With regard to the guarantee levels for the packet loss QOS requirement, we will concern ourselves with computing a bound rather than the exact value of the guarantee level. One such bound is computed as

$$G^c(T, Q^c) \geq 1 - \sum_{i=0}^{K} P_i(A_C(0,T) > 0)\pi_i \tag{25}$$

and

$$V^c(T, Q^c) \leq \sum_{i=0}^{K} P_i(A_C(0,T) > 0)\pi_i \tag{26}$$

For the small packet loss values of interest, we conjecture that the above bounds will be tight. Table 8 shows the distribution of the customer averaged loss for the $M/D/1/10$ queue. It is readily seen that this distribution is bi-modal (for small time values), i.e., there is significant mass at zero and for values of the loss greater than 10^{-3}. Hence, for loss QOS values in the range $(0, 10^{-03})$, the value of $P_m(A_C(0,t) > 0)$ should serve as a tight bound for the actual probability mass at that value.

Now, the numerical computation of equation (25) requires the computation of the probability mass at zero for the customer-averaged packet loss. The computation of the former is detailed in Section 3.1. However, it is important to note that the computation of the mass at zero in Section 3.1 is only approximate and hence the computed guarantee and violation levels will not, strictly, constitute bounds.

Next, we consider the interval-based QOS criteria. In order to compute the interval-based QOS criteria we assume a connection model that is essentially the same as that for the connection-based QOS criteria above. We assume here, however, that the connection durations are infinite. The computation of the interval-based QOS was outlined in Section 3.1. Owing to our assumption of infinite connection durations we have that (see equation (16)),

$$\begin{aligned} V^I(\Delta, Q^I) &= \sum_{n=0}^{K} P_n(A_C(0,\Delta) > Q^I)\pi_n \\ &\leq \sum_{n=0}^{K} P_n(A_C(0,\Delta) > 0)\pi_n \end{aligned} \tag{27}$$

and

$$G^I(\Delta, Q^I) \geq 1 - \sum_{n=0}^{K} P_n(A_C(0,\Delta) > 0)\pi_n \tag{28}$$

Hence the guarantee level of the interval QOS for infinite duration connections is that for the connection-based QOS value with the connection duration set equal to the size of the interval in the definition of the interval QOS. Next, we present a few numerical examples to illustrate the nature of our guarantees. First, we consider the connection-duration based criteria and then the interval-based criteria.

Table 8: Customer averaged loss in a M/D/1/10 queue

Time t (service units)	$\rho = 0.8, m = 10$	
	Range $((x_1, x_2])$	$P_m(x_1 < A_C(t) \leq x_2)$
10.000	$(0.0, 0.0]$	0.25
	$(0.05, 0.1]$	0.0226
	$(0.1, 0.2]$	0.2981
	$(0.2, 0.4]$	0.3563
	$(0.4, 0.6]$	0.0698
	$(0.6, 0.8]$	2.687e-03
	$(0.8, 1.0]$	3.2248e-03
20.000	$(0.0, 0.0]$	0.23
	$(0.01, 0.05]$	0.0076
	$(0.05, 0.1]$	0.2689
	$(0.1, 0.2]$	0.3209
	$(0.2, 0.4]$	0.1692
	$(0.4, 0.6]$	0.0028
	$(0.6, 0.8]$	0.0
50.000	$(0.0, 0.0]$	0.2072
	$(0.01, 0.05]$	0.3501
	$(0.05, 0.1]$	0.2896
	$(0.1, 0.2]$	0.1444
	$(0.2, 0.4]$	0.0087
	$(0.4, 0.6]$	0.0
500.0	$(0.0, 0.0]$	0.118
	$(1e-03, 5e-03]$	0.252
	$(5e-03, 1e-02]$	0.286
	$(1e-02, 5e-02]$	0.344
1000.0	$(0.0, 0.0]$	0.072
	$(1e-03, 5e-03]$	0.442
	$(5e-03, 1e-02]$	0.308
	$(1e-02, 5e-02]$	0.178

Table 9: Violation levels for the aggregate and individual streams in a $M/D/1/60$ queue

Offered load (Streams)	$Q^c = 2.56 \times 10^{-6}$		$Q^c = 1.0 \times 10^{-4}$	
	Aggregate	Individual	Agg.	Ind.
0.9515 (130)	18.4 ± 2.02	1.38 ± 0.27	16.5 ± 1.93	1.38 ± 0.27
0.988 (135)	93.0 ± 1.33	30.28 ± 1.07	45.98 ± 1.29	30.28 ± 1.07

5.1 $M/G/1/K$ queues: Analytical results

In all of the numerical examples, we take the QOS requirement to be a fractional packet loss of 10^{-9}. However, note that the violation and guarantee levels computed as above constitute upper and lower bounds for all values of the packet loss QOS requirement. Figure 2 shows the fraction of connections that violate the QOS requirement for various values of the offered load to the $M/M/1/20$ queue. The connection duration is taken to be 1000 units. We see that large values of the offered load, in comparison to those predicted by steady state analysis, can be supported at the expense of a small fraction of connections that will violate the QOS criteria. Similar results for the $M/D/1/40$ queue are shown in Figure 4.

Next, we examine the effect of connection duration on the violation level. In Figure 3 we see that the fraction of connections violating the QOS requirement grows relatively slowly with the connection duration for a fixed value of the offered load (the particular value in Figure 3 corresponds roughly to the knee of the curve in Figure 2 which might be a desirable operating point for this queue since only a small fraction of the connections violate the QOS requirement at this load level). Hence, we can expect low violation levels for large connection durations as well.

In the case of the interval QOS, as in the case of the connection-based QOS, with a fixed desired interval size it is possible to support larger values of the offered load at the expense of a small fraction of intervals (periods of time) that violate the QOS requirement. Also, for a given desirable value of the offered load it is feasible to bound (guarantee) the fraction of intervals that violate a given QOS requirement. In Figure 5 we show the violation level as a function of the interval size in the $M/D/1/40$ queue at an offered load of 90%. It must be fairly clear to the reader that such interval-based measures are beyond the scope of standard steady-state analysis.

Note that the analysis above was carried out by computing violation levels and other performance metrics for the aggregate stream (see Section 2). Consequently the results apply only when the workload consists of a single connection. The more interesting case is when there are multiple connections. Hence, it is of interest to question whether the analysis for the aggregate stream to the queue may still be applicable to the individual streams (at least when the streams are identical). Certainly, for long connection durations the two would be identical. For short connection durations, however, the relation between the two is not clear. Limited experimentation indicates that the violation level for the aggregate stream is higher than for the individual streams and may hence be employed as a pessimistic approximation. Tables 9 and 10 show the violation level in the $M/D/1/60$ queue for the aggregate stream and the individual streams for $T = 5$ sec.. It can be seen that the violation level for the aggregate stream is much higher than that for the individual streams.

Table 10: Violation levels for the aggregate and individual streams in a $M/D/1/60$ queue

Offered load (Streams)	$Q^c = 1.0 \times 10^{-2}$	
	Aggregate	Individual
0.9515 (130)	0.16 ± 0.09	0.26 ± 0.12
0.988 (135)	13.57 ± 0.89	11.72 ± 0.75

5.2 Packet Voice Multiplexer: A simulation study

In this section, we focus on a packet voice multiplexer and evaluate the proposed QOS criteria for this model. First, we briefly discuss the voice source model and the voice multiplexer itself. Then we consider the computation of the violation level for voice calls as a function of the offered load, connection duration, and loss QOS requirements.

The model we assume for a voice source is a standard one (see, e.g., [DL86, HL86, SW86, N$^+$91]) and has as its basic premise that an active voice source periodically generates fixed length packets when a speaker is speaking (talkspurt) and otherwise remains idle. We briefly describe this model here; the reader is referred to the above references, in particular [SW86], for additional details and discussion. The voice packetization period is assumed to be fixed at $T = 16$ msec. and the talkspurt is assumed to contain a geometrically distributed number of packets, with mean 22 packets. The mean length of a talkspurt is thus $\alpha^{-1} = 352$ msec. The period between talkspurts, known as the silence period and denoted by X, is assumed to be exponentially distributed with a mean length of $\beta^{-1} = 650$ msec.. The speech activity ratio, which is the fraction of time that the voice source is active, is thus 0.351 and each source generates on the average 22 packets every second. Given the above model, the interarrival times between packets generated by a *single* source form a renewal process. With probability 1/22, the interarrival time is 16 msec. and with probability 21/22, the interarrival time is $16 + X$ msec. [SW86].

The input traffic to the multiplexer is a superposition of a finite population of M voice sources, each of which is as characterized above. In our loss calculations in the following, we will assume 64 byte packets and that the voice sources are being multiplexed over a T-1 link. This leads to a packet transmission time of 1/3 msec.. A link utilization of 100% would then result from a superposition of approximately 136 voice sources. Also, in all of the results we assume a buffer size of 60 for the voice multiplexer.

Owing to the considerable complexity of the analysis of the voice multiplexer [HL86, N$^+$91, SW86], we resort to simulation. Before proceeding to the results, we make a few comments regarding our simulation. The simulation results obtained in this section are obtained by the method of independent replications with the transients discarded in each replication. For each run, a transient phase of 6 seconds was discarded (Experiments indicated that the ensemble packet loss probababilities stabilized around the steady state value after about 2 secs.; we choose 6 secs. as a conservative estimate). The length of each run was essentially the connection duration value. The number of independent replications ranged from 100 to 10000 depending on the connection durations involved. The 90% confidence intervals were computed in all cases.

With regard to the simulation, another point worthy of mention is that the voice connection is taken to commence with the source in a random state (talkspurt or silence, with the probability of being in talkspurt being equal to the activity factor [SW86, HL86] of the voice source) at the start of

the connection. Given the large number of sources that we consider we expect this assumption not to have a very significant effect except, perhaps, for very small connection durations. As a point of interest, we refer the reader to Heyman et al. [DHL92] in which it is shown that the initial phasing of video sources has a significant impact on the cell loss rate of individual sources. However, unlike the video source where this initial phasing among sources is preserved for the connection duration [DHL92], in the case of the voice source any such initial phasing effects are effectively destroyed by the occurence of silence periods of random duration. We will hence ignore the effect of the initial connection state in the following results.

We present two sets of results. First, we examine the violation level in the voice multiplexer as a function of the offered load, with the QOS requirement at each load being the steady state loss value at that load. Second, the loss QOS requirement is taken to be fixed and we then examine the violation level as a function of the connection duration. In most of our results, we will show the connection durations in real-time units rather than service units for a better appreciation of the time-scales involved. The corresponding value of the connection duration in service units may be obtained by multiplying the connection duration by a factor of 3000. We show in the following that the voice multiplexer exhibits severe violations of the loss QOS requirement for load values wherein steady state analysis predicts no violations of the loss requirement.

In Figure 6, we see the violation level at different load levels with the loss QOS value at each load value taken to be the steady state loss value at that load. Hence, we can for example see that as many as 52% of 5 minute long connections violate a loss QOS value of 2.71×10^{-02} at a load level of 135 sources. This is in contrast to steady state computations that predict that 100% of infinite duration connections will meet this desired loss QOS requirement.

In Figure 7 we see the violation level at different load levels as a function of the connection duration. The loss QOS value is taken to be fixed at 6.66×10^{-03}. This value of the loss is the steady state loss value with a load of approximately 125 sources. Once again, we notice that for loads of 120 and 125 sources at which steady state analysis predicts no violations we have significant violations when we account for the finite connection duration.

Figure 7 also demonstrates an interesting aspect of the violation levels. The violation level increases for all load values with connection duration up to a time of 1 minute. However, we find that the violation level has decreased for connection durations of 2 and 5 minutes with an offered load of 120 sources. Since the violation level for all values of the load below 125 sources must eventually converge to zero (since the loss value converges to the steady state value at that load level, 2.29×10^{-03}, which is lower than 6.66×10^{-03}), it appears that the violation level first increases and eventually decreases to zero for load levels below or equal to the load level at which the steady state value of the loss equals the loss QOS value under consideration.

5.3 Comparison of the $M/D/1/K$ queue and the Voice Multiplexer

In this section, we briefly compare the transient performance of the voice multiplexer to that of the $M/D/1/K$ queue. Our focus again will be on the packet loss QOS metric. Since the $M/D/1/K$ queue and the voice multiplexer differ significantly in steady state performance [N+91], it is difficult to make a fair comparison. For the same load level and buffer size, the steady state loss in the $M/D/1/K$ queue is much lower than that in the voice multiplexer. On the other hand, for comparable loss values, either the load level or the buffer size in the voice multiplexer has to be significantly smaller. To facilitate a comparison, we choose identical load levels and buffer sizes

in both cases. However, the loss QOS requirement is chosen to be different in the two cases. We choose the respective steady state loss values at a given load value as the loss QOS requirement. In another comparison, we choose the QOS requirement to be the respective values of the steady state loss at a fixed load and this value is used as the QOS criteria for all pertinent loads.

