1

Optimal Buffer and Bandwidth Allocation using Effective Bandwidth

Scott Jordan, Kalpana Jogi, Chunlin Shi, Ikhlaq Sidhu

Abstract—

We consider a single node which multiplexes a large number of traffic sources. We restrict ourselves to consideration of aggregates of i.i.d. flows that can modelled using effective bandwidth results. We are concerned with the amount of buffer and bandwidth that should be allocated to this aggregate, under a maximum overflow probability constraint. Unlike previous approaches which assume that the total buffer allocated to the class is either constant or linearly proportional to the number of sources, we wish to determine the minimum cost allocation given a cost per unit of each resource.

We first consider a class of on/off fluid flows. We find that the optimal bandwidth allocation above the mean rate and the optimal buffer allocation are both proportional to the square root of the number of sources. Correspondingly, we find that the excess cost incurred by a fixed buffer allocation or by linear buffer allocations is proportional to the square of the percentage difference between the assumed number of sources and the actual number of sources and to the square root of the number of sources.

We next consider a class of general i.i.d. sources for which the aggregate effective bandwidth is a decreasing convex function of buffer and linearly proportional to the number of sources. We find that the optimal buffer allocation is strictly increasing with the number of sources. Correspondingly, we find that the excess cost incurred by a fixed buffer allocation is an increasing convex function of the difference between the assumed number of sources and the actual number of sources.

Keywords— **Resource allocation, cost minimization, dimensioning**

I. Introduction

A. Background

THERE is now a rich literature on the use of effective bandwidth to estimate the buffer and bandwidth requirements of network traffic sources, particularly for sources with real-time loss and delay constraints.

Early results considered a single traffic flow. The typical approach estimates the loss probability of the flow by the probability that the buffer content in an infinite buffer queue will exceed a threshold. Such results establish that the resulting loss probability estimate $b(x)$ asymptotically obeys:

$$b(x) \sim \alpha e^{-\eta x} \tag{1}$$

where x is the buffer threshold, η is a positive constant called the *asymptotic decay rate*, and α is a positive constant called the *asymptotic constant*. The limit is usually taken as the buffer approaches infinity, for a fixed bandwidth (see e.g. [1]).

Later results extended (1) to a wide range of traffic sources and to multiplexed i.i.d. traffic flows, where now x is the total buffer shared by N flows (see e.g. [2], [3], [4], [5], [6], [7], [8], [9]). The limit is usually taken as the number of sources approaches infinity, with a fixed bandwidth and buffer per source. The resulting loss probability estimate is thus interpreted as the probability of exceeding a delay bound.

Such results have often been used to formulate admission control policies (see e.g. [10], [11]). If a class of flows have identical traffic characteristics, and share a common Quality of Service (QoS) requirement that the loss probability should not exceed p, then a new connection should be accepted if and only if the available bandwidth exceeds the *effective bandwidth* that results from (1).

These results have also often been interpreted in the context of dimensioning. To accommodate N flows with a maximum loss probability of p, the required bandwidth per source can be calculated if the reserved buffer per source is known. The buffer per source might be chosen based on an estimate of the typical availability of buffer versus bandwidth, and perhaps on an estimate of the average number of flows.

In this paper, we examine the assumption that buffer and bandwidth should be allocated in constant proportion. As many previous researchers have demonstrated, a set of flows can achieve a maximum loss probability using various combinations of total shared buffer and bandwidth (see e.g. [12]). Our goal in this research effort is to understand how the optimal combination of buffer and bandwidth might vary with the number of flows.

To define optimality, we assume that there are costs associated with each unit of buffer and of bandwidth. The ratio of the cost per unit bandwidth to the cost per unit buffer

Scott Jordan is with the University of California, Irvine, and can be reached at Dept. ECE, 544D Engineering Tower, University of California, Irvine, CA 92697 or at sjordan@uci.edu. This work was supported by NSF and by DARPA.

should reflect the relative demand for bandwidth to buffer from all of the traffic flowing through the router. This ratio might be based on average traffic estimates of various classes of traffic. If a pricing policy is used, then the costs can be interpreted as shadow costs (Lagrangian multipliers) that result from the pricing policy (see e.g. [12], [13], [14])

We define the optimal combination of buffer and bandwidth as the minimum cost choice that achieves the desired QoS. In this paper, we equate QoS with loss probability, but it is simple to add a limit on buffer in order to enforce a maximum delay constraint.

B. Motivating Example

As a motivating example, consider a single node which multiplexes compressed real-time voice sources, modelled as on/off fluid flows with a mean on time of 340ms, a mean off-time of 780ms, and a peak rate of 8kbps. We require that the overflow probability should not exceed 0.01. We normalize all quantities: time is represented in units equal to the mean on time, bandwidth is represented in units of the peak rate, and buffer is represented in units of the average number of arriving bits per on/off cycle. We set the ratio of cost per unit bandwidth to buffer to 1 (which due to normalization implies that 8kbps of bandwidth is equally expensive as 340 bytes of buffer).

Using an estimate of overflow probability derived by Morrison [15], we can numerically derive the minimum cost buffer and bandwidth allocations. The results are shown in figure 1, as a function of the number of sources N. The mean bandwidth has been subtracted, and the quantities have been normalized by the number of sources.

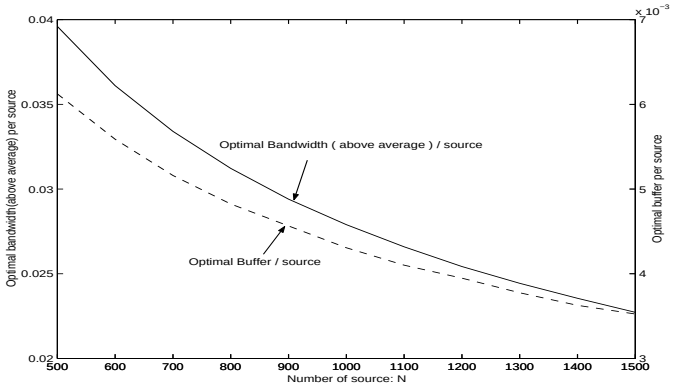

Fig. 1. Optimal buffer and bandwidth allocations versus N

We note that the optimal buffer per source and the optimal bandwidth per source (above average) appear to be decreasing convex functions of the number of sources.

Now consider two common resource allocation policies. A *Fixed Buffer* (FB) policy allocates a *fixed* amount of total buffer, and adjusts the bandwidth (depending on the number of sources) to satisfy the loss constraint. A *Incremented Buffer* (IB) policy allocates a constant amount of buffer *per source*, and adjusts the bandwidth to satisfy the loss constraint.

The results in figure 1 do not correspond to either a FB or an IB policy. The optimal resource allocation policy is neither to fix the buffer length and then add bandwidth, nor to add buffer and bandwidth in constant proportion. Indeed, we can numerically compare the optimal allocation policy to these two alternate policies. The results are shown in figure 2, where the buffer allocations for FB and IB were initially calculated for 750 sources, and then the number of sources was varied from 500 to 1000. We note that the cost difference appears to be increasing and convex with the difference between the actual and nominal number of sources ($|N - 750|$).

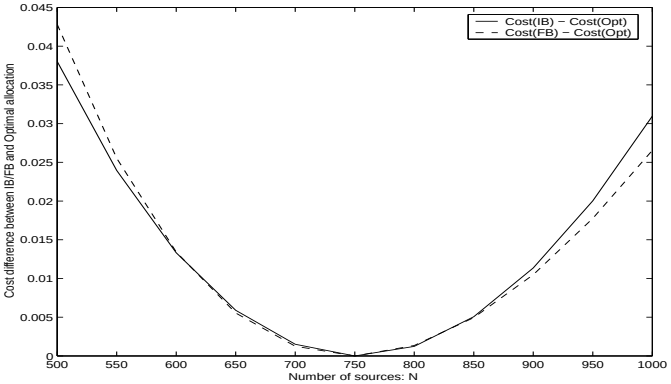

Fig. 2. Cost difference between optimal and alternate policies for a fixed $\hat{N}$

Our goal in this paper is to explain the forms of the curves in figures 1 and 2.

