

Traffic Shaping for a Loss System

George Kesidis
(**corresponding author**)
EE Dept and CS&E Dept
121 Electrical Engineering East
Pennsylvania State University
University Park, PA
16802
email: kesidis@enr.psu.edu
(814)865-0189, fax: (814)865-7065

Kaushik Chakraborty and Leandros Tassiulas
E&CE Dept, University of Maryland, College Park, MD, 20742
email: leandros@glue.umd.edu
(301)405-6620, fax: (301)314-9920

August 1, 2000

submitted to IEEE Communications Letters
EDICS no. CL1.7.3 (Performance Evaluation)

Abstract

We consider a loss system in which the connection arrival process is deterministically regulated (by leaky buckets for example) but is otherwise arbitrary. Bounds are found on connection blocking probabilities. These bounds are only in terms of the parameters of the traffic regulator, the common connection holding time distribution (which is not regulated in any way) and the server system itself.

1 Introduction

We consider the problem of reducing the number of blocked connections in an Erlang-type loss network. One extreme is not to inhibit the arriving traffic (connection set-up requests) at all; this leads to blocking probabilities given by Erlang's formula for the case of a network of $M/GI/C/C$ queues [8, 12]. The other extreme is to put a queue at each source (network access point) and have the connections wait until a circuit is available. This requires complex network monitoring of available circuits and constant communication with the network access points; also, connection set-up delay may be excessive. The approach proposed in this paper is somewhat of a compromise.

In Figure 1, a "single node" situation is depicted. Let $N(s, t]$ be the number of connection set-up requests (simply called "connections") departing the shaper over the interval of continuous time

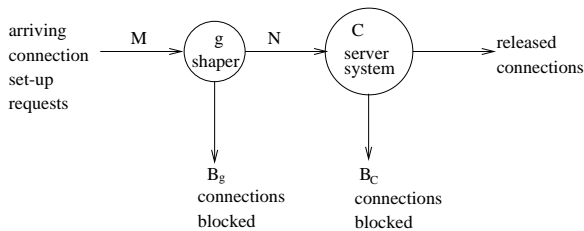


Figure 1: A traffic shaper followed by a C server loss system

$(s, t]$. The shaper operates so that, irrespective of its connection arrival process, its connection departure process satisfies

$$N(s, t] \leq g(t - s) \text{ a.s.}$$

for all $t \geq s$ where g is a strictly increasing function on \mathbb{R}^+ such that $g(0) = 1$ and $g(0-) = 0$.

The scenario of Figure 1 is also relevant to ATM virtual path (VP) switching and to label-switched routes (LSRs) in a MPLS Internet domain. In both cases, connections may be set up by “servers” in the context of limited available capacity. For a specific example, consider a LSR which has been set-up through an MPLS domain using an RSVP-type protocol so that the LSR is assigned a finite amount of associated bandwidth B . The LSR is used by a “service” provider for connections each requiring a bandwidth $b \ll B$. The number of “circuits” or “servers” of this LSR is simply $C = B/b$.

Note that the shaper does not take into account the potential holding times of the connections. When a connection is blocked by the (finite capacity) shaper, the corresponding user would quickly receive a “fast busy signal”. Typically, if resources are not available, the network responds after a small fixed amount of time with a fast busy signal. The connection capacity of the shaper can simply be sized according to this time limit.

A less desirable situation occurs if the user’s connection set-up request passes through the shaper (possibly at the expense of some delay) only to be blocked by the “ C -server system”. So, if a connection is forwarded to the C -server system by the shaper, it is desirable to have a low probability that the connection is subsequently blocked. Also, the router(s) of the C -server system can operate at only a finite rate (which is related to the “signaling bandwidth” of the system). By smoothing the connection arrival process to the C -server system, the shaper can also address this limit.

In this paper, we derive analytical bounds for the call blocking probabilities of the C -server system. A simulation study is given in [5] which explores the benefits of using a traffic shaper, as in Figure 1, to reduce the overall connection blocking probabilities subject to the network limitations described above.

