ITC 16 / P. Key and D. Smith (Editors)



# Multi-Class Elastic Data Traffic: Bandwidth Engineering Via Asymptotic Approximations

Arthur Berger[a] and Yaakov Kogan[b]

[a]Lucent Technologies, Bell Labs,
1600 Mass. Ave. #707, Cambridge MA, 02138 U.S.A.

[b]AT&T Labs,
200 Laurel Ave. Rm D5-3C05, Middletown, NJ 07748, U.S.A.

Engineering rules for dimensioning bandwidth for multi-class elastic data traffic are derived for a single bottleneck link via asymptotic approximations for a closed-queueing-network model in heavy traffic. Elastic data applications adapt to available bandwidth via a feedback control such as the Transmission Control Protocol, and we categorized them into different classes based on the file sizes and times between transfers. A quasi-potential function is found that provides the logarithmic asymptotics of the non-product-form distribution for the total number of active connections. The minimum point of the quasi-potential provides the approximation for the normalized mean, and the variance of the normal approximation is inversely proportional to the second derivative of the quasi-potential at the minimum point. The problem of dimensioning bandwidth is solved with the normal approximation. Via numerical studies we show the dependence of statistical gain on per-class parameters and show that the most stressful case is when the per-class load is balanced. Thus, if a network designer wanted to avoid the complexity of estimating per-class parameters, yet still have a conservative design, they could do so by considering a single overall class.

## 1. 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 adapt to time-varying available bandwidth via a feedback control such as the Transmission Control Protocol (TCP) or the Available Bit Rate transfer capability in ATM. Typical elastic data applications are file transfers supporting e-mail or the world wide web. We consider multiple classes of applications, generically distinguished by different mean file sizes.

The bandwidth to be dimensioned may be the capacity of a transmission facility, or may be a portion thereof, such as for a segment of a corporate Virtual Private Network (VPN), or for a Label Switched Path in Multi-Protocol Label Switching, or a Virtual Path Connection in ATM. Herein we use the generic term "link" for the object to be dimensioned.

The contribution of this paper is new asymptotic approximations and associated dimensioning rules that provide insights and guidelines for engineering bandwidth for elastic-data applications. This work can viewed as a variation on the classic capacity assignment problem in computer networks, [1], where herein the "load" is no longer the data flows (in bits/second or packets/second), but rather is the external file sizes, as we consider the resulting data flow as dependent on the state of the network via the closed-loop control. In prior work, [2], we motivated a closed queueing network (CQN) model and determined dimensioning rules for the case of a single class of connections. In the present paper, we determine engineering rules for the case of multiple classes of connections, developing new asymptotic approximations for a non-product-form distribution of the total number of active connections.

### 1.1. Motivation of Closed Queueing Network model

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 or cell loss. Although current control implementations do not necessarily meet this assumption, ongoing research efforts will lead to improved controls, see for example Heyman-Lakshman-Neidhardt [3] and Floyd [4]. We view as complementary: (1) network design that assumes well-performing closed-loop controls, and (2) control implementations that make good use of the deployed bandwidth.

Given well-performing controls and for periods of high load, which is the case of interest for engineering bandwidth, we expect that the utilization on a bottleneck link could be close to 1 for periods of minutes, particularly for smaller capacity "links" as in VPNs. We do not view this as necessarily a situation of overload, but rather feedback controls appropriately maintaining packets in queue. This leads us to consider closed queueing network (CQN) models where utilization near one does not lead to instability. Alternatively, Roberts [5] chooses to engineer for time periods where the occupancy is appreciably less than 1, and uses an open model.

We also assume that the output ports of the network nodes use some type of fair queueing across the class of elastic-data flows/connections. This assumption reflects the trend in the industry to implement such service disciplines in order to provide fairness; for a recent review of implementation of various fair queueing algorithms see Varma & Stiliadis [6]. In the model, we assume the network node uses processor sharing.

An important practical feature of the CQN models that we study is their insensitivity: the distribution of the underlying random variables influences the performance only via the mean. Thus, given the assumptions of the model, our engineering rules pertain in the topical case when the distribution of file sizes is heavy tailed [7] (though with finite mean), and the superposition of file transfers is long-range dependent [8].

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.