It is to be noted that results for the $M/D/1/K$ queueing model in this section are again obtained via simulation rather than via the analysis detailed in earlier sections. There are several reasons for this. First, in Section 5.1, we were concerned with small values of the packet loss QOS requirement only. Hence, the upper bound for the violation level was sufficient. However, here we consider a wider range of loss QOS values and hence the upper bound does not prove very useful especially when making a comparison to the voice multiplexer. Further, the upper bound tends to unity (or 100% violation) for all values of the offered load as connection durations approach infinity. Hence, this again results in the limited usefulness of the upper bound. The computation of the distribution of the customer average rather than just an upper bound is under investigation and it appears that at least for the packet loss metric the entire distribution of the customer average random variable can be evaluated. For purposes of the discussion in this section, however, we resort to simulation.

Figure 8 compares the violation levels in the voice and Poisson source multiplexers with the QOS criteria taken to be the steady state loss value at the given load (L_V^s and L_P^s respectively). We see that the violation levels in the case of the voice multiplexer are significantly higher than in the Poisson sources multiplexer except at loads close to capacity.

Figure 9 shows the violation levels in the $M/D/1/K$ queue and the voice multiplexer at differing load levels. The respective loss QOS metrics are the corresponding steady state loss values at a load level of 125 sources ($\rho = 0.915$) (Q_P^c for the $M/D/1/K$ queue and Q_V^c for the voice multiplexer). We see, as previously, that the violation levels in the voice multiplexer are significantly higher than that in the $M/D/1/K$ queue. Further, it appears that for short connections, the load levels may be increased beyond those predicted by steady state computations in the case of the Poisson sources multiplexer (with a small increase in the violation level) while in the case of the voice multiplexer the load levels may, in fact, have to be decreased to obtain a low violation level.

In the above results, we observed the comparable performance of the voice and Poisson sources multiplexers for high loads. We conjecture that the behaviour is due to the nature of the occupancy distribution, p_i, observed by arriving connections. As mentioned earlier, owing to our connection model, arriving connections observe the steady state distribution of the corresponding queueing model. In [B+91], it is shown that the steady state occupancy distribution for a multiplexer with a set of correlated input sources (identical to the voice source model with different paramaters) is *bi-modal* with mass at high occupancies. Figure 10 also shows this bi-modal behaviour with significant probability mass at high occupancies. However, in the $M/D/1/K$ queue it is uni-modal (see Figure 11) with significant mass at low occupancies. Hence, in the case of the voice multiplexer, a significant fraction of connections observe a congested queue upon arrival and this seriously impacts the QOS perceived by short connections. However, in the case of the Poisson multiplexer, most connections observe small occupancies and hence the QOS perceived by short connections is significantly better than that received by long connections. However, note that the steady state occupancy distribution in the case of the $M/D/1/K$ queue changes dramatically as the load is increased from 130 to 135 sources. The steady state distribution for the latter load value is also shown in Figure 11 and is seen to almost distribute the mass equally among the various queue occupancy values and looks similar to the steady state distribution of the voice multiplexer in Figure 10. We believe that the nature of the respective steady state occupancy distributions for

high loads is the cause for the comparable performance of the voice and Poisson multiplexers at high loads.

6 Conclusion

QOS guarantees in communication networks are typically provided by steady state performance analysis of appropriate queueing models. In this report, we examined in depth the appropriateness of steady-state queueing analysis for computing QOS criteria and most importantly in guaranteeing the QOS for envisaged applications in HSNs. We proposed new statistical QOS criteria, the *connection duration based QOS* and *interval QOS*, for high-speed networks and their applications. We argued that these new QOS criteria are more appropriate and flexible. We then showed that steady-state analysis is not entirely sufficient for computing and guaranteeing these QOS criteria. In this regard, we proposed new *cumulative transient performance measures* for queueing models and developed *analytical tools for computing these measures.* We also demonstrated as to how these new metrics can be employed to compute and guarantee the new proposed QOS criteria.

From a practical standpoint, the following results are significant:

- We showed that for a voice multiplexer, dimensioning the load based on steady state performance analysis can lead to poor user perceptions of audio quality.
- In contrast, we showed that relatively high values of the offered load, in comparison to that predicted by steady state analysis, can be supported in the case of Poisson sources while providing a very acceptable and guaranteed quality-of-service.

While we have gained considerable insight into providing statistical QOS guarantees in communication networks, the following is a brief list of interesting avenues for further research:

- Our computation of the customer average moments and mass at zero is limited to the case of a single arrival stream. Extensions to multiple heterogenous streams would be useful from the perspective of QOS guarantee computations in a realistic scenario
- Further work needs to done to compute the entire distribution for the customer average RV. It appears that at least for the loss QOS metric, simple recursive computations can be employed to compute the distribution
- Efficient schemes for computation of the proposed transient metrics are necessary in order to that QOS guarantees can be computed in on-line connection admission algorithms.

References

[As86] Harmen R. Van As. Transient analysis of markovian queueing systems and its application to congestion-control modeling. *IEEE J.Select.Areas Commun.*, SAC-4:891–904, September 1986.

[B+91] Andrea Baiocchi et al. Loss performance analysis of an ATM multiplexer loaded with high-speed on-off sources. *IEEE J.Select.Areas Commun.*, 9(3):388–393, April 1991.

[Bie91] Ernst Biersack. Error recovery in high-speed networks. In *Second International Workshop on Network and Operating System Support for Digital Audio and Video*, page 222, 1991.

[Bra90] Hugh S. Bradlow. Performance measures for real-time continuous bit-stream oriented services: Application to packet reassembly. *Computer Networks and ISDN systems*, 20:15–26, 1990.

[BZ90] Saewoong Bahk and Magda El Zarki. Routing in ATM networks. In *ITC Seminar*, page 6.4, October 1990.

[DHL92] A. Tatabai D. Heyman and T.V. Lakshmanan. Statistical analysis and simulation study of video teleconference traffic in ATM networks. In *Second ORSA Telecommunications Conference*, March 1992.

[DL86] John N. Daigle and Joseph D. Langford. Models for analysis of packet voice communications systems. *IEEE J.Select.Areas Commun.*, SAC-6:847–855, 1986.

[Fil88] Janusz Filipiak. Accuracy of traffic modeling in fast packet switching. In *Globecom'88*, pages 49.4.1–49.4.5, November 1988.

[G.880] CCITT Rec. G.821. Recommendation on error performance on an international digital connection forming part of an integrated services digital network. *CCITT Yellow Book*, III:193–195, 1980.

[GH85] Donald Gross and Carl M. Harris. *Fundamentals of Queueing Theory.* John Wiley and Sons, second edition, 1985.

[GL83] John Gruber and Nguyen Le. Performance requirements for integrated voice/data networks. *IEEE J.Select.Areas Commun.*, 1(6):981–1005, December 1983.

[Gra77] Winfried Grassman. Transient solutions in markovian queues. *European Journal of Operations Research*, 1:396–402, 1977.

[Gra87] Winfried Grassman. Means and variances of time averages in markovian environments. *European Journal of Operations Research*, 31:132–139, 1987.

[H+88] M.C. Hsueh et al. Performability modeling based on real data: A case study. *IEEE Transactions on Computers*, C-37(4):478–484, April 1988.

[HL86] Harry Heffes and David Lucantoni. A markov modulated characterization of voice and data traffic and related statistical multiplexer performance. *IEEE J.Select.Areas Commun.*, SAC-4:856–867, September 1986.

[I.188] CCITT Rec. I.121. Recommendation on broadband aspects of ISDN. 1988.

[I.388] CCITT Rec. I.350. Recommendation on broadband aspects of ISDN. 1988.

[I.390] CCITT Rec. I.311. Recommendation on broadband aspects of ISDN. 1990.

[Kot78] T. C. T. Kotiah. Approximate transient analysis of some queueing systems. *Operations Research*, 26:333–346, March 1978.

[KS60] John Kemeny and Laurie Snell. *Finite Markov Chains.* D. Van Nostrand Company Inc.,, 1960.

[KV89] Gunnar Karlsson and Martin Vetterli. Packet video and its integration into the network architecture. *IEEE J.Select.Areas Commun.*, 7(5):739–751, June 1989.

[L+90] William P. Lovegrove et al. Simulation methods for studying nonstationary behaviour of computer networks. *IEEE J.Select.Areas Commun.*, 8(9):1696–1708, December 1990.

[Moo75] S.C. Moore. Approximating the behaviour of nonstationary single-server queues. *Operations Research*, 23:1011–1032, September 1975.

[Mor58] P.M. Morse. *Queues, Inventories and Maintenance : The Analysis of Operation Systems with Variable Supply and Demand.* Wiley, New York, 1958.

[MW90] Benjamin Melamed and Ward Whitt. On arrivals that see time averages. *Operations Research*, 38:156–172, January 1990.

[N+91] Ramesh Nagarajan et al. Approximation techniques for computing packet loss in finite-buffered voice multiplexers. *IEEE J.Select.Areas Commun.*, 9(3):368–377, April 1991.

[OR83] Amedeo R. Odoni and Emily Roth. An empirical investigation of the transient behaviour of stationary queueing systems. *Operations Research*, 31:432–455, May 1983.

[Pap84] Papoulis. *Probability, Random Variables and Stochastic Processes.* McGraw Hill, 1984.

[Par92] Craig Partridge. A proposed flow specification. *Internet Research Task Force, Network Working Group, Request for Comments: 1363*, September 1992.

[PZ82] William R. Parzinsky and Philip W. Zipse. *Introduction to Mathematical Analysis.* McGraw-Hill Book Company, 1982.

[PZ93] Pramod Pancha and Magda El Zarki. Bandwidth requirements of variable bit rate MPEG sources in ATM networks. In *IFIP workshop on modeling and performance evaluation of ATM technology*, pages 5.2.1–5.2.25, January 1993.

[Ram88] V. Ramaswamy. Traffic performance modeling for packet communication whence, where and wither. In *Third Australian Teletraffic Seminar*, November 1988. Keynote Address.

[Rei82] Martin Reiser. Performance evaluation of communication systems. *Proceedings of the IEEE*, 70:171–196, February 1982.

[RT88] Andrew Reibman and Kishore Trivedi. Numerical transient analysis of markov models. *Comput. Opns. Res.*, 15:19–36, 1988.

[RT89] Andrew Reibman and Kishore Trivedi. Transient analysis of cumulative measures of markov models. *Communications in Statistics-Stochastic Models*, 5:683–710, 1989.

[RW90] V. Ramaswamy and Walter Willinger. Efficient traffic performance strategies for packet multiplexers. *Computer Networks and ISDN systems*, 20:401–407, 1990.

[S+88] R.M. Smith et al. Performability analysis: measures, an algorithm and a case study. *IEEE Transactions on Computers*, C-37(4):406–417, April 1988.

[Sha90] Nachum Shacham. Packet recovery in high-speed networks using coding and buffer management. In *INFOCOM*, pages 124–131, 1990.

[SW86] Kotikalapudi Sriram and Ward Whitt. Characterizing superposition arrival processes in packet multiplexers for voice and data. *IEEE J.Select.Areas Commun.*, SAC-4:833–846, September 1986.

[WK90] Gillian M. Woodruff and Rungroj Kositpaiboon. Multimedia traffic management principles for guaranteed ATM network performance. *IEEE J.Select.Areas Commun.*, 8:446, April 1990.

[Wol89] Ronald W. Wolff. *Stochastic modeling and the theory of queues.* Prentice Hall, 1989.

[YS89] Chin Yuan and John A. Silvester. Queueing analysis of delay constrained voice traffic in a packet switching system. *IEEE J.Select.Areas Commun.*, 7:729–738, June 1989.

[YW89] Yusaka Yamamoto and Tim Wright. Error performance in evolving digital networks including ISDNs. *IEEE Comm. Mag.*, pages 12–18, April 1989.

[Zha90] Lixia Zhang. Virtual clock: A new traffic control algorithm for packet switching networks. In *SIGCOMM*, pages 19–29, September 1990.

A Expected value of the customer average

Here we consider the computation of the mean customer average over some time interval t. We then consider the limiting behaviour, i.e., as $t \to \infty$, of the mean customer average. Finally, we illustrate the accuracy of our computation of the mean customer average by comparison to simulation.

For the first approximation, we have:

$$\begin{aligned} E_m[A_C(t)|N_t > 0] &= E_m[\frac{\sum_{i=1}^{N_t} f(X(A_i))}{N_t}|N_t > 0] \\ &\approx \sum_{n=1}^{\infty} \frac{\sum_{i=1}^{n} E_m[f(X(A_i))]}{n} P(N_t = n|N_t > 0) \end{aligned} \tag{29}$$

The second approximation in this case is:

$$E_m[A_C(t)] \approx \frac{\sum_{i=1}^{N} E_m[f(X(A_i))]}{N} \tag{30}$$

where $N = \lfloor \lambda t \rfloor$.

The computation of the above metric requires the following:

- Truncation of an infinite series for numerical computation,
- Computing $E_m[f(X(A_i))]$.

