

Dimensioning Bandwidth for Elastic Traffic in High-Speed Data Networks

Arthur W. Berger, *Senior Member, IEEE*, and Yaakov Kogan, *Senior Member, IEEE*

Abstract—Simple and robust engineering rules for dimensioning bandwidth for elastic data traffic are derived for a single bottleneck link via normal approximations for a closed-queueing network (CQN) model in heavy traffic. Elastic data applications adapt to available bandwidth via a feedback control such as the transmission control protocol (TCP) or the available bit rate transfer capability in asynchronous transfer mode. The dimensioning rules satisfy a performance objective based on the mean or tail probability of the per-flow bandwidth. For the mean objective, we obtain a simple expression for the effective bandwidth of an elastic source. We provide a new derivation of the normal approximation in CQNs using more accurate asymptotic expansions and give an explicit estimate of the error in the normal approximation. A CQN model was chosen to obtain the desirable property that the results depend on the distribution of the file sizes only via the mean, and not the heavy-tail characteristics. We view the exogenous “load” in terms of the file sizes and consider the resulting flow of packets as dependent on the presence of other flows and the closed-loop controls. We compare the model with simulations, examine the accuracy of the asymptotic approximations, quantify the increase in bandwidth needed to satisfy the tail-probability performance objective as compared with the mean objective, and show regimes where statistical gain can and cannot be realized.

Index Terms—Asymptotic approximation, asynchronous transfer mode, closed queueing networks, computer network performance, effective bandwidths, Internet, traffic engineering, transmission control protocol.

I. INTRODUCTION

THIS PAPER considers the problem of dimensioning bandwidth for elastic data applications in packet-switched communication networks, such as Internet protocol (IP) or asynchronous transfer mode (ATM) networks. Elastic data applications can adapt to time-varying available bandwidth via a feedback control such as the transmission control protocol (TCP) or the available bit rate (ABR) transfer capability in ATM. Typical elastic data applications are file transfers supporting e-mail or the World Wide Web. A contribution of this paper is to derive simple, closed-form dimensioning rules that satisfy a performance objective on per-flow or per-connection bandwidth. For the particular case of an objective on the mean per-flow bandwidth, we obtain a simple expression for the effective bandwidth of an elastic traffic source.

Manuscript received March 10, 1998; revised April 20, 1999 and June 7, 2000; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor J. Wong.

A. W. Berger was with AT&T Labs, Middletown, NJ 07748 USA. He is now with Akamai Technologies, Cambridge, MA 02139 USA (e-mail: arthur@akamai.com).

Y. Kogan is with AT&T Labs, Middletown, NJ 07748 USA (e-mail yakov@buckaroo.mt.att.com).

Publisher Item Identifier S 1063-6692(00)09122-6.

We assume that the closed-loop controls for the elastic data applications are performing well. For the present work, the key attribute of a well-performing control is that it maintains some bytes in queue at the bottleneck link with minimal packet loss. In contrast, poorly performing controls oscillate between undercontrolling (leading to excessive packet or cell loss) and overcontrolling (where, for example, TCP needlessly enters a slow-start or a time-out state). Thus, poorly performing controls limit the user’s throughput below what would have been possible, given the available bandwidth. Our assumption of well-performing controls is consistent with ongoing research efforts to improve current controls.

We focus on a single bottleneck target link and obtain a simple product-form closed-queueing network (CQN) model in heavy traffic. CQN models have been used extensively for the analysis and design of computer systems and networks, see, for example, [1] or [2], and our model is one of the simplest; though the entity modeled by a “job” is nonstandard, see Section II. Our CQN model consists of one infinite server (IS) and one processor-sharing (PS) server and was motivated by Heyman, Lakshman, and Neidhardt, [3], who also use a CQN model where the focus is on analyzing the performance of the closed-loop control TCP in the context of Web traffic. An important practical feature of these CQN models is their insensitivity: the distribution of the underlying random variables influences the performance only via the mean of the distribution. Thus, given the assumptions of the model, our engineering rules pertain in the topical case when the distribution of file sizes is heavy-tailed [4], though with finite mean, and thus the superposition of file transfers is long-range dependent [5].

The assumption of processor sharing in the CQN model is needed for the theoretical results, and corresponds to an assumption that the output ports of the network nodes use some type of fair queueing across the class of elastic data flows/connections; for a recent review of implementation of various fair queueing algorithms, see Varma and Stiliadis [6]. However, the model predictions do not seem particularly sensitive to this assumption, as illustrated in Section VI-A by a comparison with simulations that used a first-in-first-out (FIFO) discipline.

Since forecasted demand has significant uncertainty, we believe that simple ballpark dimensioning rules are appropriate. In order to attain such closed-form rules, we use a normal approximation, which pertains in the regime of interest: heavy traffic with many flows and with high-speed links. A second contribution of the paper is new derivations of the normal approximation that in particular provide an estimate on the error of this approximation.

The dimensioning problem focused on here is actually just a piece of the overall network design problem. In the present

paper, we focus on a single link and assume all of the flows/connections on the link are bottlenecked at this link. This is equivalent to the conservative procedure of sizing each link for the possibility that it can be the bottleneck for all of the connections on it. In subsequent work, we plan to extend the results to multiple links and to the overall network design problem. One aspect of this problem is the identification of bottleneck nodes in general networks; see [7] for initial results.

A. Related Work

One can view the above dimensioning problem as a variation of the well-known “capacity assignment problem” in the literature on design of computer networks, see, e.g., [8]. In the present paper, the exogenous inputs are not the offered “flows” (bits/s) but rather the file sizes, as here we consider the offered flows as dependent on the state of the network via the closed-loop control.

One can also make a comparison with traditional traffic engineering of telephone networks. Traditional dimensioning for telephone circuits uses the Erlang blocking formula, and, for multirate circuits, the generalized Erlang blocking model and associated recursive solution of Kaufman [9] and Roberts [10] and asymptotic approximations of Mitra and Morrison [11]. These formulas assume constant-rate connections, which is natural for traditional circuits. More recently, the concept of effective bandwidth extends the applicability of these formulas by assigning a constant “effective” rate to variable-rate connections, see Chang and Thomas, [12], de Veciana, Kesidis and Walrand, [13], or Kelly [14] for recent summaries. Assigning an effective bandwidth is reasonable (in some parameter regions) for classes of traffic such as packet voice, packet video, frame relay, and statistical bit rate service in ATM. However, for elastic data applications that do not have an inherent transmission rate, but rather adapt to available bandwidth, the concept of effective bandwidth seems dubious. Nevertheless, for the mean performance criterion, we do find an effective bandwidth for an elastic source, and moreover, the expression is a simple harmonic mean of two rates that are indeed naturally associated with elastic data and will occur under respective limiting network conditions.

Although we make the conservative assumption that the closed loop control is performing well, one should note that a topic of on-going research is the enhancement of closed-loop controls to improve the throughput that is indeed realized in various contexts. The implemented changes in TCP of congestion avoidance, fast retransmit, and fast recovery have this aim; see for example [15]. Further changes are also being proposed, such as explicit congestion notification, [16], [17]. The design of the feedback control for the ABR transfer capability in ATM received much attention at the ATM Forum, an industry forum; for a summary, see Fendick [18]. Also, of recent interest is the performance of TCP/IP over an ATM connection that is either unspecified bit rate (UBR) or ABR: significant degradation in throughput has been observed and various enhancements have been considered, see, for example, Romanov and Floyd [19], and Fang *et al.* [20]. The previously cited Heyman–Lakshman–Neidhardt’s model predicts behavior both when the

performance is poor and good [3]. Balakrishnan, Rahul, and Seshan propose an integrated congestion management architecture that provides feedback to the application layer and applies to user datagram protocol (UDP) as well as TCP, [21]. Thus, we view as complementary network design that assumes well-performing closed-loop controls and control implementations that make good use of the deployed bandwidth.

B. Outline of Remaining Sections

Section II describes the CQN model and performance objectives. Section III summarizes pertinent asymptotic approximations and presents new results that provide more accurate approximations and an estimate of the error of the normal approximation. Section IV presents the proposed dimensioning rules including the simple effective bandwidth formula. Section V considers the analogous problem of the number of sources that can be supported on a link of given bandwidth. Section VI contains numerical results on the accuracy of the model and on the implications of the dimensioning rules. Section VII contains the conclusions.