C. Principal Results

We first consider a single node which multiplexes a large number of i.i.d. on/off fluid flows, under a maximum overflow probability constraint on the class. We use Taylor series expansions of the overflow probability to determine a representation of the feasible combinations of buffer and bandwidth. The costs are then used to determine the optimal choice of buffer and bandwidth. Our principal result is that the optimal bandwidth is given by:

$$c^* = N(\mu + k_1^*/\sqrt{N} + O(1/N)) \tag{2}$$

and the optimal buffer is given by:

$$x^* = N(k_2^*/\sqrt{N} + O(1/N)) \tag{3}$$

where μ is the mean rate per source and k_1^* and k_2^* are positive constants that depend upon the statistics of a single

traffic source and upon the ratio of the cost per unit bandwidth to the cost per unit buffer.

These results imply that as the number of sources increase, the minimum cost solution (under fixed per unit buffer and bandwidth costs) is to not to add buffer and bandwidth in constant proportion, but instead to first add the mean bandwidth of each source, and then to add additional bandwidth and buffer in approximately constant proportion. Furthermore, we demonstrate that the the cost savings of this optimal allocation over an allocation that maintains a fixed buffer per source is proportional to the square of the percentage difference between the assumed number of sources and the actual number of sources and to the square root of the number of sources.

We base our analysis upon an estimate of overflow probability derived by Morrison [15]. This estimate predates almost all of the effective bandwidth literature, and later effective bandwidth estimates for on/off fluids are significantly more accurate, particularly with regard to the asymptotic constant (see e.g. [16], [17], [18]). However, as mentioned above, the effective bandwidth literature typically assumes that buffer and bandwidth are allocated proportionally. In contrast, Morrison derives his estimate under independently chosen buffer and bandwidth, for a wide range of buffer sizes that bracket those in (2) and (3). It is worth stressing at this point that we are not proposing that the Taylor series expansion be used to predict overflow probability, as we do not believe any Taylor series expansion would be an accurate predictor of overflow probability. Our goal in this work is to obtain an asymptotic relationship between the optimal buffer and bandwidth allocation and the number of sources. This requires a simple representation of overflow probability as a function of both buffer and bandwidth, and the Taylor series expansion serves this purpose.

We next consider a single node which multiplexes a more general class of i.i.d. flows, provided that the aggregate effective bandwidth is a decreasing convex function of buffer and linearly proportional to the number of sources. Without relying on any particular expression for effective bandwidth, our goal is to explore the variation of the optimal bandwidth and buffer allocations with respect to the number of sources for a more general class of sources than the on/off sources considered earlier.

We use the form of the aggregate effective bandwidth function to prove two principal results. First, we prove that the optimal buffer is strictly increasing in N. Second, we prove that the excess cost incurred by a fixed buffer allocation over an optimal allocation is an increasing convex function of the difference between the assumed number of sources and the actual number of sources. Both results are consistent with our results for on/off sources, but less specific.

In section II, we consider on/off sources. In sections II-A and II-B, we review our network model and Morrison's expressions for overflow probability, and illustrate buffer versus bandwidth tradeoffs with some numerical examples. In section II-C, we derive the Taylor series expansions for overflow probability. In sections II-D and II-E, we derive the minimum cost buffer and bandwidth allocations and present our principal results for on-off sources. In section III, we consider general sources.

II. On/off sources

A. The Network Model

We consider a single queue fed by N i.i.d. on/off fluid sources, as shown in figure 3. Both the on and off times are assumed to be Exponentially distributed. Without loss of generality, we measure time in units equal to the average on period of a source, and measure bandwidth in units equal to the peak rate of a source. We denote the average off time by $1/\lambda$. The mean rate per source is thus equal to $\lambda/(1+\lambda)$.

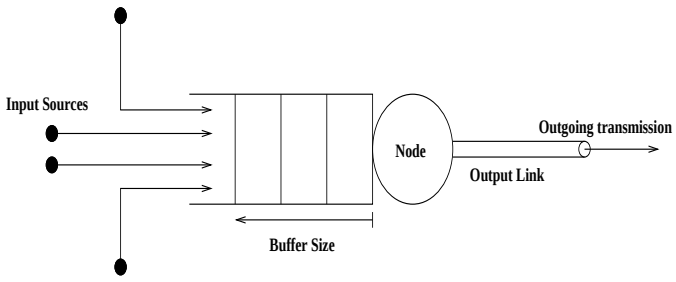

Fig. 3. Diagrammatic representation of the network model

In numerical examples, we use the parameters given in the motivating examples above. Using our normalization, with bandwidth measured in multiples of 8kpbs and buffer measured in multiples of 340B, this gives $\lambda = 0.436$ and $\frac{\lambda}{1+\lambda} = 0.3036$.

A fixed buffer x and a fixed bandwidth c is reserved for this class of traffic. We denote the buffer per source (x/N) by ξ, and the bandwidth per source (c/N) by γ, and assume that the bandwidth per source lies strictly between the mean rate and the peak rate, namely that:

$$\frac{\lambda}{1+\lambda} < \gamma < 1$$

Finally, we denote the bandwidth above the mean rate per source by δ:

$$\delta = \gamma - \frac{\lambda}{1+\lambda}$$

In numerical examples, unless explicitly mentioned we set $N = 500$, $\delta = .0410$, and the maximum probability of overflow $p = 0.01$.

We briefly restate the expressions for overflow probability derived by Morrison [15]. His derivation starts with earlier work by Anick, Mitra and Sondhi [19], which states that the equilibrium probability that the buffer content exceeds x in an infinite buffer system can be expressed as:

$$G(N, x, \gamma) = \sum_{j=0}^{N-\lfloor N\gamma \rfloor - 1} D_j e^{-\sigma_j x} \tag{4}$$

where σ_j are eigenvalues of the buffer dynamics, and D_j are constants that depend on these eigenvalues. There are a total of $N - \lfloor c \rfloor$ terms, corresponding to the range in the number of on sources for which overflow occurs. Morrison based his approximation to $G(N, x, \gamma)$ on the largest terms in (4).

Assuming that the number of sources is large ($N \gg 1$), the bandwidth per source $\gamma = O(1)$, and that the either the total buffer $x = O(1/N)$ or $x = O(1)$, Morrison shows that the main contributions arise from the largest eigenvalues. This leads to an asymptotic expression for the overflow probability:

$$\begin{aligned} G(N, x, \gamma) &= \frac{1}{2}\sqrt{\frac{r}{\pi f(\gamma)[\gamma + \lambda(1-\gamma)]N}} e^{-N\kappa(\gamma)} \\ & \quad e^{-2\sqrt{\{f(\gamma)[\gamma+\lambda(1-\gamma)]Nx\}}} e^{-g(\gamma)x} \end{aligned} \tag{5}$$

where

$$f(\gamma) = ln\left[\frac{\gamma}{\lambda(1-\gamma)}\right] - 2\frac{[\gamma(1+\lambda) - \lambda]}{[\gamma + \lambda(1-\gamma)]} \tag{6}$$

$$r = \frac{[\gamma(1+\lambda) - \lambda]}{\gamma(1-\gamma)} \tag{7}$$

$$\kappa(\gamma) = \gamma ln\gamma + (1-\gamma)ln(1-\gamma) - \gamma ln(\lambda) + ln(1+\lambda) \tag{8}$$

$$g(\gamma) = k + \frac{1}{2}[\gamma + \lambda(1-\gamma)]\frac{\rho"(1-\gamma)}{f(\gamma)} \tag{9}$$

$$\rho"(1-\gamma) = \frac{(2\gamma - 1)[\gamma(1+\lambda) - \lambda]^3}{\gamma(1-\gamma)[\gamma + \lambda(1-\gamma)]^3} \tag{10}$$

$$k = (1-\lambda) + \frac{\lambda(1-2\gamma)}{[\gamma + \lambda(1-\gamma)]} \tag{11}$$

Morrison also considered the case where $N \gg 1$, $\gamma = O(1)$, and $x = O(N)$. He develops an approximation by again expanding around most significant terms, although these no longer correspond to the largest eigenvalues. He shows that the largest terms of the resulting expression agrees with the largest terms of (5). Although it has not been proven that this approximation is uniformly accurate throughout the range from $x = O(1)$ to $x = O(N)$, we will use this expression as our starting point.