We consider the above two facets of the computation in turn.

Note that the above series is a weighted sum of probabilities. In order to truncate the above infinite series, we need to bound the weights. The weights can then be forced to values smaller than unity by appropriate scaling:

$$E_m[A_C(t) \mid N_t > 0] = L \sum_{n=1}^{\infty} \frac{\sum_{i=1}^{n} E_m[f(X(A_i))]}{nL} P(N_t = n \mid N_t > 0) \tag{31}$$

where L is the bound on the weights. This allows us to truncate, in principle, the infinite summation for numerical computation of the customer average (see Appendix D for a discussion of how L may be determined and Appendix E for details of the truncation).

Next, consider the computation of $E_m[f(X(A_i))]$,

$$E_m[f(X(A_i))] = \sum_{n=1}^{K} f(n) P^c_{mn}(i). \tag{32}$$

For the special case of the fractional loss metric we have $E_m[f(X(A_i))] = P^c_{mK}(i)$. Now $\{X(A_i), i \geq 1\}$ is a discrete-time Markov chain and its transition probabilities can be computed (see Appendix I). Let $\boldsymbol{P}$ denote this transition probability matrix. Then the customer-based transient probabilities are evaluated as,

$$\begin{aligned} P^c_{00}(1) &= 1.0 \\ P^c_{0n}(1) &= 0.0 \;\; 0 < n \leq K \\ P^c_{0n}(i) &= P^{i-1}(0,n) \;\; i > 1 \end{aligned} \tag{33}$$

with $P^k(i,j)$ being the (i,j) entry of $\boldsymbol{P}^k$. Similarly, for $0 < m \leq K$ we have,

$$P^c_{mn}(i) = P^{i-1}(m-1,n) \;\; 0 \leq n \leq K \tag{34}$$

The matrix muliplication above can be carried out by successive multiplications of the transition probability matrix, $\boldsymbol{P}$. A tree-type of matrix multiplication is, however, one simple way to minimize the loss of accuracy due to repeated matrix multiplications. A simple algorithm to implement the tree-type multiplication of matrices is outlined in Appendix G. This multiplication scheme ensures that the n-step matrix is a result of $O(log(n))$ matrix multiplications in contrast to the straightforward multiplication that requires $n-1$ matrix multiplications.

Finally, we consider some limiting properties of the mean customer average. Since $X(A_i)$ is aperiodic, irreducible and positive-recurrent, we have that

$$\lim_{i \to \infty} P_m(f(X(A_i)) \leq x) = P(X_\infty \leq x) \tag{35}$$

exists, where X_∞ is used to denote the *limiting distribution* of $f(X(A_i))$. It follows that

$$\lim_{n \to \infty} \frac{\sum_{i=1}^{n} E_m[f(X(A_i))]}{n} = E[X_\infty]. \tag{36}$$

We show below that the mean customer average converges to the steady state value, $E[X_\infty]$, for infinite time (inspite of the approximation) as is desirable. This also implies that the accuracy of the approximation need be checked only for small time.

Let

$$a_n = \frac{\sum_{i=1}^{n} E_m[f(X(A_i))]}{n} \tag{37}$$

Also, we have

$$\lim_{n\to\infty} a_n = E[X_\infty] = p \tag{38}$$

Then we have for $\epsilon > 0$ and $\forall n \geq N(\epsilon)$, $|a_n - p| < \epsilon$ or $p - \epsilon < a_n < p + \epsilon$ [PZ82]. Also

$$\lim_{t\to\infty} \sum_{n=1}^{N(\epsilon)} P(N_t = n) = 0 \tag{39}$$

since intuitively the probability of a finite number of arrivals in infinite time is zero. Now $\lim_{t\to\infty} \frac{N_t}{t} = \lambda \ \ w.p.1$. Hence for $n \in N$, we have $\lim_{t\to\infty} N_t > n \ \ w.p.1$. Define an indicator function,

$$I_{N_t=n} = \begin{cases} 1 & N_t = n \\ 0 & o.w. \end{cases} \tag{40}$$

We then have $\lim_{t\to\infty} I_{N_t=n} = 0 \ \ w.p.1$. Also, $|I_{N_t=n}| \leq 1$. Hence by the Bounded Convergence Theorem [Wol89, pp. 535] we have the desired result since

$$\begin{aligned} \lim_{t\to\infty} P(N_t = n) &= \lim_{t\to\infty} E[I_{N_t=n}] \\ &= E[\lim_{t\to\infty} I_{N_t=n}] \\ &= 0 \end{aligned} \tag{41}$$

Also, it is easy to see that $lim_{t\to\infty} P(N_t > 0) = 1.0$.

Our interest is in $\lim_{t\to\infty} E_m[A_C(t)]$,

$$\begin{aligned} \lim_{t\to\infty} E_m[A_C(t)] &= \lim_{t\to\infty} \sum_{n=1}^{\infty} a_n P(N_t = n \mid N_t > 0) \\ &= \lim_{t\to\infty} \sum_{n=1}^{\infty} a_n P(N_t = n) \\ &= \lim_{t\to\infty} (\sum_{n=1}^{N(\epsilon)} a_n P(N_t = n) + \sum_{n=N(\epsilon)+1}^{\infty} a_n P(N_t = n)) \\ &= \lim_{t\to\infty} \sum_{n=N(\epsilon)+1}^{\infty} a_n P(N_t = n) \end{aligned} \tag{42}$$

Since $\forall n \geq N(\epsilon), p - \epsilon < a_n < p + \epsilon$ we have,

$$\lim_{t\to\infty} E_m[A_C(t)] \ < \ (p+\epsilon) \lim_{t\to\infty} \sum_{n=N(\epsilon)+1}^{\infty} P(N_t = n)$$

$$
\begin{aligned}
&< (p+\epsilon) \lim_{t\to\infty} (1 - \sum_{n=0}^{N(\epsilon)} P(N_t = n)) \\
&< p + \epsilon \qquad (43)
\end{aligned}
$$

Similarly,

$$\lim_{t\to\infty} E_m[A_C(t)] > p - \epsilon \qquad (44)$$

Since ϵ is arbitrary we have

$$\lim_{t\to\infty} E_m[A_C(t)] = p \qquad (45)$$

which is as desired.

Table 11 validates our analysis for the mean of customer averaged number, i.e., customer average with $f(.)$ as defined in Equation (4), in the $M/M/1/50$ queue. We observe that the analytical values lie within the confidence intervals for most of the parameter range. However, the accuracy of the analytical approximation does deteriorate slightly with buffer size. The absolute value of the error for large buffer sizes is however minimal (the relative error is less insignificant).

Next, we evaluate the mean of the fractional loss over finite intervals. Tables 12, 13 and 14 show the predicted average loss for the $M/M/1/10$ and $M/M/1/20$ queues respectively. The analytical values are seen to be somewhat pessimistic albeit fairly accurate.

Table 17 validates the analysis for the customer averaged number in the $M/D/1/50$ queue. Tables 15, 16 and 18 show the average loss for the $M/D/1/10$ queue. The analysis is not as accurate as for the $M/M/1/K$ queue. However, the predicted values were seen to be pessimistic thus providing an upper bound (not proved). In the $M/D/1/K$ queue, the pessimistic approximation for the transition probability matrix in addition to the approximation in the computation of the mean itself leads to an increasingly pessimistic approximation. However, the approximation improves with lower traffic intensities as expected.

B Variance of the customer average

In this appendix, we detail the computation of the variance of the customer average RV. We then illustrate the accuracy of the computation through a few numerical examples.

By definition,

$$\mathrm{VAR}_m(A_c(t)|N_t > 0) = E_m[A_c^2(t)|N_t > 0] - E_m^2[A_c(t)|N_t > 0] \qquad (46)$$

As a first approximation we have,

$$
\begin{aligned}
E_m[A_c^2(t)|N_t > 0] &= E_m[(\frac{\sum_{i=1}^{N_t} f(X(A_i))}{N_t})^2 | N_t > 0] \\
&\approx \sum_{n=1}^{\infty} b_n P(N_t = n | N_t > 0) \qquad (47)
\end{aligned}
$$

where

$$b_n \equiv \frac{\sum_{i=1}^{n} E_m[f^2(X(A_i))] + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} E_m[f(X(A_i))f(X(A_j))]}{n^2} \qquad (48)$$

Table 11: Mean of the customer averaged number in a M/M/1/50 Queue

Time (Service units)	$\rho = 0.8, m = 50$		
	Simulation (90% C.I.)	Approx1	Approx2
10.0	47.36928 (4.14893e-02)	46.92173	46.89060
20.0	45.8207 (5.9852e-02)	45.36933	45.34701
30.0	44.4661 (8.00632e-02)	44.01008	43.99260
40.0	43.36587 (8.8876e-02)	42.74768	42.73318
50.0	42.05234 (9.76535e-02)	41.54469	41.53210
60.0	40.85682 (1.08781e-01)	40.38283	40.37129

Table 12: Customer averaged loss in a M/M/1/10 queue

Time (service units)	$\rho = 0.8, m = 10$			
	Mean	Sim (90 % C.I.)	Var.	Sim (90% C.I.)
10.000	0.2192	0.2247 (3.897e-03)	4.455e-02	4.1250e-02 (1.038e-03)
20.000	0.1504	0.1542 (7.302e-03)	2.172e-02	1.9704e-02 (1.78e-03)
30.000	0.1168	0.1161 (1.796e-03)	1.352e-02	1.1927e-02 (3.27e-04)
40.000	9.65e-02	9.6741e-02 (1.509e-03)	9.39e-03	8.4107e-03 (2.436e-04)
50.000	8.29e-02	8.1521e-02 (1.268e-03)	6.97e-03	5.9439e-03 (1.779e-04)
100.0	5.37e-02	5.4443e-02 (9.985e-04)	2.68e-03	2.58e-03 (1.0e-04)

Table 13: Customer averaged loss in a M/M/1/20 queue

Time (service units)	$\rho = 0.625, m = 20$			
	Mean	Sim. (90 % C.I.)	Var.	Sim. (90% C.I.)
10.000	0.1847	0.1867 (5.106e-03)	4.498e-02	4.119e-02 (1.688e-03)
20.000	1.161e-01	1.119e-01 (2.72e-03)	1.888e-02	1.543e-02 (6.21e-04)
30.000	8.43e-02	7.9778e-02 (1.629e-03)	1.041e-02	8.394e-03 (2.972e-04)
40.000	6.35e-02	6.1638e-02 (1.199e-03)	6.11e-03	5.314e-03 (1.869e-04)
50.000	5.23e-02	5.085e-02 (9.977e-04)	4.21e-03	3.678e-03 (1.327e-04)
100.0	2.681e-02	2.6237e-02 (5.18e-04)	1.14e-03	9.916e-04 (4.305e-05)

Table 14: Customer averaged loss in a M/M/1/20 queue

Time (service units)	$\rho = 0.625, m = 10$			
	Mean	Sim. (90 % C.I.)	Var.	Sim. (90% C.I.)
50.000	2.57e-04	2.6325e-04 (9.272e-05)	2.6e-05	2.6839e-05 (1.3588e-05)
60.000	2.73e-05	2.0882e-04 (6.411e-05)	2.6e-05	1.5188e-05 (6.725e-06)
80.000	2.61e-04	2.82e-04 (6.969e-05)	2.1e-05	1.7946e-05 (5.78e-06)
100.000	2.32e-04	2.2929e-04 (6.492e-05)	1.5e-05	1.5573e-05 (7.518e-06)

The approximation is similar to that for the mean of the customer average in Appendix A. Alternatively, we have the second approximation,

$$E_m[A_c^2(t)|N_t > 0] \approx \frac{\sum_{i=1}^{N} E_m[f^2(X(A_i))] + 2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} E_m[f(X(A_i))f(X(A_j))]}{N^2} \tag{49}$$

where $N = \lfloor E[N_t] \rfloor$ and

$$E_m[f(X(A_i))f(X(A_j))] = \sum_{l=0}^{K}\sum_{n=0}^{K} f(l)f(n)P(X(A_i) = l|X(0) = m)P(X(A_j) = n|X(A_i) = l) \tag{50}$$

The above is easily computed using the ith and the $(j-i)$th step transition matrices of the imbedded MC at arrival instants. For the special case of the fractional loss metric, we have,

$$E_m[f(X(A_i))f(X(A_j))] = P(X(A_i) = K|X(0) = m)P(X(A_j) = K|X(A_i) = K) \tag{51}$$

The infinite series above may be suitably truncated, as in the case of the mean in Appendix A, for numerical computation (see also Appendix E).

Tables 12, 13 and 14 show the predicted variance of the customer averaged loss for the $M/M/1/10$ and $M/M/1/20$ queues respectively. Tables 15, 16 and 18 show the same for the $M/D/1/10$ queue. The analytical values are seen to be somewhat pessimistic albeit fairly accurate.