II. CLOSED-QUEUEING NETWORK MODEL

We use the terms “source” and “flow” to apply to an IP flow of packets, which is specified by the source and destination IP addresses, or ranges thereof, and possibly the protocol field in the IP header or port numbers in the UDP and TCP headers, and with the particular choice determined by the network operator. We also use the terms source and flow to refer to the traffic on an ATM connection, or more generally a virtual circuit in a connection-oriented packet network. We use the term “link” for the object to be dimensioned, which may indeed be an entire transmission path devoted to the elastic data traffic, or may be a portion thereof, such as a label-switched path in multiprotocol label switching, or a virtual path connection in the context of ATM.

The link is to be sized to support N sources. Thus, for the dimensioning step, we take the viewpoint that there is a static number of connections present, equivalently when one source terminates another begins. Each source alternates between two phases: active and idle. The important case we have in mind is where the packets in an activity period constitute the transmission of a file (or files) across a wide-area high-speed network where multiple packets are typically in transit at a given time.

The fixed number of sources and their alternation between active and idle phases makes plausible the use of a CQN model with two types of servers. The first type is a processor sharing (PS) server that models the queueing and emission of packets on a chosen target link. The second type is an IS node (equivalently the number of servers equals to the number of sources) that models the sources while they are in the idle phase and also models all other relevant network components besides the target link. That is, the mean time in the infinite server (IS) node represents the mean time between the initiation of a file transfer, given that the target link has enough capacity that it imposes only negligible delay on the transfer of the file. A diagram of the CQN model is given in Fig. 1.

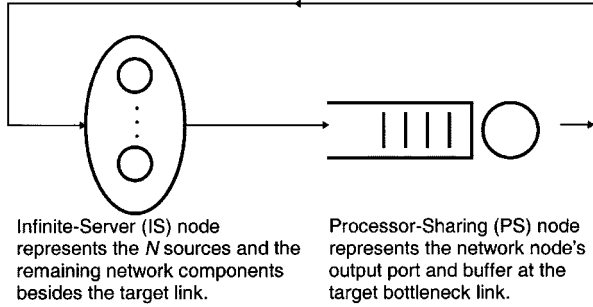


Fig. 1. Closed-queueing network model.

A. CQN Model Represents Files, Not Packets

The one nonstandard aspect of the CQN model is the entity represented by a “job.” A job in the CQN model represents a file (or multiple files). A job moving from the IS node to the PS node in the CQN represents the beginning of a file transfer in the actual network. This modeling assumption is intended for the following network scenario.

For dimensioning or admission control at the target link, we are interested in scenarios where the link is heavily loaded. That is, we take the conservative viewpoint that the target link should satisfy a per-flow bandwidth objective (see Section II-C) even when it is heavily loaded. Moreover, we make the further conservative assumption that the target link is the limiting factor on the throughput obtained for the all of the elastic data sources. The feedback controls of TCP and ABR tend to seek out and fill up the available bandwidth. A source’s feedback control, when properly designed and functioning, will attempt to keep at least one packet queued for transmission on the bottleneck link (otherwise the control is needlessly limiting the throughput). We assume that this is the case. Thus we obtain a key simplifying assumption for the model: at an arbitrary point in time, the number of sources that are currently transferring a file across the bottleneck, target link equals the number of sources that have a packet in queue at the link. Thus, in the CQN model, the number of jobs at the PS node represents the number of sources that are currently transferring a file.

As will be illustrated in Section VI-A, the CQN model begins to break down when a source’s feedback control is overconstraining, and the otherwise bottleneck node empties of packets of a source that is still in the process of transferring a file. Also, note that the CQN does not model a lightly loaded link.

Note that a job in the CQN does not capture the location of all of the *packets* of a file transfer, since at a given moment some of these packets may have reached the destination, while other packets are in transit, and others are still at the source. In particular, the queue at the PS node in the CQN node represents the number of *files* currently being transferred, and does not represent the number of *packets* queued at the egress port. Thus, the present CQN model does not attempt to capture packet losses or packet delays. However, it does attempt to model the number of sources that are in the process of transferring a file, given the assumption of well-performing controls. And this is exactly what we need to capture the desired per-flow performance objective; see Section II-C.

It is worth noting what a more detailed packet-level model could capture and its advantages and disadvantages. A model that captures the flow of individual packets can be used to predict packet queue lengths and delays, and thus could be used to dimension buffer sizes and to predict packet losses due to buffer congestion. CQN models have been used in the past to examine packet-level phenomena, and in particular have modeled window flow-control schemes where the number of jobs in the CQN equals the window size [22]. Packet-level models could be used for any traffic loading, not just heavy loads. A packet-level model is not used in the present paper as it introduces more detail and complexity than is needed. In contrast, the file-level model fits well with the performance objectives, (6) and (7), and yields what we are seeking in this paper: simple closed-form engineering rules, (25) and (30), where the former provides a simple expression for the effective bandwidth of an elastic traffic source.

B. Steady-State Solution to CQN Model

The parameters of the CQN model are: 1) the number of sources, N ; 2) the mean service at the PS node, denoted μ^{-1} ; and 3) the mean time in the IS node, denoted λ^{-1} . The mean service time at the PS node represents the mean time to transmit a file on the target link given no other files present. With the mean file size denoted f and the capacity of the target link denoted B , then $\mu^{-1} = f/B$. The mean time in the IS node, λ^{-1} , represents the mean time between initiation of file transfers by a source, in the hypothetical case that the target link imposes negligible constraint on the transfer.

Let Q_0 be the number of jobs at the IS node, and let Q_1 be the number jobs at the PS node at an arbitrary time. The steady-state distributions for Q_0 and Q_1 are well known, e.g., [8], and the following forms are useful for the sequel.

$$\Pr(Q_1 = n) = \frac{1}{G(N)} \cdot \frac{(\lambda/\mu)^n}{(N-n)!} = \frac{1}{H(N)} \frac{(\mu/\lambda)^{N-n}}{(N-n)!} \quad (1)$$

and

$$\begin{aligned} \Pr(Q_0 = n) &= \Pr(Q_1 = N-n) \\ &= \frac{1}{G(N)} \cdot \frac{(\lambda/\mu)^{N-n}}{n!} \\ &= \frac{1}{H(N)} \frac{(\mu/\lambda)^n}{n!} \end{aligned} \quad (2)$$

where

$$G(N) = \sum_{k=0}^N \frac{(\lambda/\mu)^k}{(N-k)!} \quad \text{and} \quad H(N) = \sum_{k=0}^N \frac{(\mu/\lambda)^k}{k!}. \quad (3)$$

As λ and μ appear in (1)–(3) only as a ratio and as $\lambda/\mu = \lambda f/B$, the distribution of jobs at the PS node and thus the key dimensioning equations including the effective bandwidth formula depend on λ and f only via their product. This product is the throughput of the source, in bits/s, given that the target link is imposing no restriction on the flow. A network operator could estimate the parameter λf from measurement studies during periods when its network is lightly loaded.

C. Performance Criteria

In elastic data applications, the user, and hence the network designer, is concerned with the delay in transferring a file. Thus, file transfer times are the natural performance criteria for elastic applications. (In contrast, for real-time, stream applications, such as voice, the delay and delay variation of individual packets is of concern.) Since file sizes vary greatly, a single delay objective, such as 100 ms, for all files is not sensible. Rather, the delay objective should be normalized by the file size, which yields a performance objective in units of s/bit. More conveniently, we consider the reciprocal, so that the performance objective is in terms of the bandwidth, in bits/s, that an arbitrary active source obtains. Let B_s denote the bandwidth that an arbitrary source obtains in steady state. This bandwidth per source is defined as

$$B_s = \frac{B}{\hat{Q}_1} \quad (4)$$

where B is the link bandwidth and \hat{Q}_1 is the conditional number of jobs in the PS node given that the PS node is not empty. By its definition

$$\Pr(\hat{Q}_1 = n) = \frac{\Pr(Q_1 = n)}{\Pr(Q_1 > 0)}, \quad n = 1, \dots, N. \quad (5)$$

We consider performance criteria on the mean and on the tail probability of B_s

$$E[B_s] \geq b \quad (6)$$

or

$$\Pr(B_s < b) < \alpha \quad (7)$$

for given b and α , where we have in mind typical values for b in the range of 10^4 – 10^6 bits/s and α in the range of 0.01–0.1.