B. Numerical examples

To illustrate the basic problem consider in this paper, we return to our motivating example to illustrate the effect of varying the number of sources, the buffer and the bandwidth. In figure 4, the overflow probability is plotted for a range of N for a fixed bandwidth per source γ.

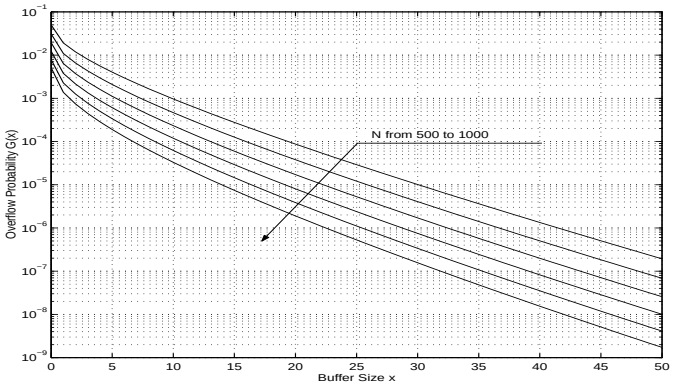

Fig. 4. Overflow probability for a range of N

The figure illustrates the relationship between overflow probability p, total buffer x, and the number of sources N, assuming that the resource allocation policy assigns bandwidth proportional to the number of sources. As discussed by many previous researchers, overflow probability decreases with N, when there is a fixed bandwidth per source and either a fixed total buffer or a fixed buffer per source. These observations represent two paths through these overflow vs. buffer curves.

An alternative view is shown in figure 5, in which the overflow probability is varied for a fixed number of sources N. Each curve represents a contour of the overflow probability function, and shows which combinations of buffer and bandwidth produce the same overflow probability. Note that there is a substantial range of slopes along each contour. The optimal resource allocation policy will choose buffer and bandwidth to equate the slope of the contour with the corresponding price ratio. Alternate policies such as fixed buffer or incremental buffer do not take into account the prices of each resource and therefore may produce quite different allocations. The range in slopes means that there is a significant achievable cost savings of the optimal resource allocation policy over fixed buffer or incremental buffer policies.

Buffer vs. bandwidth contours for fixed p but varying N are shown in figures 6 through 8. Figure 6 shows the total buffer and total bandwidth per source. The majority of the bandwidth is due to the mean rate, which must be allocated (at loss overflow probabilities) under any resource allocation policy.

Figure 7 takes out the mean rate from each contour.

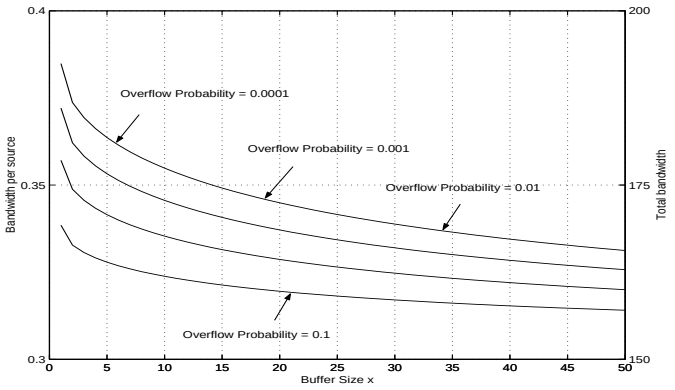

Fig. 5. Buffer vs. bandwidth contours for a range of p

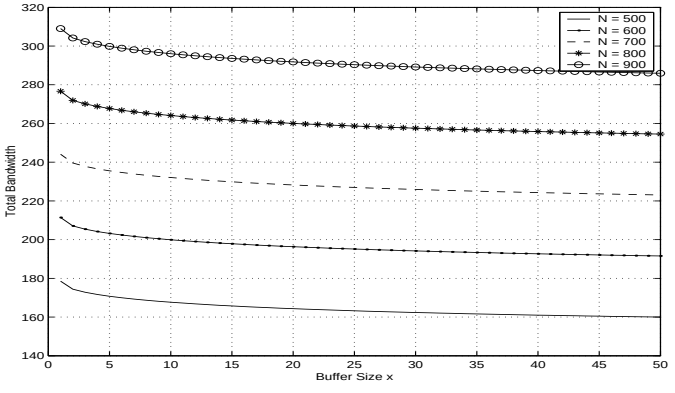

Fig. 6. Total buffer vs. total bandwidth contours for a range of N

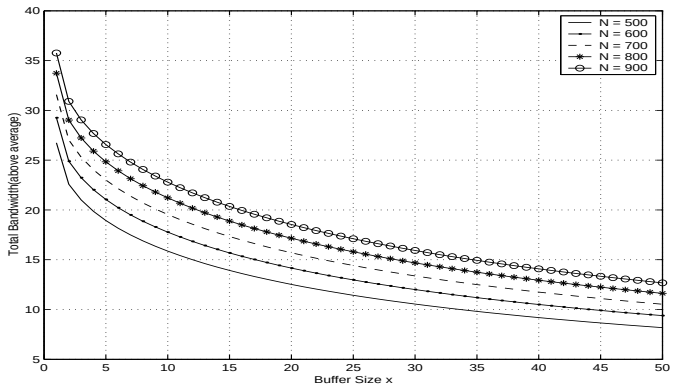

Fig. 7. Total buffer vs. total bandwidth above average contours for a range of N

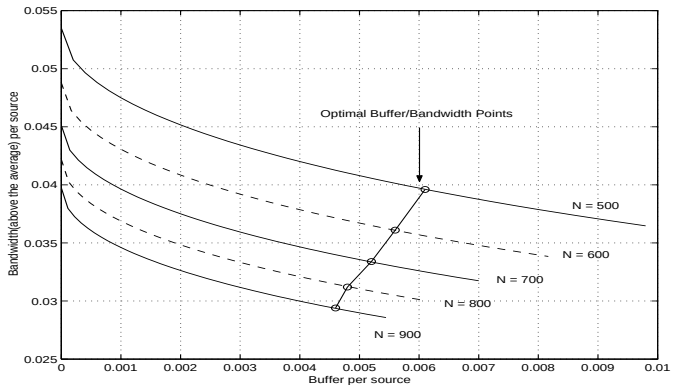

Fig. 8. Buffer per source vs. bandwidth above average per source contours for a range of N

Multiplexing gains mean that larger N correspond to *larger* bandwidth and buffers, but with decreasing increments. A fixed buffer policy would constitute a *vertical line* through the set of contours, and an incremental buffer policy would constitute a curve with *fixed horizontal increments* through the set. Note again that there is a large range of slopes, indicating that the optimal policy can adjust the allocations significantly.

