Leaky Buckets and Optimal Self-tuning Rate Control

Gustavo de Veciana*

Department of Electrical and Computer Engineering, U.T. Austin, Austin, TX 78712

ABSTRACT *Using the theory of large deviations, we consider the effect of rate control on the large buffer asymptotics of bursty arrival streams entering a communication network. We determine the "effective bandwidth" of the output stream from a rate control throttle, and propose the idea of a self-tuning rate throttle which estimates the "optimal" token or release rate in terms of minimizing the output's effective bandwidth subject to a loss or delay constraint.*

1 Introduction

In this note we consider the impact of rate control throttles used in high speed networks to regulate packetized information flows. Such devices are used as "throughput-burstiness" filters for packet streams in broadband integrated services digital networks (B-ISDN) using the asynchronous transfer mode (ATM), as well as for overload control in computer and communication systems when the arrival rate of jobs exceeds the system's capacity, see Berger et al. [3, 4] and references therein for more discussion.

We will discuss the popular "leaky bucket" rate throttle shown in Figure 1. It operates as follows: Multiple packets or jobs A_n arrive at each time slot n, while tokens arrive according to another random process, usually taken to be deterministic, with rate ρ larger than the mean arrival rate of packets. Packets consume tokens and leave instantaneously if they are available. Tokens that arrive when the token buffer of size T is full are discarded. Packets that arrive when no tokens are available are delayed until tokens come in. Here we will consider the case where the packet buffer B, is quite large, or alternatively we assume the throttle is designed for very low packet loss rates. In addition a scheme for allowing excess arrivals to enter the network as low priority traffic can also be included, via a peak rate threshold at the output of the controller or by labeling traffic arriving to a full packet buffer as low priority. When congestion occurs in the network, low priority traffic can be discarded to relieve the system. Herein we will only consider the departure process D_n, noting that the impact of thresholding is highly non-linear and traffic dependent, see [8] for further discussion.

In our formulation the network designer must select the token buffer size T and the token arrival rate ρ, given a fixed available buffer size B and possibly unknown but hopefully

*The research of G. de Veciana was supported in part by the National Science Foundation under Grant NCR-9409722.

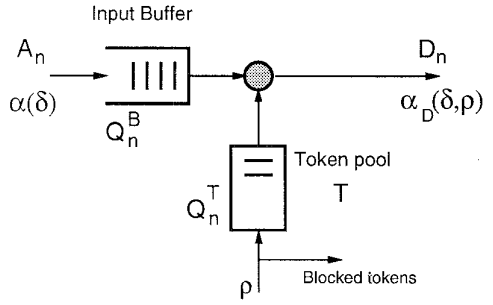

Figure 1: Leaky-bucket input control

stationary arrival traffic statistics. His goal is to reduce the "burstiness" of the output stream, so as to minimize the network resources required to buffer traffic fluctuations, while satisfying a statistical loss or delay constraint at the network edge. The resources required by a traffic stream in the network are measured via the notion of output stream's *effective bandwidth*. This roughly characterizes bandwidth requirements of the stream, see [14, 12, 5, 16, 11, 17, 8] and many references therein. Another notion of optimality is investigated in [1].

We then propose the idea of a self-tuning traffic shaper, which when confronted with unknown statistics but a known loss (or delay) constraint at the edge of the network, estimates the optimal release rate, subject to the performance constraint. This is a first step towards robust traffic control. The idea of combining bandwidth allocation, rate control, and estimation is not new, see for example [13].

2 Background - Large Deviations and Effective Bandwidths

Our arguments are based on large deviations results, so we begin with a very brief review; for a complete reference on the subject see Dembo et al. [9]. Consider the distributions $\{\mu_n\}$ of the partial sums $n^{-1}S_n = n^{-1}\sum_{i=1}^{n} A_n$, for a sequence of real-valued random variables $\{A_n\}$. We say that $\{A_n\}$ satisfies a large deviation principle (LDP) with a *good rate function* $I(\cdot)$ if for every closed set F and open set G,

$$\limsup_{n\to\infty} \frac{1}{n} \log \mu_n(F) \leq - \inf_{x\in F} I(x) \quad \text{and}$$

