Measurement Based Optimal Source Shaping and Resource Allocation for Stream Sessions with Guaranteed End-to-End Delay

Anurag Kumar, Nilesh Pastagia, Natwar Modani, Parijat Dube
Dept. of Electrical Communication Engg.
Indian Institute of Science, Bangalore 560 012, INDIA

Abstract— **We develop an approach for resource management for stream sessions based on measurements at the sources. The results in this paper assume that (i) the sessions are carried in an edge-to-edge virtual path (VP) or label switched path (LSP), that (ii) weighted fair queueing (WFQ) is used at each hop of the LSP, with one queue being used for all the traffic from this LSP, and that (iii) the stream sources are statistically equivalent; e.g., they could be packet voice sessions between two locations of an enterprise.**

The approach is based on *measurement based optimal source shaping*. **We formulate and solve the problem of selection of source parameters based on minimising the allocated bandwidth in the network, for a specified probability of violating an end-to-end delay bound;** *the end-to-end delay includes the shaping and packetisation delays*. **Our network model includes a multihop path, with WFQ at each hop. We use a statistical model for the leaky bucket shapers, and worst case delay bounds for the network with WFQ servers. Our approach yields an optimal leaky bucket (LB) rate parameter ρ^*, and the optimal sum of the shaper buffer and leaky bucket depth $(B_s + \sigma)$. We propose and study a stochastic approximation algorithm for on-line estimation of ρ^*. For fluid traffic and** *lossless multiplexing* **in the network, we find that a linear cost function in the network bandwidth and buffer is minimised by using the LB rate ρ^* and token bucket depth $\sigma = 0$. With these results, our approach for managing the bandwidth of the LSP is for each source to initially request peak bandwidth, and then renogotiate the reservation as it learns its optimal rate, ρ^*.**

We provide simulations results with on-off sources, including packet voice models, to show the bandwidth reduction possible by optimal shaping. The reduction in bandwidth relative to peak rate depends on the relative values of the end-to-end delay bound and the source burst duration. We then use simulations to explore statistical gain with *lossy multiplexing*, **for packet voice sources, when the LB rate is ρ^* and a positive value of σ is used.**

Keywords—**quality of service (QoS), resource management for Internet QoS, optimal shaping, connection admission control, parameter renegotiation, stochastic approximation, QoS for packet voice**

I. Introduction

In this paper we are concerned with the scenario depicted in Figure 1 where a number of statistically identical *stream* sources at one edge router of a packet network need to be transported to another edge router. An example of such a situation could be packet voice calls between two locations of an enterprise interconnected by the Internet or an intranet. Interactive stream traffic, such as packet telephony, requires an end-to-end delay guarantee; for example, for packet telephony the requirement can be that the mouth-to-ear (MtoE) delay is less than 200ms with a probability exceeding 95%.

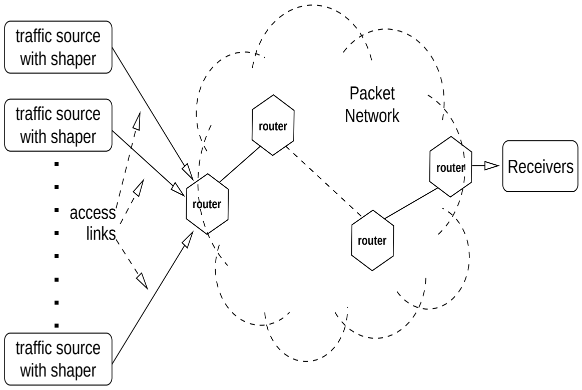


Fig. 1. The network scenario under consideration.

In order to guarantee some quality of service to a stream session in the Internet, one convenient approach is to route the session along a definite path along which resources are reserved at session initiation. Such a facility is provided by the Resource reSerVation Protocol (RSVP; see [1]). RSVP allows the source end-point of each session to reserve resources along the path to its destination. This requires per session end-to-end signalling; in addition, each router along the path of a session has to maintain a soft-state for the session. This soft-state is maintained by repeated exchange of PATH and RESV messages between the end-points of the session. It is evident that the approach of using RSVP for each session (or microflow) can lead to excessive signalling traffic and state-maintenance overheads in the routers.

An obvious alternative is to handle stream sessions as aggregates rather than as individual microflows. Such an approach has been discussed recently in the context of MultiProtocol Label Switching (MPLS); see [2]. The approach is to use an extension of RSVP (RSVP-Tunneling Extension; see [3]) that can be used to set a Label Switched Path (LSP) (with specified resource allocations) between two edge-routers. In this paper we assume that such a protocol is available, and further that Weighted Fair Queueing (WFQ) is used by the routers at

This work is based on research supported by Nortel Networks. An earlier version of this work was published in Proceedings of IEEE Infocom 2000, Tel Aviv, March 2000.

Corresponding author: Anurag Kumar; Phone: (+91)-80-360 0855, Fax: (+91)-80-360 0991; e-mail: anurag@ece.iisc.ernet.in. Nilesh Pastagia and Natwar Modani are now with IBM, India Research Lab, New Delhi, India (email: rpastagi@in.ibm.com, and mnatwar@in.ibm.com), and Parijat Dube is pursuing a PhD at INRIA, Sophia Antipolis, France (email: Parijat.Dube@sophia.inria.fr).

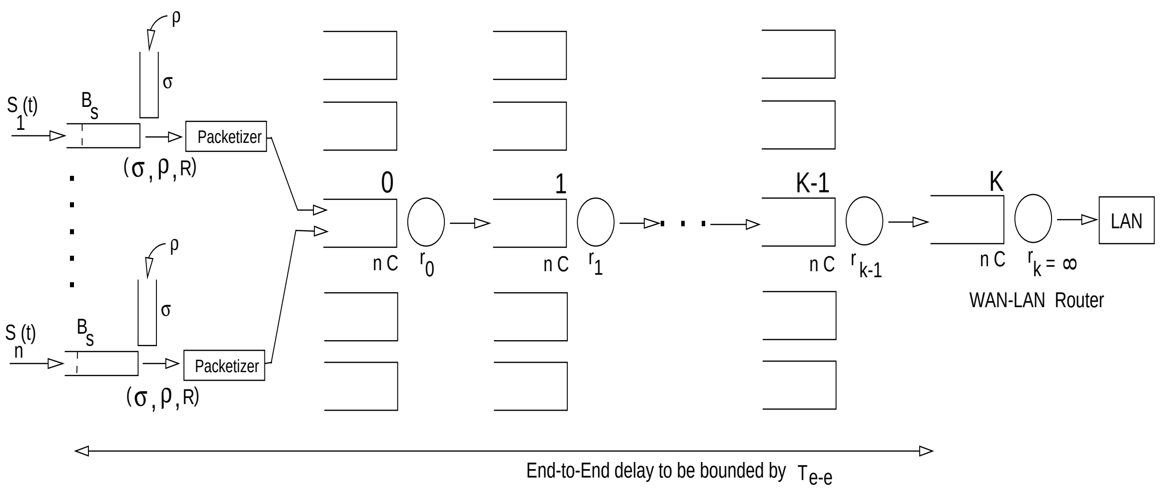


Fig. 2. The network model, showing the end-to-end LSP. Each router hop is modelled as a WFQ node. One queue at each hop is assigned to the LSP.

each hop along the path. At each hop, one queue in the WFQ scheduler is assigned to all the traffic in this LSP. Once such an LSP is set up, soft state will be maintained in the routers for the LSP, without need for per flow state maintenance. Notice that since the entire LSP is represented as one queue in the WFQ schedulers, the scheduling overhead is also reduced as compared to WFQ with per session queues. We can anticipate that the protocol for setting up the LSP will also include support for increasing or decreasing its allocated bandwidth as the stream sessions within it arrive or depart. Such a feature is automatic in RSVP, as each PATH message can potentially carry new traffic parameters[1].

Next is the question of determining an efficient bandwidth allocation for the LSP so that the stream sources it carries meet their desired end-to-end delay QoS. It is well known that for efficient resource allocation to bursty stream sources in an integrated services packet network, it is important that each source describes itself to the network in terms of some statistical parameters, and then *shapes* itself to conform to its declared parameters. A standard procedure that is used for this purpose is the Leaky Bucket (LB) algorithm [4]. There is, however, the *important question of how a source determines its leaky bucket parameters*. An on-line source (i.e., not stored; e.g., a packet voice phone call, or a live video broadcast) would need to *estimate* its LB parameters. In general, even for a stationary source these parameters would not be unique. What should be the criterion for choosing a specific set of parameters?

In this paper we develop an approach to determine a set of *optimal* leaky bucket parameters, and measurement based estimation of these parameters. The optimality is in the sense of minimising network resources, while meeting the QoS objectives for the sessions.

We consider the network scenario shown in Figure 1. There are n sources, assumed to be statistically identical (e.g., voice sources using the same coding and silence suppression scheme). Each source is shaped by a LB, and then the sources are multiplexed at the network edge node. Such a situation would arise, for example, between the packet voice "PBXs" of two enterprise locations connected by the Internet or an intranet. We assume that there is a high speed interconnection between the voice sources and the network edge router. In particular, the voice ports could be integrated into the edge router, in which case the shapers would be on a processor in the router voice card, and the interconnection medium would be the router's system bus.

Motivated by the above discussion, the detailed model that we work with is shown in Figure 2. The source outputs are modelled as being fluid. Each source is shaped by a LB, and then packetised. The LB has a token rate ρ, token bucket depth σ, and a source buffer threshold B_s, exceedance of which results in violation of the end-to-end delay bound[2]. The LSP traverses K WAN links, and hence encounters $K+1$ routers. The WAN link rates are $r_0, r_1, \ldots, r_{K-1}$. The $K+1$th router is a WAN-LAN router; assuming a high-speed interconnection between the WAN edge and the sinks of the stream sources, we take $r_K = \infty$ (see Figure 2). When there are n sources the WFQ weights at each node are set up so that the rate allocated to the LSP is nC; here C, the per source capacity required, has to be determined. The shaper parameters, taken together with the QoS requirements, determine the network resource requirements; i.e., C and the node buffer requirements.

We are interested in choosing the shaper parameters ρ, σ and B_s so as to minimise the network resources required for

[1] An alternate approach could be to make a soft reservation of some bulk bandwidth when the LSP is first set up. This soft reservation only prevents this bandwidth from being given away by the network to other *guaranteed* bandwidth flows. The network would then begin to levy a soft leasing charge for this bulk bandwidth. Firm reservation of bandwidth, and hence appropriate allocation of WFQ weights at the routers, would only be done when stream sessions actually arrive to be carried by the LSP. End-to-end signalling would then set up these weights. The network would then appropriately levy a per session charge as new calls are set up; this charge would depend on session duration. Arriving sessions would be blocked after the LSP exhausts its original soft bandwidth allocation.

[2] B_s does not represent a finite buffer; the LB source buffer is taken as infinite.

providing the desired QoS. The network resources comprise the reserved network link capacity, nC, and the buffers at the nodes. It is these network resources that are scarce and expensive (the memory on the router's WAN link interface card has to be of a much higher speed than that required for source shaping in a client computer, or on a voice card in the router), and hence we consider the minimisation of network capacity and buffer requirements. The QoS constraint is that the *shaping delay (in the source shaper buffer)* ***plus*** *the total end-to-end multiplexing delay* can exceed T_{e-e} only with a small probability q. We call this the *QoS Violation Probability (QVP)*.