2 Focus on the C -server system

Consider a system of C servers each handling connections across a communication network. This system has no associated queue. The arrival time of the i^{th} connection to this system is T_i and this connection has duration S_i , i.e., if this connection is not blocked and obtains a server, it will occupy the server for S_i seconds. Let N be the counting process of arrivals to the system, i.e., for any interval of time $(s, t]$,

$$N(s, t] = \sum_{i=-\infty}^{\infty} \mathbf{1}\{s < T_i \leq t\} \tag{1}$$

where $\mathbf{1}$ is the indicator function.

Suppose that the connection arrival process to the queueing system is deterministically constrained by a strictly increasing function g on \mathbb{R}^+ such that $g(0) = 1$ and $g(0-) = 0$; i.e., for all times $s < t$,

$$N(s, t] \leq g(t - s) \text{ a.s.} \quad (2)$$

For example, if the connection arrival process is “spaced” so that the peak connection arrival rate is π and if a leaky bucket with token buffer size $\sigma > 1$ and token rate $\rho < \pi$ is also applied, then, for $x > 0$, we take

$$g(x) \equiv \min\{1 + \pi x, \sigma + \rho x\}$$

in (2) above [1, 2, 3, 6].

If connections are not blocked in the C -server system, the number of connections in the system at time t is

$$\sum_{i=-\infty}^{\infty} \mathbf{1}\{T_i \leq t < T_i + S_i\} \equiv Q(t) \quad (3)$$

where we note that the termination (departure) time of the i^{th} connection is $T_i + S_i$.

3 A no-blocking condition for G/G/C/C queues with $C < \infty$

Clearly, if $Q(t) \leq C$ for all time t , where Q is the queue occupancy process specified by (3), then connections are not blocked. In this section, we consider the case of a G/G/C/C queue with C finite. A condition will be given on the connection arrival times and durations which imply that $Q(t) \leq C$ for all t . Suppose that all the connections have a maximum duration $S^{\max} < \infty$, i.e., for all i ,

$$S_i \leq S^{\max} \text{ a.s.} \quad (4)$$

In this case

$$Q(t) = \sum_{i=-\infty}^{\infty} \mathbf{1}\{t - S_i < T_i \leq t\} \quad (5)$$

$$\leq \sum_{i=-\infty}^{\infty} \mathbf{1}\{t - S^{\max} < T_i \leq t\} \quad (6)$$

$$= N(t - S^{\max}, t]. \quad (7)$$

Under assumption (2), (7) implies

$$Q(t) \leq g(S^{\max}). \quad (8)$$

Therefore, we have proved [7]

Theorem 1 *If*

$$g(S^{\max}) \leq C \quad (9)$$

then the occupancy of the G/G/ ∞ system will never exceed C , i.e., a connection will not be blocked by the corresponding G/G/C/C system.

4 Bounds on $P(Q(t) > C)$ for the G/GI/ ∞ queue

Note that (3) is the expression for the content of a G/G/ ∞ queue (which is obviously lossless) which we now consider. We now assume that the $\{S_i\}$ are i.i.d. and that they are independent of the arrival times $\{T_i\}$. The objective is to find a bound on $P(Q(t) > C)$ which may be used as an approximate bound on the connection blocking probability for the G/GI/C/C queue with the same arrival process.

Theorem 2

$$P(Q(t) > C) \leq \exp\left(-\sup_{\theta > 0} \left\{ \theta C - \int_0^\infty \log(\Phi(\gamma)e^\theta + 1 - \Phi(\gamma)) d\gamma \right\}\right) \quad (10)$$

where

$$\Phi(x) \equiv P(g(S_i) > x). \quad (11)$$

Proof:

Let

$$\varepsilon_k \equiv P(g^{-1}(m_k) < S_i \leq g^{-1}(m_{k+1})) \quad (12)$$

for $k \in \{0, 1, 2, \dots, K\}$ where $m_0 \equiv 0-$, $g(m_0) \equiv 0$, $m_{K+1} \equiv g(S^{\max})$ and $m_k < m_{k+1}$ for all k . Note that

$$\sum_{k=1}^K \varepsilon_k = 1$$

and that $\{m_{k+1} - m_k\}_{k=0}^K$ is a partition of the range of the random variable $g(S)$.