0-7803-1820-X/94 $4.00 © 1994 IEEE

$$\liminf_{n\to\infty} \frac{1}{n} \log \mu_n(G) \geq - \inf_{x\in G} I(x),$$

where $\{x : I(x) \leq \alpha\}$ is compact for $\alpha < \infty$. Thus for example if the sequence is stationary with mean m, the probability that the empirical mean over a long time interval $[1, n]$ exceeds $\alpha > m$ is given by

$$\mathbb{P}(\frac{S_n}{n} \geq \alpha) \approx \exp[-nI(\alpha) + o(n)],$$

where $\lim_{n\to\infty} o(n)/n = 0$. Below we briefly discuss when such bounds do indeed hold.

The Gärtner-Ellis Theorem establishes the existence of an LDP with a convex good rate function for a large class of sources. It requires that the limits

$$\Lambda(\theta) \stackrel{\Delta}{=} \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\exp[\theta S_n],$$

exist (possibly infinite) for all $\theta \in \mathbb{R}$, in addition to two important but technical assumptions, see [9]. With these conditions in place an LDP holds with a good rate function given by the convex dual $\Lambda^*(\cdot)$, of $\Lambda(\cdot)$:

$$I(x) = \Lambda^*(x) = \sup_{\theta}[\theta x - \Lambda(\theta)].$$

This result applies to i.i.d. sequences with $\mathbb{E}e^{\theta A_1} < \infty$ for all θ, which corresponds to the original large deviation estimate of Cramér. LDPs can also be found for sequences with weak dependencies, e.g., coordinate functions of Markov processes and mixing sequences, see [9].

In this note the effective bandwidth, $\alpha(\delta)$ associated with a traffic stream $\{A_n\}$, corresponds to the minimum deterministic service rate required to satisfy a quality of service δ. More specifically for a stationary discrete-time queue with input $\{A_n\}$, service rate c, and stationary queue length Q, we have that

$$\alpha(\delta) \leq c \Leftrightarrow \lim_{B\to\infty} \frac{1}{B} \log \mathbb{P}(Q > B) \leq -\delta, \qquad (1)$$

where B is to be interpreted as a large buffer. Thus in practice given that the service rate c of a queue satisfies $\alpha(\delta) \leq c$, we can conclude that the probability of overflow is roughly given by $\exp[-\delta B]$, which permits making resource management decisions subject to statistical loss (or delay constraints); e.g., the probability of buffer overflow should be no more than 10^{-9}. A further key property is that, when multiple *independent* streams share a queue, the effective bandwidth of the aggregate is the sum of the individual contributing streams; thus, it is easy to test whether the superposition of several heterogeneous traffic streams will satisfy such a constraint. In general for streams satisfying an LDP, the effective bandwidth is given by $\alpha(\delta) = \Lambda(\delta)/\delta$. For further discussion see the aforementioned references and many references therein.

3 Departure process from a leaky-bucket controller

The intuition on which our analysis is based, is that deviations in the empirical mean of the output from a "flow-conserving" traffic shaping device, necessarily correspond to those of the input unless packets are accumulating within. Thus the rate function of the output stream will equal that of the input as long as no accumulation is taking place. This intuition led to the following result characterizing the output of a deterministic discrete time queue:

Theorem 3.1 [For details see [7]] *Let $\{A_n\}$ be a stationary ergodic arrival process, such that $\mathbb{E}A_1 = m < c$, which either satisfying a "nice" LDP such that for all $\theta < \infty$*

$$\Lambda(\theta) = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\exp[\theta \sum_{i=1}^{n} A_i] < \infty,$$

and with $\Lambda^*(\cdot)$ is strictly convex[1]*. Then the Lindley process*

$$Q_{n+1} = [Q_n + A_n - c]^+ \qquad (2)$$

has a stationary distribution, say that of a random variable Q, and the associated departure process $\{D_n\}$ satisfies an LDP with with convex good rate function given by $\Lambda^*(\cdot)$ on $[0, c]$ and infinite on $[0, c]^c$.