With the above problem in mind, in this paper *we first assume that network multiplexing is lossless*. Hence the QoS (end-to-end delay of T_{e-e}) is violated only if the shaper buffer builds up to such an extent that the end-to-end delay exceeds T_{e-e}. We determine the values of ρ, σ, B_s so as to minimise the network capacity required to handle the stream flows with lossless multiplexing. This approach *characterises an optimal* ρ^*. The required value of $C = \rho^*$. We then develop *a measurement based method for on-line estimation of* ρ^*. The method makes fairly general assumptions about the source model. We then provide some analytical and simulation results for on-off Markov modulated sources. Working with an on-off Markov modulated model for voice we show how ρ^* varies with T_{e-e}. We then fix the token rate as ρ^*. Noting that for stream traffic a packet that arrives after the delay bound is as bad as a lost packet, we next consider lossy multiplexing in the edge node of the network. The QVP is split between loss due to excessive delay and buffer overflow in the edge node. We provide simulation results with the on-off voice source model that show the additional improvement in resource utilisation possible by lossy multiplexing.

There are four notable references that are related to our work in this paper. In [5] the authors study the problem of finding an optimal sustainable rate parameter based on network buffer-bandwidth cost considerations. They do not, however, consider any delay constraint, as we do in our paper. Another related paper is [6]. The objectives of the research reported in this paper are similar to ours, i.e., to choose optimal leaky bucket parameters subject to a QoS constraint. The approach and results are different, however. Whereas in [6] the author only considers delay in the LB buffer, we consider the problem of choosing LB parameters under an end-to-end delay constraint. We derive the LB parameters that minimise the network bandwidth required. In addition, we demonstrate the efficacy of a stochastic approximation based technique for estimating the optimal sustainable rate parameter, and for tracking slow changes in the source statistics. Another related work is reported in [7]. The authors minimise a network cost function, but have only put a constraint on the shaping delay. Also their network cost is simply the capacity required for a given network buffer, whereas we have considered the capacity-buffer tradeoff.

This paper is organized as follows. In Section II, we review the leaky bucket shaper. In Section III we develop the end-to-end delay QoS requirement. In Section IV, we formulate and solve the problem of finding the optimal sustainable rate parameter ρ^* in the model of Figure 2. We show that for a fluid source model, lossless multiplexing and a linear buffer-bandwidth cost function, ρ^* and $\sigma = 0$ are the optimal LB parameters. In Section V, we provide an on-line estimation scheme to determine ρ^*. In Section VI we show how ρ^* can be analytically estimated. Section VII provides extensive numerical results, including simulation results for an on-off packet voice model, and for lossy multiplexing. We conclude in Section VIII. Two proofs are provided in the Appendix.

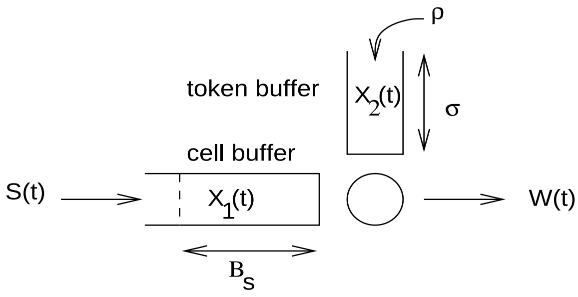


Fig. 3. The Leaky Bucket shaper. The buffer is infinite, but we want the probability of exceeding the buffer level B_s to be small.

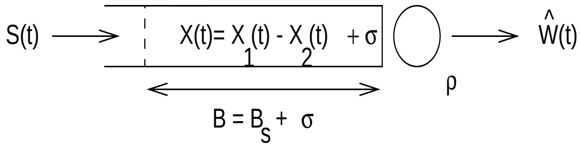


Fig. 4. A single server queuing system, with service rate ρ and infinite buffer, that is equivalent to the (σ, ρ) leaky bucket shaper from the QVP (see text) point of view.

II. THE LEAKY BUCKET SHAPER: A REVIEW

Figure 3 shows the leaky bucket (LB) controller/shaper, and the associated notation that we shall use. We shall not concern ourselves with peak rate control, *assuming* that the input is already peak rate controlled to the rate R (e.g., a PCM voice coder, with activity detection, would emit bits at 64Kbps during active periods). The processes $S(t)$ and $W(t)$ shown in Figure 3 are to be viewed as fluid rate processes. The process $W(t)$ is packetised before being offered to the network (see Figure 2). We note here that while in the analysis we assume a fluid model at the input to the LB, the simulations will be done with the source generating discrete fixed length blocks (e.g., voice coders emit code frames).

When there is data in the LB source buffer, since tokens are arriving at the rate ρ, the buffer is depleted at the rate ρ. If the source buffer level exceeds B_s, and since the source would not lose its own data, we will view this as a QoS violation; i.e., B_s *does not represent a memory limitation, but a delay bound of* $\frac{B_s}{\rho}$. Thus our view is that the source buffer "behind" the LB is infinite but the buffer level exceeds B_s with a small probability; QoS Violation Probability (QVP). This idea will be formally developed in Section III.

With reference to Figure 3, define $X(t) = X_1(t) - X_2(t) +$

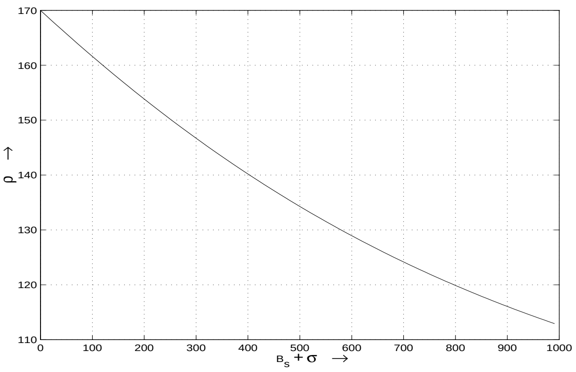
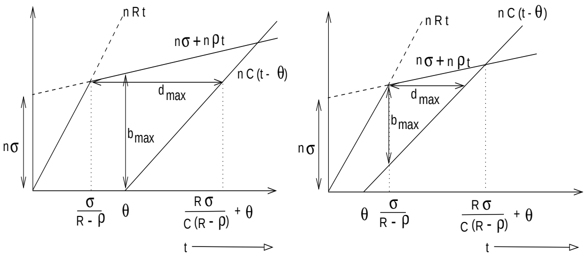


Fig. 5. A typical ρ vs. $B = (B_s + \sigma)$ trade-off curve; i.e., the function $h_{p_s}(B)$. Two state Markovian on-off fluid source with mean on-time $5/3$, mean off time $5/2$, and peak rate $R = 170$; $p_s = 10^{-3}$.



(a) $\theta \geq \frac{\sigma}{R-\rho}$ (b) $\theta < \frac{\sigma}{R-\rho}$

Fig. 6. Upper bound on delay d_{max}, and on queue length b_{max}, for $n\ (\sigma, \rho, R)$ controlled processes when served by a latency-rate service element with rate nC and latency θ.

σ. Observe that $X_1(t)$ and $X_2(t)$ are never both positive, and with probability 1 at least one is positive. The QVP is then just $P(X > B_s + \sigma)$, where X is the stationary random variable for $\{X(t)\}$. For a fluid model, the queue shown in Figure 4 is equivalent to the leaky bucket shown in Figure 3 from the QVP point of view (see also [8], [4]). Thus for a given ρ, the QVP depends on σ and B_s only through their sum. We will use the notation p_s to denote $P(X > B_s + \sigma)$; the subscript s is mnemonic for "source". Writing $B := B_s + \sigma$, for fixed p_s we denote the B vs ρ tradeoff function by $g_{p_s}(\rho) = B$. Let $h_{p_s}(\cdot)$ denote the inverse of $g_{p_s}(\cdot)$ (an example of $h_{p_s}(\cdot)$ is in Figure 5).

III. The End-to-End Delay QoS Requirement

We recall the end-to-end model of the LSP (or VP) from Figure 2. It is assumed that the n sources are statistically equivalent, that they are peak rate controlled to R, and are shaped by the LBs with the same shaper parameters ρ and σ. Define L_{max} as the maximum packet length from the packetisers. Each link is shared by other sessions apart from the above considered stream sessions; define V_{max} as the maximum packet length over all links and over all sessions. Since all sessions will carry best-effort traffic as well, typically V_{max} will be the maximum TCP segment size; e.g., 1500 bytes.

We have assumed that WFQ (or PGPS [9]) scheduling is used at each link. It is then well known that from the point of view of the LSP, each link can be modeled as a *latency rate server* (see [10], and [11]). Since the WFQ weights at each link are chosen such the minimum service rate for the traffic in the LSP is nC, hop $i, 0 \leq i \leq K$ has a minimum service curve with rate nC and latency θ_i. It is easily seen from Figure 2 that

$$\theta_i = \begin{cases} \frac{L_{max}}{\rho} + \frac{V_{max}}{r_i} & \text{for } i = 0 \\ \frac{L_{max}}{nC} + \frac{V_{max}}{r_i} & \text{for } i \in \{1, 2, \ldots, K-1\} \\ \frac{L_{max}}{nC} & \text{for } i = K \end{cases} \qquad (1)$$

Note that hop 0 includes the packetiser.

It is also well known that the tandem of latency rate servers above, is equivalent to a latency rate server with rate nC and latency $\theta = \sum_{i=0}^{K} \theta_i$. The (worst case) envelope of the superposition of the n LB controlled processes is given by

$$E(t) = \min\{nRt, n\sigma + n\rho t\} \qquad (t \geq 0) \qquad (2)$$

We seek a bound on the network delay after the source shaper.

If an arrival process with envelope $E(t)$ is served by a network element with minimum service curve $G(t)$, the upper bound on delay is given by (see [11])

$$d_{max}(E, G) := \inf\{\delta \geq 0 : G(u) - E(u - \delta) \geq 0,\ \forall u \geq 0\} \qquad (3)$$

For the above envelope $E(t)$, and a latency-rate server of latency θ and rate nC, it is easy to verify (see Figure 6) that *there is a finite delay bound only if* $C \geq \rho$, and then the upper bound on the packetiser and network delay is given by

$$d_{network} \leq \frac{n\sigma}{nC}\left(\frac{nR - nC}{nR - n\rho}\right) + \theta$$

i.e.,

$$d_{network} \leq \frac{\sigma}{C}\left(\frac{R - C}{R - \rho}\right) + \sum_{i=0}^{K} \theta_i$$

Substituting the values of θ_i

$$d_{network} \leq \frac{\sigma}{C}\left(\frac{R - C}{R - \rho}\right) + \frac{L_{max}}{\rho} + \frac{V_{\max}}{r_0} + \sum_{i=1}^{K-1}\left(\frac{L_{max}}{nC} + \frac{V_{\max}}{r_i}\right) + \frac{L_{max}}{nC}$$

We now turn to the delay in the shaper. There is nonzero delay in the shaper only if fluid from the source arrives to find the LB source buffer nonempty. When there is fluid in the LB source buffer it is drained at the rate ρ. With reference to Figure 3, denote by X_1 the stationary random variable for the amount of fluid in the LB source buffer. It follows that the delay in the LB is bounded by $\frac{X_1}{\rho}$. Note that this bound is a random variable.