Using (5) and applying Jensen's inequality to (6), the performance criteria (6) and (7) are satisfied, respectively, if the following conditions in terms of Q_1 pertain:

$$\frac{E[Q_1]}{\Pr(Q_1 > 0)} \leq B/b \quad (8)$$

$$\frac{\Pr(Q_1 > B/b)}{\Pr(Q_1 > 0)} < \alpha. \quad (9)$$

In the following sections, we consider approximations in the heavy traffic region where N is large and $\Pr(Q_1 > 0)$ is exponentially close to 1.

III. ASYMPTOTIC APPROXIMATIONS

Although the exact probability mass function (pmf) for Q_1 , (1), can be used in a numerical iteration to compute exactly the bandwidth B such that the performance criterion (6) or (7) is satisfied, it does not yield the simple closed-form dimensioning rules that we are seeking. These closed-form expressions yield insights, including the simple effective bandwidth formula, that are not evident in (1), where the dependence of Q_1 on B is obscured. In addition, even for off-line computations, such as in design tools, the closed-form expressions for B can be preferred

to a numerical iteration, if the computation is to be done many times and the tool is being used interactively for what-if studies. To obtain these closed-form expressions, we use the asymptotic approximations summarized in the present section.

A. Asymptotic Regime

In our application, N is large, and the service rate μ is also large and of the same order as N (a regime of many sources and of high-speed links). We assume that $\rho \equiv N\lambda/\mu$ is constant. Under these assumptions, the asymptotics of the normalization constant $N!G(N)$ and performance measures that can be derived from it (such as queue length, moments, and utilization) is given by Ferdinand [23]. In particular [23]

$$E[Q_1] \approx \frac{\rho}{1-\rho}, \quad \text{for } \rho < 1 \quad (10)$$

and

$$E[Q_1] \approx N(1-\rho^{-1}), \quad \text{for } \rho > 1 \quad (11)$$

for $N \gg 1$. Moreover, the PS-node utilization approaches ρ when $\rho < 1$, while its utilization is exponentially close to 1 when $\rho > 1$. Thus, when $\rho < 1$, the mean queue length and PS-node utilization is approximately the same as in an open queueing system with Poisson arrivals, FIFO service discipline, and exponential service times (M/M/1) with offered load ρ . In fact, the queue length distribution at the PS node converges to that of the M/M/1/ queue. For $\rho > 1$, there is no corresponding M/M/1 system. However, ρ still can be considered as a characterization of the load at the PS node because the mean queue length increases with ρ , as shown in (11). Note that (11) makes sense only for finite N , while the M/M/1 approximation, for $\rho < 1$, is obtained as $N \rightarrow \infty$. Here we are interested only in the heavy traffic regime defined by the condition

$$\rho > 1. \quad (12)$$

Thus we consider a regime where the utilization at the target link is exponentially close to 1. This is not the usual practice. More typically, one would try to engineer the link for an occupancy qualitatively less than 1, say 80%. However, to do so, particularly for data sources, one should have some understanding of how the occupancy varies over time scales. For example, the occupancy measured over an hour might be 80%, but within the hour, the five-minute occupancies could vary significantly and with multiple scattered seconds of occupancy near 1. In which case, it may be difficult to know whether the performance objective is indeed being met. Here, we avoid this difficulty by considering a more extreme case. Effectively, we consider the target link during its busy periods. This is consistent with our modeling assumption of well-performing end-system controls that keep packets in queue at the bottleneck link. Given that the per-source bandwidth objectives are satisfied at occupancies near one, the network operator is using a conservative design.

B. Asymptotic Results

For the tail performance criterion, we need asymptotics for the right-hand tail of the distribution for Q_1 given by $\Pr(Q_1 - N(1 - \rho^{-1}) > k)$. An asymptotic for the general product-form

CQN has been derived by Pittel [24, Lemma 1]. For our simple CQN, this asymptotic can be written in the following form:

$$\Pr(Q_1 = n) = c(N) \cdot \exp\{-N[F(x) - F(x^*)]\},$$

$$\rho > 1, \quad x \equiv n/N, \quad x \in (0, 1) \quad (13)$$

where

$$F(x) = (1-x) \ln(1-x) + x(1 - \ln \rho) \quad (14)$$

and

$$x^* = 1 - \rho^{-1} = \operatorname{argmin}_{x \in (0, 1)} F(x). \quad (15)$$

Here

$$F(x^*) = 1 - \rho^{-1} - \ln \rho \quad (16)$$

and $c(N)$ has the property that

$$\frac{\ln c(N)}{N} \rightarrow 0 \quad \text{as } N \rightarrow \infty.$$

Thus, the asymptotic (13) is logarithmic in the sense that

$$\frac{\ln(\Pr(Q_1 = n))}{N} \rightarrow -(F(x) - F(x^*)) \quad \text{as } N \rightarrow \infty.$$

Equation (13) implies that Q_1/N converges to x^* in probability as $N \rightarrow \infty$, as shown in Pittel [24, Theorem 2].

The function $F(x)$ is known as a quasipotential in the theory of large deviations, and it can be formally obtained as a solution of a nonlinear differential equation, [25]. By approximating $F(x)$ by its second order Taylor expansion about x^* , one obtains a normal approximation for Q_1 with mean Nx^* and variance N/ρ . This result is proved by Pittel [24] for general CQNs. The following two lemmas provide more accurate approximations for the distributions of Q_0 and Q_1 , and from which the normal approximation easily follows. Moreover, the second lemma provides an explicit estimate for the error in the normal approximation. The proofs are provided in the Appendix.

Lemma 1: For $\rho > 1$ and $N \gg 1$, $H(N)$ defined in (3) has the following asymptotic expansion:

$$H(N) = e^{N/\rho} \left(1 + o\left(e^{NF(x^*)}\right)\right) \quad (17)$$

where $F(x)$ is the quasipotential, and $F(x^*)$, given in (16), is negative, and

$$\Pr(Q_0 = n) = e^{-N/\rho} \cdot \frac{(N/\rho)^n}{n!} \left(1 + o\left(e^{NF(x^*)}\right)\right). \quad (18)$$

Thus, Q_0 is asymptotically Poisson with parameter N/ρ . In the sequel, we call this approximation the “ Q_0 -Poisson” approximation.

A Poisson distribution with parameter N/ρ can be considered as the distribution of the sum of independent random variables each with Poisson distribution with parameter $1/\rho$. The application of the central limit theorem results in:

Corollary 1: Under the conditions of Lemma 1

$$\frac{Q_0 - N/\rho}{\sqrt{N/\rho}} \Rightarrow Z \quad \text{as } N \rightarrow \infty \quad (19)$$

where Z has a normal distribution with mean zero and variance 1.

Lemma 2: Let $k = xN$ where x is constant, $\rho > 1$ and $N \gg 1$, and

$$0 < (1 - \rho(1 - x)) < \frac{c}{\sqrt{N}} \quad (20)$$

where c is a constant independent of N . Then

$$\Pr(Q_1 > k) = \frac{1}{\sqrt{\pi}} \int_{\beta(x)\sqrt{N}}^{\infty} e^{-y^2} dy - e^{-N\beta^2(x)}$$

$$\cdot \left[\frac{1}{3\sqrt{2\pi N(1-x)}} + O(N^{-3/2}) \right] \quad (21)$$

where

$$\beta^2(x) = F(x) - F(x^*)$$

$$= \rho^{-1} \{1 - \rho(1 - x) + \rho(1 - x) \ln[\rho(1 - x)]\} \quad (22)$$

where $F(x)$, which is the quasipotential, and $F(x^*)$, are given in (14) and (16) respectively.

Lemma 2 provides a numerically convenient and accurate calculation of the tail probabilities of Q_1 in the region of interest. This can be used in an iterative calculation for the bandwidth B that satisfies the performance criterion. To derive the normal approximation from (21), we consider only the first term, which provides

$$\Pr(Q_1 > k) = \Pr\left(Z > \beta(x)\sqrt{2N}\right) (1 + O(N^{-1/2})) \quad (23)$$

where Z has a normal distribution with mean zero and variance 1, and $\beta(x)$ is given by (22). If we in turn approximate $\beta(x)$ by expanding the log in (22) to second order [i.e., $\ln(1 - y) \approx -y - (1/2)y^2$], we obtain $\beta(x) \approx (2\rho)^{-0.5}(1 - \rho(1 - x))$, which results in

Corollary 2: Under the conditions of Lemma 2

$$\Pr(Q_1 > k) \approx \Pr\left(Z > \frac{k - N(1 - \rho^{-1})}{\sqrt{N/\rho}}\right). \quad (24)$$

Remark 1: In contrast with (13), asymptotic approximations (18) and (21) are exact in the sense that the ratio of the left-hand side to the first term in the right-hand side tends to 1 as $N \rightarrow \infty$.

