Title
The variation of optimal bandwidth and buffer allocation with the number of sources

Permalink
https://escholarship.org/uc/item/65z2d4rd

Journal
IEEE-ACM Transactions on Networking, 12(6)

ISSN
1063-6692

Authors
Jordan, Scott
Jogi, Kalpana
Shi, Chunlin
et al.

Publication Date
2004-12-01

DOI
10.1109/TNET.2004.838603

Peer reviewed

The Variation of Optimal Bandwidth and Buffer Allocation With the Number of Sources

Scott Jordan, *Member, IEEE*, Kalpana Jogi, Chunlin Shi, and Ikhlaq Sidhu, *Member, IEEE*

***Abstract*—We consider a single node which multiplexes a large number of traffic sources. We ask a simple question: how do the optimal allocations of bandwidth and buffer vary with the number of sources? We investigate this issue using previous results on the probability of overflow for an aggregate of i.i.d. flows, e.g., overflow resulting from effective bandwidth models. We wish to determine the variation of the minimum cost allocations of bandwidth and buffer with the number of sources, 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.**

***Index Terms*—Cost minimization, dimensioning, resource allocation.**

I. Introduction

A. Background

There is now a rich literature on the use of effective bandwidth to estimate the bandwidth and buffer requirements of network traffic sources, particularly for sources with real-time loss and delay constraints. In this paper, we ask a simple question: how do the optimal allocations of bandwidth and buffer vary with the number of sources? We investigate this issue using previous results on the probability of exceeding a delay bound for an aggregate of i.i.d. flows sharing a single queue, e.g., that resulting from effective bandwidth models. We wish to determine the variation of the minimum cost allocations of bandwidth and buffer with the number of sources, given a cost per unit of each resource.

For background, we briefly outline the use of effective bandwidth models for resource allocation and dimensioning. Many effective bandwidth models considered a single traffic flow or the flow resulting from the multiplexing of multiple identical sources. Typically, the loss probability of the flow is estimated by the probability that the buffer content in an infinite buffer queue will exceed a threshold. The resulting loss probability estimate is thus interpreted as the probability of exceeding a delay bound. Many papers have shown that the loss probability estimate $b(x)$ asymptotically obeys

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

as the buffer approaches infinity for a fixed bandwidth, 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*. Similar asymptotic expressions have been proven for a wide variety of source models, c.f. [1]–[7]. Other effective bandwidth models directly considered multiplex i.i.d. traffic flows in the many sources regime, in which the number of sources N, the total bandwidth allocation c, and the total buffer allocation x all approach infinity with fixed ratios. Similar asymptotic results to (1) have been shown, c.f. [8]–[16].

The most common use of such results is to predict the loss probability given a specific bandwidth and buffer allocation. In addition, many papers have used these results to formulate admission control policies, c.f. [17], [18]. A typical approach, 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, is to accept a new connection if and only if the available bandwidth exceeds the *effective bandwidth* that results from (1).

Our focus, however, is in on dimensioning. We are interested in the minimum cost allocation of bandwidth and buffer that can accommodate N flows with a maximum loss probability of p. As many previous researchers have demonstrated, a set of flows can achieve a maximum loss probability using various combinations of total shared bandwidth and buffer, c.f. [19]. Only a few algorithms, however, have been proposed to solve the joint bandwidth and buffer allocation problem, c.f. [20], [21]. Our goal in this research effort is to understand how the optimal combination of bandwidth and buffer might vary with the number of flows.

To define optimality, we assume that there are costs associated with each unit of bandwidth and buffer. The ratio of the cost per unit bandwidth to the cost per unit buffer should reflect the relative demand for bandwidth to buffer from all of the traffic flowing through the router. This ratio might be based on

Manuscript received March 18, 2002; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor S. Low. This work was supported by the National Science Foundation under Grant 9523580, Grant 9903055, and Grant 0137103, and by the Defense Advanced Research Projects Agency under Grant N66001-00-8935.

S. Jordan is with the Department of Electrical Engineering and Computer Science, University of California, Irvine, CA 92697 USA (e-mail: sjordan@uci.edu).