C Computing the mass at zero for the customer average

Here we consider the computation of the mass at zero for the customer average. We also consider the limiting behaviour of the mass at zero. Numerical examples of the computation were presented in Section 4.

In Section 3.1, our interest was in computing,

$$a_n = P_m(f(X(A_1)) = 0, f(X(A_2)) = 0, \cdots, f(X(A_n)) = 0) \tag{52}$$

which was the joint probability term in the infinite series for the mass at zero. We outline its computation below.

Let $\boldsymbol{P}$ be the transition matrix for the DTMC embedded at the arrival instants i.e., $\{X(A_i),\ i > 0\}$, of the $GI/M/1/K$ queue. In the following, we adopt the terminology in [KS60].

Since the DTMC is irreducible the states of the DTMC belong to a single equivalence class and hence the MC has a single ergodic set. Consider the states of the DTMC, values of $X(A_i)$, for which $f(X(A_i)) \neq 0$. Modify the transition matrix such that these states are now absorbing states. Note that this has does not effect the behaviour of the MC while not in the absorbing states. Now, the states of the DTMC for which $f(X(A_i)) = 0$, form an equivalence class and the individual states for which $f(X(A_i)) \neq 0$ form equivalence classes by themselves. Following the classification of Markov chains in [KS60], we now have a chain with a transient set (states wherein $f(X(A_i)) = 0$) and the ergodic sets being unit sets (states wherein $f(X(A_i)) \neq 0$). We are now

Table 15: Customer averaged loss in a M/D/1/10 queue

Time (service units)	$\rho = 0.8, m = 10$			
	Mean	Sim. (90% C.I.)	Var. (Approx2)	Sim. (90% C.I.)
10.000	0.18959	0.1933 (2.475e-03)	2.934e-02	2.152e-02 (1.732e-03)
20.000	0.12032	0.1133 (1.5229e-03)	1.275e-02	8.4802e-03 (1.891e-04)
30.000	8.937e-02	8.4270e-02 (1.157e-03)	7.42e-03	4.946e-03 (1.26e-04)
40.000	7.133e-02	6.5401e-02 (9.1748e-04)	4.91e-03	3.111e-03 (8.639e-05)
50.000	5.946e-02	5.4001e-02 (7.773e-04)	3.50e-03	2.233e-03 (6.464e-05)
100.00	3.343e-02	2.989e-02 (4.45e-04)	1.13e-03	7.319e-04 (4.18e-04)

Table 16: Customer averaged loss in a M/D/1/10 queue

Time (service units)	$\rho = 0.8, m = 4$			
	Mean	Simulation (90 % C.I.)	Var.	Simulation (90 % C.I.)
10.000	4.5e-04	1.322e-03 (2.052e-04)	6.0e-05	2.537e-04 (4.864e-05)
20.000	2.39e-03	2.547e-03 (2.652e-04)	3.2e-04	2.599e-04 (3.747e-05)
30.000	3.8e-03	3.261e-03 (2.83e-04)	4.5e-04	2.96e-04 (3.64e-05)
40.000	4.54e-03	3.368e-03 (2.696e-04)	4.7e-04	2.761e-04 (1.117e-04)
50.000	4.93e-03	3.537e-03 (2.562e-04)	4.4e-04	2.198e-04 (2.429e-05)
100.000	5.57e-03	3.4914e-03 (1.984e-04)	2.9e-04	1.454e-04 (1.34e-05)

Table 17: Mean of the customer averaged number in a M/D/1/50 Queue

Time (Service units)	$\rho = 0.8, m = 0$		
	Simulation (90% C.I.)	Approx1	Approx2
10.000	1.0707 (1.3182e-02)	0.9057	0.9367
20.000	1.4219 (1.5515e-02)	1.3098	1.3282
30.000	1.6028 (1.6888e-02)	1.5489	1.5623
40.000	1.7342 (1.7572e-02)	1.7135	1.7239
50.000	1.8352 (1.7993e-02)	1.8351	1.8435

Table 18: Customer averaged loss in a M/D/1/10 queue

Time (service units)	$\rho = 0.6, m = 10$			
	Mean	Sim. (90 % C.I.)	Var. (Approx2)	Sim. (90% C.I.)
10.000	0.15169	0.1509 (2.693e-03)	2.844e-02	2.3978e-02 (7.46e-04)
20.000	8.603e-02	8.158e-02 (1.396e-03)	9.82e-03	7.1743e-03 (1.94e-04)
30.000	5.946e-02	5.622e-02 (9.787e-04)	4.86e-03	3.5423e-03 (3.19e-04)
40.000	4.519e-02	4.197e-02 (7.44e-04)	2.85e-03	2.0467e-03 (6.073e-05)
50.000	3.635e-02	3.361e-02 (6.01e-04)	1.86e-03	1.3366e-03 (4.138e-05)
100.000	1.832e-02	1.714e-02 (3.13e-04)	4.8e-04	3.8225e-04 (1.387e-04)

ready to write the transition matrix in the aggregated canonical form i.e., we group all the ergodic sets and the transient sets. We then have,

$$\boldsymbol{P} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{O} \\ \boldsymbol{R} & \boldsymbol{Q} \end{pmatrix} \tag{53}$$

where the region $\boldsymbol{O}$ consists entirely of 0's, $\boldsymbol{Q}$ concerns the Markov chain in the transient states, $\boldsymbol{R}$ concerns the transitions of the Markov chain from the transient states to the ergodic states. The identity matrix $\boldsymbol{I}$ is a consequence of the fact that the ergodic states are simply absorbing states. Consider next the n-fold product of $\boldsymbol{P}$,

$$\boldsymbol{P}^n = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{O} \\ \boldsymbol{R}' & \boldsymbol{Q}^n \end{pmatrix} \tag{54}$$

The values of a_n are now completely determined via the matrix $\boldsymbol{Q}^n$. Indeed, if q_{ij}^n are the elements of $\boldsymbol{Q}^n$, we have,

$$a_n = \sum_{j \in T} q_{mj}^n \tag{55}$$

where T is the set of transient states. Finally, we consider the behaviour of a_n as $n \to \infty$. By [KS60, theorem 3.1.1], we have that $\boldsymbol{Q}^n \to \boldsymbol{O}$ and hence $a_n \to 0$.

D Boundedness of weights in customer average computation

In Appendix A and B, the mean and variance of the customer average were expressed as a weighted sum of probabilities. Here we compute bounds on these bounds which are essential for the truncation of the inifinite series arising in these computations.

Define

$$a_n \equiv \frac{\sum_{i=1}^{n} E[f(X(A_k)]}{n} \quad n = 1, 2, 3, \cdots \tag{56}$$

and

$$b_n \equiv \frac{\sum_{i=1}^{n} E_m[f^2(X(A_i))] + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} E_m[f(X(A_i))f(X(A_j))]}{n^2}. \tag{57}$$

Clearly, $\{a_n\}$ and $\{b_n\}$ are bounded since $f(\cdot)$ is bounded. We can, hence, choose a bound for a_n as

$$L_1 = sup\{a_n, \;\; n \geq 1\} \tag{58}$$

and for b_n as

$$L_2 = sup\{b_n, \;\; n \geq 1\}. \tag{59}$$

These bounds, however, may not be easily computable. We will show, in the following, that looser bounds may be easily determined for the QOS metrics of interest in this paper.

First, consider the loss QOS metric. In this case, we show that both a_n and b_n are bounded by unity and hence the truncation of the infinite series requires no additional computational effort.

For the loss metric, we have

$$\begin{aligned} E_m[f(X(A_i))] &\leq 1 \\ E_m[f^2(X(A_i))] &\leq 1 \\ E_m[f(X(A_i))f(X(A_j))] &\leq 1. \end{aligned} \tag{60}$$

It is then easily established that $a_n,\ b_n \leq 1$.

Next, consider the average number metric. In this case, we have

$$\begin{aligned} E_m[f(X(A_i))] &\leq K \\ E_m[f^2(X(A_i))] &\leq K^2 \\ E_m[f(X(A_i))f(X(A_j))] &\leq K^2. \end{aligned} \tag{61}$$

Similarly, one can establish that $a_n \leq K$ and $b_n \leq K^2$.

E Truncation of infinite series in customer average computations

The evaluation of the mass at zero for the customer average in Section 3.1 and moments in Appendices A and B required the appropriate truncation of an infinite series. We now outline the procedure to truncate the series so that the resulting value is within a specified tolerance of the exact value.

The infinite sum to be approximated is, in general, of the form

$$\sum_{n=1}^{\infty} a_n P(N_t = n | N_t > 0) \tag{62}$$

From the discussion in Appendix D note that $a_n \leq L$, where L is the appropriate bound on the weights.

Let δ be the specified error tolerance and l and r be the left and right truncation points for the infinite series. First, consider left truncation. The error on left truncating the infinite series at l is

$$\begin{aligned} \mathrm{E_l} &= \sum_{n=1}^{l-1} a_n P(N_t = n | N_t > 0) \\ &= L \sum_{n=1}^{l-1} \frac{a_n}{L} P(N_t = n | N_t > 0) \\ &\leq L \sum_{n=1}^{l-1} P(N_t = n | N_t > 0) \end{aligned} \tag{63}$$

Hence, in order that $E_l \leq \delta/2$, choose l such that

$$l = \max(i \ : \ \sum_{n=1}^{i-1} P(N_t = n | N_t > 0) \leq \frac{\delta}{2L}) \tag{64}$$

Next, consider the error due to right truncation.

$$\begin{aligned} E_r &= \sum_{n=r+1}^{\infty} a_n P(N_t = n | N_t > 0) \\ &= L \sum_{n=r+1}^{\infty} \frac{a_n}{L} P(N_t = n | N_t > 0) \\ &\leq L \sum_{n=r+1}^{\infty} P(N_t = n | N_t > 0) \\ &\leq L(\sum_{n=1}^{\infty} P(N_t = n | N_t > 0) - \sum_{n=1}^{r} P(N_t = n | N_t > 0)) \\ &\leq L(1 - \sum_{n=1}^{r} P(N_t = n | N_t > 0)) \end{aligned} \tag{65}$$

Choose r such that

$$r = \min(i \ : \ 1 - \sum_{n=1}^{i} P(N_t = n | N_t > 0) \leq \frac{\delta}{2L}) \tag{66}$$

Then $E_r \leq \delta/2$ and the total error is bounded by δ.

We make a few comments on the computation of the left and right truncation points in the case of Poisson arrivals to the queue. Double precision computation leads to underflows for large time values. This in turn causes the computation to considerably slow down, presumably due to the underflow handling, on the Microvax used to implement the computation. A switch to quadruple precision is necessary in order to carry out computations for large time values. This results in a small slow down compared to the double precision implementation. If speed of computation were critical one could keep both the double and quadruple precision implementations and use the former for small times and the latter for large times. As a thumb of rule use double precision for $\lambda t \leq 85.0$ and quadruple precision otherwise. Alternatively, one could adopt the normal approximation for the Poisson probabilities and avoid the underflows [Gra87].

F Numerical complexity of customer average computation

In this appendix, we investigate the complexity (in arithmetic operations) of the analysis for computing the mass at zero and moments of the customer average in Section 3.1 and Appendices A and B.

There are two major steps in the computation of the performance metrics. First, we truncate an infinite summation according to the desired numerical accuracy, δ. Second, we carry out matrix multiplications to evaluate desired performance metrics. We focus on the complexity of the former since the complexity of matrix multiplications is well known.

As in Appendix E, we let r be the right-truncation point. Then we require that

$$\sum_{n=r+1}^{\infty} P(N_t = n | N_t > 0) \leq \epsilon = \frac{\delta}{L} \tag{67}$$

Note that the above inequality is too stringent when we perform left truncation and hence will yield a upper bound for r. Hence, we require

$$P(N_t > r | N_t > 0) \leq \epsilon. \tag{68}$$

We, now, conjecture on the value of r that satisfies the above inequality. Let

$$r = \frac{K(\epsilon)\lambda t}{P(N_t > 0)} \tag{69}$$

where $K(\epsilon) = 1/\epsilon$. Substituting in equation (68) and employing Tchebycheff's inequality [Pap84], we have

$$\begin{aligned} P(N_t > r | N_t > 0) &\leq \frac{E[N_t | N_t > 0]}{r} \\ &\leq \frac{\lambda t}{P(N_t > 0)} \frac{P(N_t > 0)}{K(\epsilon)\lambda t} \\ &\leq \epsilon. \end{aligned} \tag{70}$$

Finally, we have $r \in O(\lambda t)$ since $\lim_{t \to \infty} P(N_t > 0) = 1.0$.

G Matrix multiplication algorithm

Here we present a simple algorithm which implements the tree-type of matrix multiplication used in Section 3 and Appendices A and B. The algorithm is in a Pascal type of language.