We use the following as a rough model for the dynamics of a leaky bucket controller:

$$\tilde{Q}_{n+1} = \max[-T,\ \tilde{Q}_n + A_n - \rho], \qquad (3)$$

where the token queue is given by $\tilde{Q}_n^T = \max[0, -\tilde{Q}_n]$, while the cell buffer queue is given by $\tilde{Q}_n^B = \max[0, \tilde{Q}_n]$. Observe that the underlying dynamics are those of a random walk reflected at $-T$, while the dynamics of the queue in Theorem 3.1 were those of a random walk reflected at 0.

We will show that the departures from a leaky bucket controller have the same rate function as those from a discrete-time queue with service rate $c = \rho$. Let S_n^A, S_n^D and $\tilde{S}_n^A, \tilde{S}_n^D$ denote the cumulative arrivals and departures for the discrete-time queue and leaky bucket (respectively) over the time interval $[1, n]$.

Let $\tilde{Q}_n = Q_n - T$ and note that if Q_n satisfies Eq. 2, then $\tilde{Q}_n$ satisfies Eq. 3. Moreover, cumulative departures of the leaky bucket are given by $\tilde{S}_n^D = S_n^A - \tilde{Q}_{n+1}^B$ while the cumulative departures from a deterministic queue are given by $S_n^D = S_n^A - Q_{n+1}$. Starting both systems empty they converge to steady state, and it should be clear that $\tilde{S}_n^D \geq S_n^D$, hence

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\tilde{S}_n^D > n\alpha) \geq \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n^D > n\alpha) \geq -\Lambda^*(\alpha).$$

[1] See reference for details. A typical application would be to sources with bounded arrivals satisfying the Gärtner-Ellis Theorem.

Where the last inequality follows from Theorem 3.1 as long as $\alpha \in [0, \rho]$. We have shown a large deviations lower bound for open sets of the form (α, ∞), the same argument follows for open balls from which we get the lower bound for open sets.

The large deviations upper bound is a little more difficult. One seeks an upper bound not by starting with an empty system but by bounding the possible content of the system in steady state. Indeed, in steady state the cumulative output of the controller is bounded above: $\tilde{S}_n^D \leq S_n^A + Q^B \leq S_n^A + Q$, where Q^B and Q denote the stationary distributions in the leaky bucket's cell buffer and that of the deterministic queue with the same service rate. The result now follows the proof of Theorem 3.1 where the fact that Q has finite exponential moments $\mathbb{E}\exp[\theta Q] < \infty$ for some range of θ, and the Chebychev bound are used to establish that the LDP of $\tilde{S}_n^D$ is that of S_n^A in the interval if interest, i.e,,

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\tilde{S}_n^D \geq n\alpha) \leq -\Lambda^*(\alpha), \tag{4}$$

see [7] for the details of a similar argument used to prove Theorem 3.1. Clearly the output rate from leaky bucket cannot exceed ρ so the rate function is infinite outside of the interval $[0, \rho]$. As for the lower bound we can extend the upper bound for closed sets $[\alpha, \infty)$ to arbitrary closed sets by considering F contains the mean, see Dembo et al. [9].

Thus the rate function for the stationary departure process from a leaky bucket controller, is simply that of the input stream on $[0, \rho]$ and infinite otherwise. The effective bandwidth for the output process is given by $\alpha_D(\delta, \rho) = \Lambda_D(\delta)/\delta$ where $\Lambda_D(\cdot)$ is the convex dual of the departure process' rate function, see §2. This gives

[Output Effective Bandwidth for Leaky Bucket]

$$\boxed{\alpha_D(\delta, \rho) = \begin{cases} \alpha(\delta) & \text{if} \quad \alpha^*(\delta) < \rho, \\ \rho - \frac{1}{\delta}\,\Lambda^*(\rho) & \text{otherwise,} \end{cases}} \tag{5}$$

where $\alpha^*(\delta)$ is defined implicitly by the convex duality relationship $\Lambda(\delta) = \alpha^*(\delta)\delta - \Lambda^*(\alpha^*(\delta))$. Note that $\rho - \frac{1}{\delta}\,\Lambda^*(\rho) \leq \alpha(\delta)$, so for some values of δ the throttle can reduce the effective bandwidth of the arrival stream. The quantity $\alpha^*(\delta)$, introduced in [7] as the *decoupling bandwidth*, turns out to be smaller than the peak rate of the stream (if it exists) and always larger than the effective bandwidth, that is

$$\text{peak rate} > \alpha^*(\delta) > \alpha(\delta). \tag{6}$$

Based on this input-output relationship for the effective bandwidth of a stream passing through a leaky bucket controller we can consider several design scenarios.