K. Jogi is a consultant in Foster City, CA 94404 USA (e-mail: kalpanajogi@yahoo.com).

C. Shi is with Maxspeed Corporation, Palo Alto, CA 94303 USA (e-mail: rshi@maxspeed.com).

I. Sidhu is with the University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (e-mail: isidhu@uiuc.edu).

Digital Object Identifier 10.1109/TNET.2004.838603

1063-6692/04$20.00 © 2004 IEEE

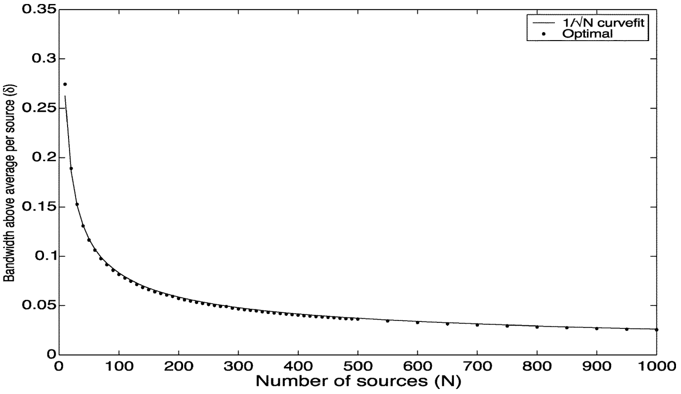

Fig. 1. Optimal bandwidth allocations versus N.

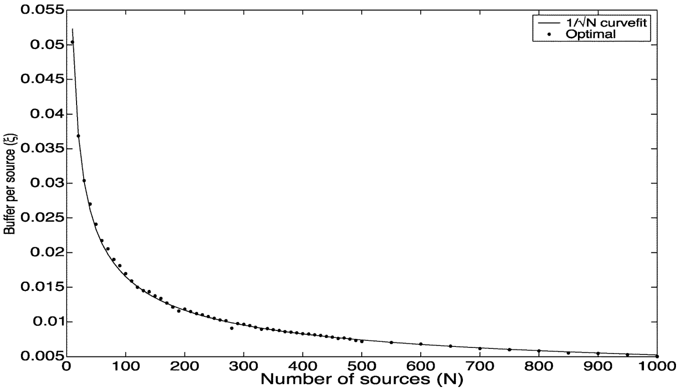

Fig. 2. Optimal buffer allocations versus N.

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, c.f. [19], [22], [23]. The ratio of the prices can also be interpreted directly as the slope of the buffer–bandwidth tradeoff curve at the desired operating point. It has been shown that this tradeoff curve is convex for a wide range of sources with effective bandwidths [21]. We define the optimal combination of bandwidth and buffer as the minimum cost choice that achieves the desired QoS. In this paper, we equate QoS with estimated loss probability, but it is simple to add a limit on buffer in order to enforce a maximum delay constraint. This approach, of course, does not address other possible QoS measures.

The many-sources regime might lead one to believe that total bandwidth and buffer allocated to a single class should be in constant proportion as the number of sources vary. We show, however, for a class of ON/OFF fluid flows, 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 and thus that bandwidth and buffer should not be allocated in constant proportion. We believe that these results are the first to characterize the variation of optimal bandwidth and buffer allocations with the number of sources.

B. Motivating Example

As a motivating example, consider a single node which multiplexes compressed real-time voice sources, modeled as ON/OFF fluid flows with exponentially distributed ON and OFF times with respective means of 340 and 780 ms and a peak rate of 8 kb/s. 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 8 kb/s of bandwidth is equally expensive as 340 bytes of buffer).

Using analytic expressions for overflow probability derived by Anick *et al.* [24], we can numerically derive the minimum cost bandwidth and buffer allocations. The results are shown in Figs. 1 and 2, 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.

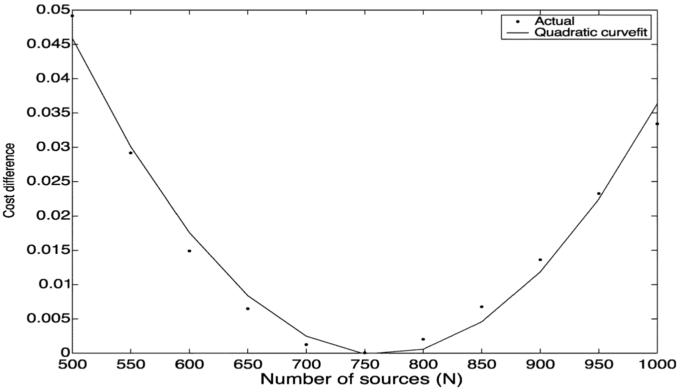

Fig. 3. Cost difference between optimal policy and FB policy for a fixed $\hat{N}$.