### 1.2. Outline of Paper

Section 2 describes our CQN model and Section 3 presents the performance criteria. Section 4 contains the new asymptotic approximations and associated engineering rules, and Section 5 presents numerical examples of dimensioning bandwidth.

## 2. CLOSED QUEUEING NETWORK MODEL

For simplicity we use the term "connection" to apply to either an ATM connection, or more generally a virtual circuit in a connection orientated packet network, or, with some blurring of meaning, to an IP flow of packets, where a flow is defined at the discretion of the network designer and would include the source and destination IP address, or sub-net thereof, and possibly additional descriptors such as the protocol field in the IP header or the port numbers in the TCP header. More informally, one can think of a connection as representing a source or user.

The link is to be sized to support $N$ connections. Thus, for the dimensioning step, we take the viewpoint that there is a static number of connections present. Each connection alternates between two phases: idle and active, where during the active phase packets are transmitted from the source to the destination. The fixed number of connections and their alternation between active and idle phases makes plausible the use of a closed queueing network (CQN) model with two types of servers. The first type, referred to as a source, is an infinite server (IS) (equivalently the number of servers equals to the number of connections) that models connections while they are in the idle phase. The second type is a processor-sharing (PS) server that models the queueing and emission of packets on the link.

A non-standard aspect of the CQN model is the entity represented by a "job." For network dimensioning, we are interested in scenarios where the data network is heavily loaded. During such times, network resources will tend to be the limiting factor on the throughput obtained for the elastic-data connections. Moreover, the feedback controls of TCP and ABR will tend to seek out and fill up the available bandwidth. At heavily loaded links, a connection's feedback control, when properly designed and functioning, will attempt to keep at least one packet queued for transmission on the link (otherwise the control is needlessly limiting the throughput). We assume that this is the case. Thus, at an arbitrary point in time, the number of connections that are in the active phase equals the number of connections that have a packet in queue at the bottleneck node, which equals the number of connections that have a packet in queue under the hypothetical scenario that the stream of packets of an active phase arrived as a batch to the network node. A "job" in the CQN model represents this hypothetical batch arrival. Thus, a job represents all of the packets of an active phase of a connection. 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. Clearly, with this notion of job, the CQN can not model packet queue lengths or packet losses. However, it does model the number of connections that are in the process of transferring a file, given the assumption of well-performing controls. And this latter entity is just what we need to model the per-connection performance objective, described in Section 3.

In the present paper we are particularly concerned with the case of multiple connection types. We consider $K$ connection classes, with $N_k$ connections in class $k$, $k = 1, \ldots, K$. Let $\lambda_k^{-1}$ be the mean service time of a class-$k$ job at the IS node, and $\mu_k^{-1}$ be the mean service time for a class-$k$ job at the PS node, assuming no other jobs present. $\mu_k^{-1} = f_k/B$, where $f_k$ is the mean file size of a class-$k$ connection and $B$ is the engineered bandwidth. Thus, a class-$k$ connection is characterized by $\lambda_k$ and $f_k$. Moreover, for the distribution of jobs at the PS node only the product $u_k = \lambda_k f_k$ is pertinent. Heyman et al. [3] in their modified Engset models also find that the impact on performance by the distribution of file sizes and think times is only via the product $u_k$. $u_k$, in bits-per-second, represents the throughput averaged over both active and idle periods, for a class-$k$ source, given that the target link is imposing no restriction on the flow. Thus, the rate $u_k$ includes the effects of all of the factors other than the target link, such as the actual thinking time by the user, and access-line speed. Likewise, $\lambda_k$ represents the throughput in file-transfers per second, given that the target link is imposing no restrictions.

## 3. PERFORMANCE CRITERIA

In elastic data applications, the user, and hence the network designer, is concerned with the delay in transfering a file. Since file sizes vary greatly, a single delay objective, such as 100ms, 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 seconds/bit. More conveniently, we take the reciprocal, so that the performance objective is in terms of the bandwidth, in bits/second, that an arbitrary, active connection obtains.