3.1 Minimizing output effective bandwidth subject to an overflow or loss constraint

Suppose we are given a statistical *overflow* constraint δ_o on the leaky bucket controller of the form

$$\lim_{B\to\infty} \frac{1}{B} \mathbb{P}^\rho(Q^B > B) \leq -\delta_o$$

where B denotes a reasonably large input cell buffer, and Q^B has the steady state distribution of cells in in a leaky bucket with token rate ρ. Thus, roughly, the constraint might ensure a small probability of overflow

$$\mathbb{P}^\rho(Q^B > B) \leq \exp[-\delta_o B] \approx 10^{-9},$$

for an appropriate choice of δ_o.

The goal is to determine the "optimal" token rate in terms of minimizing the departure processes' effective bandwidth subject to this performance constraint. Mathematically can we express the optimal control ρ^* as the solution to,

$$\begin{aligned} \min_{\rho>0} \quad & \alpha_D(\delta, \rho) \\ \text{such that} \quad & \lim_{B\to\infty} \tfrac{1}{B} \log \mathbb{P}^\rho(Q^B > B) \leq -\delta_o \end{aligned} \tag{7}$$

Let us express the size of the token buffer T as a fraction t of the job buffer, that is $T = tB$. Recall from our earlier derivation that

$$\mathbb{P}^\rho(Q^B > B) = \mathbb{P}^\rho(Q > B + T) = \mathbb{P}^\rho(Q > B(1+t)),$$

where Q has the steady state queue length distribution of a discrete-time queue with service rate ρ. The constraint in Eq. 7 will be satisfied as long as

$$\lim_{B\to\infty} \frac{1}{B} \log \mathbb{P}^\rho(Q > B(1+t)) \leq -\delta_o,$$

now comparing with Eq. 1 we see that we need only require that $\alpha(\delta_o/(1+t)) \leq \rho$.

Since the cost, $\alpha_D(\delta, \rho)$ is non decreasing in ρ for all δ (see Eq. 5) the optimal token rate is $\rho^* = \alpha(\delta_o/(1+t))$; this is the slowest release rate that satisfies the overflow constraint. Thus traffic arrives into the controller gets queued as much as possible and is released smoothly into the network. For this release rate the effective bandwidth of the output stream is given by:

$$\boxed{\begin{aligned} &\rho^* = \alpha(\delta_o/(1+t))) && \text{[Optimal Token Rate];} \\ &\alpha_D(\delta, \rho^*) && \text{[Output Eff. Band.],} \end{aligned}} \tag{8}$$

where the argument δ corresponds to a desired QoS in the network and δ_o is a QoS constraint on overflows in the throttle's cell buffer. Further consideration of Eqs. 5, 6 along with remarks following these equations, we find that in the usual scenario where $\delta = \delta_o$, the effective bandwidth of the output can in fact be reduced. Note that we have ignored the effects of actual losses in the leaky bucket on the premise that they are indeed small, and to be avoided at all costs.

3.2 Minimizing the output effective bandwidth subject to a statistical delay constraint.

When a statistical *delay* constraint at the traffic shaper is specified rather than an overflow (or loss) constraint the problem is significantly more complex. As in the previous case a change in the token rate will affect the asymptotic tail disribution of the occupancy in the job buffer. We can then determine how long it will take to empty the work therein by dividing by the token rate. Thus the token rate affects the delay performance via two mechanisms and the optimal rate of release represents a compromise between these two mechanisms. We will see that unfortunately the compromise reduces to solving a nonlinear equation.