Figure 8 shows the same information, but with each axis normalized by the number of sources. Multiplexing gains mean that larger N correspond to *lower* bandwidth and buffers per source. A fixed buffer policy would now constitute a curve through the set, while an incremental buffer policy would constitute a vertical line. The cost minimizing choices of buffer and bandwidth are also shown.

C. Taylor Series Expansions

In this section, we develop Taylor series approximations for the overflow probability (5). We start by expanding the constituent parts of (5) expressed in (6) through (11). The general approach is to expand the expression using:

$$\gamma = \frac{\lambda}{1+\lambda} + \delta \tag{12}$$

for $\delta \ll 1$.

Substituting (12) into the first term of (6), we get:

$$ln\left[\frac{\gamma}{\lambda(1-\gamma)}\right] = ln\left[1+\frac{\delta(1+\lambda)}{\lambda}\right] - ln[1-\delta(1+\lambda)]$$

Using the Taylor series expansion

$$ln(1+z) = z - \frac{z^2}{2} + \frac{z^3}{3} - \cdots \tag{13}$$

this reduces to

$$\left[(1+\lambda)(1+\frac{1}{\lambda})\right]\delta + \frac{1}{2}\left[(1+\lambda)^2(1-\frac{1}{\lambda^2})\right]\delta^2 + \frac{1}{3}\left[(1+\lambda)^3(1+\frac{1}{\lambda^3})\right]\delta^3 + O(\delta^4) \tag{14}$$

provided that $\left|\frac{\delta(1+\lambda)}{\lambda}\right| < 1$.

The second term in $f(\gamma)$ similarly reduces to

$$\frac{(1+\lambda)^2}{\lambda}\delta + \frac{(1+\lambda)^2(\lambda^2-1)}{2\lambda^2}\delta^2 + \frac{(\lambda^2-1)^2(1+\lambda)^2}{4\lambda^3}\delta^3 + O(\delta^4)$$

provided that $|\delta(1+\lambda)| < 1$.

Together these two terms give

$$f(\gamma) = \frac{(1+\lambda)^6}{12\lambda^3}\delta^3 + O(\delta^4) \tag{15}$$

We next consider (7). The numerator reduces to $\delta(1+\lambda)$. The denominator can be expressed as

$$\left[\frac{\lambda}{1+\lambda} + \delta\right]\left[1 - \frac{\lambda}{1+\lambda} - \delta\right]$$

Together we find

$$r = \left[\frac{\delta(1+\lambda)^3}{\lambda}\right]\left[\frac{1}{1 - \left(\frac{(1+\lambda)(\lambda-1)}{\lambda}\right)\delta + O(\delta^2)}\right]$$

Using the Maclaurin expansion

$$\frac{1}{1-z} = 1 + z + z^2 + z^3 + \cdots$$

we can express this as

$$r = \left[\frac{(1+\lambda)^3}{\lambda}\delta\right]\left[1 + \left(\frac{(1+\lambda)(\lambda-1)}{\lambda}\right)\delta + O(\delta^2)\right] \tag{16}$$

provided that $\left|\frac{(1+\lambda)(\lambda-1)}{\lambda}\delta\right| < 1$.

We continue with (8). We can combine terms to express this as

$$\kappa(\gamma) = \gamma ln\left[\frac{\gamma}{(1-\gamma)\lambda}\right] + ln[(1-\gamma)(1+\lambda)] \tag{17}$$

An approximation for the first log term was found above to be (14). Multiplying by γ we have

$$\left(\frac{\lambda}{1+\lambda} + \delta\right)\left[(1+\lambda)(1+\frac{1}{\lambda})\right]\delta + \frac{1}{2}\left[(1+\lambda)^2(1-\frac{1}{\lambda^2})\right]\delta^2 + \frac{1}{3}\left[(1+\lambda)^3(1+\frac{1}{\lambda^3})\right]\delta^3 + O(\delta^4)$$

which after some simple manipulation results in

$$(\lambda+1)\delta + \frac{(\lambda+1)^3}{2\lambda}\delta^2 + O(\delta^3)$$

The second term in (17) can be expressed as

$$ln[1 - \delta(1+\lambda)]$$

Using (13), this becomes

$$-\delta(1+\lambda) - \delta^2\frac{(1+\lambda)^2}{2} - \delta^3\frac{(1+\lambda)^3}{3} + O(\delta^4)$$

Putting together these expressions for the two terms in (17) we find

$$\kappa(\gamma) = \frac{(1+\lambda)^2}{2\lambda}\delta^2 + O(\delta^3) \tag{18}$$

We continue with (10). A similar use of Taylor series results in:

$$\rho"(1-\gamma) = \frac{1}{8}\left[\frac{(1+\lambda)^7(\lambda-1)}{\lambda^4}\right]\delta^3 + \left[\frac{1}{4}\left(\frac{(1+\lambda)^8}{\lambda^4}\right) + \frac{5}{16}\left(\frac{(1+\lambda)^8(\lambda-1)^2}{\lambda^5}\right)\right]\delta^4 + O(\delta^5)$$

We continue with (11). A similar approach results in:

$$k = \frac{3}{2}(1-\lambda) - \left[(1+\lambda)\left(1 + \frac{(\lambda-1)^2}{4\lambda}\right)\right]\delta + O(\delta^2)$$

Finally, we substitute these two expressions into (9) to find:

$$g(\gamma) = \frac{(\lambda+1)}{4\lambda}[11\lambda^2 - 26\lambda + 11]\delta + O(\delta^2) \tag{19}$$

This completes the development of Taylor series expansions for (6) through (11). We now use these expressions to derive the Taylor series expansion for the overflow probability (5). Using (15), the first term can be expressed as

$$\frac{1}{2}\sqrt{\frac{r}{\pi f(\gamma)[\gamma + \lambda(1-\gamma)]N}} = \sqrt{\frac{3\lambda}{2\pi N}}\left[\frac{1}{(1+\lambda)}\delta^{-1} + O(\delta^0)\right]$$

Using (18), the second term can be expressed as

$$e^{-N\kappa(\gamma)} = e^{-N\frac{(1+\lambda)^2}{2\lambda}\delta^2 + O(N\delta^3)}$$

Similarly the third term can be expressed as

$$e^{-2\sqrt{\{f(\gamma)[\gamma+\lambda(1-\gamma)]Nx\}}} = e^{-2\sqrt{\frac{(1+\lambda)^5}{6\lambda^2}\delta^3 Nx + O(\delta^4 Nx)}}$$

Using (19), the fourth term can be expressed as

$$e^{-g(\gamma)x} = e^{-\frac{(\lambda+1)}{4\lambda}(11\lambda^2 - 26\lambda + 11)\delta x + O(\delta^2 x)}$$

Finally, combining these four terms we get

$$G(N,x,\delta) = \left(\frac{c_1}{\sqrt{N}\delta} + O(1/\sqrt{N})\right) e^{-(c_2 N\delta^2 + c_3\sqrt{\delta^3 Nx} + c_4\delta x + O(N\delta^3) + O(\delta^{\frac{5}{2}}N^{\frac{1}{2}}x^{\frac{1}{2}}) + O(\delta^2 x))} \tag{20}$$

where

$$\begin{aligned} c_1 &= \sqrt{\frac{3\lambda}{2\pi}}\frac{1}{(1+\lambda)} \\ c_2 &= \frac{(1+\lambda^2)}{2\lambda} \\ c_3 &= 2\sqrt{\frac{(1+\lambda)^5}{6\lambda^2}} \\ c_4 &= \frac{\lambda+1}{4\lambda}(11\lambda^2-26\lambda+11) \end{aligned} \tag{21}$$

We will use this expression for overflow probability to derive the optimal resource allocation scheme in the following sections. The benefit of this Taylor series representation is that it is amenable to analysis.