Thus the end-to-end delay is bounded by

$$\frac{X_1}{\rho} + \frac{L_{max}}{\rho} + \frac{\sigma}{C}\left(\frac{R - C}{R - \rho}\right) + K\left(\frac{L_{max}}{nC}\right) + \sum_{i=0}^{K-1} \frac{V_{max}}{r_i} \qquad (4)$$

Define,

$$H := K\left(\frac{L_{max}}{n}\right)$$

Then from Equation 4 the end-to-end delay bound is written as:

$$\frac{X_1}{\rho} + \frac{L_{max}}{\rho} + \frac{\sigma}{C}\left(\frac{R-C}{R-\rho}\right) + \frac{H}{C} + \sum_{i=0}^{K-1}\frac{V_{max}}{r_i} \quad (5)$$

For a given path *the propagation delay is a constant*. So subtract it from the given maximum allowed end-to-end delay and define the resultant value as T_{e-e}. Now $\frac{V_{max}}{r_i}$ is a constant term in Equation 4 (i.e., it does not depend on the design parameters ρ, C, σ etc.), so we define

$$T := T_{e-e} - \sum_{i=0}^{K-1}\frac{V_{max}}{r_i}$$

Finally, based on the above delay bound we can write our QoS requirement as

$$Prob\left(\frac{X_1}{\rho} + \frac{L_{max}}{\rho} + \frac{\sigma}{C}\left(\frac{R-C}{R-\rho}\right) + \frac{H}{C} > T\right) \le q \quad (6)$$

where q is the QVP. Since we have been working with delay bounds, this is a conservative representation of our original QoS requirement. We will use this as it leads to a tractable analysis.

We now recall the notation and concepts introduced in Section II. Notice that, for some given Δ, such that $T > \Delta$, the requirement $Prob(\frac{X_1}{\rho} + \Delta > T) \le q$ is equivalent to $Prob(X > (T-\Delta)\rho + \sigma) \le q$ (here the random variable X is as defined in the last paragraph of Section II). Recalling that $B = B_s + \sigma$, this requirement is satisfied by taking $g_q(\rho) = B$, with $B \le (T-\Delta)\rho + \sigma$. Hence the QoS specification in Equation 6 is met if we require that

$$\begin{aligned} g_q(\rho) &= B \quad \text{and} \\ B &\le \left(T - \left(\frac{L_{max}}{\rho} + \frac{\sigma}{C}\left(\frac{R-C}{R-\rho}\right) + \frac{H}{C}\right)\right)\rho + \sigma \end{aligned} \quad (7)$$

This finally will be the QoS requirement that we will work with.

IV. Shaping for Minimum Bandwidth Allocation with Lossless Multiplexing

We now formulate the problem of obtaining the LB parameters that minimise the per source bandwidth C while meeting the above developed QoS requirement without loss in the network. Noting that, for lossless multiplexing, it is necessary that $C \ge \rho$ (see Section III), we consider the following optimisation problem.

Optimisation Problem P1:

$$\min_{(B,\sigma,\rho)} \quad C$$

subject to:

$$\begin{gathered} C \ge \rho \\ g_q(\rho) = B \\ B \le \left(T - \left(\frac{L_{max}}{\rho} + \frac{\sigma}{C}\left(\frac{R-C}{R-\rho}\right) + \frac{H}{C}\right)\right)\rho + \sigma \\ \rho \ge 0, \text{ and } 0 \le \sigma \le B \end{gathered}$$

Problem P1 can be rewritten in a more convenient form. The inequality in Equation 7 can be rearranged to obtain

$$C\left[1 + \left(T - \frac{B-\sigma+L_{max}}{\rho}\right)\left(\frac{R-\rho}{\sigma}\right)\right] \ge R + H\left(\frac{R-\rho}{\sigma}\right)$$

So for $T \ge \left(\frac{B-\sigma+L_{max}}{\rho} = \frac{B_s+L_{max}}{\rho}\right)$ and $R \ge \rho$, we obtain the following lower bound on C,

$$C \ge \frac{R + \left(\frac{R-\rho}{\sigma}\right)H}{1 + \left(T - \frac{B-\sigma+L_{max}}{\rho}\right)\left(\frac{R-\rho}{\sigma}\right)}$$

Note that the condition $T \ge \frac{B_s+L_{max}}{\rho}$ is necessary; it comes from the constraint on delay in the shaper buffer and in the packetizer. Consider now the following optimisation problem.

Optimisation Problem P2:

$$\min_{(B,\sigma,\rho)} \; C = \max\left\{\rho,\ \frac{R + \left(\frac{R-\rho}{\sigma}\right)H}{1 + \left(T - \frac{B-\sigma+L_{max}}{\rho}\right)\left(\frac{R-\rho}{\sigma}\right)}\right\} \quad (8)$$

subject to:

$$\begin{gathered} \frac{B-\sigma+L_{max}}{\rho} \le T \\ B = g_q(\rho) \\ \rho \ge 0, \quad \text{and} \quad 0 \le \sigma \le B \end{gathered}$$

Lemma IV.1: The optimal values of problems P1 and P2 are the same.

Proof: Let C_1 and C_2 be the optimal values of P1 and P2; let $\sigma_i, \rho_i, B_i, i \in 1,2$ denote the corresponding optimising variables. Note that σ_1, ρ_1, B_1 are feasible for Problem P2. Observe from Problem P1 that $C_1 = \max\left\{\rho_1,\ \frac{R+\left(\frac{R-\rho_1}{\sigma_1}\right)H}{1+\left(T-\frac{B_1-\sigma_1+L_{max}}{\rho_1}\right)\left(\frac{R-\rho_1}{\sigma_1}\right)}\right\}$. Hence $C_1 \ge C_2$. It is also easily seen that $C_2, \sigma_2, \rho_2, B_2$ are feasible for Problem P1, and hence $C_2 \ge C_1$. Hence $C_1 = C_2$, as was required to be proved. □

The solution to Problem P2 is provided by the following theorem.

Theorem IV.1: For $g_q(\rho)$ a convex and decreasing function and $\frac{H+L_{max}}{R} < T$, the optimal value of the Problem P2 is the unique ρ^* that solves the equation

$$T\rho - H - L_{max} = g_q(\rho) \quad (9)$$

Further, the optimal $B^* = T\rho^* - H - L_{max}$, and $0 \le \sigma \le B^*$.

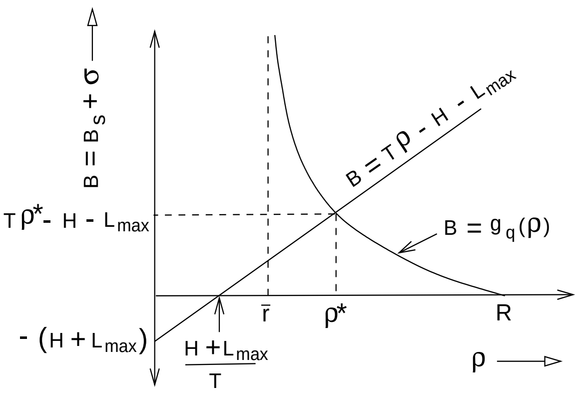


Fig. 7. Characterization of the optimal token rate ρ^*. $\bar{r}$ is the mean rate of the source.

Proof : See Appendix I. The geometry of the solution is depicted in Figure 7. Note that while the solution precisely fixes B^*, the value of σ is determined only to an interval; this is because, as discussed in Section II, the probability of the LB exceeding a buffer level depends on σ and B_s only through their sum B. □

Discussion of Theorem IV.1 : The only assumption we have used is that the function $g_q(\rho)$ is convex and decreasing. Observe that the optimal token rate ρ^* depends on the source process, the QVP q, the maximum allowed end-to-end delay T_{e-e}, the fixed propagation delay, the number of hops K in the path, the number of sessions n sharing that path, the total capacity of each server on that path (i.e., $r_i, 0 \leq i \leq K-1$), V_{max} and L_{max}. In practice, the network path parameters (K, V_{max}, L_{max}, and the r_is) can be obtained at the source when RSVP-TE sets up the LSP. The number of sessions sharing the path, n, can be obtained if several statistically identical raw voice sources, generated from analog phones, are being shaped and multiplexed at a VoIP PBX. Thus there is a possibility that ρ^* can be determined by the source in real-time by making measurements. We have proposed and studied an approach for measurement based estimation of the optimal token rate in Section V.

Observe from Figure 7 that if $\frac{H+L_{max}}{R} > T$, i.e., $\frac{H+L_{max}}{T} > R$, then this problem does not have a solution. Note that as the value of T_{e-e} or total server capacity r_i, $0 \leq i \leq K-1$, decreases the value of T and hence the slope of $B = T\rho - H - L_{max}$ decreases. This makes the value of ρ^* increase (see Figure 7). The increase in the number of hops or maximum packet length L_{max}, increases the value H and so the value of ρ^*. If the number of sessions being multiplexed is increased then the value of ρ^* decreases because of decrease in the value of H. Also observe that for optimal value of per source capacity $C = \rho^*$, if the source is shaped by (σ, ρ^*), $0 \leq \sigma \leq T\rho^* - H - L_{max}$, to ensure a lossless service to the shaped source at the multiplexer, we must use a buffer corresponding to b_{max} in Figure 6 (with $\theta = \theta_0 = \frac{L_{max}}{\rho} + \frac{V_{max}}{r_0}$).

The function $\mathbf{g_q}(\cdot)$: With reference to Figure 4, an important approximation approach to determine the service rate ρ, so that the overflow probability $Prob(X > B) < q$, is to use the asymptotic approximation developed in [12]. Such an approach would be particularly applicable, for example, to VBR voice sources for which an on-off Markov model with exponential state sojourn-time is a standard model. Details about an analytical approach to estimate the value ρ^* are given in Section VI. Here if we write the negative of the slope of the tail of the $\ln(P(X > B))$ vs B curve as $\eta(\rho)$, then the approach [12] is to design the shaper by taking

$$\eta(\rho) = \frac{-\ln q}{B} \tag{10}$$

Hence with this approach, we have

$$g_q(\rho) = \frac{-\ln q}{\eta(\rho)} \tag{11}$$

Lemma IV.2: If the source process is a Markov modulated fluid process, then the function $g_q(\rho)$, as defined in Equation 11, is convex and decreasing.

Proof: The proof based on results in [12] is provided in [13]. □

Minimum Cost Lossless Multiplexing: If n statistically identical sources, each shaped according to (σ,ρ,R) are being multiplexed in a lossless manner in the LSP path shown in Figure 2, then for a per source bandwidth allocation of C, the total network buffering required is given by (see Figure 6; $\theta \geq 0$ is as defined in Section III):

$$b_{max} = \begin{cases} \sigma + \rho\,\theta & \text{if } \theta \geq \frac{\sigma}{R-\rho} \\ \frac{(R-C)\sigma}{(R-\rho)} + C\,\theta & \text{if } \theta < \frac{\sigma}{R-\rho} \end{cases}$$

For lossless multiplexing $C \geq \rho$, hence observe that $b_{max} \geq \rho\,\theta$. Note also that, for any given values of ρ and C, b_{max} is minimised by taking $\sigma = 0$. Thus if we take $C = \rho = \rho^*$, then $\sigma = 0$ minimises the required buffering in the network, i.e., $b_{max} = \rho^*\,\theta$. Also, for the case $L_{max} = 0$ (fluid model for traffic in the network), we can consider the linear buffer-bandwidth cost function $\gamma C + \beta b_{max}$ (where γ and β are the per unit cost of capacity and buffers, respectively; $\gamma > 0$ and $\beta > 0$) over all choices of the shaper parameters, and under the QoS constraint: end-to-end delay $\leq T_{e-e}$ with QVP $= q$. It is then easy to infer that the cost function is minimized for leaky bucket parameters $\sigma = 0$ and $\rho = \rho^*$. Note that for a fluid source we interpret $\sigma = 0$ to mean that all fluid arrival from a source queues up at its shaper buffer, and is served at the rate ρ (i.e., as the fluid "tokens" arrive). In practice, with a packetized source, σ needs to be at least the maximum packet size.