The statisitcal delay constraint is specified as

$$\lim_{D\to\infty} \frac{1}{D} \log \mathbb{P}^{\rho}(V > D) \leq -\delta_d$$

where V denotes stationary *virtual delay* in the leaky bucket and δ_d is to be understood as a constraint on the tail distribution of delays.[2] D is assumed to be relatively large so we are aiming to have a rough estimate on the probability of large delays of the form $\mathbb{P}(V > D) \leq \exp[-\delta_d D]$. We consider minimizing the output effective bandwidth subject to this new constraint, that is we replace Eq. 7 with the above delay constraint.

In the spirit of these approximations we compute

$$\mathbb{P}^{\rho}(V > D) \approx \mathbb{P}^{\rho}(Q > D/\rho + T) = \mathbb{P}^{\rho}(Q > B(1+t))$$

where $B = D/\rho$. The delay constraint is satisfied if

$$\lim_{D\to\infty} \frac{1}{D} \log \mathbb{P}^{\rho}(Q > B(1+t)) \leq -\delta_d.$$

By comparing with Eq. 1 we see that this will be the case if $\alpha(\delta) \leq \rho$ where $\delta = \delta_d \rho/(1+t)$ So the optimal release rate ρ^* should satisfy

$$\begin{array}{ll} \alpha(\delta_d \rho^*/(1+t)) = \rho^* & [\text{Optimal Token Rate}], \\ \alpha_D(\delta, \rho^*) & [\text{Output Eff. Band.}]. \end{array} \quad (9)$$

4 Towards Self-tuning Optimal Traffic Shaping

It is unclear whether the statistical measurements required to obtain the effective bandwidth of a source can be carried out prior to the transmission of a traffic stream. However, it is likely that a QoS constraint on the delays or losses required by a particular application, e.g., video or audio, can be determined a-priori. A self-tuning traffic shaper monitors the arrival process and selects the token or release rate such that the effective bandwidth at the output is minimized subject to a known buffering constraint, see Figure 2. The advantage is that only the constraint needs to be specified while the actual traffic statistics may in fact be unknown. Moreover in practice such a scheme might track slowly varying traffic statistics.

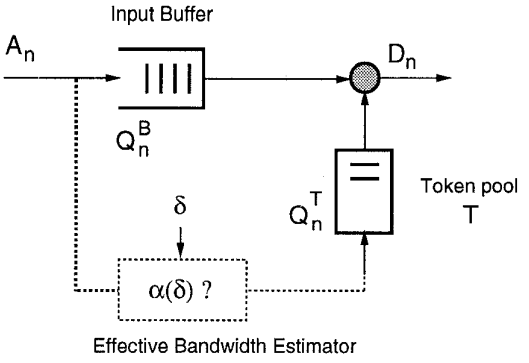

Figure 2: Self-tuning traffic shaper

In the previous section we found that, given a δ_o constraint on cell buffer overflows, the optimal token rate corresponded to the effective bandwidth $\alpha(\delta)$ where $\delta = \delta_o/(1+t)$). Recall that $\alpha(\cdot)$ is is given by

$$\alpha(\delta) = \frac{1}{\delta} \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\exp[\delta \sum_{i=1}^{n} A_i].$$

The tuning part of the traffic shaper monitors the input and estimates the optimal release rate $\alpha(\delta)$.

In order for this setup to work we require a consistent online estimator for this quantity. If the arrivals were i.i.d. then the following would work

$$\alpha_n(\delta) = \frac{1}{\delta} \log(\frac{1}{n} \sum_{i=1}^{n} \exp[\delta A_i])$$

If the arrival process is approximately regenerative at random times $\{T_i\}_{i=1}^{\infty}$, we need only estimate the regeneration rate $1/\mathbb{E}T$, and estimate the average value of $\mathbb{E}\exp[\delta \sum_{i=1}^{T} A_i]$ where T denotes the typical time to regenerate. Thus,

$$\lambda_n^{-1} = \frac{1}{n} \sum_{j=1}^{n} T_j \qquad C_n = \frac{1}{n} \sum_{j=1}^{n} \exp[\delta \sum_{i=1}^{T_j} A_{i,j}]$$

where $A_{i,j}$ denotes the arrivals in the i^{th} slot of the j^{th} regeneration cycle, see Asmussen [2][page 132]. The following is then a consistent estimator of the desired quantity

$$\alpha_n(\delta) = \frac{\lambda_n}{\delta} \log C_n \to \alpha(\delta).$$

[2] The virtual delay is that which an external observer would see if he came in at some typical time. In general, we are interested in the typical delay of actual customers, however this quantity is very difficult to work with, outside of the Poissonian framework.