However, we stress that our goal in this paper is to explain the forms of the curves in figures 1 and 2. We do not expect that any Taylor series expansion would be an accurate predictor of overflow probability. To underscore this point, we numerically compare the Taylor series expansion (20) with Morrison's expression for overflow probability (5) in figure 9.

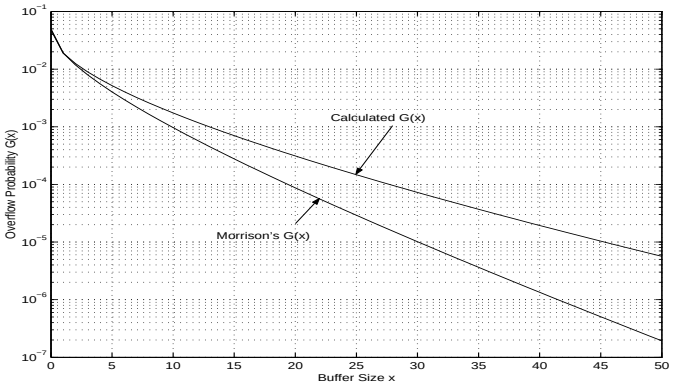

Fig. 9. Taylor series expansion vs. Morrison's expression

Although the Taylor series approximation to the leading constant is good, the approximation to the exponential terms is rough, since only the first term was retained. The error can be greatly reduced by incorporating additional terms into the expansion, but these additional terms do not affect the principal results given in (2) and (3) and therefore we do not include them in our analysis.

D. Optimal Resource Allocation

In this section, we will derive the optimal allocation of buffer and bandwidth to a class of on/off fluid flows under a maximum overflow constraint. Our principal result is:

Theorem 1: Suppose that each unit of buffer incurs a cost p_x and each unit of bandwidth incurs a cost p_c. Assume that $G(N,x,\delta)$ is decreasing and jointly convex in x and δ. The buffer and bandwidth allocation that minimize cost subject to a maximum overflow probability of p are:

$$\begin{aligned} \delta^* &= \frac{k_1^*}{\sqrt{N}} + O(1/N) \\ x^* &= mk_1^*\sqrt{N} + O(1) \end{aligned}$$

where $m = p_c/p_x$ and k_1^* is the solution to

$$k_1^{*2}(c_2 + \sqrt{m}c_3 + mc_4) + ln(\frac{pk_1^*}{c_1}) = 0 \tag{22}$$

where c_1, c_2, c_3, and c_4 are the constants given above.

Proof:

We start with the constraint $G(N,x,\delta) = p$, with $G(N,x,\delta)$ given by (20). Taking logarithms on both sides and rearranging:

$$\begin{aligned} c_2N\delta^2 + c_3\sqrt{\delta^3 Nx} + c_4\delta x \\ +O(N\delta^3) + O(\delta^{\frac{5}{2}}N^{\frac{1}{2}}x^{\frac{1}{2}}) + O(\delta^2 x) \\ = -ln\left(\frac{p}{c_1/(\sqrt{N}\delta) + O(1/\sqrt{N})}\right) \end{aligned} \tag{23}$$

Now suppose that $\delta = k_1/\sqrt{N}$ and $x = k_2\sqrt{N}$ for some $k_1 = O(1)$ and $k_2 = O(1)$, and furthermore suppose that $k_2 = mk_1 + O(1/\sqrt{N})$. By substitution into (23), we get

$$k_1^2(c_2 + \sqrt{m}c_3 + mc_4) + ln(pk_1/c_1) = O(1/\sqrt{N}) \tag{24}$$

Let k_1^* be the solution to (22), and let k_1 be the solution to (24). Then $k_1 = k_1^* + \Delta k$, where:

$$\begin{aligned} \Delta k &\approx \frac{O(1/\sqrt{N})}{\frac{d}{dk_1}\left[k_1^2(c_2+\sqrt{m}c_3+mc_4) + ln(pk_1/c_1)\right]|_{k_1=k_1^*}} \\ &= \frac{O(1/\sqrt{N})}{2k_1^*(c_2+\sqrt{m}c_3+mc_4) + 1/k_1^*} \\ &= O(1/\sqrt{N}) \end{aligned}$$

It follows that:

$$\begin{aligned} k_1 &= k_1^* + O(1/\sqrt{N}) \\ \delta^* &= \frac{k_1^* + O(1/\sqrt{N})}{\sqrt{N}} = \frac{k_1^*}{\sqrt{N}} + O(1/N) \\ x^* &= \left[m(k_1^* + O(1/\sqrt{N})) + O(1/\sqrt{N})\right]\sqrt{N} \\ &= mk_1^*\sqrt{N} + O(1) \end{aligned}$$

This establishes that δ^* and x^* as stated in the theorem satisfy $G(N,x^*,\delta^*) = p$.

Furthermore, this solution minimizes the cost if and only if the slope of the $G(N, x, c)$ contour at a fixed p is equal to the price ratio, namely iff

$$-\frac{\partial G/\partial x}{\partial G/\partial c} = -\frac{p_x}{p_c} = -\frac{1}{m}$$

It can be easily verified that δ^* and x^* satisfy this condition, since $k_2 = mk_1 + O(1/\sqrt{N})$. The theorem follows. ■

These results imply that as the number of sources increase, the minimum cost solution (under fixed per unit buffer and bandwidth costs) is to not to add buffer and bandwidth in constant proportion, but instead to first add the mean bandwidth of each source, and then to add additional bandwidth and buffer in approximately constant proportion.

For numerical illustration, the optimal buffer and bandwidth (above average) allocations per source were shown in figure 1, as a function of the number of sources N (for $m = 1$). The optimal bandwidth per source follows the predicted $1/\sqrt{N}$ form (22) very accurately. The optimal buffer per source also follows the predicted $1/\sqrt{N}$ form, but with a small error that indicates the presence of a smaller order term.

In figure 10, we plot the optimal buffer and bandwidth (above average) allocations per source versus each other. As illustrated in figure 8, the optimal allocations per source decrease with increasing N. If the price ratio of bandwidth to buffer is decreased from $m = 1$ to $m = 0.8$, then the optimal allocation shifts to a higher bandwidth and lower buffer. However, the form of $1/\sqrt{N}$ remains true.

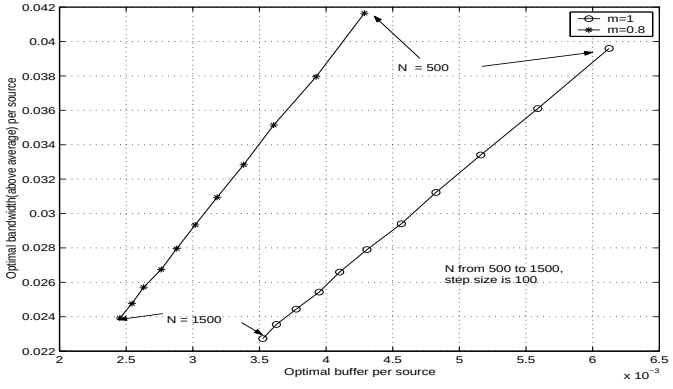

Fig. 10. Optimal buffer versus optimal bandwidth

E. Comparison to Alternative Schemes

In this section, we compare the costs of the optimal resource allocation to methods in which the total buffer allocated to the class is either constant or linearly proportional to the number of sources. We demonstrate that the the cost savings of the optimal allocation over either of these alternative resource allocation policies is proportional to the square of the percentage difference between the assumed number of sources and the actual number of sources and to the square root of the number of sources.

We define our two alternatives formally as follows. Define $\hat{N}$ as the nominal number of sources upon which the initial buffer and bandwidth allocation is calculated. Correspondingly, denote $\hat{x}$ and $\hat{c}$ as the minimum cost allocation of buffer and bandwidth such that $G(\hat{N}, \hat{x}, \hat{c}) = p$.