The basic data structure is a 3-dimensional array of matrices. At any given time, the data structure contains a entire level of the tree (e.g., the first set of multiplications yields level 1 of the tree and the data structure contains level 1 i.e., matrices $\boldsymbol{P}^2$ and $\boldsymbol{P}^3$). Let mstra and mstrb be two sets of the aforementioned data structure. The 0th entry of the data structure contains the transition probability matrix i.e., mstra(0) and mstrb(0) contain $\boldsymbol{P}$. Also, let mstrb(1) contain $\boldsymbol{P}$ to start with. Note that the n-step transition matrix appears in level $\lfloor log_2(n) \rfloor$ of the tree. Also, assume that a procedure to multiply two matrices, mult(a,i,j,k,b), exists which multiplies matrices a(i) and a(j) and returns the result in matrix b(k).

```
even ← false
(* Upto the n-step transition matrix *)
for (j = 1 to j = ⌊log2(n)⌋) {
    leaves = 2^(j-1) (* number in level j-1 *)
    k = 1
    for (i = 1 to i = leaves) {
        if (even) then {
        mult(mstra,i,i,k,mstrb)
        k = k + 1
        if (i ≠ leaves) {
        mult(mstra,i,i+1,k,mstrb)
        k = k + 1
```

```
        }
        else {
        mult(mstrb,i,i,k,mstra)
        k = k + 1
        if (i ≠ leaves) {
        mult(mstrb,i,i+1,k,mstra)
        k = k + 1
        }
    }
(* Last leaf of level j *)
if (even) mult(mstrb,k-1,0,k,mstrb)
  else mult (mstra,k-1,0,k,mstra)
(* Performance using level j of tree *)
if (even) then calcperf (mstrb,j)
  else calcperf (mstra,j)
(* Toggle *)
even ← not(even)
}
```

H Analysis of the $M/E_r/1/K$ queue

Here, we consider the $M/E_r/1/K$ queue and detail the computation of the customer averaged metrics for this queue. The $M/E_r/1/K$ queue is characterized by Poisson arrivals and a r-stage Erlangian service-time distribution. The queue length process, $X(t)$, is not Markovian. However, the number of phases of service in the system, $X_p(t)$, is Markovian. We assume, as in the case of $X(t)$ previously, that the sample paths of $X_p(t)$ are left continuous. Consider this process at arrival instants. We first outline the evaluation of the transition probabilities of the DTMC $\{X_p(A_i),\ i > 0\}$. Given the transition probability matrix, the evaluation of the customer average is similar to that for the $M/M/1/K$ queue in Section 3.

Define

p_n : Probability of n phases of service completed in an interarrival time.

Note that the state space of $X_p(A_i)$ is $0 \leq X_p(A_i) \leq rk$. Also note that service phases are completed according to a Poisson process of rate $r\mu$, (when the system is non-empty) with μ^{-1} being the mean service time. Hence

$$p_n = \int_0^\infty \frac{e^{-r\mu t}(r\mu t)^n}{n!} dF(t) \tag{71}$$

and the transition probability matrix is

$$\boldsymbol{P} = \begin{pmatrix} 1 - \sum_{i=0}^{r-1} p_i & p_{r-1} & p_{r-2} & \cdots & p_0 & 0 & \cdots & 0 & 0 \\ 1 - \sum_{i=0}^{r} p_i & p_r & p_{r-1} & \cdots & p_1 & p_0 & 0 & \cdots & 0 \\ 1 - \sum_{i=0}^{rk-1} p_i & p_{rk-1} & p_{rk-2} & p_{rk-1} & \cdots & p_3 & p_2 & p_1 & p_0 \end{pmatrix} \tag{72}$$

Since $F(t)$ is an exponential distribution, the p_n can be solved for recursively as follows:

$$\begin{aligned} p_0 &= \frac{\lambda}{\lambda + r\mu} \\ p_n &= \frac{r\mu}{r\mu + \lambda} p_{n-1} \\ &= (\frac{r\mu}{r\mu + \lambda})^n \frac{\lambda}{\lambda + r\mu} \end{aligned} \tag{73}$$

The customer-based transient probabilities for the $M/E_r/1/K$ queue are computed as:

$$\begin{aligned} P^c_{00}(1) &= 1.0 \\ P^c_{0n}(1) &= 0.0 \ 0 \\ P^c_{0n}(i) &= \sum_{k=(n-1)r+1}^{nr} P^{i-1}(0,k) \ \ i > 1 \end{aligned} \tag{74}$$

and for $m > 0$,

$$P^c_{mn}(i) = \sum_{k=(n-1)r+1}^{nr} P^{i-1}((m-1)r, k) \tag{75}$$

where the matrix entries on the right hand side of the above equations are the corresponding entries in the transition probability matrix for $X_p(t)$ computed above. Finally, note that the above formulation for $m > 0$ assumes that the customer in service at $t = 0$ is in its first stage of service. Given the above customer-based transient probabilities, the rest of the analysis for computing the moments of the customer average is identical to that in Appendices A and B.

For illustration, we compute the mean of the customer averaged number and loss in the $M/E_r/1/K$ queue. Tables 19, 20 and 21 show the results of our analysis and simulation for the customer averaged number. We see that the approximation performs well except for small time. Table 22 validates our analysis for the mean customer averaged loss. The analysis is again seen to be extremely accurate.

I Transition probabilities for the $M/D/1/K$ queue

Here we compute the transition probability matrix for the $M/D/1/K$ queue (see Section 4.2). We take τ to be the deterministic service time.

Define

b_i : Probability of i service completions in an interarrival time.

Then the transition probability matrix is

$$\boldsymbol{P} = \begin{pmatrix} e^{-\rho} & 1 - e^{-\rho} & 0 & \dots & 0 & 0 \\ 1 - \sum_{n=0}^{1} b_n & b_1 & b_0 & 0 & \dots & 0 \\ \vdots & & & & & \\ 1 - \sum_{n=0}^{K-1} b_n & b_{K-1} & b_{K-2} & \dots & b_1 & b_0 \end{pmatrix} \tag{76}$$

Table 19: Mean of the customer averaged number in a $M/E_r/1/5$ queue

Time (Service units)	$\rho = 0.5, r = 20, m = 0$		
	Simulation (90% C.I.)	Approx1	Approx2
10.0	0.492 (7.88e-03)	0.406	0.440
20.0	0.600 (7.17e-03)	0.563	0.579
30.0	0.643 (6.50e-03)	0.626	0.634
40.0	0.668 (6.80e-02)	0.658	0.663
50.0	0.684 (5.46e-03)	0.677	0.680

Table 20: Mean of the customer averaged number in a $M/E_r/1/5$ queue

Time (Service units)	$\rho = 0.5, r = 20, m = 5$		
	Simulation (90% C.I.)	Approx1	Approx2
10.0	2.3167 (1.795e-02)	2.2770	2.1729
20.0	1.5479 (1.214e-02)	1.6200	1.5518
30.0	1.2906 (1.073e-02)	1.3299	1.2922
40.0	1.1522 (7.965e-03)	1.1794	1.1573
50.0	1.0712 (6.890e-03)	1.0899	1.0758

Table 21: Mean of the customer averaged number in a $M/E_r/1/10$ queue

Time (Service units)	$\rho = 0.5, r = 10, m = 0$		
	Simulation (90% C.I.)	Approx1	Approx2
10.0	0.4877 (7.961e-03)	0.4065	0.4406
20.0	0.6061 (7.671e-03)	0.5682	0.5841
30.0	0.6618 (7.030e-03)	0.6354	0.6438
40.0	0.6917 (6.603e-03)	0.6707	0.6758
50.0	0.7042 (6.257e-03)	0.6921	0.6954

Table 22: Mean of the fractional loss in a $M/E_r/1/10$ queue

Time (Service units)	$\rho = 0.8, r = 10, m = 10$		
	Simulation (90% C.I.)	Approx1	Approx2
10.0	1.9509e-01 (2.5818e-03)	2.0449e-01	1.9180e-01
20.0	1.1956e-01 (1.5945e-03)	1.2493e-01	1.2045e-01
30.0	8.8172e-02 (1.2257e-03)	9.139e-02	8.897e-02
40.0	7.0444e-02 (1e-03)	7.229e-02	7.073e-02
50.0	5.7722e-02 (8.3936e-04)	5.988e-02	5.877e-02

The b_i are computed as

$$\begin{aligned} b_i &= \frac{1}{\tau}\int_0^{\tau}\int_{s+(i-1)\tau}^{s+i\tau}\lambda e^{-\lambda u}duds, \;\; i>0 \\ &= \frac{1}{\rho}e^{-(i-1)\rho}(1-e^{-\rho})^2 \end{aligned} \tag{77}$$

and

$$\begin{aligned} b_0 &= \frac{1}{\tau}\int_0^{\tau}(1-e^{-\lambda u})du \\ &= 1-\frac{(1-e^{-\rho})}{\rho} \end{aligned} \tag{78}$$

The first arrival sees a distribution of the queue state that is not governed by the above transition probability matrix when there is a non-empty queue at $t=0$ and service for the first customer starts at $t=0$ (rather than being sampled from a uniform distribution). In this case, we have

$$\begin{aligned} b_i &= \int_{i\tau}^{(i+1)\tau}\lambda e^{-\lambda u}du \\ &= e^{-i\rho}(1-e^{-\rho}) \end{aligned} \tag{79}$$