This may however take some time to converge if regenerations are lengthy. Ideally, we wish to make even weaker assumptions about the input process, e.g., mixing, and still find an approximate estimate of the desired quantity. In such case we might use block methods as suggested in [10].

In the case where a delay constraint has been placed at the leaky bucket a combination of a numerical method to solve Eq. 9 and estimation would be required to determine the optimal release rate without prior knowlege of the arrival statistics.

We are currently investigating the tradeoffs and effecitveness of several possible estimation methods. We believe that direct measurements of the effective bandwidth, as proposed above, can result in good estimates, rather than modelling of traffic statistics combined with a numerical or analytical exploration of bandwidth requirements. Other approachs to this problem might use indirect estimation methods [15] combined with interpolation, see for example [6].

Acknowledgement: We thank an anonymous reviewer for a detailed (and enthusiastic) critique of a previous version of this note.

References

[1] V. Anantharam and T. Konstantopoulos. An optimal flow control scheme that regulates the burstiness of traffic subject to delay constraints. In *Proceedings of 32nd CDC*, 1993.

[2] S. Asmussen. *Applied Probability and Queues.* John Wiley & Sons, 1987.

[3] A. W. Berger. Overload control using rate control throttle: Selecting token bank capacity for robustness to arrival rates. *IEEE Trans. Automatic Control*, 32(2), 1991.

[4] A. W. Berger and W. Whitt. The impact of a job buffer in a token-bank rate-control throttle. *Stochastic Models*, 8:685–717, 1992.

[5] C.S. Chang. Stability, queue length and delay of deterministic and stochastic queueing networks. *IEEE Automatic Control*, 39(5):913–931, May 1994.

[6] C. Courcoubetis, G. Kesidis, A. Ridder, J. Walrand, and R. Weber. Call acceptance and routing using inferences from measured buffer occupancy. *to appear in IEEE Trans. Comm.*, 1994.

[7] G. de Veciana, C. Courcoubetis, and J. Walrand. Decoupling bandwidths for networks: A decomposition approach to resource management for networks. *IEEE Infocom Proceedings '94, also submitted to IEEE/ACM Trans. Networking*, 1993.

[8] G. de Veciana and J. Walrand. Effective bandwidths: Call admission, traffic policing and filtering for ATM networks. *To appear Queueing Systems*, 1994.

[9] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications.* Jones & Bartlett, Boston, 1992.

[10] N. G. Duffield, J.T. Lewis, and N. O'Connell. The entropy of an arrivals process: A tool for estimating QoS parameters of ATM traffic. In *Proceedings of the 11th IEE Teletraffic Symposium*, March 1994.

[11] A. I. Elwalid and D. Mitra. Effective bandwidth of general Markovian traffic sources and admission control of high speed networks. *IEEE/ACM Trans. Networking*, 1(4), June 1993.

[12] R. Guérin, H. Ahmadi, and M. Naghshineh. Equivalent capacity and its application to bandwidth allocation in high-speed networks. *IEEE JSAC*, 9:968–981, 1991.

[13] R. Guérin and L. Gün. A unified approach to bandwidth allocation and access contro in fast packet-switched networks. In *IEEE Infocom Proceedings*, 1992.

[14] F.P. Kelly. Effective bandwidths at multi-class queues. *Queueing Systems*, 9:5–16, 1991.

[15] G. Kesidis. Personal communication.

[16] G. Kesidis, J. Walrand, and C.S. Chang. Effective bandwidths for multiclass Markov fluids and other ATM sources. *IEEE/ACM Trans. Networking.*, 1(4), August 1993.

[17] W. Whitt. Tail probabilities with statistical multiplexing and effective bandwidths in multi-class queues. *Telecommunication Systems*, 2:71–107, 1993.