Denote the current number of sources as N, and the error in the estimate of the number of sources as $\Delta N = N - \hat{N}$. The fixed buffer (FB) resource allocation policy allocates a buffer of $x' = \hat{x}$ and a bandwidth of c', where c' is the value that satisfies $G(N, x', c') = p$. The incremental buffer (IB) resource allocation policy allocates a buffer of $x' = N/\hat{N}\hat{x}$ and a bandwidth of c', where c' is the value that satisfies $G(N, x', c') = p$.

The cost of the optimal policy is $C^* = p_x x^* + p_c c^*$ where x^* and c^* are the optimal buffer and bandwidth allocations as shown above. Expressing the bandwidth allocation as $c^* = N(\lambda/(1+\lambda) + \delta^*)$, we can break out the cost as

$$C^* = p_c N \frac{\lambda}{1+\lambda} + p_x x^* + p_c N \delta^*$$

Similarly, the cost of an alternate policy is

$$C' = p_c N \frac{\lambda}{1+\lambda} + p_x x' + p_c N \delta'$$

where $c' = N(\lambda/(1+\lambda) + \delta')$.

The cost savings is therefore

$$\Delta C = C' - C^* = p_x(x' - x^*) + p_c N(\delta' - \delta^*)$$

We should expect that the cost savings will be a function both of the nominal number of sources, $\hat{N}$, and of the error in the estimate of the number of sources, ΔN. Our principal result is:

Theorem 2: Consider either the FB or IB policy given above, with $\hat{N}$ as the nominal number of sources upon which the initial buffer and bandwidth allocation is calculated. Let C' represent the associated cost when the number of sources is N, as given above. Then the cost savings of the optimal policy over the alternate policy is:

$$\Delta C \sim \left(\frac{\Delta N}{\hat{N}}\right)^2 \sqrt{\hat{N}} \tag{25}$$

Proof:

For the FB policy,

$$x' - x^* \sim \sqrt{N} - \sqrt{\hat{N}} \sim \frac{\Delta N}{\sqrt{\hat{N}}}$$

provided that $\frac{\Delta N}{\hat{N}} << 1$.

For the IB policy,

$$x' - x^* \sim \sqrt{N} - \frac{N}{\hat{N}}\sqrt{\hat{N}} \sim \frac{\Delta N}{\sqrt{\hat{N}}}$$

provided that $\frac{\Delta N}{\hat{N}} << 1$.

Now all these policies (the optimal, FB, and IB) lie on the same buffer vs. bandwidth curve ($G(N, x, c) = p$). Furthermore, the optimal allocation is tangent to the minimum cost line. We use a Taylor series expansion about x^* for $C' - C^*$:

$$\Delta C(N) \approx \frac{\partial^2 c}{\partial x^2} \mid_{x^*} \frac{(x' - x^*)^2}{2} \tag{26}$$

The cost savings thus depends on the shape of the buffer versus bandwidth curve. We approximate this contour, by starting with the representation of it expressed in (23). Dropping the $O()$ terms, and substituting $y = \sqrt{x}$, we can restate this as

$$ay^2 + by + d \approx 0$$

where

$$\begin{aligned} a &= c_4\delta \\ b &= c_3\sqrt{\delta^3 N} \\ d &= c_2 N\delta^2 + ln(\frac{p\sqrt{N}\delta}{c_1}) \end{aligned}$$

Assuming that $\delta = O(1/\sqrt{N})$ and $x = O(\sqrt{N})$, we find that $a = O(1/\sqrt{N})$, $b = O(N^{-1/4})$, and $d = O(1)$. Since $y > 0$, it follows that

$$\begin{aligned} y &\approx \frac{-b + \sqrt{b^2 - 4ad}}{2a} \\ &\approx \frac{-b + b(1 - \frac{1}{2}\frac{4ad}{b^2})}{2a} \\ &\approx -\frac{d}{b} \end{aligned}$$

and thus

$$x \approx \frac{d^2}{b^2}$$

Differentiating x twice with respect to δ, and using $\delta = O(1/\sqrt{N})$ gives (after a lot of algebra)

$$\frac{\partial^2 x}{\partial \delta^2} = O(N^{3/2})$$

Consequently,

$$\frac{\partial^2 x}{\partial c^2} = O(\frac{1}{\sqrt{N}})$$

and thus

$$\frac{\partial^2 c}{\partial x^2} = O(\frac{1}{\sqrt{N}})$$

since $\frac{\partial c}{\partial x} = O(1)$.

Substituting this back into the Taylor series (26) and using $N \sim \hat{N}$,

$$\Delta C(N) \sim \frac{1}{\sqrt{\hat{N}}}(x' - x^*)^2$$

Finally using $x' - x^* \sim \frac{\Delta N}{\sqrt{\hat{N}}}$,

$$\Delta C(N) \sim \left(\frac{\Delta N}{\hat{N}}\right)^2 \sqrt{\hat{N}}$$

■

For numerical illustration, in figure 11, we plot the total buffer and bandwidth allocations for the optimal policy, for the fixed buffer policy, and for the incremental buffer policy.

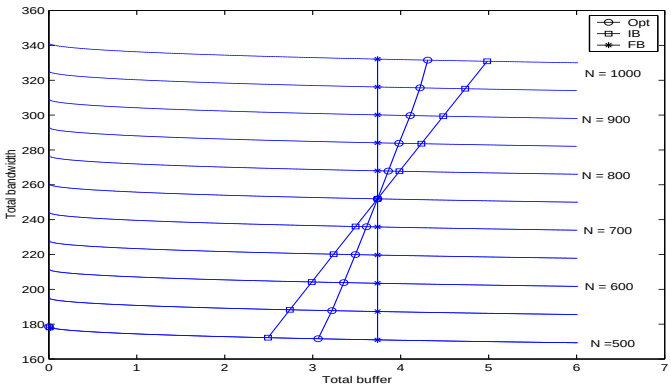

Fig. 11. Buffer versus bandwidth allocations for alternative policies

As mentioned above, the fixed buffer policy constitutes a *vertical line* through the set of contours, and the incremental buffer policy constitutes a curve with *fixed horizontal increments* through the set. We have set the nominal number of sources upon which the initial buffer and bandwidth allocation is calculated ($\hat{N}$) to be 750, and then varied the actual number of sources about this value. Correspondingly, when $N = 750$ all allocations are identical by definition. When N varies from this nominal value, the fixed buffer policy changes *only the bandwidth* so that the new allocation is on the new buffer vs. bandwidth contour. The incremental buffer policy varies the buffer *linearly*, and sets the bandwidth so that the new allocation is also on the new contour.

A close examination of figures 6 and 7 shows that for our set of parameters the slope of each contour, at a fixed total buffer, is increasing in magnitude with increasing N.

It follows that the optimal policy will increase the total buffer allocation with N in order to maintain a constant slope equal to the price ratio. Similarly, an examination of figure 8 shows that the slope of each contour, at a fixed buffer per source, is *decreasing* with increasing N. It follows that the optimal policy will decrease the buffer allocation per source with N in order to maintain a constant slope. Thus, for our set of parameters, the optimal policy lies strictly between the fixed buffer and incremental buffer policies.

The analysis for the cost comparison explains figure 2, which shows the cost differences between the optimal policy and FB and IB. As in figure 11, N is varied about the nominal value of $\hat{N} = 750$. All three policies are generated directly using Morrison's expression for overflow probability. The Taylor series analysis above (25) predicts that the resulting cost savings should be quadratic in ΔN for a fixed $\hat{N}$ (for small values of ΔN). The plot agrees well with this form. The asymmetry can be attributed to the presence of a third order term, which was neglected in the analysis.