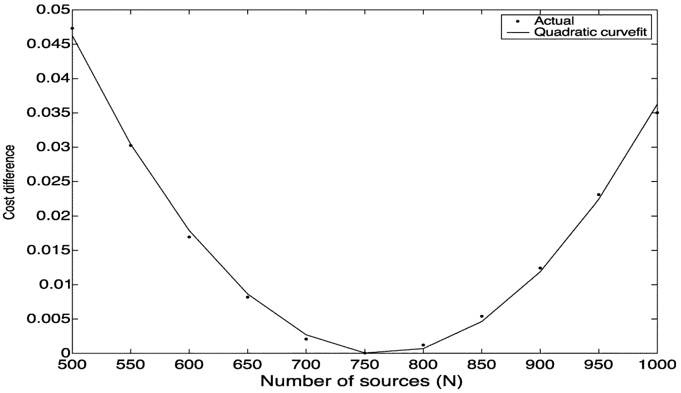

Fig. 4. Cost difference between optimal policy and IB policy for a fixed $\hat{N}$.

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. An *Incremented Buffer* (IB) policy allocates a constant amount of buffer *per source* and adjusts the bandwidth to satisfy the loss constraint.

The results in Figs. 1 and 2 do not correspond to either an FB or an IB policy. The optimal resource allocation policy is neither to fix the buffer length and then add bandwidth nor to add bandwidth and buffer in constant proportion. Indeed, we can numerically compare the optimal allocation policy to these two alternate policies. The results are shown in Figs. 3 and 4, 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 ($|N - 750|$).

Our goal in this paper is to explain the forms of the curves in Figs. 1–4.

C. Principal Results

We first consider a single node which multiplexes a large number of i.i.d. ON/OFF fluid flows with exponentially distributed ON and OFF times, 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 bandwidth and buffer. The costs are then used to determine the optimal choice of bandwidth and buffer. Our principal result is that the optimal bandwidth is given by

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

and the optimal buffer is given by

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

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 bandwidth and buffer costs) is to not add bandwidth and buffer in a constant proportion, but instead to first add the mean bandwidth of each source and then to add additional bandwidth and buffer in an approximately constant proportion. Furthermore, we demonstrate that 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 [25], which is a direct exploitation of analytic expressions presented in [24]. These results can be viewed as refinements to the early large deviations results presented in [8]. Large deviations results in the many-sources regime can produce more accurate estimates of loss probability, particularly with regard to the asymptotic constant, c.f. [15], [16], [26]. However, our goal in this study is to obtain an asymptotic relationship between the optimal bandwidth and buffer allocation and the number of sources. This requires a simple representation of overflow probability as a function of both bandwidth and buffer. In contrast to large deviations results in the many-sources regime, in which bandwidth and buffer are scaled linearly with the number of sources, Morrison's results apply to independently chosen bandwidth and buffer, for a wide range of buffer sizes that bracket those in (2) and (3).

It is worth stressing at this point that we are certainly not proposing that Morrison's expression for the overflow probability, or our Taylor series expansion of it, be used to predict overflow probability, as we do not believe any Taylor series expansion would be an accurate predictor of overflow probability. We validate our results using numerical investigations which show that the errors introduced by either the Morrison approximation or the Taylor series do not affect our principal results.

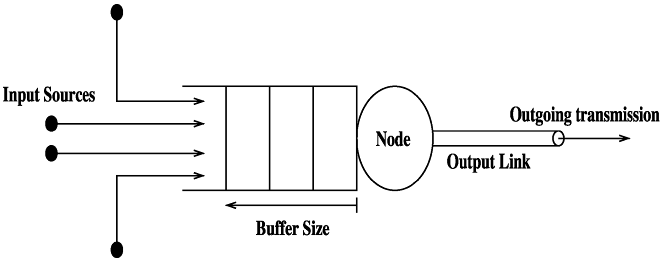

Fig. 5. Network model.

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 bandwidth and buffer allocations and present our principal results for ON/OFF sources. In Section III, we consider general sources.

II. ON/OFF SOURCES

A. Network Model

We consider a single queue fed by N i.i.d. ON/OFF fluid sources, as shown in Fig. 5. 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 $\lambda/(1+\lambda)$.

In numerical examples, we use the parameters given in the motivating examples above. Using our normalization, with bandwidth measured in multiples of 8 kp/s and buffer measured in multiples of 340 B, this gives $\lambda \approx 0.4359$ and $\lambda/(1+\lambda) \approx 0.3036$.

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 as

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

In numerical examples, unless explicitly mentioned we set $N = 750$, $\delta = .03$ and the maximum probability of overflow $p = 0.01$.