V. Measurement Based Estimation of ρ^*

We use the *Robbins–Monro (RM)* stochastic approximation algorithm to obtain the optimal value of ρ, i.e., ρ^* (see [14]). The RM algorithm addresses the problem of finding the root of a function when we can only observe the function values corrupted by noise. It is an iterative algorithm that uses noisy measurements of the function for given values of the argument, and iteratively obtains an estimate of the root.

Consider a function $f(\rho)$, whose root ρ^* needs to be found. Suppose that, given the argument ρ we can observe $f(\rho) + v$ where v is measurement noise. In the RM algorithm, at the kth iteration, the current estimate ρ_k is updated as follows

$$\rho_{k+1} = \rho_k - a_k(f(\rho_k) + v_{k+1}) \tag{12}$$

where $\{a_k\}$ is a "gain" sequence. For a suitably nice function $f(\cdot)$, sufficient conditions for the convergence of the RM algorithm are [14]: (i) The gain coefficient sequence a_k should be such that $\sum_{k=0}^{\infty} a_k = \infty$ and $\sum_{k=0}^{\infty} a_k^2 < \infty$; (ii) conditions on noise: for all $k \geq 0$ $E(V_{k+1}|\rho_0, (v_i, \rho_i), 1 \leq i \leq k) = 0$ and $E(V_{k+1}^2|\rho_0, (v_i, \rho_i), 1 \leq i \leq k) < s^2$, for some s^2 finite. In the model of Figure 4, with X the stationary queue length, define $p(\rho, B) = P(X > B)$. Then we define $d(\rho, B)$ as

$$d(\rho, B) = -\ln p(\rho, B) + \ln q \tag{13}$$

where q is the desired QVP. Then, recalling Theorem IV.1, our problem is to find the root ρ^* of the function $f(\rho) = d(\rho, T\rho - H - L_{max})$. An update interval is chosen (we study the effect of choices of this interval in Section VII), $p(\rho_k, T\rho_k - H - L_{max})$ is measured in the kth interval (details about measuring the loss probability are given below), and then a new value ρ_{k+1} is computed according to the RM algorithm in Equation 12. In the RM algorithm we have found it useful to take the gain sequence to be of the form

$$a_k = \frac{R}{(k+J)(-\ln q)D} \tag{14}$$

with J an integer, and D a real number. J and D can be used to control the transient behaviour and the convergence of the algorithm. Also, R (the peak rate) and $-ln(q)$ are used to scale the gain properly. It is easy to verify that $\sum_{k=0}^{\infty} a_k = \infty$ and $\sum_{k=0}^{\infty} a_k^2 < \infty$. The conditions on the noise sequence hold approximately. The first condition requires the measurement to be conditionally unbiased. It can be argued that if we obtain the estimate of packet loss using the Virtual Buffer technique that we will describe below, the measurements are asymptotically unbiased as the measurement interval becomes large. Also, we are making a heuristic modification to take care of unbounded values. Whenever we encounter a zero loss (leading to an unbounded function value), the function value is artificially bounded (e.g., by taking the loss to be a small nonzero value a few orders of magnitude smaller than q). This modification ensures that the RM algorithm steps are executed only on bounded values of the function. Hence the conditional second moment of the observations *used* by the RM algorithm is bounded.

Measuring $p(\rho_k, T\rho_k - H - L_{max})$: Since the target QVP of interest can be very small, we need to use special techniques to measure $p(\rho_k, T\rho_k - H - L_{max})$, a rare event probability. We have used a virtual buffer approach based on large deviation asymptotics (see [15]).

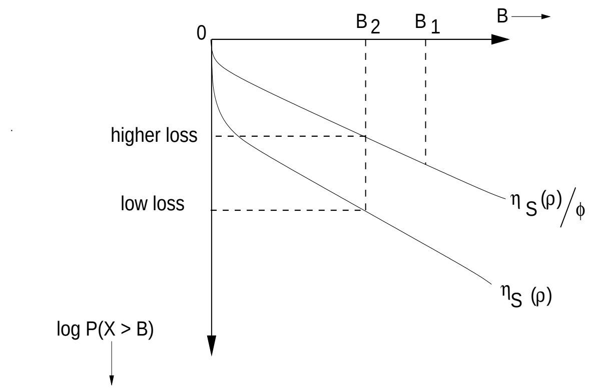


Fig. 8. Scaling the arrival process and the service rate by ϕ scales the asymptotic slope of $\ln P(X > B)$ by $1/\phi$, thus increasing the probability of exceedance of a buffer level.

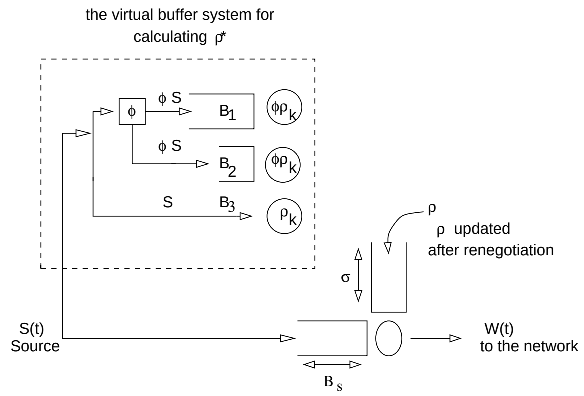


Fig. 9. A virtual buffer system for estimating ρ^*. Three virtual buffers at each source $S(t)$, are used to obtain an estimate of $-\ln p(\rho_k, T\rho_k - H - L_{max})$, which is used for finding the next iterate of ρ^* using the RM algorithm. The actual LB parameters are updated only periodically after renegotiation.

We use an affine approximation for $\ln p(\rho, B)$ (see [16] and [17]). Writing $\eta_S(\rho)$ as the negative of the asymptotic slope of the $\ln p(\rho, B)$ vs. B curve (the subscript S denotes that fact that the source $S(t)$ feeds the buffer), we approximate

$$\ln p(\rho, B) \approx \ln P(S > \rho) - \eta_S(\rho)B$$

where the random variable S is the marginal of the source rate process $\{S(t)\}$. Also, it is easy to see that if the source is scaled by a positive multiplier ϕ (i.e., each arrival actually brings ϕ arrivals)

$$\eta_{\phi S}(\phi\rho) = \frac{\eta_S(\rho)}{\phi},$$

i.e., scaling the source and the service rate results in an asymptotic slope that is scaled by $1/\phi$ (see [18]). The usefulness of this for measurement of small overflow probabilities is shown in Figure 8. For example, with $\phi = 4$ an overflow probability of 10^{-5} becomes roughly $10^{\frac{-5}{4}}$, thus making a rare event relatively frequent. With this we can write the approximation as

$$\ln p(\rho, B) \approx \ln P(S > \rho) - \phi\eta_{\phi S}(\phi\rho)B$$

Hence if $\ln P(S > \rho)$ and $\eta_{\phi S}(\phi\rho)$ can be estimated, we can then approximate

$$\ln p(\rho, T\rho - H - L_{max}) \approx \ln P(S > \rho) - \phi\eta_{\phi S}(\phi\rho)(T\rho - H - L_{max}) \quad (15)$$

Virtual buffers in the source can now be used to measure the terms $\ln P(S > \rho)$ and $\eta_{\phi S}(\phi\rho)$, using the arrangement shown in Figure 9. The scaled source is fed to two virtual buffers (B_1 and B_2) that are served by $\phi\rho_k$; since the approximation for $\ln p(\rho, B)$ has a linear form with slope $\phi\eta_{\phi S}(\phi\rho)$, this yields an estimate of $\eta_{\phi S}(\phi\rho_k)$. The bufferless component B_3 (= 1 for a discrete source) yields an estimate of $\ln P(S > \rho_k)$. Equation 15 is then used to get a measurement of $\ln p(\rho_k, T\rho_k - H - L_{max})$. Details of the approach are available in a technical report [19] by the authors.

VI. An Analytical Estimate of ρ^* for a 2-State MMFP

In order to verify the measurement based approach described in Section V it is important to be able to analytically estimate ρ^* in some cases. While the measurement based approach is more general, here we use standard asymptotic techniques for estimating ρ^* in the case of a Markov Modulated Fluid Process (MMFP).

With reference to the notation introduced in Section V we need to solve the equation

$$\ln p(\rho, T\rho - H - L_{max}) = \ln q$$

for $\rho = \rho^*$. We will use a linear approximation for $\ln p(\rho, B)$, i.e.,

$$\ln p(\rho, B) = -\eta_s(\rho)B$$

where $\eta_s(\rho)$ is the negative of the asymptotic slope of the $\ln P(X > B)$ vs. B curve, where X is the stationary queue length in the model of Figure 4. With this approximation we need ρ^* that solves

$$-\eta_s(\rho) = \frac{\ln q}{T\rho - H - L_{max}} \quad (16)$$

Since the linear approximation of $\ln p(\rho, B)$ ignores the small buffer behaviour, this approach will overestimate ρ^*.

For a Markov Modulated Fluid Process (MMFP) feeding a queue with a constant rate server, the functional relation between the asymptotic slope, z, of $\ln P(X > B)$, and the service rate ρ is well known; see [12]. Suppose the MMFP is characterized by $(M, \vec{\lambda})$ where M is the irreducible generator matrix (transition rate matrix) of the controlling Markov chain. The source with state space $\mathcal{S}$ generates fluid at the constant rate λ_s when in state $s \in \mathcal{S}$; let $\vec{\lambda} = \left(\lambda_1, \lambda_2, \ldots, \lambda_{|\mathcal{S}|}\right)$. Define the matrix $\Lambda = \mathrm{diag}(\vec{\lambda})$. Then for given z, the corresponding ρ is the maximum real eigenvalue of the matrix $\left[\Lambda - \frac{1}{z}M\right]$.