Let $Q_k$ be the random variable for the steady-state number of class-$k$ jobs at the PS node. Let $Q \equiv \sum_{k=1}^{K} Q_k$ be the total number of jobs at the PS node, and let $\hat{Q}$ denote the conditional total number of jobs in the PS node given that the PS node is not empty. In steady state, the bandwidth per active connection, denoted $B_c$. is defined as $B_c = B/\hat{Q}$.

$$\text{By its definition,} \quad Pr(\hat{Q} = n) = \frac{Pr(Q = n)}{Pr(Q > 0)}, \quad n = 1, \ldots, N. \tag{1}$$

We consider performance criteria on the mean and on the tail probability of $B_c$:

$$\mathrm{E}[B_c] \geq b \qquad \text{or} \tag{2}$$

$$Pr(B_c < b) < \alpha, \tag{3}$$

for given $b$ and $\alpha$, where typical values for $\alpha$ are in the range 0.01 to 0.1.

Consider first the performance criterion on the mean (2). Given $B_c = \frac{B}{\hat{Q}}$, and applying Jensen's inequality, then (2) is satisfied if

$$E[\hat{Q}] \leq B/b. \tag{4}$$

We use (4) in the heavy traffic region, where $Pr(Q > 0)$ is exponentially close to 1 for large $N$, i.e. the distribution of $\hat{Q}$ is close to $Q$ and $Q/N$ is approximately an atom, in which case (4) holds if and only if (2) does.

As for the tail performance criterion, since $Pr(B_c < b) = Pr(\hat{Q} > B/b)$, (3) is satisfied if and only if $Pr(\hat{Q} > B/b) < \alpha$. Using (1), the performance criteria (2) and (3) are satisfied, respectively, if the following conditions in terms of $Q$ pertain:

$$\mathrm{E}[Q]/Pr(Q > 0) \leq B/b, \tag{5}$$

$$Pr(Q > B/b)/Pr(Q > 0) \leq \alpha. \tag{6}$$

## 4. ASYMPTOTIC APPROXIMATIONS

It is known that the steady state probability distribution $Pr\{Q_1 = n_1, \ldots, Q_K = n_K\}$ has a product form:

$$Pr\{Q_1 = n_1, \ldots, Q_K = n_K\} = \frac{1}{G} \prod_{k=1}^{K} \frac{N_k!}{(N_k - n_k)!} n! \frac{r_k^{n_k}}{n_k!}, \tag{7}$$

where $n \equiv \sum_{k=1}^{K} n_k$, $r_k \equiv \lambda_k/\mu_k$ and $G$ is the normalization constant. The probability mass function for the total number of customers at the PS node is

$$P(n) \equiv Pr\{Q = n\} = \sum_{n_1 + \cdots + n_K = n} Pr\{Q_1 = n_1, \ldots, Q_K = n_K\}. \tag{8}$$

Distribution $P(n)$ does not have a product form. However the generating function $\mathcal{P}(z)$ for the sequence $\hat{P}(n) = P(n)/n!$ has the following simple expression:

$$\mathcal{P}(z) = \sum_{n=0}^{N} \hat{P}(n) z^n = G^{-1} \prod_{k=1}^{K} (1 + r_k z)^{N_k} \tag{9}$$

which is easily derived from (7) and definitions of $P(n)$ and $\mathcal{P}(z)$.

We are interested in asymptotic approximations for $P(n)$ to obtain engineering insights and closed-form expessions for dimensioning bandwidth. We study the asymptotics of $P(n)$ under the following two assumptions.

1. The total number of connections in the network $N = \sum_{k=1}^{K} N_k$ is large, i.e. $N \gg 1$ and moreover

   $$\rho_k = N r_k \quad \text{and} \quad \alpha_k = N_k/N, \tag{10}$$

   where $\rho_k$ and $\alpha_k$, $k = 1, \ldots, K$, remain constant as $N \to \infty$.

2. The PS station is saturated which is expressed by the following heavy usage condition [9]

   $$\sum_{k=1}^{K} \alpha_k \rho_k > 1. \tag{11}$$

Starting with (9) and using the Cauchy formula, we obtain for $P(n)$ an integral representation in complex space. Evaluating the integral by the saddle-point method and using

Stirling's formula for $n!$ and the asymptotic expansion for the normalization constant $G$ in [9], we obtain the following result.