Remark 2: In Lemma 2 in condition (20), note that $0 < 1 - \rho(1 - x)$ if and only if $k > N(1 - \rho^{-1})$, i.e., the quantile k is greater than the asymptotic mean. Also, $1 - \rho(1 - x) < c/\sqrt{N}$ if and only if $k - N(1 - \rho^{-1}) < (c/\rho)\sqrt{N}$, and thus the difference between the quantile k and the asymptotic mean is $O(\sqrt{N})$.

Remark 3: Asymptotic expansion (21) is obtained using the uniform asymptotic approximation for the partition function, which was initially derived in [26], see also [27]. A new element in (21) is dependence of the lower integration limit β on the parameter $x = k/N$, which can be viewed as a variable.

Remark 4: Asymptotic expansion (21) reveals two sources of inaccuracy of the normal approximation. First, we take only the first term in the expansion (21). Then, we approximate the quasipotential $F(x)$ by the second-order Taylor expansion. Numerical results in Section VI-B show that in our application the inaccuracy of the normal approximation is mainly caused by the

second step, however, when N is relatively small, both steps affect the accuracy.

Remark 5: Asymptotic approximation (21) can be similarly derived for the multiclass Erlang and Engset models where the uniform asymptotic expansion for the partition function is also available [27], [28]. Moreover, the form of (21), where the lower integration limit is explicitly expressed in (22) through the quasipotential, suggests its generalization for product-form multiclass CQNs and loss networks, by taking advantage of explicit expressions for the quasipotential that are available in many important particular cases (see, e.g., [7], [29]).

IV. DIMENSIONING RULES

In this section we use the asymptotic approximations to derive engineering rules for dimensioning the bandwidth.

The dimensioning problem is simply stated as

$$\text{Minimize } B \text{ such that the chosen performance criterion (6) or (7) is satisfied.} \quad (\text{P})$$

A. Mean Criterion and Effective Bandwidth

From the asymptotic limit (13)–(15), where $\rho > 1$, we obtain the approximate solution to (P) for the mean performance criterion. In this asymptotic limit, the distribution of Q_1/N is concentrated at a single point mass x^* . Thus, Q_1 (and $E[Q_1]$) is approximately $N(\rho-1)/\rho$, which equals $N - \mu/\lambda = N - B/\lambda f$, and (8) and (6) are satisfied if

$$B = N \cdot h \quad \text{where } h = \left(\frac{1}{b} + \frac{1}{\lambda f} \right)^{-1}. \quad (25)$$

Equation (25) is a simple approximate solution to (P). Equation (25) has the classic effective-bandwidth form where each source has an effective rate of h . As mentioned in the Introduction, the concept of effective bandwidths seems unlikely to be suitable for elastic data traffic as the inherent objects are the file sizes, and not the rates that packets traverse network. However, the parameter h is the harmonic mean of two rates that are indeed naturally associated with elastic data and will occur under respective limiting network conditions. Suppose the sources have a lot to send and for a given bandwidth B the link is significantly restraining on the potential throughput, such that all sources have a job at the PS node almost surely, and Q_1 (and thus \hat{Q}_1) equals N , and from (4), $B_s = B/N$. Given that the mean performance criterion (6) holds as an equality, then each source would be obtaining the average rate b . At the other extreme, suppose the target link, or the given service provider's network, is very lightly loaded and imposes no constraint on the traffic from the sources. This occurs when other factors limit the flow, such as other networks or the time to transfer a file is negligible compared to the user's think time. Then each source would be transmitting at an average rate of λf .

B. Tail Criterion and Increase in Dimensioned Bandwidth

Now consider the performance criterion based on quantiles: $\Pr(B_s < b) < \alpha$, (7). Note that one can numerically compute the exact solution to (P) by various methods. One could iterate

over B for the smallest value that satisfies (7), where for a given B , the tail probability of Q_1 is computed and the inequality (9) is tested. For this iteration, the asymptotic approximation (15) provides a numerically stable method to compute the pmf of Q_1 when N is large. Start at $n^* = \lfloor x^*N \rfloor$, set $\Pr(Q_1 = n^*) = 1$ and iteratively compute the log of the numerators in (1), incrementing up from and down from n^* , and then normalize the terms to sum to 1. If the computation is to be done many times as part of a larger design problem, one could increase the speed by using the standard error function and Lemma 1 above.

One can also express the exact solution, implicitly, in terms of the partition function $H(\cdot)$, (3). Since $\Pr(Q_1 > B/b)/\Pr(Q_1 > 0) = [1 - \Pr(Q_1 \leq \lfloor B/b \rfloor)]/[1 - \Pr(Q_1 = 0)]$, from (3) we have that the tail performance criterion (7) is satisfied if

$$\frac{1 - H(\lfloor B/b \rfloor)/H(N)}{1 - H(0)/H(N)} < \alpha. \quad (26)$$

Thus, the solution to (P) is the smallest B that satisfies (26). However, this is not just a simple iteration over the argument of $H(\cdot)$ as the terms in the sum $H(\cdot)$ depend on μ , which in turn depends on the unknown B .

A more explicit solution uses the asymptotic approximation (23), and setting $\text{Prob}(Q_1 > 0)$ equal to 1, note that (7) is satisfied if

$$\beta \sqrt{2N} \geq q_\alpha \quad (27)$$

where q_α is the $(1 - \alpha)$ -quantile of a Normal (mean = 0, variance = 1) random variable, i.e., $\Pr(Z > q_\alpha) = \alpha$ for Z distributed Normal (0,1). Substituting in the expression for β , (22), we have that (7) is satisfied if

$$2N\{\rho^{-1} - (1-x) + (1-x) \ln[\rho(1-x)]\} \geq q_\alpha^2. \quad (28)$$

Thus, one can numerically solve for the $\rho \in (1, (1-x)^{-1})$ where (28) holds as an equality, and hence determine B , since $\rho = N\lambda f/B$.

If we make the further approximation of expanding the log to second order, and use the normal approximation (24), i.e., Q_1 is Normal with mean $N - \mu/\lambda$ and variance μ/λ , we obtain that (7) is satisfied if

$$\frac{B/b - (N - B/\lambda f)}{\sqrt{B/\lambda f}} \geq q_\alpha. \quad (29)$$

The minimum B that satisfies (29) is

$$B = h \left[N + \gamma + \sqrt{2\gamma N + \gamma^2} \right] \quad (30)$$

where $\gamma = (1/2)q_\alpha^2 h/\lambda f = (1/2)q_\alpha^2 \cdot b/(b + \lambda f)$.

C. Summary of Dimensioning Rules

Summarizing, we have

Proposition 1: An approximate solution to (P) is

$$B = h \cdot N \quad (31)$$

given the mean performance criterion (6) and

$$B = h \cdot \left[N + \gamma + \sqrt{2\gamma N + \gamma^2} \right] \quad (32)$$

given the tail performance criterion (7) where

$$h = \left(\frac{1}{b} + \frac{1}{\lambda f} \right)^{-1}$$

and

$$\gamma = \frac{1}{2} q_\alpha^2 h / \lambda f = \frac{1}{2} q_\alpha^2 \cdot b / (b + \lambda f)$$

and where the input parameters are such that

$$\sqrt{N} \lambda f / b > q_\alpha. \quad (33)$$

Note that when B is chosen by the mean criterion, then $B < N\lambda\bar{f}$ and we have the self-consistency that $\rho > 1$, and thus the asymptotic approximation and the CQN model apply. However, when B is chosen by the tail-criterion guideline (32) without restrictions on the possible input parameters, then one can have the inconsistent outcome where $\rho < 1$ and the asymptotic approximation does not apply. However, if the possible input parameters are restricted by (33), then the desired self-consistency pertains. Note also that (33) holds in cases of interest, as q_α is $O(1)$ and $N \gg 1$; for a discussion of the magnitude of the ratio $\lambda f / b$, see Section VI-D.

Proposition 1 is our proposed dimensioning rule. It provides explicit simple closed-form expressions for the dimensioned bandwidth. It is, of course, less accurate than an exact calculation, or the asymptotic approximation (28), but given the uncertainty in the forecasted input parameters, greater accuracy does not seem needed. (We use the more accurate calculations in the next section to obtain a perspective on the inaccuracy of the normal approximation.)