We briefly restate the expressions for overflow probability derived by Morrison [25]. His derivation starts with earlier work by Anick *et al.* [24], which states that the equilibrium probability that the buffer content exceeds x in an infinite buffer system can be expressed as

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

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 $G_A(N, x, \gamma)$ 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 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 as follows:

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

where

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

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

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

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

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

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

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.

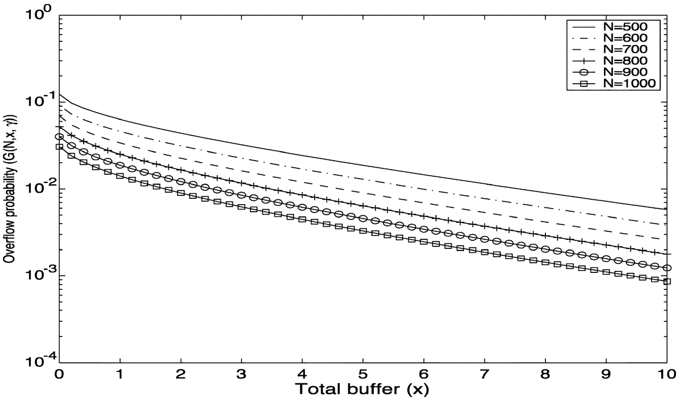

Fig. 6. Overflow probability for a range of N.

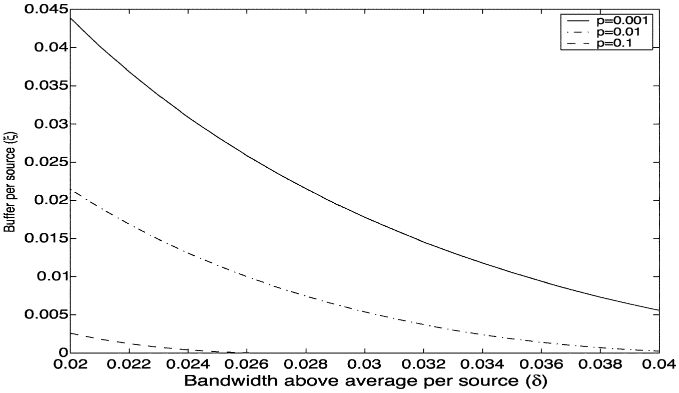

Fig. 7. Buffer versus bandwidth contours for a range of p.

B. Numerical Examples

To illustrate the basic problem considered 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 Fig. 6, the overflow probability $G_A(N, x, \gamma)$ is plotted for a range of N for a fixed bandwidth per source γ.

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 versus buffer curves.

An alternative view is shown in Fig. 7, 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 bandwidth and buffer produce the same overflow probability. Note that there is a substantial range of slopes along each contour. The optimal resource allocation policy will choose bandwidth and buffer to equate the slope of the contour with the corresponding price ratio. Alternate policies such as FB or IB do not take into account the

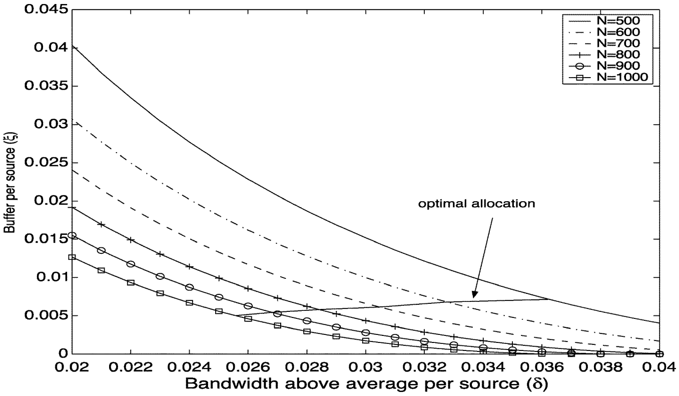

Fig. 8. Buffer per source versus bandwidth above average per-source contours for a range of N.

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 FB or IB policies.

Buffer versus bandwidth contours for fixed p but varying N are shown in Fig. 8. The majority of the bandwidth allocation is due to the mean rate, which must be allocated (at loss overflow probabilities) under any resource-allocation policy. We therefore plot the bandwidth per source above the mean rate. Multiplexing gains mean that larger N correspond to *larger* bandwidth and buffers, but with decreasing increments. When plotted per source, multiplexing gains mean that larger N correspond to *lower* bandwidth and buffers per source. Note again that there is a large range of slopes, indicating that the optimal policy can adjust the allocations significantly. A fixed buffer policy would constitute a curve through the set, while an incremented buffer policy would constitute a horizontal line. The cost minimizing choices of bandwidth and buffer are also shown.