Define the job indexes here so that

$$T_{-1} \leq t < T_0.$$

Thus,

$$\begin{aligned} Q(t) &\leq \sum_{i=-\infty}^{-1} \mathbf{1}\{t - S_i < T_i \leq t\} \\ &= \sum_{i=-\infty}^{-1} \sum_{k=0}^K \mathbf{1}\{t - S_i < T_i \leq t\} \mathbf{1}\{g^{-1}(m_k) < S_i \leq g^{-1}(m_{k+1})\} \\ &= \sum_{k=0}^K \sum_{i=-\infty}^{-1} \mathbf{1}\{t - S_i < T_i \leq t\} \mathbf{1}\{g^{-1}(m_k) < S_i \leq g^{-1}(m_{k+1})\} \\ &\leq \sum_{k=0}^K \sum_{i=-\infty}^{-1} \mathbf{1}\{t - g^{-1}(m_{k+1}) < T_i \leq t\} \mathbf{1}\{g^{-1}(m_k) < S_i \leq g^{-1}(m_{k+1})\}. \end{aligned}$$

By (2), $N(t - g^{-1}(m_{k+1}), t] \leq m_{k+1}$ a.s. Therefore,

$$Q(t) \leq \sum_{k=0}^K \sum_{i=-m_{k+1}}^{-1} \mathbf{1}\{g^{-1}(m_k) < S_i \leq g^{-1}(m_{k+1})\}.$$

Switching the order of summation again, we get

$$\begin{aligned}
Q(t) &\leq \sum_{i=-m_1}^{-1} \sum_{k=0}^K \mathbf{1}\{g^{-1}(m_k) < S_i \leq g^{-1}(m_{k+1})\} \\
&\quad + \sum_{i=-m_2}^{-m_1-1} \sum_{k=1}^K \mathbf{1}\{g^{-1}(m_k) < S_i \leq g^{-1}(m_{k+1})\} \\
&\quad + \sum_{i=-m_3}^{-m_2-1} \sum_{k=2}^K \mathbf{1}\{g^{-1}(m_k) < S_i \leq g^{-1}(m_{k+1})\} \\
&\quad \dots + \sum_{i=-m_{K+1}}^{-m_K-1} \mathbf{1}\{g^{-1}(m_K) < S_i \leq g^{-1}(m_{K+1}) \equiv S^{\max}\}
\end{aligned}$$

and, consequently,

$$\begin{aligned}
Q(t) &\leq \sum_{j=0}^K \sum_{i=-m_{j+1}}^{-m_j-1} \mathbf{1}\{g^{-1}(m_j) < S_i\} \\
&= \sum_{j=0}^K \sum_{i=-m_{j+1}}^{-m_j-1} \mathbf{1}\{g(S_i) > m_j\}
\end{aligned} \tag{13}$$

Note that the summands of (13) are all *independent* random variables.

Now recall that the S_i are i.i.d. and the definition of Φ in (11). Let $q_j = \mathbf{P}(g(S_i) > m_j)$. For all real $\theta > 0$, (13) implies that

$$\begin{aligned}
\mathbf{E}e^{\theta Q(t)} &\leq \prod_{j=0}^K \prod_{i=-m_{j+1}}^{-m_j-1} (q_j e^\theta + 1 - q_j) \\
&= \prod_{j=0}^K (q_j e^\theta + 1 - q_j)^{m_{j+1} - m_j}
\end{aligned}$$

Thus, for all real $\theta > 0$,

$$\log \mathbf{E}e^{\theta Q(t)} \leq \sum_{j=0}^K (m_{j+1} - m_j) \log(q_j e^\theta + 1 - q_j). \tag{14}$$

Finally note that, as the partition $\{m_{k+1} - m_k\}_{k=0}^K$ becomes infinitely fine (i.e., $K \rightarrow \infty$ and $m_{k+1} - m_k \rightarrow 0$ for all k), the right-hand-side of (14) converges to (the Lebesgue integral)

$$\int_0^\infty \log(\Phi(\gamma)e^\theta + 1 - \Phi(\gamma)) \, d\gamma.$$

Thus,

$$\log \mathbf{E}e^{\theta Q(t)} \leq \int_0^\infty \log(\Phi(\gamma)e^\theta + 1 - \Phi(\gamma)) \, d\gamma. \tag{15}$$