An illuminating form for (31) and (32) is to normalize B by what would have been a full allocation of capacity, namely b times the number of sources.

$$\frac{B}{Nb} = \frac{\lambda f}{b + \lambda f} \quad (34)$$

given the mean performance criterion (6), and

$$\frac{B}{Nb} = \frac{\lambda f}{b + \lambda f} \left[1 + \gamma/N + \sqrt{2\gamma/N + (\gamma/N)^2} \right] \quad (35)$$

given the tail performance criterion (7).

V. SELECTING NUMBER OF SOURCES FOR GIVEN BANDWIDTH

In the previous section, the bandwidth B is determined given the number of sources N . One might be interested in the reverse viewpoint: determining N given B .

A topical example would be an edge router with a link to a core router, or enterprise LAN switch or router with a link to a router of an internet service provider. Suppose the bandwidth of the up link is given, say an OC-3 or OC-12. In terms of the CQN model, let “a source” represent the aggregate traffic on an access line (or an LAN), that is routed to the up link. The task is to determine the maximum number of sources (equivalently, access line cards or Ethernet ports) N^* that can be supported by the given bandwidth B of the up link.

Equation (31) gives N^* directly for the mean performance criterion (6). For the tail performance criterion (7) solving (32) or (29) for N yields

$$N^* = \left\lfloor B/h - q_\alpha \sqrt{B/\lambda f} \right\rfloor \quad (36)$$

where $\lfloor x \rfloor$ is the integer part of x . The stricter performance criterion reduces the number of sources that could be supported, but this reduction is relatively small for the pertinent case of a large link bandwidth where the standard deviation of Q_1 , $\sqrt{B/\lambda f}$, is relatively small compared with its mean.

Once a network is in service, the network operator may exercise no connection/flow admission control (CAC), as is the case in the present best-effort-service IP-based networks. In which case, the performance objectives (6) and (7) should be viewed as design objectives. The network operator could advertise that the network has been designed based on such objectives. However, since the realized traffic will differ from the forecast, it could be imprudent for the network operator to offer a per-flow service commitment to individual users. If such a service commitment is desired as part of a business strategy, the network operator would likely wish to exercise a CAC policy on the realized traffic. If each TCP session represents a source, then the CAC could be based on denial of TCP synchronization (SYN) packets. In a simple implementation, a counter could track the current number of established TCP sessions by incrementing and decrementing with each establishment (SYN) and tear-down (finish, FIN packet). When the counter equals a threshold, N^* , TCP SYN packets could be discarded, and possibly a RESET packet could be generated.

VI. NUMERICAL EXAMPLES

A. Comparison of CQN Model to Simulation of TCP/IP Sessions

In this section we compare the simple CQN model, (1), with simulations of file transfers control by TCP. Heyman *et al.* [3] also compare their more detailed “TCP-modified Engset” model with simulations of TCP transfers, and they have kindly shared their simulation results with us. Heyman *et al.* simulate sources that transfer files across access links and then across a shared T1 link of 1.5 Mb/s. A source alternates between an idle state and an active state, where during the latter the source transfers a file using TCP/IP. The TCP code used in the simulations closely parallels the prevalent Tahoe with fast recovery and Reno implementations. One of the outputs from the simulation is the number of sources A that are in the active state at an arbitrary time. In cases where the TCP control is working well, then the number of active sources equals the number of sources that have packets queued at the T1 link, and thus would equal the number of jobs at the processor sharing node in the CQN model. Thus in this case, A should equal Q_1 . However, in cases where the TCP control is not working as well, and packet losses are causing the sources to enter the slow-start phase, or to time-out, then there would be intervals where the number of the active sources (the number of sources presently attempting to transfer a file) would be greater than the number of sources that have packets queued at the bottleneck link; in which case, the random variable

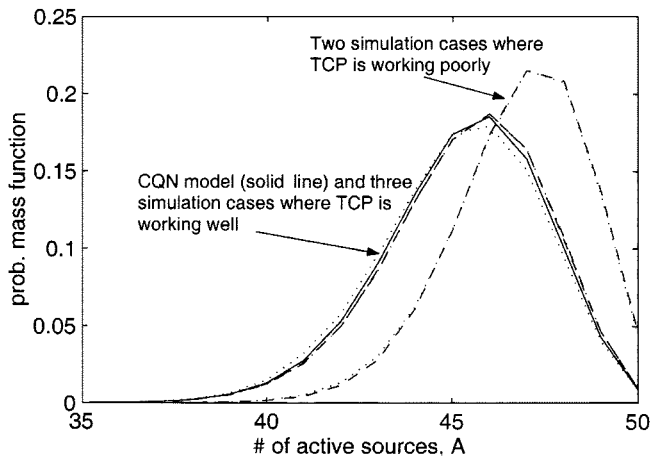


Fig. 2. Comparison of CQN model with simulation.

A would be greater than Q_1 , in some sense such as the mean. Of course, by design the CQN does not distinguish between these two cases, and assumes that the former pertains. Thus, it is of interest to compare the CQN model with simulated scenarios where the TCP is working well and where it is not.

For all of the simulations in [3], the mean idle time λ^{-1} is assumed to be 5 s, and the mean file size f (including overhead) is 200 K bytes. Thus, in the CQN model, the mean service time at the PS node μ^{-1} is 1.0667 s (f divided by the link speed of 1.5 Mb/s). The number of sources N is either 25 or 50. This then specifies the distribution for Q_1 in the CQN model (1). Further pertinent aspects of the simulation include the assumption that the buffer feeding the T1 links is served FIFO and the choice of buffer management strategy; when the buffer is full and an additional packet arrives, discard: 1) the new packet; or 2) the oldest packet, i.e., the packet at the front of the queue. Heyman *et al.* also consider different distributions for the idle times and the file sizes, including the Pareto distribution. The simulations show that these distributions influence the distribution of A only via their means, as predicted by their Engset models, as well as the CQN herein. Fig. 2 shows the pmf for five of the scenarios considered in [3], as well as the prediction for Q_1 in the CQN model. The five scenarios have various combinations of deterministic, exponential, and Pareto distributions for the file sizes and idle times, as well as discard policies of drop-from-front and drop-from-tail, and they all have 50 sources. Note that the prediction from the CQN model is close to three of the scenarios. These three scenarios have a TCP degradation factor of between 0.95–0.99, where this factor in the Heyman *et al.* model captures “the performance impairment due to TCP not accurately tracking the bottleneck link rate,” and whose ideal value is 1. Thus, in these three scenarios, TCP is working excellently, and the match to the CQN model is surprisingly good. In contrast, the two remaining scenarios consider TCP/IP running over ATM and the TCP degradation factor is 0.80. As expected, the number of active sources increases (the distribution shifts to the right) and the CQN model is no longer a good match. Note however that from the viewpoint of dimensioning, since there are more active sources than there are sources with bytes queued at the bottleneck link, those sources that do have bytes queued will still be receiving (better than) the bandwidth objective for which

TABLE I
COMPARISON OF EXACT AND APPROXIMATE CALCULATIONS OF (1)

Input parameters			% Error in approx. for $\Pr(Q_1 > n)$, as compared with exact computation (1)			
# of connections, N	$\rho \equiv N\lambda/\mu$	Quantile n	“ Q_0 -Poisson” approx.	Asymp. approx. (21)	First-term of asymp. approx. (21)	Normal approx.
25	2.5	21	0.17%	-14.0%	48%	180%
50	$3\frac{1}{3}$	42	0.00%	-6.0%	31%	96%
100	$3\frac{1}{3}$	81	0.00%	-2.6%	20%	72%
500	10	465	0.00%	-1.5%	15%	57%
1,000	10	922	0.00%	-0.7%	10%	39%

the link was dimensioned. Of course, some users are nevertheless obtaining a degraded goodput, but the cause is primarily the interaction of TCP with ATM and buffer size and buffer management policy. This degradation in throughput ought to be corrected by improvements in TCP and/or in ATM layer controls.