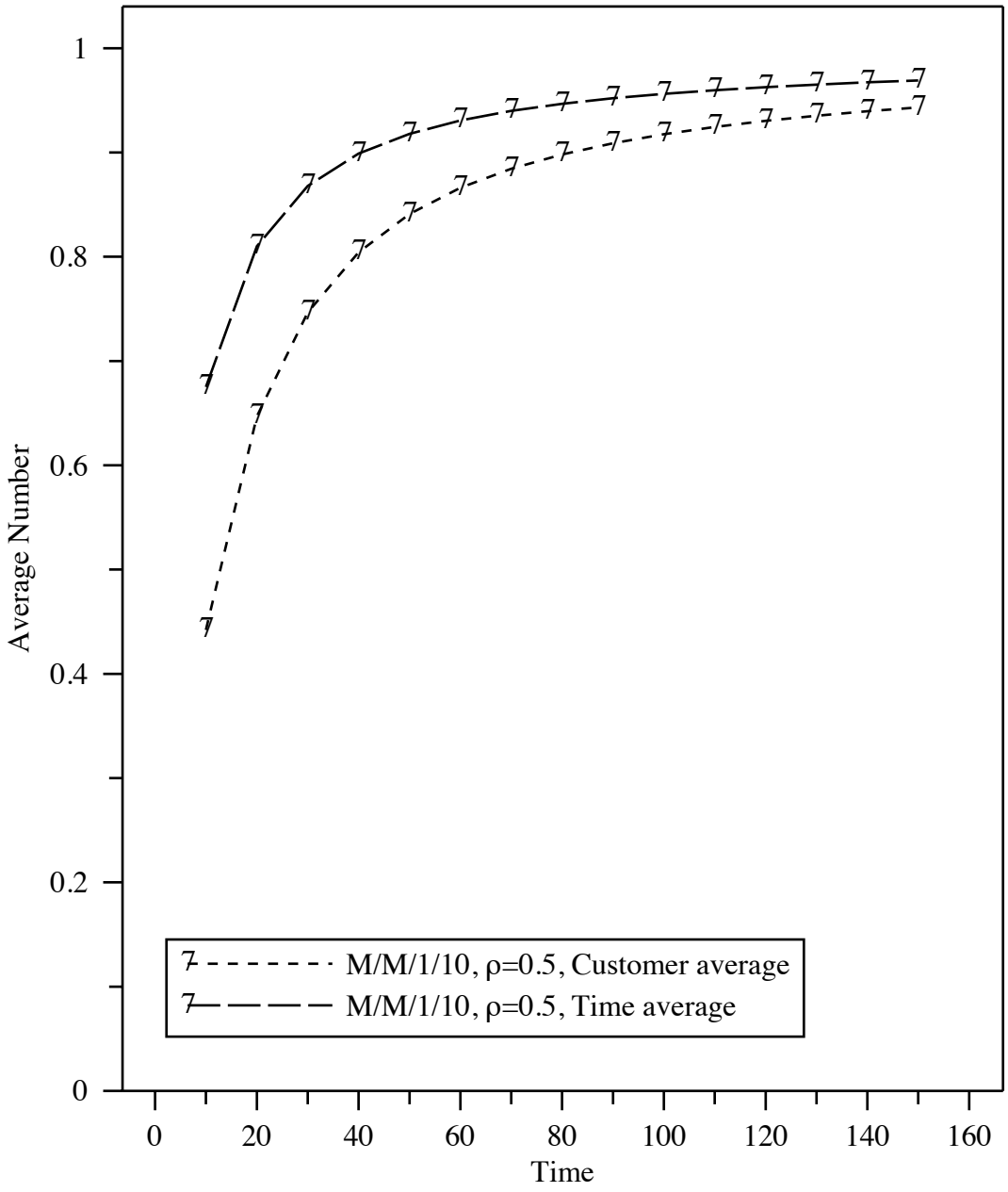

Figure 1: Mean customer and time-averaged number in the $M/M/1/10$ queue

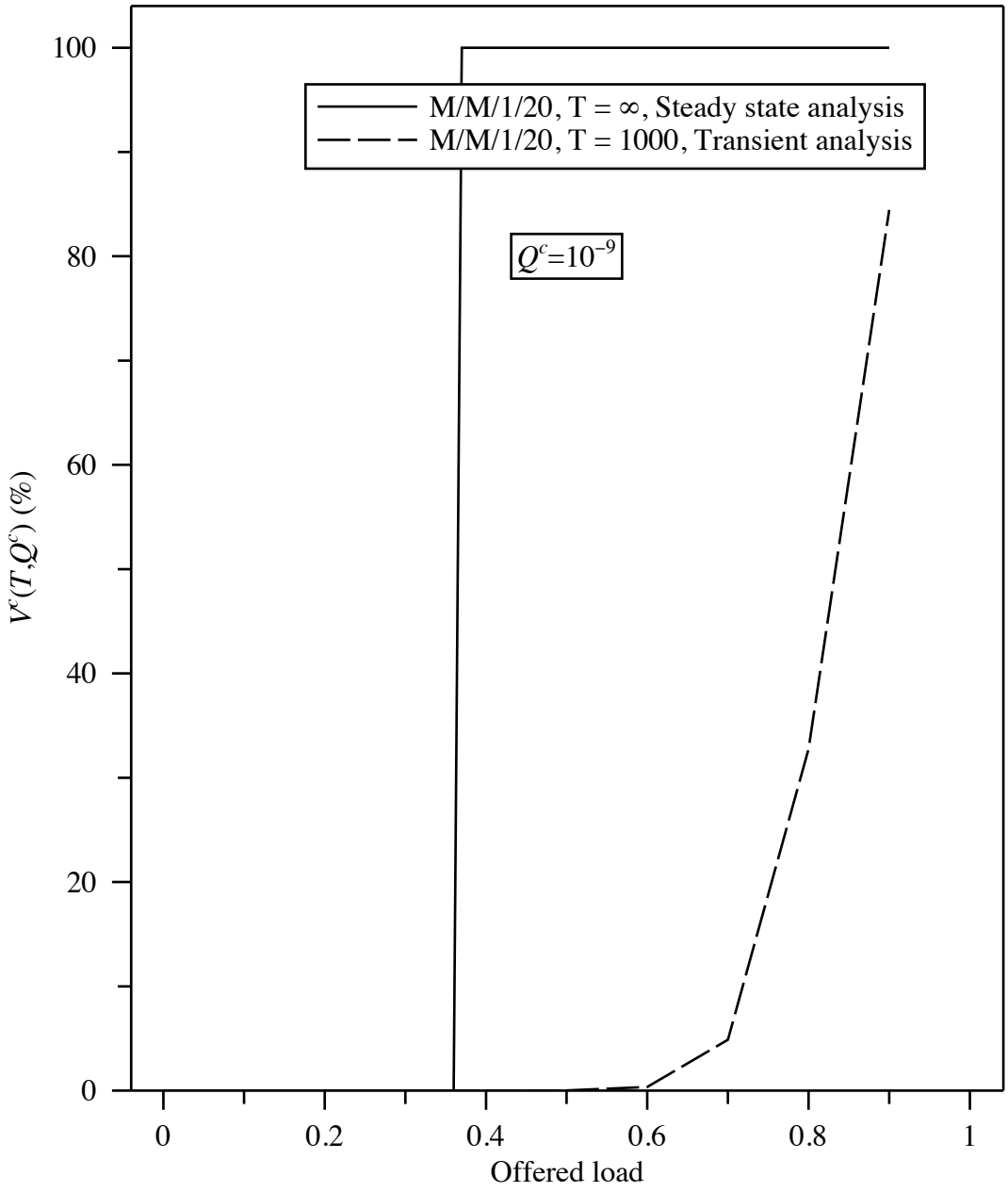

Figure 2: Violation level in the M/M/1/20 queue: Effect of Offered load

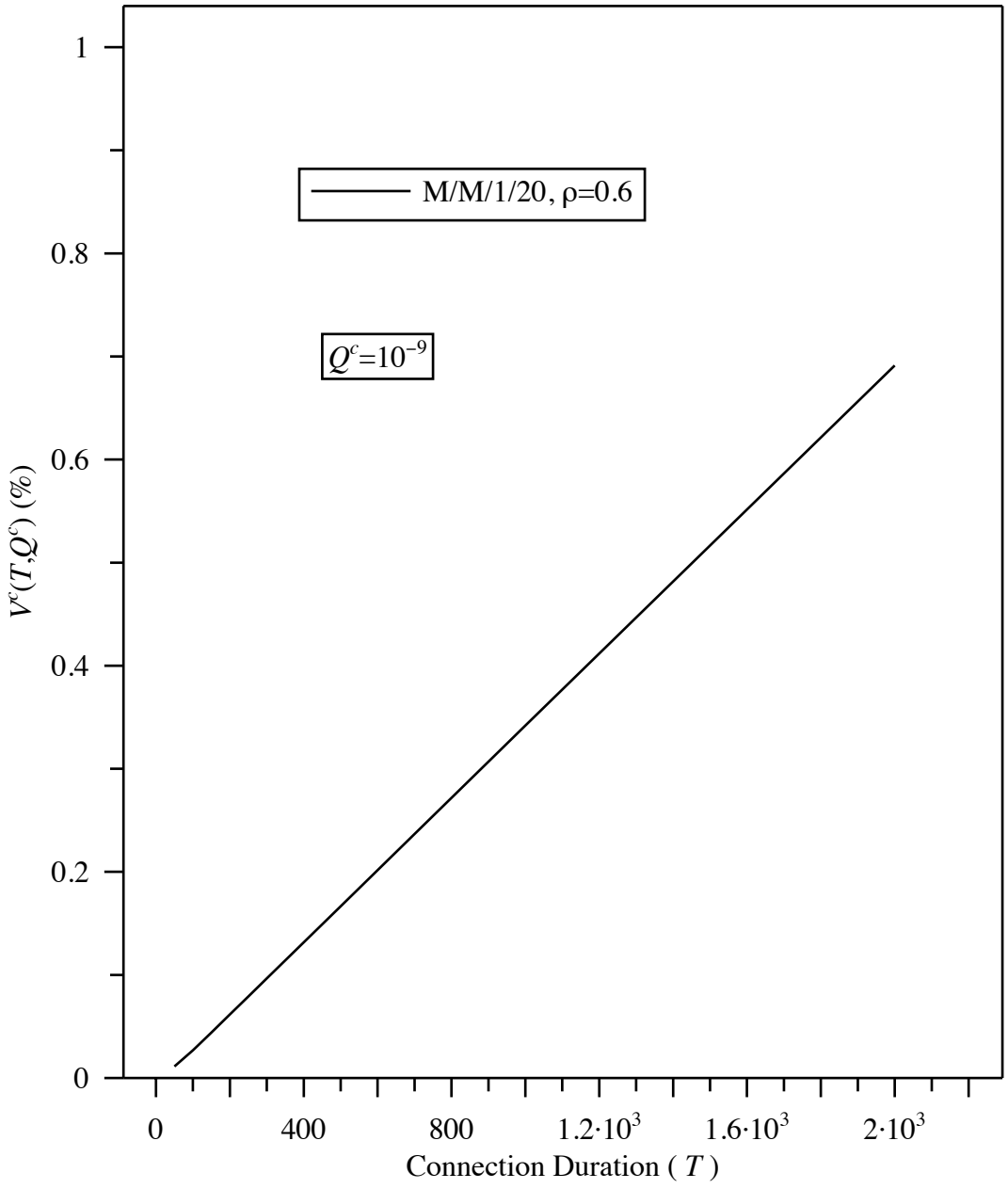

Figure 3: Violation level in the M/M/1/20 queue: Effect of connection duration

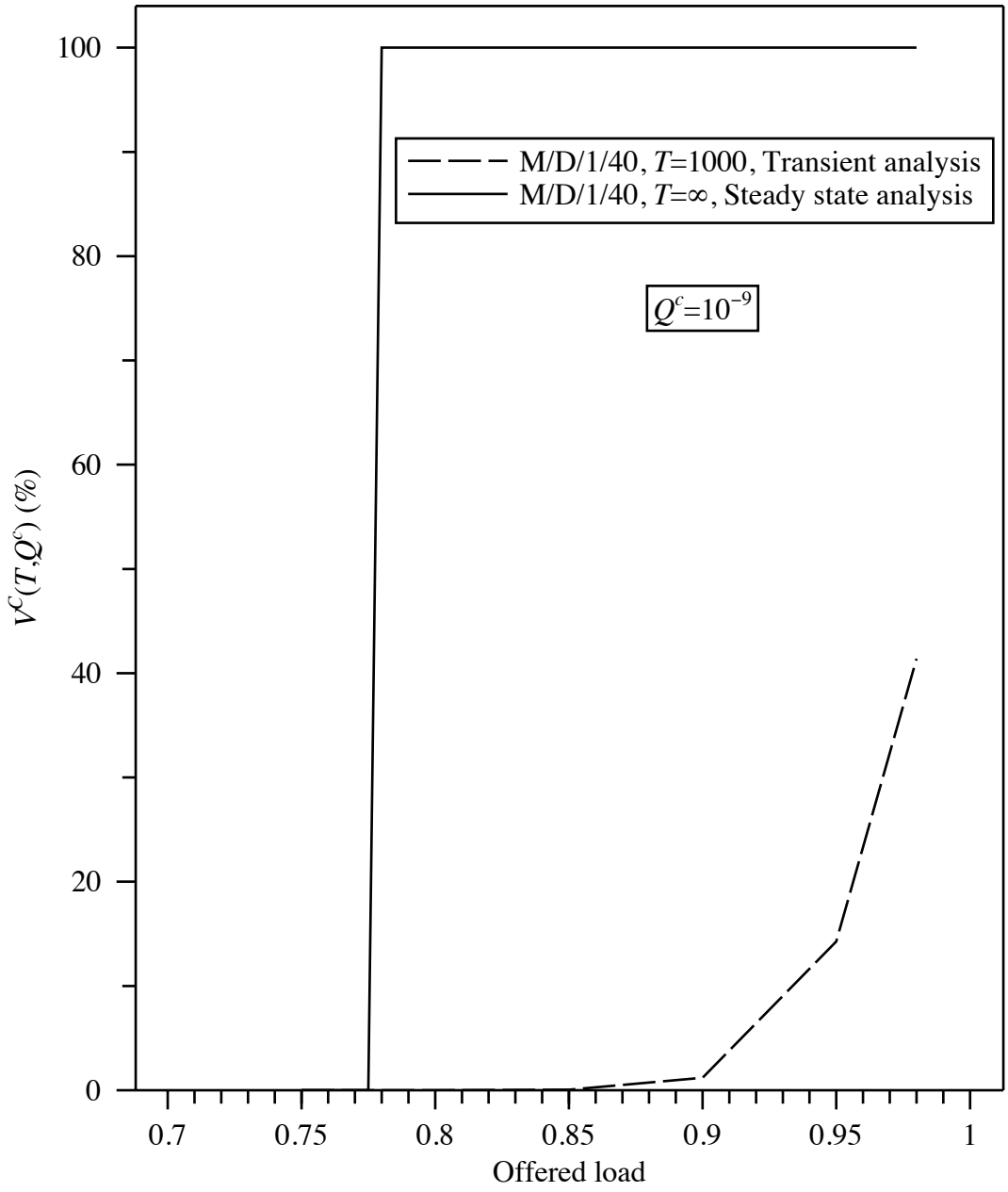

Figure 4: Violation level in the M/D/1/40 queue: Effect of Offered load

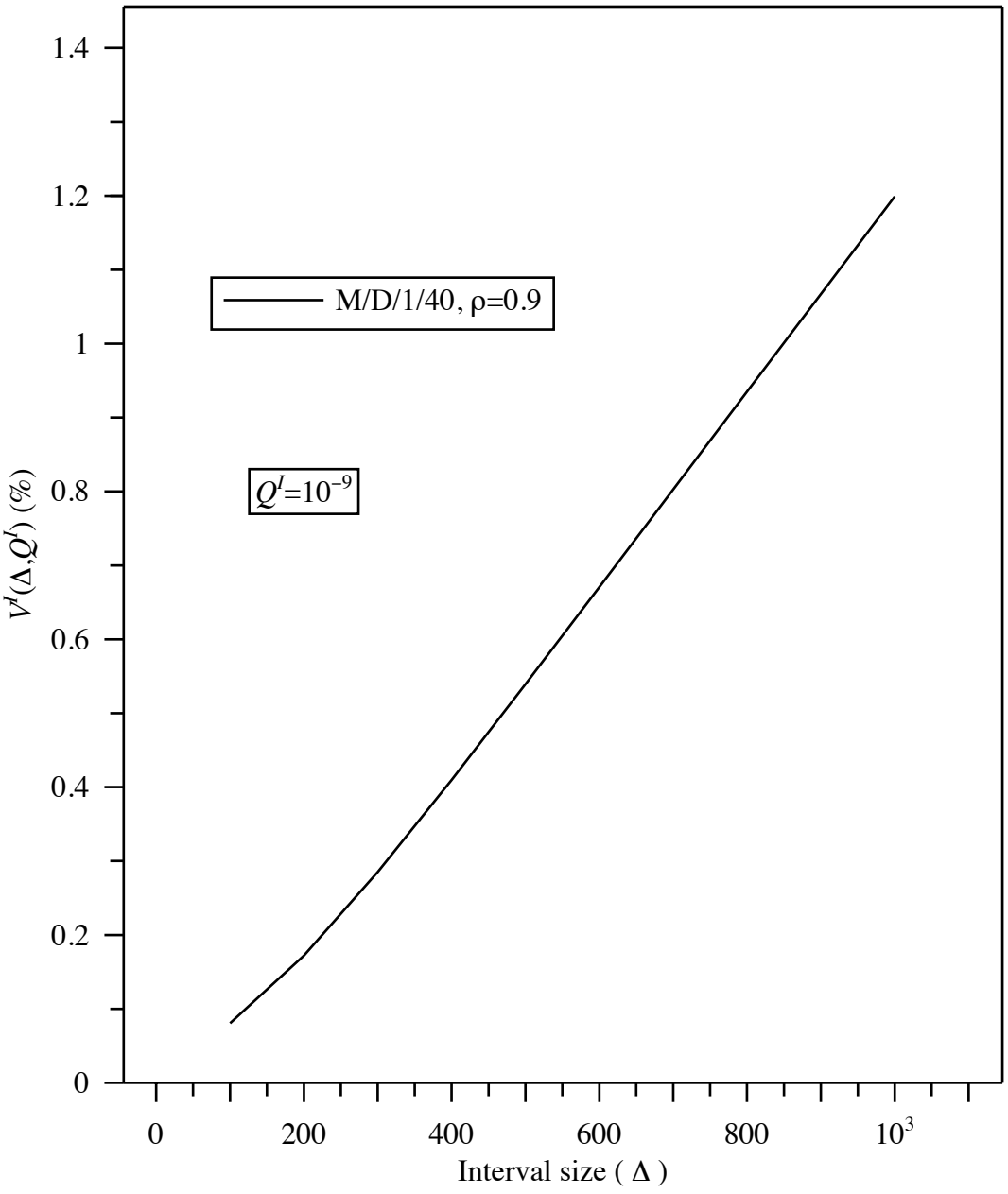