**Proposition 1.** *Under conditions (10) and (11) the probability distribution of the total number of jobs at the PS station has the following asymptotic expansion*

$$Pr\{Q=n\} = \sqrt{\frac{\Delta}{2\pi N}} f(n/N) \exp\{-N(F(n/N) - F(x^*))\}(1 + O(1/N)), \tag{12}$$

where

$$F(x) = x - x\ln x - S(v_o(x)), \qquad f(x) = \frac{1}{v_o(x)}\sqrt{\frac{x}{S''(v_o(x))}}, \tag{13}$$

$$S(v) = \sum_{k=1}^{K} \alpha_k \ln(1+\rho_k v) - x\ln v, \tag{14}$$

*and where $v_o(x)$ is the single positive root of the equation*

$$\sum_{k=1}^{K} \frac{\alpha_k \rho_k}{1+\rho_k v} - \frac{x}{v} = 0. \tag{15}$$

*The function $F(x)$ has its minimum at $x^*$ which is the single positive root of*

$$1 - \sum_{k=1}^{K} \frac{\alpha_k \rho_k}{1+\rho_k x} = 0, \quad \textit{and} \quad \Delta = \sum_{k=1}^{K} \frac{\alpha_k \rho_k^2}{(1+\rho_k x^*)^2}. \tag{16}$$

**Corollary 1.** *The function $F(x)$ defines the logarithmic asymptotics of the probability distribution $P(n) = Pr\{Q=n\}$ in the following sense: $\lim_{N\to\infty} \frac{\ln P(n)}{N} = -(F(x) - F(x^*))$. Moreover, $F(x^*)$ defines the logarithmic asymptotics of the normalization constant $G = G(N)$: $\lim_{N\to\infty} \frac{\ln G(N)}{N} = -F(x^*)$.*

**Corollary 2.** *The normalized total number of processor sharing customers $Q/N$ converges to $x^*$ with probability 1 and*

$$\frac{Q - Nx^*}{\sqrt{N}} \textit{ is asymptotically normal with mean 0 and variance} \tag{17}$$

$$\sigma^2 = \frac{1}{F''(x^*)} = \Delta^{-1} - x^*. \tag{18}$$

Note that $\sigma^2$ can also be calculated from the multidimensional normal approximation for the $K$ classes at the PS station given in [10], [11, Chapter 4]. However, such a calculation produces a very cumbersome expression.

In the following remarks, we comment on Proposition 1 and its usage for bandwidth dimensioning.

**Remark 1.** Following [12] the function $F(x)$ is referred to as the quasi-potential for the distribution $P(n)$. For a birth and death process, the quasi-potential satisfies a non-linear differential equation (see [12]) which has an explicit solution only in some particular cases, e.g., for single-class product form closed queueing networks. Proposition 1 provides an

example of a non-Markov process, where the quasi-potential can be found explicitly or easily computed.

**Remark 2.** There are two cases $K = 1$ and $K = 2$, where the equations for $v_o(x)$, (15), and $x^*$, (16), are respectively linear and quadratic, which leads to explicit expressions for the quasi-potential $F(x)$ and function $f(x)$.

For arbitrary $K$, when $\rho_1 = \ldots = \rho_K = \rho$, herein called 'balanced load,'

$$\begin{aligned} F(x) &= (1-x)\ln(1-x) + x(1-\ln\rho), \\ x^* &= 1-\rho^{-1}, \qquad F(x^*) = 1-\rho^{-1}-\ln\rho. \end{aligned}$$

These expressions equal those obtained for the single-class case, [2].

**Remark 3.** Note that $x^*$ and $\Delta$, and thus $\sigma$ depend on $\rho_1, \ldots, \rho_K$, (16), (18), and thus depend on $B$, which we denote as $x^*(B)$ and $\sigma(B)$. From the normal approximation of Corollary 2, (17), we obtain the following simple, implicit equations for $B$ that respectively satisfy the mean and tail performance criteria, (2), (3), for an arbitrary number of classes:

$$B = Nbx^*(B), \tag{19}$$

$$B = Nbx^*(B) + q_\alpha b\sqrt{N}\sigma(B), \tag{20}$$

where $q_\alpha$ is the $1-\alpha$ quantile of the normal distribution with mean zero and variance one, and the $Pr(Q > 0)$ is taken to be one.