B. Accuracy of Asymptotic Approximations

Table I illustrates the accuracy of the “ Q_0 -Poisson” approximation, the asymptotic approximation (21), the first term of (21), and the normal approximation. For various numbers of sources N , and for ρ chosen so that the PS node is relatively heavily loaded, the quantile n was then chosen so that the exact tail probability is around 0.01. Note that although ρ is greater than 1, the CQN is stable since N is finite; see Section III for further discussion. Table I gives the percent error in the approximate calculation of $\Pr(Q_1 > n)$, as compared with the exact calculation. The “ Q_0 -Poisson” approximation uses $\exp(\mu/\lambda)$ for the partition function $H(N)$, (3), and is clearly the most accurate approximation of the four listed. The asymptotic approximation (21) also uses $\exp(\mu/\lambda)$ for $H(N)$ and in addition approximates the summation in (1). Table I shows that the asymptotic approximation (21) is qualitatively closer to the exact calculation than is the first-term alone of the asymptotic approximation (21). The normal approximation gives a rather poor estimate of the tail probability for these parameter values, although, as illustrated in Fig. 3, the overall match to the pmf appears reasonable, even for the case of $N = 25$. In any case, as shown and discussed in the subsequent sections, the bandwidth dimensioned via the normal approximation is rather close to the correct value.

Fig. 4(a) and (b) illustrate the relative accuracy of the approximations for two of the cases in Table I. For a relatively few sources, $N = 25$, Fig. 4(a) visually shows that the correction term (the second term) in (21) significantly adds to the accuracy of the approximation. For a larger number of sources, $N = 100$, Fig. 4(b) shows that the asymptotic approximation is accurate to tail probabilities as small as 10^{-6} .

C. Impact of Choice of Performance Criterion

As compared with the mean criterion, the increased tightness of the tail criterion impacts the dimensioning rule (32) via the parameter γ , which in turn depends on q_α , the $(1-\alpha)$ -quantile of a Normal(0, 1) random variable. Note that in the case of the normal approximation, the mean performance criterion

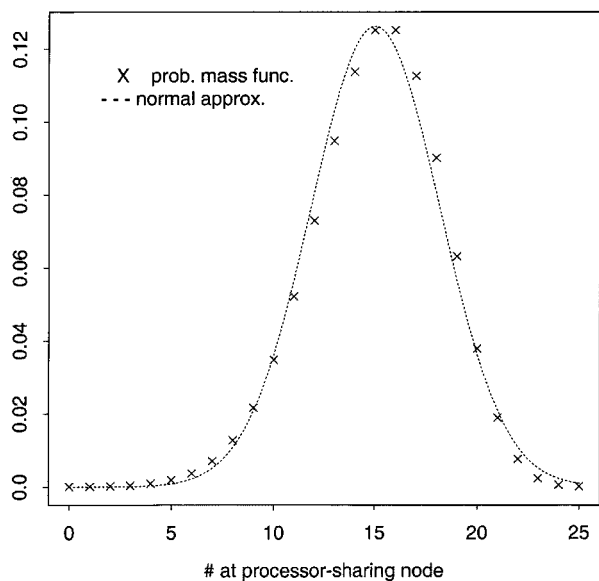


Fig. 3. Probability mass function for # at PS node, and associated normal approximation. $N = 25, \mu/\lambda = 10$.

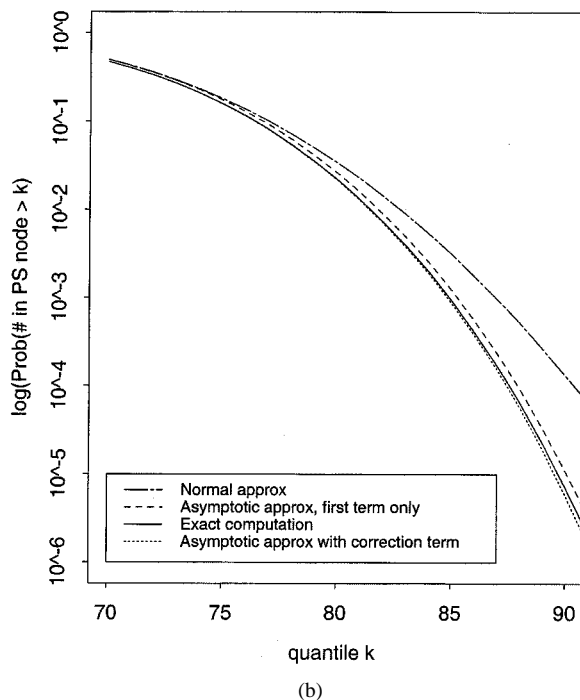
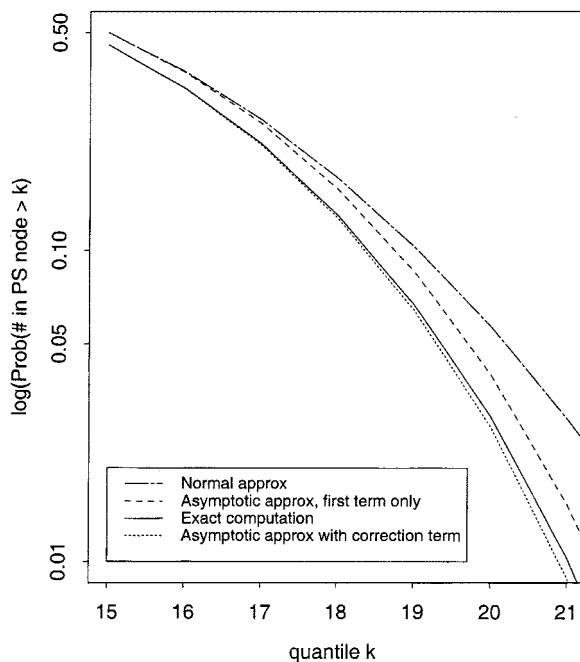


Fig. 4. (a) Comparison of asymptotic approximations with exact calculation, for $N = 25, \rho = 2.5$. (b) Comparison of asymptotic approximations with exact calculation, for $N = 100, \rho = 3.33$.

TABLE II
REQUIRED BANDWIDTH, FOR GIVEN NUMBER OF SOURCES N , WHERE $\lambda f = b = 100$ Kb/s. THE PERCENTS NOT-IN-PARENTHESES ARE FROM THE EXACT CALCULATION OF BANDWIDTH, WHILE THE PERCENTS IN PARENTHESES ARE FROM THE NORMAL APPROXIMATION, PROPOSITION 1, (31) AND (32)

Number of sources, N	Required bandwidth [Mbps], given mean performance criterion, (6)	Percent increase in required bandwidth, relative to that for the mean criterion (7) with α equal to:		
		0.10	0.05	0.01
50	2.5	12.7% (13.7%)	16.3% (17.9%)	24.0% (26.1%)
100	5.0	9.8% (9.5%)	12.0% (12.3%)	17.1% (17.9%)
200	10.0	6.6% (6.6%)	8.3% (8.6%)	12.0% (12.3%)
500	25.0	4.0% (4.1%)	5.2% (5.3%)	7.6% (7.6%)
1,000	50.0	2.9% (2.9%)	3.7% (3.7%)	5.2% (5.3%)
5,000	250.0	1.3% (1.3%)	1.6% (1.7%)	2.3% (2.4%)

is given by the 50% quantile, i.e., when q_α is zero. Table II illustrates the increase in the required bandwidth as α decreases. The key observation is that although tighter performance criteria (smaller values of α) increase the required bandwidth B the percentage increase is relatively minor for the relevant cases of a few hundred sources. This is evident from (32) where $N \gg \gamma$ and the square-root term is dominated by the linear term of N , for large N . Heuristically, one would expect this, since the pmf for Q_1 is fairly concentrated around the mean; see Fig. 2. Moreover, in the asymptotic limit where $N \rightarrow \infty$ and $\rho \equiv N\lambda/\mu > 1$ (many sources and fast service at the PS node), the distribution of Q_1/N becomes a single point mass. Thus, the model suggests that a network operator can design for a rather strong-sounding objective—with 99% probability the per-flow throughput is greater than a threshold—without requiring the deployment of much more bandwidth than for an objective on

the mean. If the network operator also implements a connection admission control (see Section V), then the design objective could be strengthened to be a service objective.

Also note that the percentage increase based on the normal approximation is close to that from the exact calculation. This is due to the fact that the normal approximation describes deviations from the mean of the order of \sqrt{N} .

If a network operator would like to use the tail performance criterion, but would prefer a simpler dimensioning rule than (32), then the dimensioning rule for the mean criterion could

TABLE III
RATIO OF FULL ALLOCATION TO ENGINEERED BANDWIDTH, Nb/B , FOR
 $N = 1000$, AND $\alpha = 0.01$

Input parameter: $b/\lambda f$	Statistical Gain, Nb/B exact computation	simple normal approx.
0.1	1.076	1.076
0.2	1.166	1.165
0.5	1.439	1.438
1.0	1.901	1.899
2.0	2.826	2.825
5.0	5.617	5.610
10.0	10.229	10.255

again be used with a larger value of the effective bandwidth parameter h , where this increase is chosen to bound the percentage increases for parameter values of interest, such as in Table II.