Let the generator matrix M for a two state MMFP source be given by

$$M = \begin{bmatrix} -\alpha & \alpha \\ \beta & -\beta \end{bmatrix}$$

and $\vec{\lambda} = [\lambda_1 \quad \lambda_2]$. The eigenvalues of the matrix $\left[\Lambda - \frac{1}{z}M\right]$ are solutions of the following equation:

$$\rho^2 z^2 - \rho\left\{J_1 z^2 + J_2 z\right\} + J_3 z^2 + J_4 z = 0 \quad (17)$$

where,

$$J_1 = \lambda_1 + \lambda_2, \quad J_2 = \alpha + \beta, \quad J_3 = \lambda_1\lambda_2, \quad J_4 = \alpha\lambda_2 + \beta\lambda_1$$

Returning to the problem of estimating ρ^*, from Equation 16 it is clear that we need

$$z = \frac{\ln q}{T\rho - H - L_{max}} = \frac{k}{\rho - l} \quad (18)$$

where $k = \frac{\ln q}{T}$ and $l = \frac{H + L_{max}}{T}$. Substituting this value of z in Equation 17 we get the following quadratic equation in ρ:

$$\rho^2(k - J_2) - \rho(J_1 k - J_2 l - J_4) + (J_3 k - J_4 l) = 0 \quad (19)$$

We obtain the value of ρ^* from the above equation using the following theorem.

Theorem VI.1: The desired value of optimal token rate ρ^* is the larger of the two roots of Equation 19.

Proof : See Appendix II. □

Now in Equation 18, $k < 0$. So ρ^* is given by the following larger of the two roots of Equation 19.

$$\rho^* = \frac{(J_4 - J_1 k + J_2 l)}{2(J_2 - k)} + \frac{\sqrt{(J_4 - J_1 k + J_2 l)^2 + 4(J_2 - k)(J_3 k - J_4 l)}}{2(J_2 - k)}$$

For an on-off MMFP by setting $\lambda_1 = 0$ and $\lambda_2 = R$ (peak rate) in the above equation,

$$J_1 = R, \quad J_2 = \alpha + \beta, \quad J_3 = 0, \quad J_4 = \alpha R$$

and we get

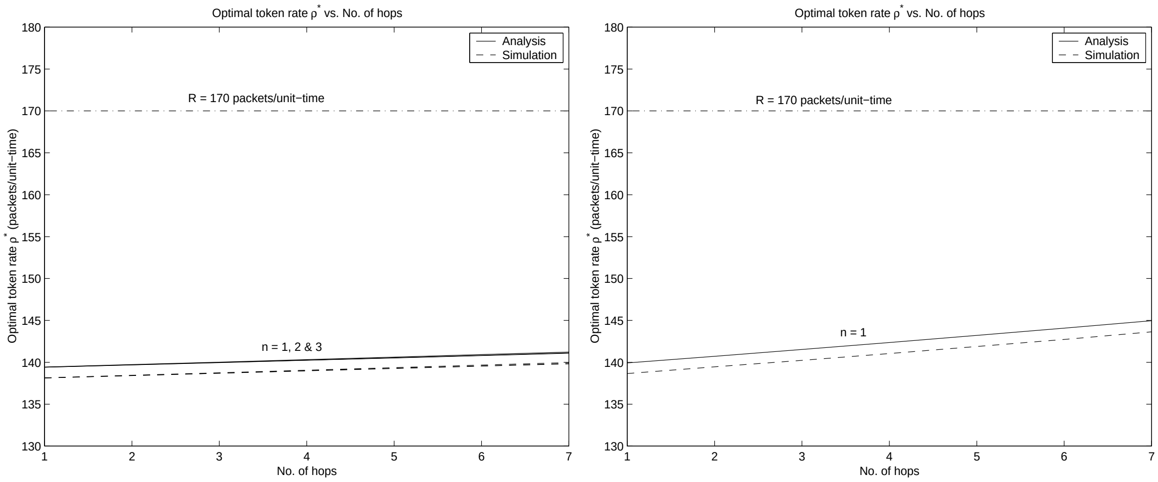
$$\rho^* = \frac{R(\alpha - k) + l(\alpha + \beta)}{2(\alpha + \beta - k)} + \frac{\sqrt{\left[R(\alpha - k) - l(\alpha + \beta)\right]^2 - 4Rkl\beta}}{2(\alpha + \beta - k)} \quad (20)$$

where $k = \frac{\ln q}{T}$ and $l = \frac{H + L_{max}}{T}$.

VII. Simulation Results

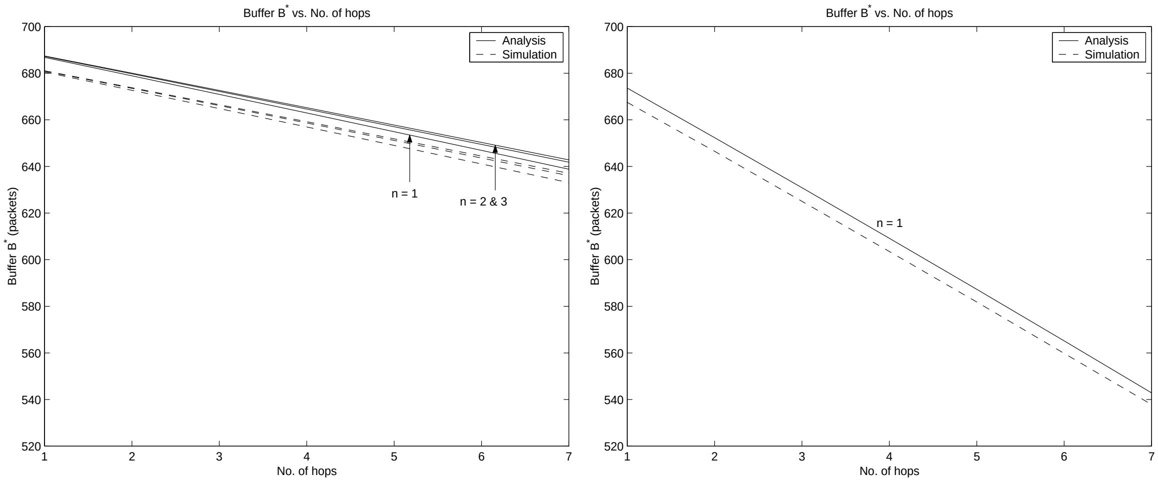
A. Results for a 2-State On-Off Source

We first consider the on-off source used in [20], [13]. The process has a mean on-time of $\frac{5}{3}$ time units and a mean off-time of $\frac{5}{2}$ time units. The other parameters are: peak rate R = 170 packets/unit time, delay bound T_{e-e} = 5 time units, and QVP $q = 10^{-5}$. If we take the unit of time to be 10ms, and 48 bytes of payload per packet (ATM cell), then these parameters will correspond to a mean on-time of 16.67ms, mean off-time of 25ms, peak rate of about 6.5 Mbps, mean rate of 2.6 Mbps, and delay constraint of 50ms.



(a) (b)

Fig. 10. Optimal token rate ρ^* vs the number of hops curve for link capacity of (a) 200 Kbits/unit_time and (b) 70 Kbits/unit_time. For three different values of number of sessions the curves in (a) are overlapping.



(a) (b)

Fig. 11. The buffer B^* vs the number of hops curve for link capacity of (a) 200 Kbits/unit_time and (b) 70 Kbits/unit_time.

For the estimation of ρ^* the gain sequence parameters (see Equation 14) used are: $J = 3$ and $D = 4$. The virtual buffer values $B_1 = \frac{2B}{3}$ and $B_2 = \frac{B}{2}$ are used along with scaling factor $\phi = 4$.

The total capacity of each link (i.e., r_i) is 200 Kbits/unit time. So for unit time of 10ms the link rate is 20 Mbps. We consider a K hop path in a network of PGPS servers. We have assumed $V_{\max}$ corresponding to a TCP packet of maximum size 1500 bytes. With $r_i = 200$ Kbits/unit time this gives an additional delay of 0.06 time units. Thus for unit time of 10ms, PGPS service system introduces an additional delay of 0.6ms. Figure 10(a) and (b) show the values of ρ^* obtained from simulation and from the approximate analysis, for link capacities 200Kbits/unit time and 70Kbits/unit time. Results are shown for the number of sessions $n = 1, 2$, and 3 for 200Kbits/unit time; only $n = 1$ is possible with 70Kbits/unit time.

Observe that the value of ρ^* increases with the increase in number of hops but it does not change much with the number of sessions. The analysis results and the simulation results for three different values of number of sessions are almost similar and so the three different graphs are overlapping in Figure 10(a). The primary effect of change in number of hops comes through the addition of $\frac{V_{\max}}{r_i}$ terms in Equation 4. This term takes care of the additional delay introduced by PGPS servers and it adds up as the number of hops traversed by a session is increased. So the end-to-end delay budget T decreases and the value of ρ^* increases (see Figure 7). The effect of change in number of hops or change in number of sessions through the H term is negligible. In Figure 10(a) the increase in ρ^* with increase in the number of hops is not very significant because of the high link capacity of 200Kbps/unit time (20 Mbps). As shown in the Figure 10(b) with the increase in the number of hops the value of ρ^* increases faster for a smaller link capacity of 70Kbits/unit time (7Mbps).

Note that the analytically obtained values and those obtained from the simulation are quite close. As expected, the analysis over-estimates the value of ρ^*. For the parameters chosen for these experiments, for 200Kbits/unit-time, for three sessions and a five hop path the optimal LB rate, ρ^* is about 140 packets/unit time, i.e., almost 18% saving in the bandwidth requirement compared to the peak rate of 170 packets/unit time.

The value of buffer B^* decreases as the number of hops K increases (see Figure 11). For a given number of hops if the number of sessions is increased then because $B^* = T\rho^* - H - L_{max}$, even though the value of ρ^* remains almost constant, the decrease in the H term increases the value of B^* slightly, and this effect is visible in Figure 11(a). Since the value of ρ^* increases with K more rapidly for a smaller capacity link, the value of B^* decreases more rapidly for a smaller capacity link as shown in Figure 11(b).

B. Results for Packet Voice

A single voice source with silence removal is well represented by a two-state process. Telephony speech consists of an alternating sequence of active, or *talk-spurt*, intervals, typically averaging 0.4-1.2 second in length, followed by silence (inactive) intervals averaging 0.6-1.8 second in length. To a reasonably good approximation, the sojourn times in the two states may be assumed to be exponentially distributed. This gives rise to a two state Markov modulated source process, with the source emitting data at the peak rate of R packets/second while in talk-spurt and no data generation while in silence.

We take a packetization time of 20ms. So the voice source generates a periodic stream of packets at the peak rate $R = 50$ packets/second while in talk-spurt. The speech parameters which we have used for our analysis and simulation purpose are the following (see [21], [22], [23], [24]). The mean talk spurt length α = 400ms and mean silence length β = 600ms. Thus the average voice activity factor is 40%. Usually the voice activity factor lies between 35% to 48%. PCM coding with sampling rate of 8KHz and 8 bits/sample gives a bit-rate of 64 Kbps. Usually a packet loss fraction up to 0.1%- 1% is found to be acceptable. Instead of dropping whole packets if only selected bits of the packets are dropped then a loss fraction up to 10% is also found to be acceptable (see [22]).

B.1 Lossless Multiplexing for Voice Packets

With T_{e-e}= 100ms, r_i = 2Mbps, $V_{\max}$ = 1500 bytes (corresponds to the usual maximum TCP packet size), $q = 1\%$, α = 400ms, β = 600ms, and the peak rate R = 50 packets/second, we show the optimal token rate ρ^* vs number of hops, for different values of the number of sessions ($n = 1, 10, 30$) in Figure 12(a). For the estimation of ρ^* the gain sequence parameters used are: $J = 6$ and $D = 2$. Here since we are operating in the small buffer region we do not need the arrangement shown in Fig 9; we can directly measure the loss probability. Observe that the value of ρ^* is only slightly smaller than the peak rate of the source. As shown in Figure 13(a) the maximum buffer requirement for lossless multiplexing with T_{e-e} = 100ms is also very small, i.e., 3.5 packets. With lossless multiplexing, there is not much saving in bandwidth in this case as the value of T_{e-e} is much smaller than the burst duration.