Finally, the Chernoff bound is

$$\begin{aligned}
\mathbf{P}(Q(t) > C) &\leq \exp\left(-\sup_{\theta > 0} \{\theta C - \log \mathbf{E}e^{\theta Q(t)}\}\right) \\
&\leq \exp\left(-\sup_{\theta > 0} \left\{\theta C - \int_0^\infty \log(\Phi(\gamma)e^\theta + 1 - \Phi(\gamma)) \, d\gamma\right\}\right)
\end{aligned} \tag{16}$$

as desired. □

Note that an immediate consequence of (13) is

$$EQ(t) \leq Eg(S) \tag{17}$$

and therefore, by Markov's inequality,

$$P(Q(t) > C) \leq \frac{Eg(S)}{C}. \tag{18}$$

This bound is simpler to compute but cruder than the Chernoff bound of Theorem 2.

5 Summary and final remarks

In this note, we consider the use of traffic shaping for *connection*-level management in communication networks. A probabilistic bound on the connection blocking probability was derived at a single boundary node consisting of C servers which receives a deterministically shaped connection arrival process.

There is easily enough time between admission control decisions (on the order of hundreds of μ s or more) to numerically compute the Chernoff bound. In [5], the bound was quickly computed using Simpson's rule to evaluate the integral and a simple linear search (or Newton's method) to find the maximizing θ . We observed that the bound was accurate for some ranges of parameters of the traffic models studied (e.g., on-off Markov fluids) but was quite conservative (sometimes by an order of magnitude) for other ranges of parameters. We reiterate that the derived Chernoff bound holds for a very wide variety of connection arrival process N output from the traffic shaper g ; in modern communication network settings, the incident traffic M is often unknown and difficult to predict. Given a desired upper bound on the connection blocking probability, the derived bound can also be used to numerically compute the "traffic capacity" of the shaper/ C -server system in terms of the (traffic) parameters of the shaper and the service-time distribution Φ [9, 11]. Finally, the Chernoff bound can be used to approximate how the connection blocking probability depends on the holding time distribution Φ and the parameters of the traffic shaper g .

References

- [1] Internet Engineering Task Force (IETF) documents concerning MPLS and RSVP at URL www.ietf.org
- [2] ITU-T Study Group 13. Traffic control and congestion control in B-ISDN. Technical Report I.371, ITU-T, Geneva, Apr. 29 - May 10, 1996.
- [3] ATM Forum Technical Committee. Traffic management specification version 4.0. Technical Report af-tm-0056.000, The ATM Forum, Draft version 3.0, April 1996.
- [4] Baccelli, F. and Brémaud, P. *Elements of Queueing Theory*. Springer-Verlag, New York, 1994.
- [5] K. Chakraborty, L. Tassiulas and G. Kesidis. Reducing connection blocking likelihoods by traffic shaping. University of Maryland I.S.R. Technical Report, in preparation.
- [6] R. L. Cruz. A calculus for network delay, Part 1: Network elements in isolation. *IEEE Trans. Inform. Theory*, 37:114–131, 1991.

- [7] R. L. Cruz. Quality of Service Guarantees in Virtual Circuit Switched Networks. *IEEE JSAC*, 13(6):1048–1056, Aug. 1995.
- [8] F.P. Kelly. Loss networks. *Ann. Appl. Prob.*, Vol. 1:pp. 317–387, 1991.
- [9] G. Kesidis and T. Konstantopoulos. Extremal Shape-Controlled Traffic Patterns in High-Speed Networks. *IEEE Trans. Comm.*, May 2000.
- [10] G. Kesidis and T. Konstantopoulos. Worst-case performance of a buffer with independent shaped arrival processes. *IEEE Communications Letters*, Vol. 4, No. 1, p. 26–28, Jan. 2000.
- [11] G. Kesidis and T. Konstantopoulos. Extremal traffic and worst-case performance for a queue with shaped arrivals. In *Proc. Workshop on Analysis of Communication Networks: Call Centres, Traffic and Performance, Fields Institute, Toronto*, Nov. 9-13, 1998.
- [12] K.W. Ross. *Multiservice Loss Models for Broadband Telecommunication Networks*. Springer-Verlag, London, 1995.