For the case $K = 2$, we derive from (19) the following compact, explicit expression for the engineered bandwidth, given the mean performance criterion,

$$B = N \cdot \frac{b^{-1} + (\alpha_1 u_1 + \alpha_2 u_2)/(u_1 u_2)}{\left(b^{-1}+u_1^{-1}\right)\left(b^{-1}+u_2^{-1}\right)}. \tag{21}$$

**Remark 4.** For the case of balanced load, $\rho_1 = \ldots = \rho_K = \rho$ (equivalently $u_1 = \ldots = u_K = u$), $\sigma^2 = \rho^{-1}$, and using the normal approximation (17) the engineered bandwidth has the closed-form expression of the single-class case, [2]:

$$B = hN \quad \text{given the mean performance criterion (2)} \tag{22}$$

$$B = h\left[N + \gamma + \sqrt{2\gamma N + \gamma^2}\right] \quad \text{given the tail performance criterion (3)} \tag{23}$$

where $h = (b^{-1}+u^{-1})^{-1}$, the harmonic mean of $b$ and $u$, and $\gamma = \frac{1}{2}q_\alpha^2 b/(b+u)$, and where $q_\alpha$ is the $1-\alpha$ quantile of the normal distribution with mean zero and variance one. For self-consistency with the condition $\rho > 1$, (11), the input parameters satisfy $\sqrt{N}u/b > q_\alpha$.

**Remark 5.** Starting with (12) and using the Euler-Maclauren summation formula and then evaluting the resulting integral by the Laplace method, one can refine the normal approximation in Corollary 2 and obtain two terms of the asymptotic expansion for the complemetary distribution function $Pr\{Q > m\}$, where $0 < m - Nx^* = O(\sqrt{N})$.

**Remark 6.** The results of this paper can be generalized. First, Propositions 1 can be generalized for a marginal distribution at bottleneck stations, see e.g. [13], in closed queueing networks with two or more PS stations. Second, similar results can be obtained in the context of loss networks [14] for the distribution of the number of busy circuits for a link with normal load. We are pursuing this work and hope to report on the results subsequently. A more difficult problem is to derive multidimensional asymptotic expansions that are pertinent to the case of several bottlenecks.

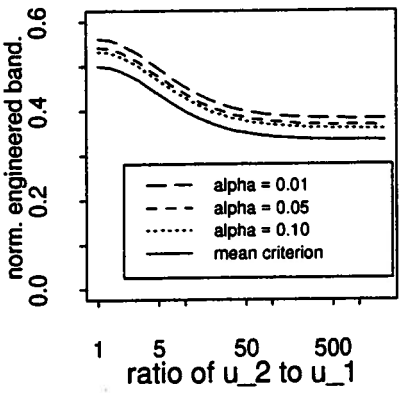


Figure 1. The engineered bandwidth, $B$, normalized by $Nb$, versus the ratio $u_2/u_1$, given $N_1 = N_2 = 100$, $b = 1$, and $N_1u_1 + N_2u_2$ held constant at 200.

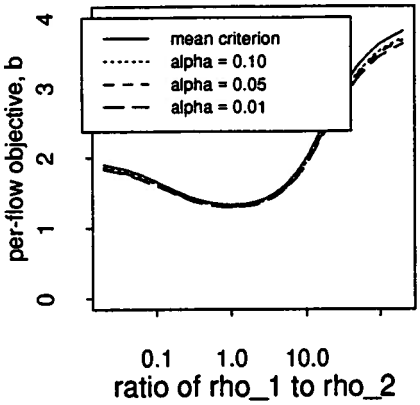


Figure 2. $b^*$ versus the ratio $\rho_1/\rho_2$, given $\rho = 4$, $N_1 = 1,000$, $N_2 = 2,000$, bandwidth $B = 3,000$.