Figure 5: Effect of interval size on the interval QOS

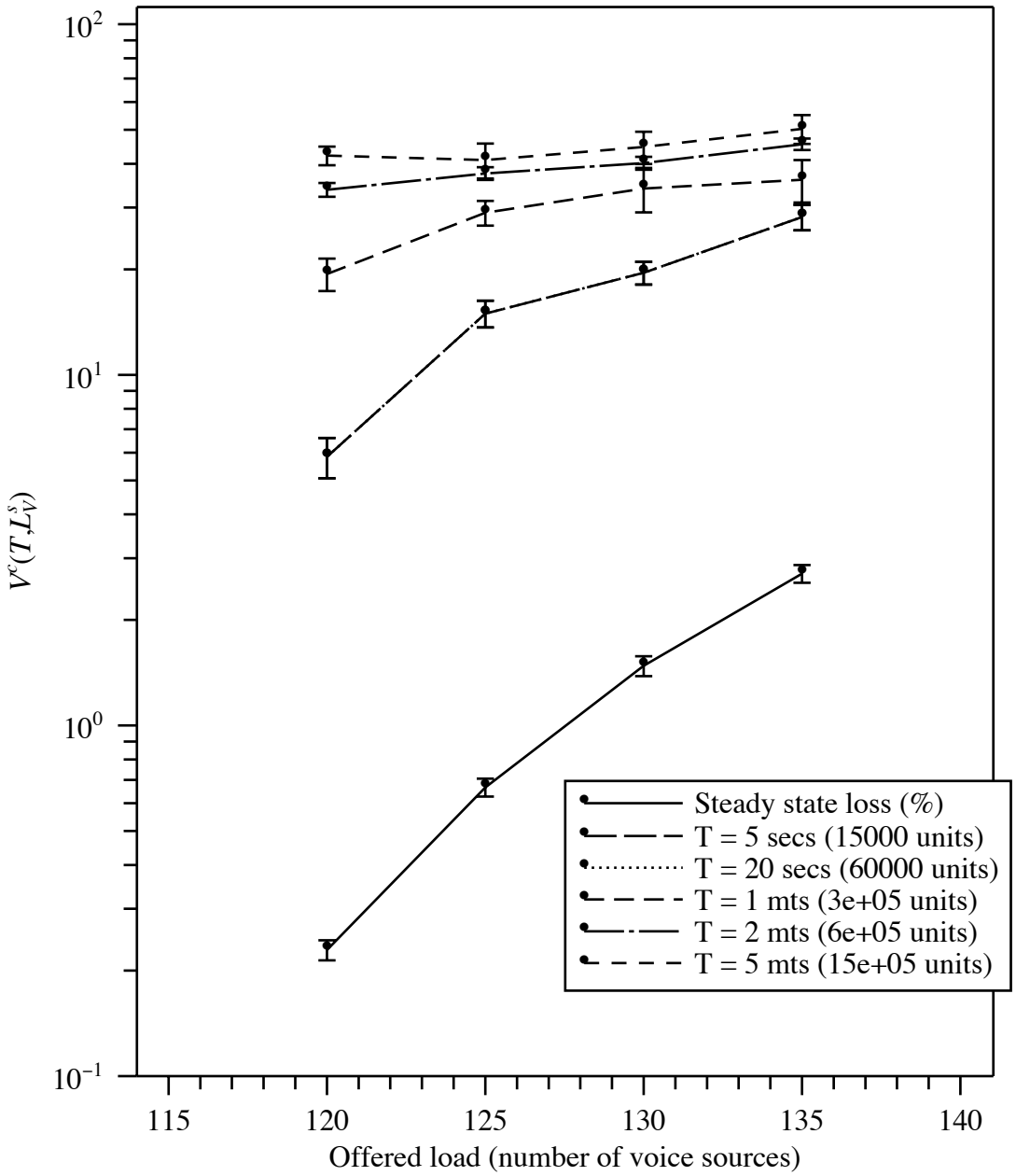

Figure 6: Violation levels in the voice multiplexer with steady state loss QOS values

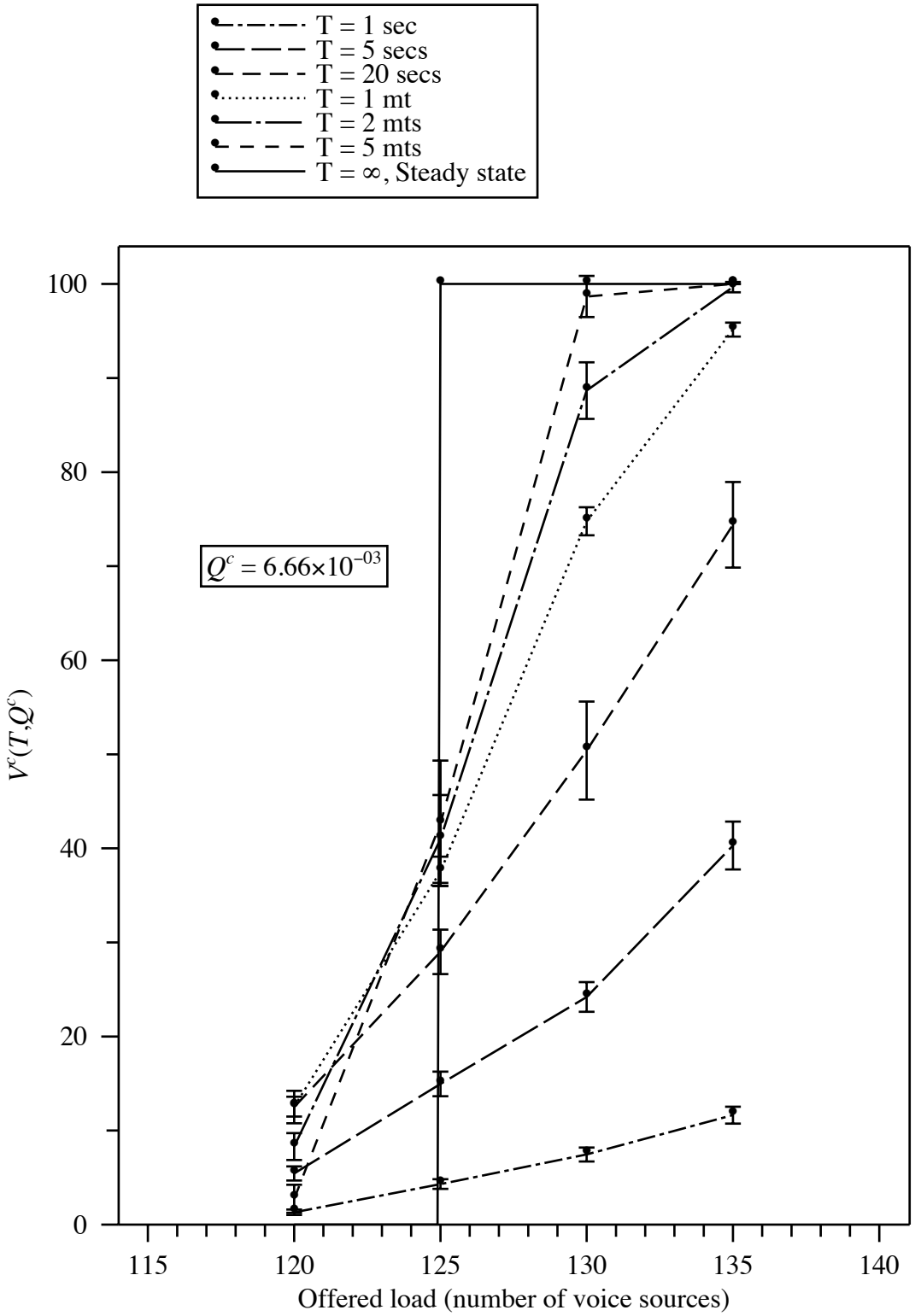

Figure 7: Violation levels in the voice multiplexer - Effect of connection duration

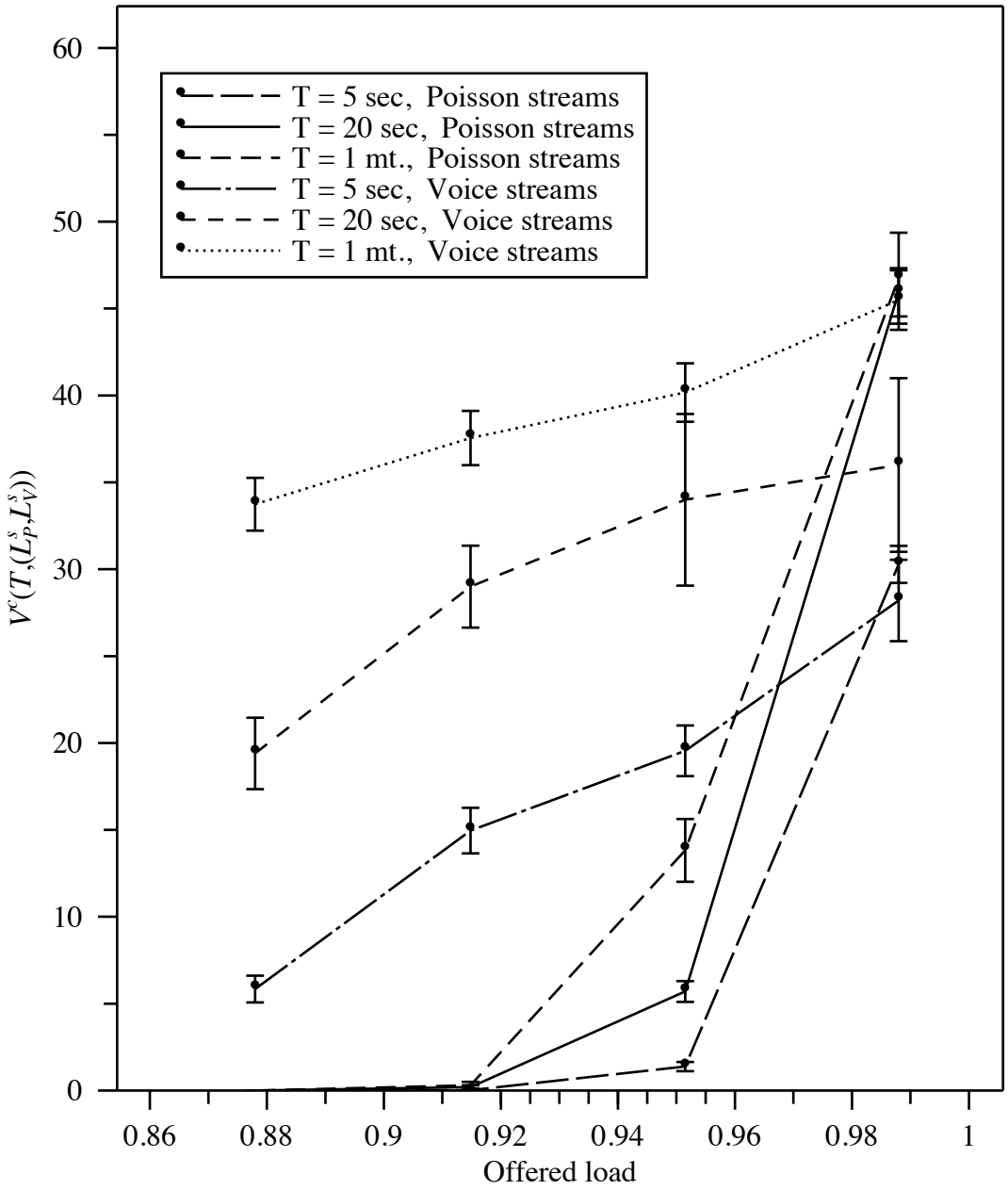

Figure 8: Violation levels in the Voice and Poisson sources multiplexer at steady state loss QOS values - A Comparison