In figure 12, we plot the cost differences between the optimal policy and FB and IB, but with a fixed percentage error between the nominal and actual number of sources $\frac{N-\hat{N}}{N} = 0.2$.

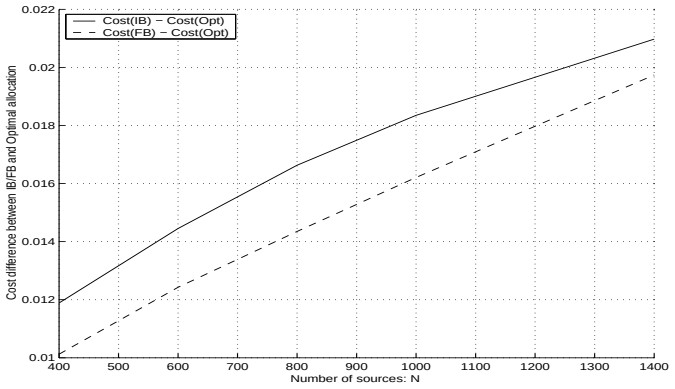

Fig. 12. Cost difference between optimal and alternate policies for a fixed $\frac{N-\hat{N}}{N}$

The Taylor series analysis (25) predicts that the resulting cost savings should be proportional to the square root of N for a fixed percentage error. The plot agrees quite well with this form.

III. General sources

Our goal in this section is to explore the shape of the variation of the optimal bandwidth and buffer allocations with respect to the number of sources for a more general class of sources.

A. The Network Model

We again consider a single queue fed by N sources. In contrast to the assumption in previous sections that the sources are i.i.d. on/off fluid flows, we now allow any general form provided that the aggregate effective bandwidth is a decreasing convex function of buffer and linearly proportional to the number of sources.

The convexity property is satisfied by many effective bandwidth derivations in the literature. The assumption that effective bandwidth scales linearly with respect to the number of sources, however, is clearly inaccurate, as demonstrated in the literature on effective bandwidth and in the previous section. The literature on multiplexing, however, has often proposed the view that multiplexing gains come from two sources. First, variance in the distribution of the rate of sources *at a fixed time* give rise to efficiencies when multiple sources share a common bandwidth (even with no buffer). Second, variation *over time* in the rate of a single source give rise to efficiencies when that source is buffered (and therefore smoothed). We view the results in this section as descriptive of the second type of multiplexing gain (smoothing).

As above, we denote the aggregate bandwidth by c, the aggregate buffer by x, the overflow probability by $G(N, x, c/N)$, and the maximum acceptable overflow probability by p. We denote the effective bandwidth for a single flow by

$$eb(x) \equiv c \mid [G(1, x, c/N) = p]$$

and the effective bandwidth for N multiplexed flows by

$$eb(N, x) \equiv N eb(x)$$

B. Optimal Resource Allocation

As above, we assume that each unit of bandwidth incurs a cost p_x and each unit of buffer incurs a cost p_c. We denote the optimal buffer allocation by

$$x^*(N) = \arg\min_x [p_x x + p_c eb(N, x)]$$

and the resulting optimal cost by

$$C^*(N) = p_x x^*(N) + p_c N eb(x^*(N))$$

It follows that the slope of the aggregate effective bandwidth with respect to the allocated buffer, at the optimal point, must be equal to the price ratio:

$$\frac{\partial N eb(x)}{\partial x} \mid_{x^*(N)} = N \frac{\partial eb(x)}{\partial x} \mid_{x^*(N)} = -\frac{p_x}{p_c} = -1/m$$

whenever $x^*(N) > 0$.

The constant cost contour and optimal allocation are illustrated in figure 13.

Our principal result in this section is:

Theorem 3: The optimal buffer assignment is strictly increasing with the number of sources, N, when $x^*(N) > 0$.

Proof: The proof is by contradiction. Suppose that $x^*(N) \geq x^*(N+1)$. It follows that

$$N\frac{\partial eb(x)}{\partial x}\mid_{x^*(N)} = (N+1)\frac{\partial eb(x)}{\partial x}\mid_{x^*(N+1)} = -1/m$$

and therefore that

$$\frac{\frac{\partial eb(x)}{\partial x}\mid_{x^*(N+1)}}{\frac{\partial eb(x)}{\partial x}\mid_{x^*(N)}} = \frac{N}{N+1} < 1$$

However if $x^*(N) \geq x^*(N+1)$, then this violates the assumption that $eb(x)$ is a decreasing convex function. ■

This theorem can be compared to Theorem 1, which states that for on/off sources the optimal buffer allocation is proportional to $\sqrt{N}$. Theorem 3 considers a wider class of flows, but is weaker than Theorem 1 in that it only guarantees that the buffer allocation is increasing.

C. Comparison to Alternative Schemes

In this section, we compare the cost of the optimal resource allocation to a Fixed Buffer policy. The cost of the Fixed Buffer policy, using $\hat{N}$ as the nominal number of sources upon which the initial buffer and bandwidth allocation is calculated, is

$$\begin{aligned} C_{FB}(N) &= p_x x^*(\hat{N}) + p_c N eb(x^*(\hat{N})) \\ &= C^*(\hat{N}) + p_c(N-\hat{N})eb(x^*(N)) \end{aligned}$$

Denote the cost savings of the optimal policy over the FB policy by:

$$\Delta C_{FB}(N,\hat{N}) \equiv C_{FB}(N) - C^*(N)$$

Our principal result in this section is:

Theorem 4: $\Delta C_{FB}(N,\hat{N})$ is increasing and convex in $|N-\hat{N}|$, when $x^*(N) > 0$.

Proof:

Substituting expressions for $C_{FB}(N)$ and $C^*(N)$ from above,

$$\begin{aligned} \Delta C_{FB}(N,\hat{N}) = &-p_x[x^*(N) - x^*(\hat{N})] \\ &+p_c N[eb(x^*(\hat{N}) - eb(x^*(N))] \end{aligned}$$

Without loss of generality, assume that $N > \hat{N}$. This expression can be written as:

$$\begin{aligned} \Delta C_{FB}(N,\hat{N}) &= p_c\{-N[eb(x^*(N)) - eb(x^*(\hat{N}))] \\ &- \frac{1}{m}[x^*(N) - x^*(\hat{N})]\} \\ &= p_c \int_{x^*(\hat{N})}^{x^*(N)} \left[-N\frac{\partial eb(x)}{\partial x} - \frac{1}{m}\right] dx \end{aligned}$$

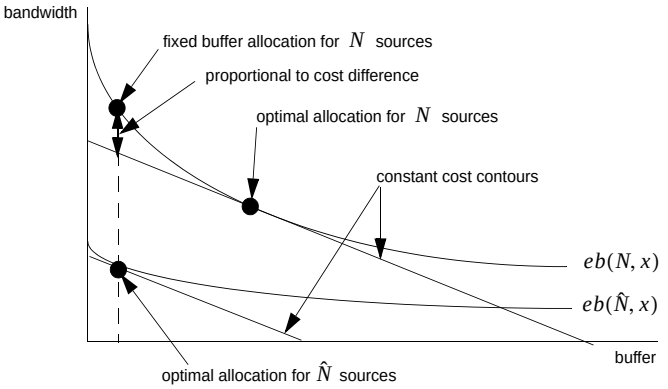

Fig. 13. Illustration of cost difference

This last expression can be viewed as p_c times the vertical distance between the aggregate effective bandwidth curve and the tangent line to the curve at the nominal allocation, evaluated at N sources. This vertical distance is illustrated in figure 13.