## 5. DIMENSIONING BANDWIDTH

In this section we examine the dimensioned bandwidth for various input parameters. Suppose there are two classes and 100 sources per class, $N_1 = N_2 = 100$, and the per-connection bandwidth objective, $b$ equals 1. Consider various values of $u_1$ and $u_2$, while holding $N_1u_1 + N_2u_2$ constant and equal to 200. Lastly, consider the mean performance criterion and the tail-performance criterion for three values of $\alpha$, 0.01, 0.05, and 0.10. Figure 1 shows the normalized engineered bandwidth versus the ratio $u_2/u_1$, indexed by the four performance criteria. The normalized engineered capacity, is the engineered capacity, $B$, divided by $b(N_1+N_2)$, where the latter is the full-allocation of bandwidth given the per-bandwidth objective is reserved for each source. The reciprocal of the normalized engineeried bandwidth can be viewed as the statistical gain. A natural interpretation of $u_2/u_1$ is the ratio of the mean file sizes of the two classes, where the arrival rates, $\lambda_1$ and $\lambda_2$, are equal. The plots in Figure 1 used the normal approximation, Corollary 2, (21), (20). Complementing Figure 1, Table 1 compares the engineered bandwidth based on the normal approximation with that from the exact calculation, (7), (8), for sample values of $u_2/u_1$ and the mean criterion and the tail criterion with $\alpha = 0.01$. For the present example, Table 1 shows that the normal approximation is accurate to two - three significant figures. This also holds for the other two values of $\alpha$ in the tail criteria of Figure 1. Two to three signifcant figures is more than sufficient for the candidate engineering context where the estimates for the input parameters might have only one or two signifcant digits.

A first observation from Figure 1 is that the normalized engineered bandwidth is appreciably less than one, indicating statistical gain as compared with the full allocation. Second, the required bandwidth increases only modestly as the performance criterion becomes more strict, and the increase is fairly constant across the ratios of $u_2/u_1$. For example, the needed increase in bandwidth to satisfy the tail-performance criterion with $\alpha = 0.01$ as compared with the mean-performance criterion is only 12% to 14%. This suggests that the network designer can use a rather strong sounding objective – with 99%

Table 1
Normalized engineered bandwidth, $B/(Nb)$ for selected ratios of $u_2/u_1$, given $N_1 = N_2 = 100$, $b = 1$, and $N_1u_1 + N_2u_2$ held constant at 200.

| | Normalized Enginered Bandwidth | | | |
|---|---|---|---|---|
| | Mean Criterion | | Tail Criterion $\alpha = 0.01$ | |
| $u_2/u_1$ | Normal | Exact | Normal | Exact |
| 1 | 0.5000 | 0.5002 | 0.5618 | 0.5598 |
| 5 | 0.4375 | 0.4377 | 0.4944 | 0.4949 |
| 10 | 0.3995 | 0.3996 | 0.4538 | 0.4534 |
| 50 | 0.3500 | 0.3503 | 0.3996 | 0.3987 |
| 100 | 0.3419 | 0.3420 | 0.3909 | 0.3899 |
| 500 | 0.3351 | 0.3352 | 0.3830 | 0.3801 |

probability the per-connection throughput is greater than a threshold – without requiring the deployment of much more bandwith than for an objective on the mean. Note that if the network operator used a connection admission control policy, or a routing policy that limited the number of connections on a path, then these (otherwise internal) per-connection bandwidth objectives could become performance commitments as part of a service level agreement.

Lastly, for a given performance criterion, the engineered bandwidth is greatest when the ratio $u_2/u_1$ is one. This suggests that "balanced" load is the most stressful case. This would be a useful result as the network designer, in ignorance of what are the true per-class parameters, could then use the assumption of balanced load to obtain a conservative design However, $\rho$ is not constant in Figure 1 since $\rho = (N_1u_1 + N_2u_2)/B$, and the bandwidth $B$ is clearly varying, causing $\rho$ to vary between 2 and 3 for the mean performance criterion, with somewhat lower values for the tail criteria. To substantiate the suggestion that balanced load is most stressful, we need to hold $\rho$ constant.

To obtain a well-formed conjecture we need to hold $B$ fixed, and examine the maximum per-connection bandwidth objective that can be supported. For given $\rho$, $N$, and bandwidth $B$, let $b^*$ denote the largest $b$ that satisfies the mean performance criterion or a tail performance criterion with given $\alpha$.