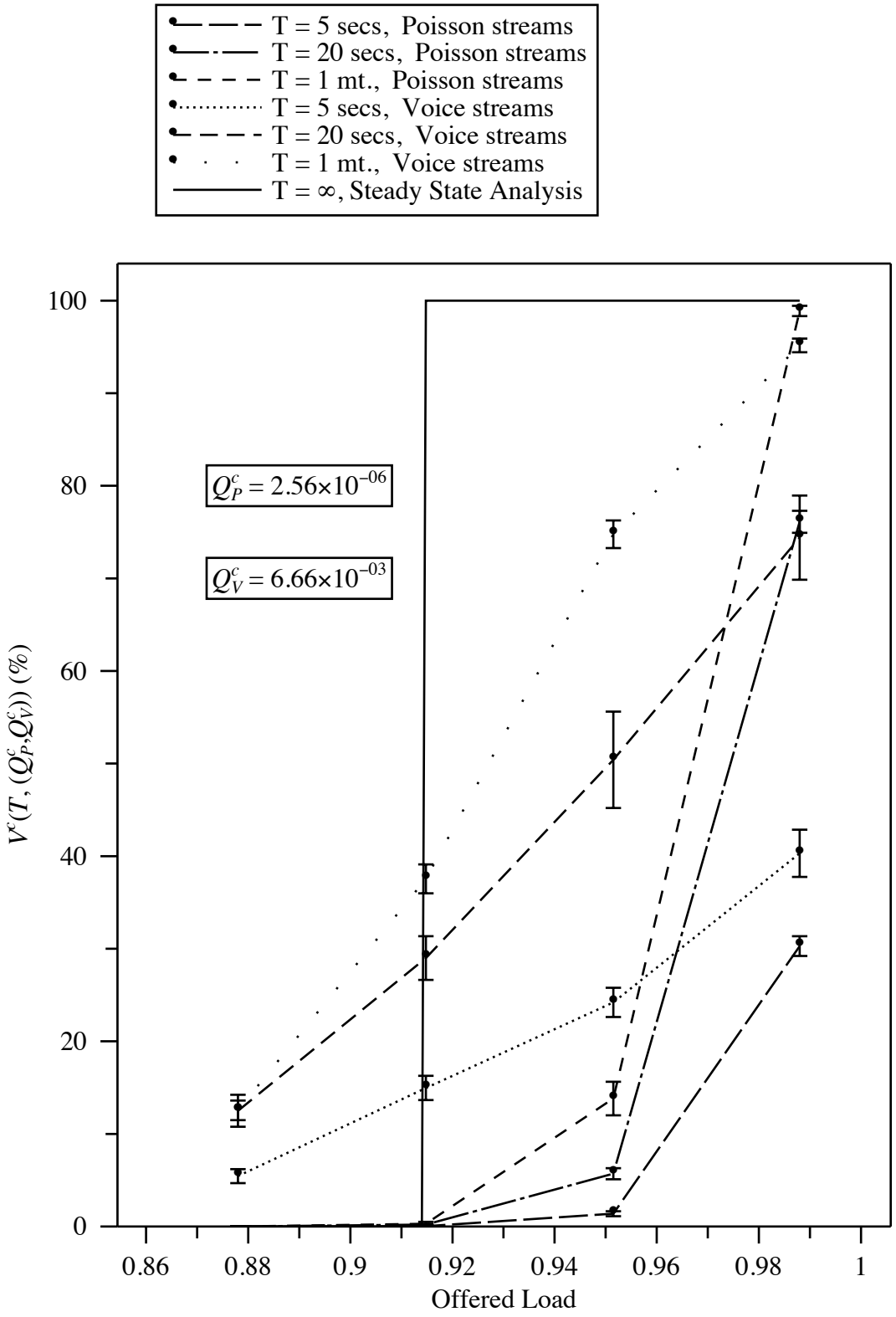

Figure 9: Violation levels in the Voice and Poisson sources multiplexer at a fixed QOS value - A Comparison

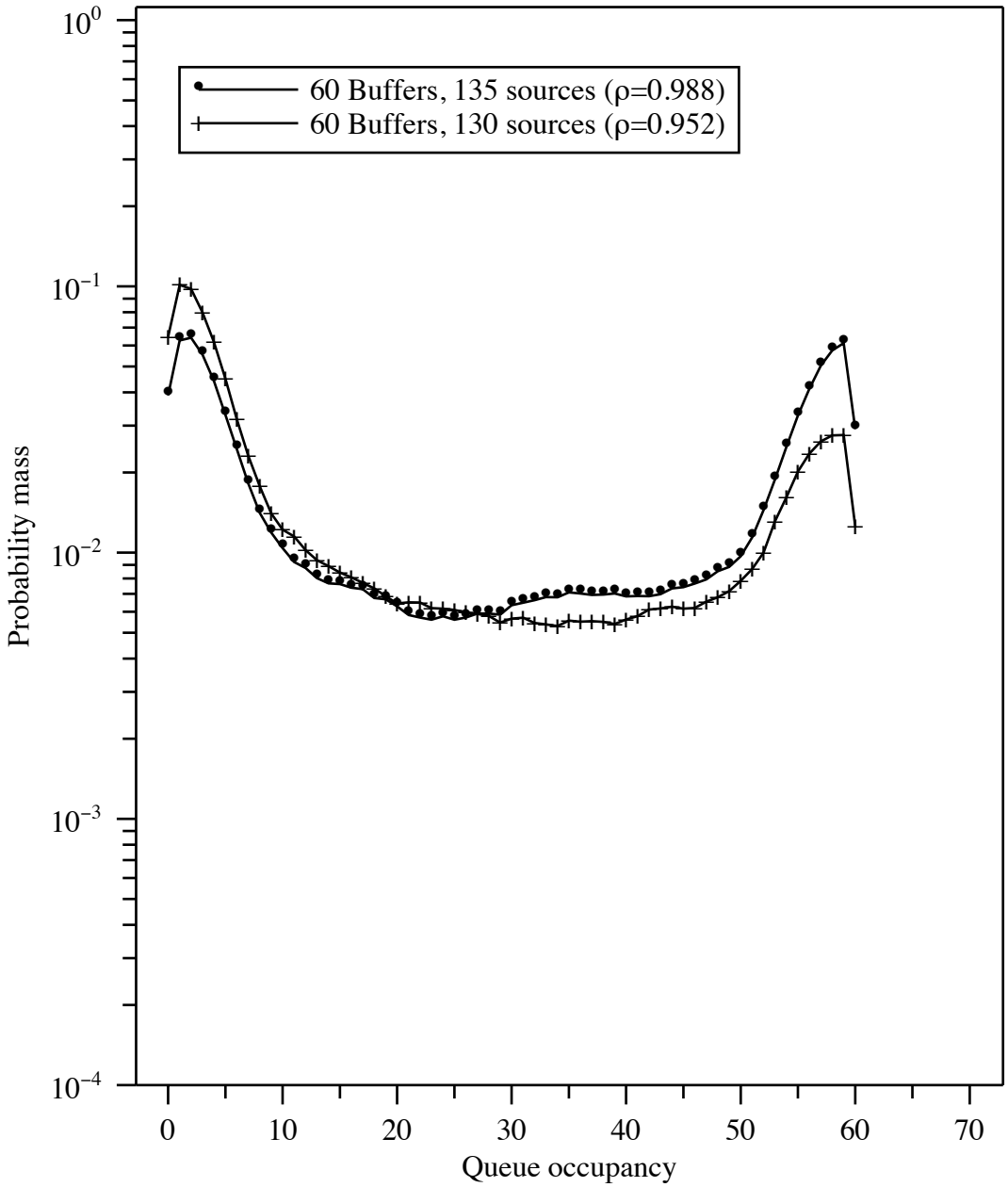

Figure 10: Steady state occupancy distribution in the voice multiplexer

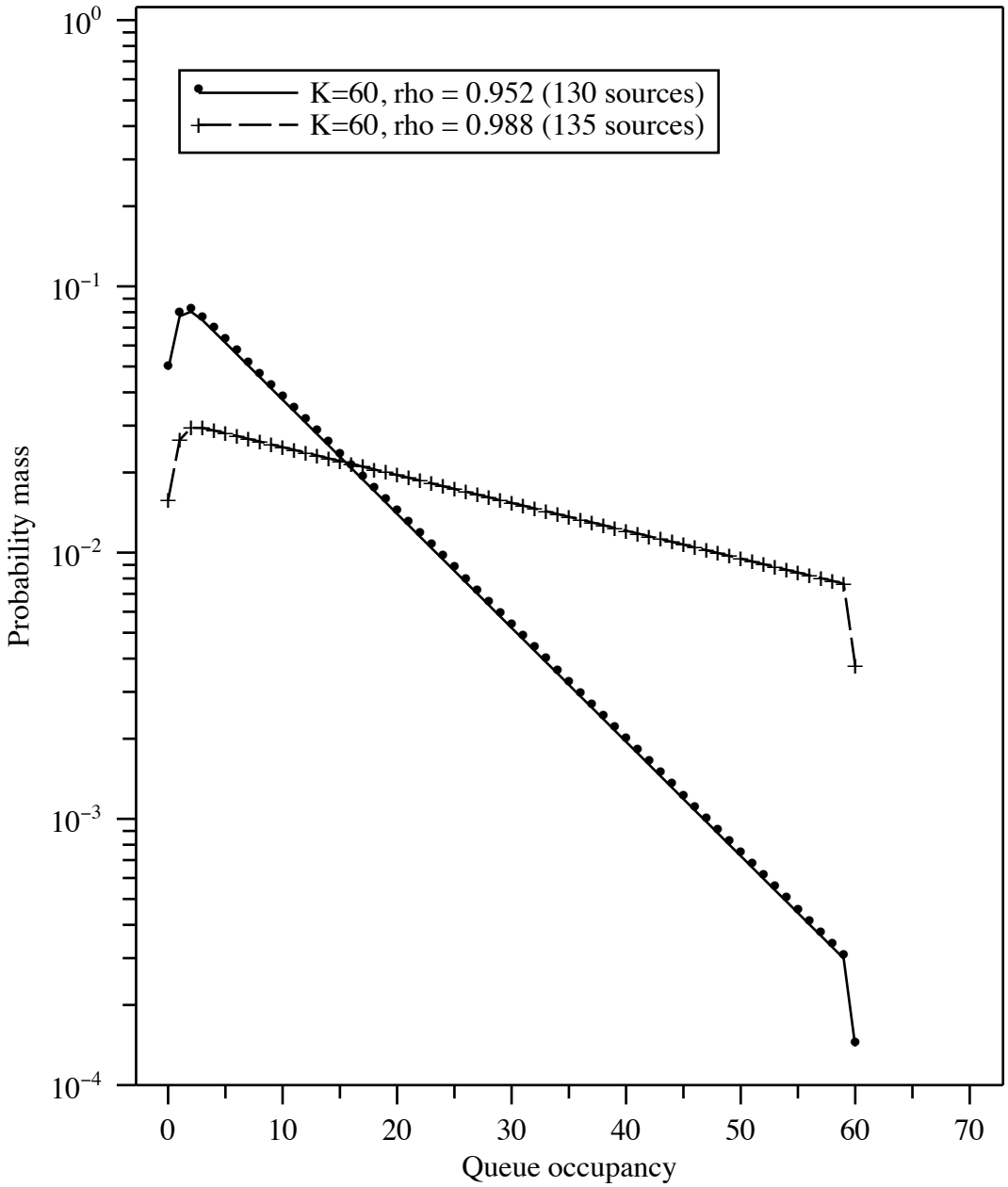

Figure 11: Steady state occupancy distribution in the Poisson multiplexer