Observe from Figure 12(a) and Figure 13(a) that for number of sessions n = 1 and end-to-end delay requirement of 100ms the source can traverse at most two hops. For the number of hops $K > 2$, $\frac{H+L_{max}}{T} > R$ (see Figure 7), i.e., the end-to-end delay bound cannot be satisfied (at least using the worst case delay bounding approach used in this paper).

If we relax the maximum allowed end-to-end delay for the voice source, e.g., say 1 second for a streaming lecture transmission then, as shown in Figure 12(b), $\rho^* \approx 36$ packets/second compared to $R = 50$ packets/second, i.e., there is a 28% saving in the required network bandwidth for lossless multiplexing. As shown in the figure the value of ρ^* does not increase much with the increase in number of hops. Similarly for different values of number of sessions, except for the $n = 1$, the curves of ρ^* vs number of hops are overlapping. Similar kind of results are shown for buffer B^* vs number of hops in Figure 13(b), where the value of B^* is 28 to 35 packets as compared to 1.5 to 3.5 packets for an end-to-end delay of 100ms.

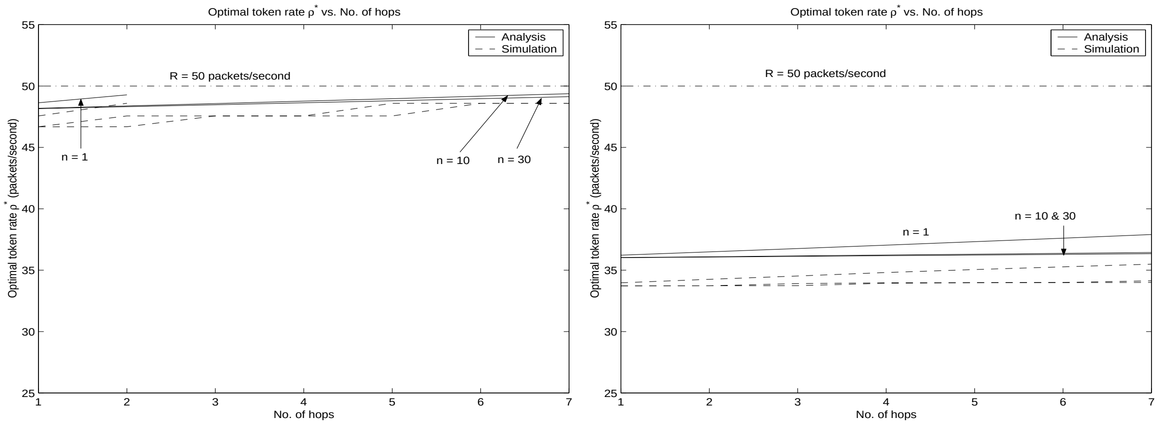
Effect of Measurement Period

The effect of measurement period on the convergence of the ρ^* estimation algorithm is shown in Figure 14. The plots are for $n = 20, K = 5$, with $T_{e-e} = 1$ sec; the algorithm gain parameters are $J = 6$ and $D = 2$. There are two columns of plots in the figure; the first column shows the iterates of ρ and the second column shows the measured values of the delay violation probability. As the update time period is increased the measurements are less noisy. We could obtain a reasonable convergence in 3-4 iterations for a measurement period of 10 seconds.

Remark on Renegotiation: With reference to the discussion in the Introduction, we can expect that a protocol such as RSVP-TE can be used to renegotiate the *aggregate rate* required by the sources as the measurement updates proceed and better estimates of ρ^* are obtained at each source. The soft-state of RSVP-TE provides the possibility that PATH messages can carry the aggregate rate requirements of the sources being multiplexed into the LSP, and thus will serve as renegotiation requests.

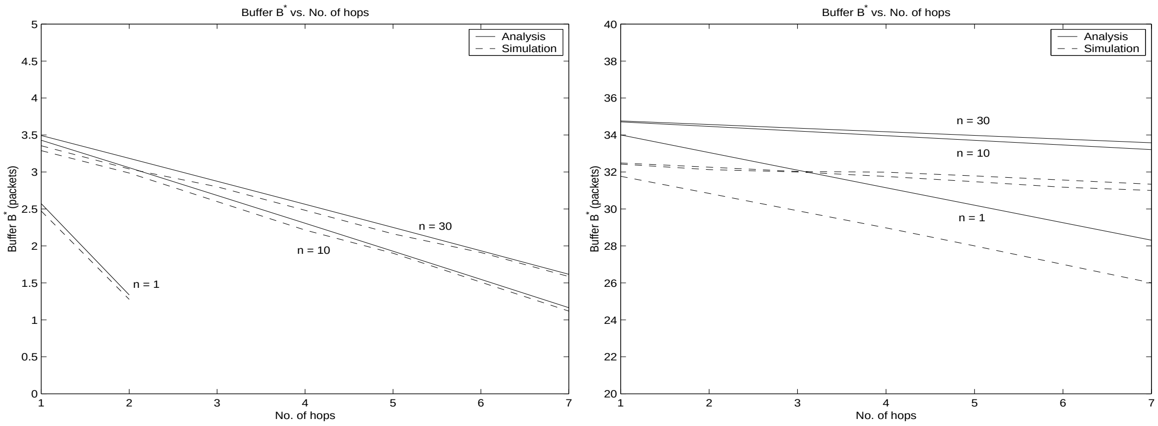
B.2 Lossy Multiplexing for Packet Voice

Motivated by the result in Section IV, it is reasonable for the source to use a token rate of ρ^* for each of the sources. For lossless multiplexing, and an end-to-end delay constraint of T_{e-e} with a QVP = q, the source can use any σ such that $0 \leq \sigma \leq T\rho^* - H - L_{max}$; then $B_s = T\rho^* - H - L_{max} - \sigma$, and the network sets the per source bandwidth to $C = \rho^*$. Thus this approach does not yield a unique value of σ (if the linear buffer-bandwidth cost function is used then, for a fluid model, $\sigma^* = 0$ minimises such cost, as explained at the end of Section IV). Clearly, a positive σ would facilitate statistical multiplexing, and *if a packet loss probability comparable to*



(a) (b)

Fig. 12. Optimal value of ρ^* vs number of hops for maximum allowed end-to-end delay of (a) 100ms and (b) 1 second, for an on-off voice source.



(a) (b)

Fig. 13. The buffer B^* vs number of hops for maximum allowed end-to-end delay of (a) 100ms and (b) 1 seconds for an on-off voice source.

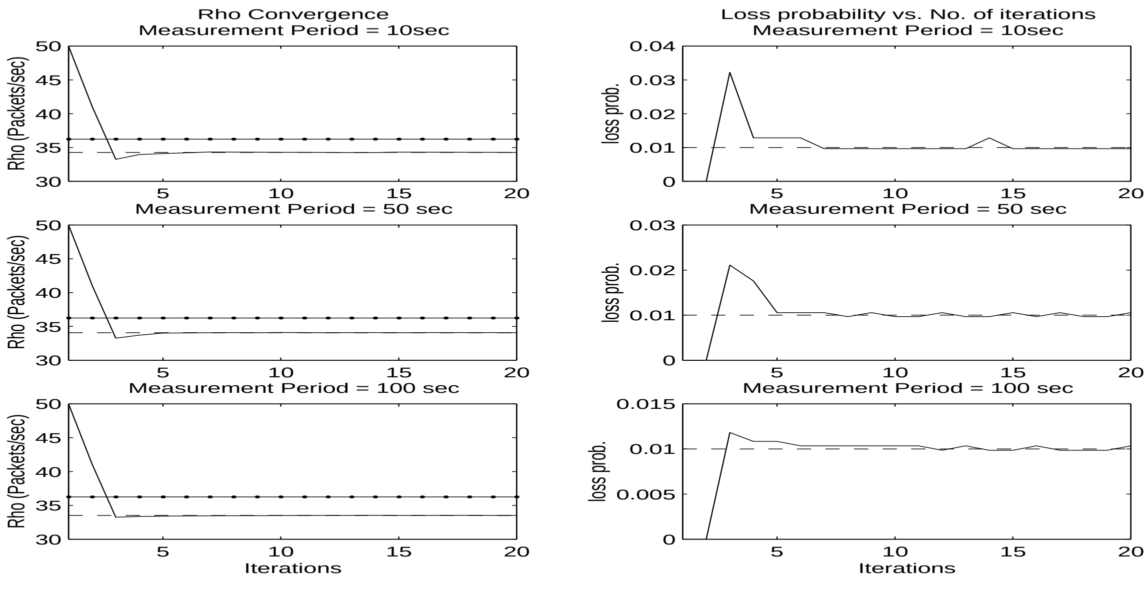


Fig. 14. Effect of measurement interval on the algorithm for estimating ρ^*. The plots on the left show the convergence of ρ_k (the flat solid line is the value of ρ^* computed from the analytical approximation), and the plots on the right show the measured QoS violation probability, the target being 1%. Each row of plots is for a different measurement period.

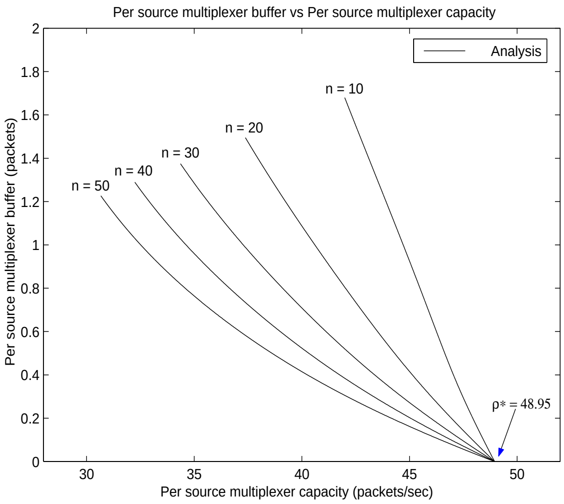


Fig. 15. Fluid voice sources, optimally shaped and fed to a multiplexer. Plot shows analytical results for multiplexer's per source buffer vs. per source capacity for a multiplexer loss probability of 1%. Plots are shown for various numbers of multiplexed sources. Along each curve the value of σ increases from 0 (bottom of the curve) to $T\rho^*$ (top of the curve).

the QVP[3] *is permitted then the network resource requirement could be reduced.* We denote the loss probability in the first network node by p_m.

First consider the single hop case, $K = 1$, and a fluid model; i.e., $H = 0, L_{max} = 0$ and we get the model of our earlier paper [20]. The input to the shaper is a two-state MMFP. When this kind of a fluid source is fed to the shaper, the output process is approximated by a three state MMFP (see [12]). To analytically estimate the value of the per source capacity required at the multiplexer, we use an asymptotic approximation with the three state MMFP as the input source; see [13]. We use this analysis without giving its details here, owing to space constraint. Since we are considering a fluid model we have $B^* = T\rho^*$. We take T_{e-e} = 40ms, $p_s = 1\%$, p_m = 1%, $\alpha = 400$ms, $\beta = 600$ms and peak rate $R = 50$ packets/second for voice packets. For the above voice parameters the effective bandwidth approach gives the per source capacity vs the per source buffer requirement curves as shown in Figure 15. These curves are obtained as follows. Note that for each value of $\sigma, 0 < \sigma < T\rho^*$, the network delay permitted for the source is $\frac{\sigma}{\rho^*}$. Hence, with σ fixed at a value in the range, the per source capacity $c(\sigma)$ is found such that probability of the multiplexer buffer exceeding $\frac{\sigma n c(\sigma)}{\rho^*}$ is p_m (see [20]).