**Conjecture.** *$b^*$ is minimized when the $u_i$ $i = 1, \ldots, K$ (equivalently $\rho_i$) are equal and the $N_i$ (equivalently $\alpha_i$) are arbitrary, subject to being non-negative and summing to $N$ or 1 respectively.*

Consider the case of two classes, $\rho = 4$, $N_1 = 1,000$, $N_2 = 2,000$, and bandwidth $B = 3,000$. Note that if one made a dedicated equal partition of bandwidth across the sources, each source would receive 1 unit of bandwidth. Thus, the extent that $b^*$ is greater than 1 indicates statistical gain. Using the normal approximation, Figure 2 shows a plot of $b^*$ versus the ratio $\rho_1/\rho_2$. As expected, $b^*$ is minimal when the ratio is one, and equals 1.33 for the mean criterion. If a network designer wanted to avoid the complexity of estimating per-class parameters yet still have a conservative design, they could do so by considering a single overall class and just use estimates for $N$ and a single $u$. However,

Figure 2 also illustrates that substantial gain could be obtained if indeed the classes were highly asymmetric and the designer took this fact into account.

For example, at the extremes when $\rho_1$, respectively $\rho_2$, is zero, then $b^*$ is 2 and 4 (for the mean criterion), which is subtantially higher than the 1.33 obtained in the balanced case. Thus, depending on circumstances, the network designer might choose the additional complexity of making per-class estimates in order to obtain a better service objective, for given deployed bandwidth.

Note also, as expected, the stricter performance criteria cause a decrease in $b^*$. However, this effect is small compared with the impact of changes in the ratio $\rho_1/\rho_2$. This suggests that if the network designer plans on further optimization, more benefit can be obtain from considering the per-class asymmetries than from adjusting the performance criteria.

## REFERENCES


1. D. Bertsekas and R. Gallager, *Data Networks* 2nd Edition, Prentice Hall, Englewood Cliffs NJ, 1992.
2. A. Berger and Y. Kogan, *Dimensioning bandwidth for elastic traffic in high-speed data networks*, submitted for publication in IEEE/ACM Transactions on Networks.
3. D. P. Heyman, T. V. Lakshman, & A. L. Neidhardt, *A new method for analyzing feedback-based protocols with applications to engineering web traffic over the internet*, Performance Evaluation Review, Vol. 25, no. 1, Proc. ACM-SIGMETRICS'97, 1997, pp. 24-38.
4. S. Floyd, *TCP and explicit congestion notification*, ACM Computer Communication Review, Vol. 24, October, 1994, pp. 10-23.
5. J. Roberts, *Realizing quality of service guarantees in multiservice networks*, Proc. of IFIP Conference PMCCN'97, Chapman and Hall, 1998.
6. A. Varma and D. Stiliadis, *Hardware implementation of fair queueing algorithms for ATM networks*, IEEE Communications Magazine, Vol. 35, Dec. 1997, pp. 54-68.
7. V. Paxson and S. Floyd, *Wide-area traffic: the failure of Poisson modeling*, IEEE/ACM Trans. on Networking, Vol. 3, 1995, pp. 226-244.
8. W. Willinger, M. S. Taqqu, R. Sherman, & D.V. Wilson, *Self-similarity through high-variability: statistical analysis of ethernet LAN traffic at the source level*, IEEE/ACM Trans. on Networking, Vol. 5, 1997, pp. 71-86.
9. J. McKenna, D. Mitra and K. G. Ramakrishnan, *A class of closed Markovian queueing networks: Integral representations, asymptotic expansions and generalizations*, Bell Syst. Tech. J., 60 (1981), pp. 599-641.
10. B. Pittel, *Closed exponential networks of queues with saturation: The Jackson-type stationary distribution and its asymptotic analysis*, Math.Oper.Res., 6 (1979), pp. 357-378.
11. G. P. Basharin, P. P. Bocharov and Y. A. Kogan, *Queueing Analysis in Computer Networks*, Nauka, Moskow, 1989 [in Russian].
12. M. Freidlin and A. Wentzell, *Random Perturbation of Dynamical Systems*, Springer, New York, 1984.
13. A. Berger, L. Bregman and Y. Kogan, *Bottleneck analysis in multiclass closed queueing networks and its application*, to appear in Queueing Systems.
14. F. P. Kelly, *Loss networks*, Ann. Appl. Prob. 1 (1991) 319-378.