D. Comparison to Full Allocation of Bandwidth

It is instructive to compare the engineered bandwidth B with what would have been required if a full allocation of the per-flow objective b were provided for each of the N sources. Setting $x = b/\lambda f$, the equation (35) becomes

$$B/Nb = \frac{1}{1+x} \left[1 + \gamma/N + \sqrt{2\gamma/N + (\gamma/N)^2} \right] \quad (37)$$

where $\gamma = (1/2)q_{\alpha}^2 \cdot x/(1+x)$. As γ/N is typically small, B/Nb is roughly a simple hyperbola in x . For an analogous viewpoint, one can define the “statistical gain” to be the ratio of the full allocation of bandwidth to the engineered bandwidth, Nb/B . For the mean performance criterion, Nb/B is approximately $1 + b/\lambda f$. For a numerical example, Table III shows the statistical gain, given the tail performance criterion with $\alpha = 0.01$ and $N = 1000$, for various values of λf and b . For the given parameter values, ρ is greater than one, and the normal approximation pertains. For comparison, the engineered bandwidth is computed both iteratively using (1) and via the normal approximation (32). As shown in Table III, the normal approximation is quite close to the exact numerical computation. Also, the statistical gain is minor when the ratio $b/\lambda f$ is small, but is significant when $b/\lambda f \geq 1$.

Recall that λf is the throughput a source would obtain assuming the target link is not constraining the flow. Note that λf is determined by the constraints of other network components and includes the idle times at the source, whereas the per-flow bandwidth objective b applies only during active periods of file transfers. Thus, a service provider could reasonably choose a bandwidth objective that is equal to or greater than λf . When b is larger than λf , the service provider can realize significant savings in the engineered bandwidth B as compared with the full allocation.

VII. CONCLUSION

We have derived simple and robust engineering rules for dimensioning bandwidth for elastic data traffic for a single bottleneck link. The derivation is based on normal approximations

for a CQN model in heavy traffic. For a mean performance criterion, we obtain the effective bandwidth of an elastic data source. We believe that simple ballpark dimensioning rules are appropriate because of the uncertainty in the forecasts of the traffic demands. The robustness of the dimensioning rules follows from the insensitivity property of the CQN model, whereby the distribution of the underlying random variables is pertinent only via the mean, and of particular interest, the mean of the file-sizes, and not their heavy-tail characteristics.

We compared our CQN with simulations of file transfers regulated by TCP. Despite the simplicity of our CQN model, it accurately predicted the distribution for number of active sources at the bottleneck link, given the condition that the feedback control was performing well.

The dimensioning rules satisfy a performance measure based on the mean or the tail-probability of the per-source bandwidth. In the case of the mean performance measure, the dimensioning rule has the linear effective-bandwidth form. For the tail performance measure, the dimensioning rule is still in closed-form, though no longer linear in the number of sources. If the network designer wishes to use a linear rule for the sake of simplicity, then the dimensioning rule for the mean criterion could again be used with an increased value of the effective bandwidth parameter, where the increase is estimated for parameter values of most interest. The dimensioning rules are easily inverted to obtain the number of sources that can be supported on a link of given bandwidth.

We illustrated the increase in bandwidth needed to satisfy the tail performance objective as compared with the mean objective. The key observation is that although tighter performance criteria increases the required bandwidth, the percentage increase is relatively minor, particularly for the relevant case of at least a few hundred sources. This occurs since the pmf of Q_1 is rather concentrated around the mean. Thus, the model suggests that a network operator can design for a rather strong-sounding objective—with 99% probability the per-source throughput is greater than a threshold—without requiring the deployment of much more bandwidth than for an objective on the mean. We also showed regimes where statistical gain can and cannot be realized.

We provide a new derivation of the normal approximation in CQNs using more accurate uniform asymptotic approximations and give an explicit estimate of the error in the normal approximation. For the region of applicability, the uniform asymptotic expansion is accurate to within a few percentage points, and the “ Q_0 -Poisson” approximation is accurate to within a tenth of a percentage point. Although the normal approximation is less accurate than the other approximations, the bandwidth dimensioned based on this approximation is rather close to the correct value in the region of interest of some hundreds of sources, as the pmf for the number of active sources, Q_1 , is then relatively concentrated around the mean.

The present work has focused on a single bottleneck link and a single class of traffic. The generalization to multiple classes is examined in [30] and [31]. In future work, we intend to extend the results to derive dimensioning methods for general network topologies and multiple classes of elastic traffic. The methods will depend on additional asymptotic results.

APPENDIX
PROOF OF LEMMAS

Proof of Lemma 1: The idea of the proof is to determine an asymptotic approximation for $G(N)$, (3), and then use the relation

$$H(N) = (N/\rho)^N G(N) \quad (\text{A.1})$$

where $H(N)$ is given in (3) and $\rho = N\lambda/\mu$. For $\rho > 1$, the asymptotic approximation for $G(N)$ is a particular case of (16) in [27]. To apply (16) of [27], note that the z -transform of $G(N)$ is $e^z/(1 - (\rho z/N))$, and from the inverse Cauchy formula, $G(N)$ can be expressed as

$$\begin{aligned} G(N) &= \frac{1}{2\pi i} \oint_{C_1} \frac{e^z}{(1 - z\rho/N)z^{N+1}} dz \\ &= \frac{1}{2\pi i} \oint_{C_1} N^{-N} \frac{e^{-N(-t+\ln t)}}{t(1-\rho t)} dt. \end{aligned}$$

In the notation of [27], $p(t) = -t + \ln t$, $q(t) = t^{-1}(1 - \rho t)^{-1}$, and $c(N) = N^{-N}$. From (16) in [27], for $\rho > 1$ and $N \gg 1$

$$G(N) = (\rho/N)^N e^{N/\rho} [1 + o(e^{aN})] \quad (\text{A.2})$$

where $a = 1 - \rho^{-1} - \ln \rho$, which is negative for $\rho > 1$ and equals $F(x^*)$ in (16). Lemma 1 follows from (A.2) and (A.1).

Proof of Lemma 2: From the pmf for Q_1 , (1), and for $\rho = N\lambda/\mu$

$$\begin{aligned} \Pr(Q_1 > k) &= \Pr(Q_1 \geq k + 1) \\ &= \frac{1}{G(N)} \cdot \sum_{n=k+1}^N \frac{(\lambda/\mu)^n}{(N-n)!} \\ &= \frac{(\rho/N)^N}{G(N)} \cdot \sum_{j=0}^{N-k-1} \frac{(N/\rho)^j}{j!} \\ &= \frac{(\rho/N)^N H(N-k-1)}{G(N)} \end{aligned}$$

where $H(\cdot)$ is given in (3). From the conditions of Lemma 2 that $\rho > 1$ and $N \gg 1$, $G(N)$ can be replaced with (A.2) yielding

$$\Pr(Q_1 \geq k + 1) = \frac{H(N-k-1)}{e^{N/\rho} [1 + o(e^{-cN})]}. \quad (\text{A.3})$$

The remainder of the proof is to obtain the uniform asymptotic approximation for $H(N-k-1)$. The z -transform of the partition function $H(\cdot)$ is

$$h(z) = \sum_{m=0}^{\infty} H(m)z^m = \frac{e^{z/\rho}}{1-z}, \quad |z| < 1.$$

Following [27], and using the inverse Cauchy formula, we have

$$\begin{aligned} H(N-k-1) &= \frac{1}{2\pi i} \oint_{C_1} \frac{e^{t/\rho}}{(1-t)t^{N-k}} dt \\ &= \frac{1}{2\pi i} \oint_{C_1} e^{-Np(t)} q(t) dt \end{aligned} \quad (\text{A.4})$$

where

$$p(t) = -t/\rho + (1-x) \ln t \quad \text{where } x = k/N \quad (\text{A.5})$$

and

$$q(t) = (1-t)^{-1}. \quad (\text{A.6})$$

The saddle point of $p(t)$, denoted t_o , is $\rho(1-x)$, and the nonzero pole of the integrand in (A.4) is $\gamma_1 = 1$. From the condition of Lemma 2 that $0 < (1 - \rho(1-x)) < c/\sqrt{N}$ for some constant c , we know that the saddle point t_o is less than the pole 1, and that the difference $1 - t_o$ becomes small for large N . Thus we need the uniform asymptotic approximation given by (17) in the preliminaries of [27]; from which one obtains

$$\begin{aligned} H(N-k-1) &= e^{N/\rho} \left\{ \frac{1}{2} \operatorname{erfc}(\beta(x)\sqrt{N}) - e^{-N\beta^2(x)} \right. \\ &\quad \left. \cdot \left[\frac{1}{3\sqrt{2\pi N(1-x)}} + O(N^{-3/2}) \right] \right\} \end{aligned} \quad (\text{A.7})$$

where $\operatorname{erfc}(x) = (2/\sqrt{\pi}) \int_x^{\infty} e^{-y^2} dy$ and

$$\beta^2(x) = \rho^{-1} \{1 - \rho(1-x) + \rho(1-x) \ln[\rho(1-x)]\}. \quad (\text{A.8})$$

From (14), (15) note that $F(x) - F(x^*)$ equals the right-hand side of (A.8). Substituting (A.7) into (A.3) yields (21) of Lemma 2.