We obtain the results for number of sessions n ranging from 10 to 50. In Figure 15, for a given value of n as the value of σ is increased from 0 to $T\rho^*$ the value of per source capacity requirement decreases from the value ρ^*, and the value of per source buffer requirement increases. The value of optimal token rate is ρ^*= 48.95 packets/second. For $\sigma = 0$ the required capacity is equal to ρ^*; thus there is not much bandwidth saving if lossless multiplexing is used. Denote the minimum value of the per source required capacity C, corresponding to $\sigma = T\rho^*$ by $C_{\min}$. Observe that, with lossy multiplexing, for the case of n = 50, $C_{\min}$ = 31 packets/second and B^* = 1.2 packets. Thus, there is 38% saving in the bandwidth requirement with a correspondingly very small buffer requirement. These results are indicative of the *additional bandwidth savings if optimal shaping is combined with lossy multiplexing*. With these curves, we can ask the question of *minimising a linear capacity-buffer cost function for lossy multiplexing*. For each value of n the optimal per source buffer and per source capacity will be found to lie on one of these curves; the corresponding value of $\sigma, 0 \leq \sigma \leq T\rho^*$ will be the optimal token bucket size, thus yielding the optimal LB parameters (ρ^*, σ^*); these ideas have been developed in [20], [13].

Finally, we report the results of a simulation study with packet voice, optimal shaping and lossy multiplexing. We consider the parameters: T_{e-e} = 100ms and 1sec, $p_s = 1\%$, p_m = 1%, $\alpha = 400$ms, $\beta = 600$ms and peak rate $R = 50$ packets/second for voice packets. We take the link capacity r_i equal to 2Mbps; there are $K = 5$ hops and $V_{max} = 1500$bytes. We first determine the value of ρ^*; this limits σ to $0 \leq \sigma \leq T\rho^* - H - L_{max}$. We set $\sigma = T\rho^* - H - L_{max}$, the max-

[3]Note that if we think of a packet that is delayed more than T as being equivalent to packet loss, then a packet loss ratio of q at the network node still yields a QVP of $1 - (1 - q)^2 \approx 2q \approx q$ for q small.

Peak Rate (pps)	n	ρ^* (pps)		B (pkts)		$C_{\min}$ (pps)			Buffer Bm (pkts)	
		anal	sim	anal	sim	$\sigma_{\max}$	anal	sim	anal	sim
50	5	49.22	48.56	1.44	1.40	1	46.47	45.47	6.52	4.36
	10	48.97	48.56	1.93	1.90	2	42.09	39.31	15.85	15.24
	15	48.88	48.10	2.09	2.03	2	38.96	36.41	23.94	21.49
	20	48.84	48.10	2.17	2.12	2	36.61	34.64	31.24	27.40

TABLE I

RESULTS FOR PACKETISED ON-OFF VOICE SOURCES: ANALYSIS AND SIMULATION RESULTS FOR OPTIMAL TOKEN RATE ρ^*, $B = B_s + \sigma$, AND PER SOURCE REQUIRED CAPACITY $C_{\min}$, AND BUFFER B_m, AT THE MULTIPLEXER WITH $\sigma = \sigma_{max}$. PARAMETERS: T_{e-e} = 100MS, $p_s = 1\%$, $p_m = 1\%$, K = 5 HOPS, $V_{max} = 1500$BYTES, AND C_{link} = 2 MBPS.

Peak Rate (pps)	n	ρ^* (pps)		B (pkts)		$C_{\min}$ (pps)			Buffer Bm (pkts)	
		anal	sim	anal	sim	$\sigma_{\max}$	anal	sim	anal	sim
50	5	36.41	34.54	33.32	31.50	31	25.77	23.08	116.46	101.92
	10	36.30	34.49	33.71	31.96	32	23.14	20.96	213.04	186.46
	15	36.27	34.36	33.84	32.00	32	22.15	20.24	308.02	280.62
	20	36.25	34.28	33.91	32.01	32	21.62	19.67	402.60	365.14

TABLE II

RESULTS FOR PACKETISED ON-OFF VOICE SOURCES: ANALYSIS AND SIMULATION RESULTS FOR OPTIMAL TOKEN RATE ρ^*, $B = B_s + \sigma$, AND PER SOURCE REQUIRED CAPACITY $C_{\min}$, AND BUFFER B_m, AT THE MULTIPLEXER WITH $\sigma = \sigma_{max}$. PARAMETERS: T_{e-e} = 1SEC, $p_s = 1\%$, $p_m = 1\%$, K = 5 HOPS, $V_{max} = 1500$BYTES, AND C_{link} = 2 MBPS.

imum value σ_{max}; this will result in the minimum possible value of C, the per source capacity. We then determine the value of C and B by analysis and simulation. The analytical results reported are from the approximations discussed above (C_{min} is obtained by using the 3-state Markov model for the LB output, as also discussed above in this section). The virtual buffer technique as developed earlier in the paper is used in the simulation. In this simulation we are feeding the LB by a packetized source process, and we have rounded off the value of σ to the nearest integer value. Hence, $\sigma_{\max}$ in the tables denotes the maximum possible integer value for σ, and correspondingly the minimum value of the per source capacity is denoted by C_{min}. The results obtained through simulations for different values of n are summarized in Tables I and II. Notice that with $T_{e-e} = 1$sec, we get a substantial reduction in ρ^* (ρ^* is about 49 with $T_{e-e} = 100$ms, but reduces to about 36 for $T_{e-e} = 1$sec), and also since larger values of σ can be used *the additional reduction in network capacity from statistical multiplexing* is also greater (C_{min} is about 36 to 47 for $T_{e-e} = 100$ms, but drops to about 21 to 26 for $T_{e-e} = 1$sec).

VIII. CONCLUSION

In this paper, we have considered statistically identical, peak rate controlled, and leaky bucket shaped sources feeding a multiplexer. For a shaping plus multiplexing delay constraint, and constraint violation probability, we have considered lossless multiplexing in the network, and have formulated an optimisation problem that leads to a network bandwidth minimising choice for the token rate parameter (ρ). For the optimal sustainable rate parameter so obtained, we have studied a stochastic approximation technique for on-line computation of this parameter at the source.

We showed that for a stringent end-to-end delay requirement (meaning that the delay bound is small compared to the source burst duration), as in the case of packet telephony, the optimal token rate ρ^* is only slightly less than the peak rate. But if we relax our end-to-end delay requirement for the voice source, e.g., say 1 second for a streaming lecture transmission, then using optimal shaping we obtained 28% reduction in the bandwidth requirement. Since in the case of stringent end-to-end delay requirement we could not get much capacity gain using optimal shaping and lossless multiplexing, we experimented with lossy multiplexing. For the lossy multiplexing case we showed that there is a significant reduction in the bandwidth requirement (relative to the peak rate) as the value of σ is increased from 0 to $T\rho^* - H - L_{max}$. This reduction in bandwidth is more if the end-to-end delay bound is large compared to the source burst duration.

The homogeneous source and QoS (same T and q) model is appropriate for IP telephony sources being multiplexed at an "IP PBX". Further work is needed to relax the assumption of source homogeneity, and the requirement that all the sources need the same QoS. Our treatment of the statistical QoS case in this paper is entirely by simulation; recent results on statistical analysis with LB shaped sources can be used to develop an analysis for lossy multiplexing as well.

APPENDIX

I. PROOF OF THEOREM IV.1

Proof: Using the objective function first we obtain the condition for the following inequality to hold

$$\rho \leq \frac{R + \left(\frac{R-\rho}{\sigma}\right) H}{1 + \left(T - \frac{B-\sigma+L_{max}}{\rho}\right)\left(\frac{R-\rho}{\sigma}\right)}$$

i.e.,

$$1 \leq \frac{R(\sigma+H) - \rho H}{\rho\sigma + (T\rho - B - L_{max} + \sigma)(R-\rho)}$$

Since $T \geq \left(\frac{B-\sigma+L_{max}}{\rho} = \frac{B_s+L_{max}}{\rho}\right)$ and $R \geq \rho$, we have

$$\rho\sigma + (T\rho - B - L_{max} + \sigma)(R-\rho) \leq R(\sigma+H) - \rho H$$

from which we get

$$T\rho \leq B + H + L_{max}$$

Thus, independent of σ, $\rho > \frac{R\left(1+\frac{H}{\sigma}\right)-\rho\left(\frac{H}{\sigma}\right)}{1+\left(T-\frac{B-\sigma+L_{max}}{\rho}\right)\left(\frac{R-\rho}{\sigma}\right)}$ for $B < T\rho - H - L_{max}$. Recall that $h_q(\cdot)$ is the inverse function of $g_q(\cdot)$ as discussed in Section II. Since $g_q(\cdot)$ is decreasing and convex, the same properties also hold for $h_q(\cdot)$. Now ρ and B are related by $\rho = h_q(B)$, and ρ^* is defined by the solution of $\rho = h_q(T\rho - H - L_{max})$, i.e., $B^* = T\rho^* - H - L_{max}$. Since $h_q(B)$ is decreasing in B, $Th_q(B) > B + H + L_{max}$ for $B < (T\rho^* - H - L_{max})$. Thus, for $\rho = h_q(B)$, we can write (see Figure 16; see also Figure 7),

$$C(B) = \begin{cases} h_q(B) & \text{if } B < (T\rho^* - H - L_{max}) \\ \frac{R\left(1+\frac{H}{\sigma}\right)-\rho\left(\frac{H}{\sigma}\right)}{1+\left(T-\frac{B-\sigma+L_{max}}{\rho}\right)\left(\frac{R-\rho}{\sigma}\right)} & \text{if } B \geq (T\rho^* - H - L_{max}) \end{cases} \quad (21)$$

We see that the C vs B curve is just the $h_q(B)$ curve up to $B^* = T\rho^* - H - L_{max}$, and is hence decreasing up to $B^* = T\rho^* - H - L_{max}$. We will now show that for $B > T\rho^* - H - L_{max}$, $C(B) = \frac{R\left(1+\frac{H}{\sigma}\right)-\rho\left(\frac{H}{\sigma}\right)}{1+\left(T-\frac{B-\sigma+L_{max}}{\rho}\right)\left(\frac{R-\rho}{\sigma}\right)} \geq \rho^*$. This will establish the result.