Using similar expressions for $\Delta C_{FB}(N+1,\hat{N})$ and $\Delta C_{FB}(N+2,\hat{N})$, we can represent second order differences as:

$$\begin{aligned} &\Delta C_{FB}(N+1,\hat{N}) - \Delta C_{FB}(N,\hat{N}) = \\ &p_c \int_{x^*(N)}^{x^*(N+1)} \left[-(N+1)\frac{\partial eb(x)}{\partial x} - \frac{1}{m}\right] dx \\ &\qquad + p_c \int_{x^*(\hat{N})}^{x^*(N)} \left[-\frac{\partial eb(x)}{\partial x}\right] dx \end{aligned} \tag{27}$$

and as:

$$\begin{aligned} &\Delta C_{FB}(N+2,\hat{N}) - \Delta C_{FB}(N+1,\hat{N}) = \\ &p_c \int_{x^*(N+1)}^{x^*(N+2)} \left[-(N+2)\frac{\partial eb(x)}{\partial x} - \frac{1}{m}\right] dx \\ &\qquad + p_c \int_{x^*(\hat{N})}^{x^*(N+1)} \left[-\frac{\partial eb(x)}{\partial x}\right] dx \end{aligned} \tag{28}$$

In equation 27, the first integral is an integral of a positive quantity (since $eb(x)$ is decreasing and convex and $\frac{\partial (N+1)eb(x)}{\partial x}\mid_{x^*(N+1)} = -1/m$) over a positive range (from Theorem 3). The second integral is also an integral of a positive quantity (since $eb(x)$ is decreasing) over a positive range. The sum therefore is positive, establishing that $\Delta C_{FB}(N,\hat{N})$ is increasing in N when $N > \hat{N}$, or more generally increasing in $|N-\hat{N}|$.

We can establish convexity by considering the third order differences. Subtracting the second order differences

(equation 27 from equation 28) and collecting terms, we obtain:

$$[\Delta C_{FB}(N+2,\hat{N}) - \Delta C_{FB}(N+1,\hat{N})] - [\Delta C_{FB}(N+1,\hat{N}) - \Delta C_{FB}(N,\hat{N})] = \quad (29)$$

$$p_c \int_{x^*(N)}^{x^*(N+1)} \left[\frac{1}{m} + N\frac{\partial eb(x)}{\partial x}\right] dx$$

$$+p_c \int_{x^*(N+1)}^{x^*(N+2)} \left[-(N+2)\frac{\partial eb(x)}{\partial x} - \frac{1}{m}\right] dx \quad (30)$$

Similar to the previous integrals, these can be shown to be positive, using the decreasing convexity property of $eb(x)$ and using Theorem 3. The sum is therefore positive, and it follows that $\Delta C_{FB}(N,\hat{N})$ is convex in $|N-\hat{N}|$ when $x^*(N) > 0$. ■

This theorem can be compared to Theorem 2, which states that for on/off sources the equivalent expression for the cost difference is proportional to the square of $|N-\hat{N}|$. Theorem 4 considers a wider class of flows, but is weaker than Theorem 2 in that it only guarantees that the cost difference in increasing and convex in $|N-\hat{N}|$.

IV. Conclusion

We first considered a single node which multiplexes a large number of on/off fluid flows. Under a maximum overflow probability, we proved that the optimal bandwidth allocation above the mean rate and the optimal buffer allocation are both proportional to the *square root of the number of sources*. This is in contrast to current approaches which often allocate either a *fixed total buffer* or a *fixed buffer per source*. We compared the optimal allocation to these alternative allocations, and proved that the excess cost incurred by a fixed buffer allocation or by linear buffer allocations is proportional to the square of the percentage difference between the assumed number of sources and the actual number of sources and to the square root of the number of sources. These properties were verified by numerical results.

We next considered a class of general i.i.d. sources for which the aggregate effective bandwidth is a decreasing convex function of buffer and linearly proportional to the number of sources. We proved that the optimal buffer allocation is strictly increasing with the number of sources. We also proved that the excess cost incurred by a fixed buffer allocation is an increasing convex function of the difference between the assumed number of sources and the actual number of sources. Both results are consistent with, but weaker than, the corresponding on/off sources, but hold for a wider class of flows.

*** merge reference lists and update ??? ***

References

[1] F.P. Kelly, "Effective bandwidths at multi-class queues," *Queueing Systems*, 1991.

[2] G. Kesidis, J. Walrand, and C.-S. Chang, "Effective bandwidths for multiclass markov fluids and other atm sources," *IEEE/ACM Transactions on Networking*, vol. 1, no. 4, pp. 424–428, Aug. 1993.

[3] A.I. Elwalid and D. 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. 329–344, June 1993.

[4] G.L. Choudhary, D. M. Lucantoni, and W. Whitt, "On the effectiveness of effective bandwidths for admission control in ATM networks," in *International Teletraffic Congress*, 1994.

[5] Cheng-Shang Chang, "Effective bandwidth in high-speed digital networks," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 1091–1100, Aug. 1995.

[6] Anwar Elwalid, Debasis Mitra, and Robert H. Wentworth, "A new approach for allocating buffers and bandwidth to heterogeneous, regulated traffic in an atm node," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 1115–1127, Aug. 1995.

[7] Alain Simonian and Jacky Guibert, "Large deviations approximation for fluid queues fed by a large number of on/off sources," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 1017–1027, Aug. 1995.

[8] David N.C. Tse and Robert G. Gallager, "Statistical multiplexing of multiple time-scale markov streams," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 1028–1038, Aug. 1995.

[9] Arthur W. Berger and Ward Whitt, "Effective bandwidths with priorities," *TN*, vol. 6, no. 4, pp. 447–460, August 1998.

[10] Debasis Mitra, John A. Morrison, and K.G. Ramakrishnan, "Atm network design and optimization: a multirate loss network framework," *IEEE/ACM Transactions on Networking*, vol. 4, no. 4, pp. 531–543, Aug. 1996.

[11] Harry G. Perros and Khaled M. Elsayed, "Call admission control schemes: a review," *IEEE Communications Magazine*, vol. 34, no. 11, pp. 82–91, Nov. 1996.

[12] Steven Low and Pravin Varaiya, "A new approach to service provisioning in atm networks," *IEEE/ACM Transactions on Networking*, vol. 1, no. 5, pp. 547–553, Oct. 1993.

[13] Hong Jiang and Scott Jordan, "The role of price in the connection establishment process," *European Transactions on Telecommunications*, vol. 6, no. 4, pp. 421–429, July-August 1995.

[14] Jeffrey K. MacKie-Mason, Liam Murphy, and John Murphy, "On the role of responsive pricing in the internet," in *Internet Economics*, J. Bailey and L. McKnight, Eds., pp. 279–304. MIT Press, 1996.

[15] J.A. Morrison, "Asymptotic analysis of a data-handling system with many sources," *SIAM J. Appl. Math*, vol. 49, no. 2, pp. 617–637, April 1989.

[16] Anwar Elwalid, D. Heyman, T.V. Lakshman, D. Mitra, and A. Weiss, "Fundamental bounds and approximations for atm multiplexors with applications to video teleconferencing," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 1004–1016, Aug. 1995.

[17] M. Montgomery and G. DeVeciana, "On the relevance of time scales in performance oriented traffic characterizations," in *Proceedings of the Conference on Computer Communications (IEEE Infocom)*, San Francisco, CA, Mar. 1996, pp. 513–520.

[18] Ness B. Shroff and Mischa Schwartz, "Improved loss calculations

at an ATM multiplexor," *TN*, vol. 6, no. 4, pp. 411–421, August 1998.

[19] D. Anick, D. Mitra, and M.M. Sondhi, "Stochastic theory of a data-handling system with multiple sources," *Bell System Tech. J.*, vol. 61, pp. 1871–1894, 1982.