ACKNOWLEDGMENT

The authors would like to thank D. Heyman for the simulation results from his joint studies with T. Lakshman and A. Neidhardt. They also would like to thank W. Whitt for his thoughtful comments on a prior draft of this paper, and to the anonymous reviewers for their constructive suggestions.

REFERENCES

- [1] S. Lavenberg, Ed., *Computer Performance Modeling Handbook*. New York, NY: Academic, 1983.
- [2] I. Mitrani and E. Gelenbe, *Analysis and Synthesis of Computer Systems*. New York, NY: Academic, 1980.
- [3] D. P. Heyman, T. V. Lakshman, and A. L. Neidhardt, "A new method for analyzing feedback-based protocols with applications to engineering web traffic over the internet," in *Proc. ACM SIGMETRICS'97, Performance Evaluation Rev.*, vol. 25, 1997, pp. 24–38.
- [4] V. Paxson and S. Floyd, "Wide-area traffic: The failure of Poisson modeling," *IEEE/ACM Trans. Networking*, vol. 3, pp. 226–244, 1995.
- [5] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson, "Self-similarity through high variability: statistical analysis of ethernet LAN traffic at the source level," *IEEE/ACM Trans. Networking*, vol. 5, pp. 71–86, 1997.
- [6] A. Varma and D. Stiliadis, "Hardware implementation of fair queueing algorithms for ATM networks," *IEEE Commun. Mag.*, vol. 35, pp. 54–68, Dec. 1997.
- [7] A. Berger, L. Bregman, and Y. Kogan, "Bottleneck analysis in multiclass closed queueing networks and its application," *Queueing Systems*, vol. 31, pp. 217–237, 1999.
- [8] D. Bertsekas and R. Gallager, *Data Networks*, 2nd ed. Englewood Cliffs, NJ: Prentice Hall, 1992.
- [9] J. S. Kaufman, "Blocking in a shared resource environment," *IEEE Trans. Commun.*, vol. 29, pp. 1474–1481, 1981.
- [10] J. W. Roberts, "A service system with heterogeneous user requirements," in *Performance of Data Communications Systems and Their Applications*, G. Pujolle, Ed. Amsterdam, The Netherlands: North Holland, 1981, pp. 423–431.

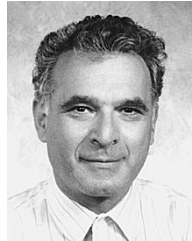
- [11] D. Mitra and J. Morrison, "Erlang capacity and uniform approximations for shared unbuffered resources," *IEEE/ACM Trans. Networking*, vol. 2, pp. 581–587, 1994.
- [12] C.-S. Chang and J. A. Thomas, "Effective bandwidth in high-speed digital networks," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 1091–1100, Aug. 1995.
- [13] G. de Veciana, G. Kesidis, and J. Walrand, "Resource management in wide-area ATM networks using effective bandwidths," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 1081–1090, Aug. 1995.
- [14] F. P. Kelly, "Notes on effective bandwidths," in *Stochastic Networks*. Oxford, U.K.: Clarendon, 1996, pp. 141–168.
- [15] W. R. Stevens, *TCP/IP Illustrated, Volume 1, The Protocols*. Reading, MA: Addison-Wesley, 1994.
- [16] S. Floyd, "TCP and explicit congestion notification," *ACM Comput. Commun. Rev.*, vol. 24, pp. 10–23, Oct. 1994.
- [17] K. Ramakrishnan and S. Floyd, "A proposal to add explicit congestion notification (ECN) to IP," Internet Engineering Task Force, Request for Comments 2481, Jan. 1999.
- [18] K. W. Fendick, "Evolution of controls for the Available Bit Rate service," *IEEE Commun. Mag.*, vol. 34, pp. 35–39, Nov. 1996.
- [19] A. Romanov and S. Floyd, "Dynamics of TCP traffic over ATM networks," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 641–633, 1995.
- [20] C. Fang, H. Chen, and J. Hutchins, "A simulation study of TCP performance in ATM networks," in *Proc. IEEE Globecom'94*, Nov. 1994, pp. 1217–1223.
- [21] H. Balakrishnan, H. Rahul, and S. Seshan, "An integrated congestion management architecture for internet hosts," in *ACM SIGCOMM'99*, Cambridge, MA, 1999, pp. 175–187.
- [22] M. Schwartz, *Telecommunication Networks: Protocol, Modeling, and Analysis*. Reading, MA: Addison-Wesley, 1986.
- [23] A. E. Ferdinand, "An analysis of the machine interference model," *IBM Syst. J.*, vol. 10, pp. 129–142, 1971.
- [24] B. Pittel, "Closed exponential networks of queues with saturation: The Jackson-type stationary distribution and its asymptotic analysis," *Math. Oper. Res.*, vol. 4, pp. 357–378, 1979.
- [25] M. Friedlin and A. Wentzell, *Random Perturbations of Dynamical Systems*. New York, NY: Springer-Verlag, 1984.
- [26] A. Birman and Y. Kogan, "Asymptotic evaluation of closed queueing networks with many stations," *Commun. Statistics, Stochastic Models*, vol. 8, pp. 543–564, 1992.
- [27] —, "Error bounds for asymptotic approximations of the partition function," *Queueing Syst.*, vol. 23, pp. 217–234, 1996.
- [28] Y. Kogan and M. Shenfield, "Asymptotic solutions of generalized multiclass Engset model," in *Proc. 14th Int. Teletraffic Cong.—The Fundamental Role Teletraffic in the Evolution of Telecommunications Networks*, J. Labetoule and J. W. Roberts, Eds, 1994, pp. 1239–1249.
- [29] F. P. Kelly, "Loss networks," *Ann. Appl. Probability*, vol. 1, pp. 319–378, 1991.
- [30] A. W. Berger and Y. Kogan, "Multiclass elastic data traffic: Bandwidth engineering using asymptotic approximations," in *Proc. 16th Int. Teletraffic Cong.—Teletraffic Engineering in a Competitive World*, P. Key and D. Smith, Eds, 1999, pp. 77–86.
- [31] —, "Distribution of processor-sharing customers for a large closed system with multiple classes," *SIAM J. Appl. Math.*, vol. 60, pp. 1330–1339, 2000.



Arthur W. Berger (S'82–M'83–SM'97) received the Ph.D. degree in applied mathematics from Harvard University, Cambridge, MA, in 1983.

He joined AT&T Bell Laboratories and subsequently AT&T Labs in 1996 and Lucent Bell Labs in 1998. Currently, he is a Senior Research Scientist with Akamai Technologies, Cambridge. His research interests include: content delivery over the Internet, predicting Internet performance, design and performance of high-speed data networks and equipment, quality of service in the Internet, IP over

optical transport networks, traffic engineering, congestion controls, stochastic models.



Yaakov Kogan (SM'91) received the Candidate of Sciences (the Soviet equivalent of the Ph.D. degree) and Doctor of Sciences degrees from the U.S.S.R. Academy of Sciences in 1968 and 1987, respectively.

He is a Principal Technical Staff Member at AT&T Labs, Middletown, NJ, where he works in the Network Design and Performance Analysis Department. After receiving the Ph.D., he worked on performance analysis of computer and communication systems and developed nonparametric and asymptotic methods for solving stochastic models of large dimension. His recent activities include performance and reliability analysis of corporate IP backbone networks.

Dr. Kogan is a member of IFIP Working Group on Computer System Modeling.