Notice that for $B \geq T\rho^* - H - L_{max}$, and $B = g_q(\rho)$, we have $B \geq (T\rho - H - L_{max})$ as can be seen from Figure 7. Also we have, from the constraints, that $B + L_{max} - T\rho \leq \sigma \leq B$. It can then be shown that (some detailed algebraic manipulations are needed), for $B \geq T\rho^* - H - L_{max}$ and $R \geq \rho$,

$$\frac{R\left(1+\frac{H}{\sigma}\right)-\rho\left(\frac{H}{\sigma}\right)}{1+\left(T-\frac{B-\sigma+L_{max}}{\rho}\right)\left(\frac{R-\rho}{\sigma}\right)} \geq \frac{R\left(1+\frac{H+L_{max}}{B}\right)-\rho\left(\frac{H+L_{max}}{B}\right)}{1+\frac{T}{B}(R-\rho)}$$

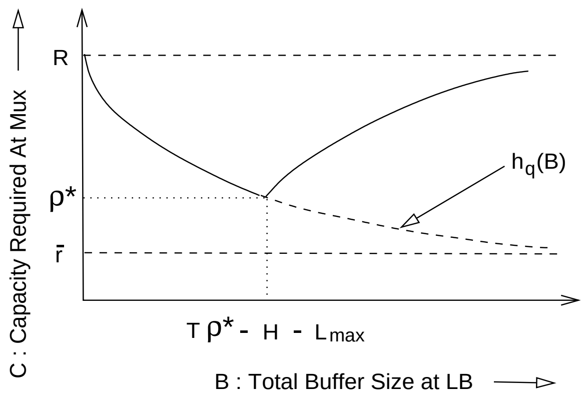


Fig. 16. Capacity requirement C at multiplexer (first network node) as a function of total buffer B at the shaper. $\bar{r}$ is the mean rate of the source.

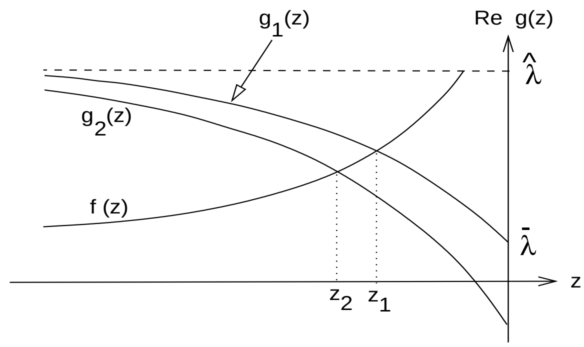


Fig. 17. The maximum real eigenvalue $g_1(z)$ is a concave decreasing function of z, while $f(z)$ is a convex increasing function of z. The other eigenvalue, i.e., $g_2(z)$, is shown to be concave decreasing but it can be of any shape which satisfies the condition $g_1(z) > g_2(z)$, $\forall z \in (-\infty, 0]$.

Hence it suffices to show that, for $B \geq T\rho^* - H - L_{max}$,

$$\frac{R\left(1+\frac{H+L_{max}}{B}\right)-\rho\left(\frac{H+L_{max}}{B}\right)}{1+\frac{T}{B}(R-\rho)} \geq \rho^*$$

But this follows since, by the convex decreasing nature of $g_q(\cdot)$, we have, for $\bar{r} < \rho < \rho^*$, and $B = g_q(\rho)$ (see Figure 7),

$$\frac{T\rho^* - H - L_{max}}{R-\rho^*} \leq \frac{B}{R-\rho}$$
$$\Rightarrow \quad \frac{RB + (R-\rho)(H+L_{max})}{B + T(R-\rho)} \geq \rho^*$$
$$\Rightarrow \quad \frac{R\left(1+\frac{H+L_{max}}{B}\right)-\rho\left(\frac{H+L_{max}}{B}\right)}{1+\frac{T}{B}(R-\rho)} \geq \rho^*$$

In Figure 16 the increasing solid curve to the right of $B^* = T\rho^* - H - L_{max}$ is a sketch of the lower bound to C, i.e., $\frac{R\left(1+\frac{H+L_{max}}{B}\right)-\rho\left(\frac{H+L_{max}}{B}\right)}{1+\frac{T}{B}(R-\rho)}$. Note that the condition $B^* = T\rho^* - H - L_{max} > 0$ implies $\frac{H+L_{max}}{T} < \rho^* < R$.

II. PROOF OF THEOREM VI.1

Proof: It is shown in [12] that there exists a real eigenvalue, say $g_1(z)$, of the matrix $\left[\Lambda - \frac{1}{z}M\right]$, such that if $g(z)$ is any other eigenvalue then $\mathrm{Re}(g(z)) < g_1(z)$. For a 2-state MMFP, we denote the maximal real eigenvalue by $g_1(z)$ and the real part of the other eigenvalue by $g_2(z)$. So we have

$$g_1(z) > g_2(z), \quad \text{for all } z \in (-\infty, 0] \tag{22}$$

Since the maximal real eigenvalue of the nonnegative, irreducible matrix $\left[\Lambda - \frac{1}{z}M\right]$ is a *strictly concave decreasing function* of z [12], we have

$$g_1(z_1) > g_1(z_2) \;\; \forall\, z_1, z_2 \text{ such that } z_1 < z_2 \tag{23}$$

Now from the asymptotic slope constraint shown in Equation 18, $z = \frac{k}{\rho - l}$, where $k = \frac{\ln q}{T}$ and $l = \frac{H + L_{max}}{T}$. Define $f(z) = \frac{k}{z} + l$. Observe that $k < 0$ and $f(z)$ is a *strictly convex increasing function* of $z,\ z \in (-\infty, 0]$, as shown in Figure 17, where sketches of $g_1(z)$ and $g_2(z)$ are also shown. Here $\bar{\lambda} \leq g_1(z) \leq \hat{\lambda}$, where $\bar{\lambda}$ is the mean rate and $\hat{\lambda}$ is the peak rate of the source process. It is clear that $f(z)$ will intersect $g_1(z)$ at a unique point, say z_1. If $f(z)$ does not intersect $g_2(z)$ (as an example if $g_2(z) < 0\ \forall z \in (-\infty, 0]$) then we have unique root $g_1(z_1)$ of Equation 17 with asymptotic slope z_1 and we are done. Consider the other case that $f(z)$ intersects $g_1(z)$ at z_1 and $g_2(z)$ at z_2. It is now easily seen that, with the properties of $f(z)$ and $g_1(z)$ described above, the only possibility is that $z_1 > z_2$, as shown in Figure 17. Then

$$g_1(z_1) = f(z_1) > f(z_2) = g_2(z_2) \tag{24}$$

Thus we have $g_1(z_1) > g_2(z_2)$ and the larger root corresponds to the maximal real eigenvalue $g_1(z)$ with $z = z_1$, the intersection point of $g_1(z)$ with $f(z)$. $\square$

REFERENCES

[1] R. Braden et al, "Resource reservation protocol (rsvp) – version 1, functional specification," IETF RFC 2205, September 1997.

[2] M. Gibson, "The management of mpls lsps for scalable qos service provision; draft-gibson-manage-mpls-qos-01.txt," IETF Internet Draft, March 2000.

[3] D.O. Awduche, Alan Hannan, and Xipeng Xiao, "Applicability statement for extensions to rsvp for lsp-tunnels; draft-ietf-mpls-rsvp-tunnel-applicability-00.txt," IETF Internet Draft, September 1999.

[4] Arthur Berger, "Performance analysis of a rate-control throttle where tokens and job queue," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 2, pp. 165–170, Feb 1991.

[5] Fabrice Guillemin, Catherine Rosenberg, and Josee Mignault, "On characterizing an ATM source via sustainable cell rate traffic descriptor," in *IEEE INFOCOM*, 1995.

[6] Gustavo de Veciana, "Leaky buckets and optimal self-tuning rate control," in *IEEE GLOBECOM*, 1994.

[7] Brian L. Mark and Gopalakrishnan Ramamurthy, "Real-time estimation and dynamic renegotiation of UPC parameters for arbitrary traffic sources in ATM networks," *IEEE/ACM Transactions on Networking*, vol. 6, no. 6, pp. 811–827, December 1998.

[8] Anwar I. Elwalid and Debasis Mitra, "Analysis and design of rate based congestion control of high speed networks, i: stochastic fluid models, access regulatiom," *Queueing Systems*, vol. 9, pp. 29–64, 1991.

[9] A. K. Parekh and R. G. Gallager, "A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: The Single Node Case," *IEEE/ACM Transactions on Networking*, vol. 1, pp. 344–357, June 1993.

[10] D. Stiliadis and A. Varma, "Latency-rate Servers: A General Model for Analysis of Traffic Scheduling Algorithms," *IEEE/ACM Transactions on Networking*, vol. 6, pp. 611–624, October 1998.

[11] Rajeev Agrawal and Rajendra Rajan, "Performance Bounds for Guaranteed and Adaptive Services," *IBM Research Report, IBM Research Division, T. J. Watson Research Center*, May 1996.

[12] Anwar I. Elwalid and Debasis Mitra, "Effective Bandwidth of General Markovian Traffic Sources and Admission Control of High Speed Networks," *IEEE/ACM Transactions on Networking*, vol. 1, no. 3, pp. 329–343, June 1993.

[13] Parijat Dube, "Measurement Based Optimal Source Shaping in Integrated Services Packet Networks," in *M.Sc.(Engg.) Thesis, Indian Institute of Science, India*, October 1999.

[14] Harold J. Kushner and Dean S. Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer-Verlag, 1978.

[15] C. Courcoubetis, G. Kesidis, A. Ridder, J. Walrand, and R. Weber, "Admission control and routing in ATM networks using inferences from measured buffer occupancy," *IEEE Transactions on Communications*, vol. 43, pp. 1778–1784, April 1995.

[16] Anwar Elwalid, Daniel Heyman, T. V. Lakshman, Debasis Mitra, and Alan Weiss, "Fundamental bounds and approximations for ATM multiplexers with applications to video teleconferencing," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 1004–1016, August 1995.

[17] G. de Veciana and J. Walrand, "Effective bandwidths: Call admission, traffic policing & filtering for ATM networks," *Queueing Systems (QUESTA)*, vol. 20, pp. 37–59, 1995.

[18] Santosh Abraham and Anurag Kumar, "A new approach for asynchronous distributed rate control of elastic sessions in integrated packet networks," *IEEE Transactions on Networking*, submitted.

[19] Parijat Dube and Anurag Kumar, "Measurement based selection of token buffer size for leaky bucket controlled sources: A simulation study," Tech. Rep., Indian Institute of Science, November 1998.

[20] Natwar Modani, Parijat Dube, and Anurag Kumar, "Measurement Based Optimal Source Shaping with a Shaping+Multiplexing Delay Constraint," in *Proc. of IEEE INFOCOM*, March 2000.

[21] Kotikalapudi Sriram and David M. Lucantoni, "Traffic Smoothing Effects of Bit Dropping in a Packet Voice Multiplexer," *IEEE Transactions on Communications*, vol. 37, no. 7, pp. 703–712, July 1989.

[22] Nanying Yin, San-Qi Li, and Thomas E. Stern, "Congestion Control for Packet Voice by Selective Packet Discarding," *IEEE Transactions on Communications*, vol. 38, no. 5, pp. 674–683, May 1990.

[23] Kotikalapudi Sriram and Yung-Terng Wang, "Voice over ATM Using AAL2 and Bit Dropping : Performance and Call Admission Control," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 1, pp. 18–28, January 1999.

[24] Nanying Yin and Michael G. Hluchyj, "A Dynamic Rate Control Mechanism for Source Coded Traffic in a Fast Packet Network," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 7, pp. 1003–1012, September 1991.