C. Taylor Series Expansions

In this section, we develop Taylor series approximations for the overflow probability (5).

Theorem 1: A Taylor series representation of $G_M(N, x, \gamma)$, in N, δ, and x, is given by (12), shown at the bottom of the page, where

$$\begin{aligned}\delta &= \gamma - \frac{\lambda}{1+\lambda}\\ c_1 &= \sqrt{\frac{3\lambda}{2\pi}}\frac{1}{(1+\lambda)}\\ c_2 &= \frac{(1+\lambda)^2}{2\lambda}\\ c_3 &= 2\sqrt{\frac{(1+\lambda)^5}{6\lambda^2}}\\ c_4 &= \frac{\lambda+1}{2\lambda}(5\lambda^2 - 3\lambda + 2).\end{aligned} \tag{13}$$

Proof: We start by expanding the constituent parts of (5) expressed in (6)–(11). The general approach is to expand the expression using

$$\gamma = \frac{\lambda}{1+\lambda} + \delta \tag{14}$$

for $\delta \ll 1$.

Substituting (14) into the first term of (6), we obtain

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

Using the Taylor series expansion

$$ln(1+z) = z - \frac{z^2}{2} + \frac{z^3}{3} - \frac{z^4}{4} + \cdots \tag{15}$$

this reduces to

$$\begin{aligned}&(1+\lambda)\left(1+\frac{1}{\lambda}\right)\delta + \frac{1}{2}(1+\lambda)^2\left(1-\frac{1}{\lambda^2}\right)\delta^2 + \frac{1}{3}(1+\lambda)^3\\ &\quad\times\left(1+\frac{1}{\lambda^3}\right)\delta^3 + \frac{1}{4}(1+\lambda)^4\left(1-\frac{1}{\lambda^4}\right)\delta^4 + O(\delta^5)\end{aligned} \tag{16}$$

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

Using the Taylor series expansion

$$\frac{1}{1+z} = 1 - z + z^2 - z^3 + \cdots \tag{17}$$

the second term in $f(\gamma)$ reduces to

$$\begin{aligned}&\frac{(1+\lambda)^2}{\lambda}\delta - \frac{(1+\lambda)^3(1-\lambda)}{2\lambda^2}\delta^2 + \frac{(1+\lambda)^4(1-\lambda)^2}{4\lambda^3}\delta^3\\ &\qquad - \frac{(1+\lambda)^5(1-\lambda)^3}{8\lambda^4}\delta^4 + O(\delta^5).\end{aligned}$$

Together these two terms give

$$f(\gamma) = \frac{(1+\lambda)^6}{12\lambda^3}\delta^3 - \frac{(1+\lambda)^7(1-\lambda)}{8\lambda^4}\delta^4 + O(\delta^5). \tag{18}$$

We next consider (7). Substituting (14), this gives

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

which, using (17), reduces to

$$r = \frac{(1+\lambda)^3}{\lambda}\delta + O(\delta^2). \tag{19}$$

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

$$\kappa(\gamma) = \gamma ln\left[\frac{\gamma}{(1-\gamma)\lambda}\right] + ln\left[(1-\gamma)(1+\lambda)\right]. \tag{20}$$

An approximation for the first log term was found above to be (16). Multiplying by γ and reducing, we obtain

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

$$G_M(N, x, \gamma) = \left(\frac{c_1}{\sqrt{N}\delta} + O\left(\frac{1}{\sqrt{N}}\right)\right) e^{-\left(c_2 N\delta^2 + c_3\sqrt{\delta^3 N x} + c_4\delta x + O(N\delta^3) + O\left(\delta^{\frac{5}{2}} N^{\frac{1}{2}} x^{\frac{1}{2}}\right) + O(\delta^2 x)\right)} \tag{12}$$

The second term in (20) can be expressed as

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

Using (15), this becomes

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

Putting together these expressions for the two terms in (20), we find

$$\kappa(\gamma)=\frac{(1+\lambda)^2}{2\lambda}\delta^2+O(\delta^3). \tag{21}$$

We continue with (10). The numerator reduces to

$$-(1-\lambda)(1+\lambda)^2\delta^3+2(1+\lambda)^3\delta^4$$

and the denominator reduces to

$$\frac{8\lambda^4}{(1+\lambda)^5}+\frac{20\lambda^3(1-\lambda)}{(1+\lambda)^4}\delta+O(\delta^2).$$

Using (17), this can be expressed as

$$\begin{aligned}\rho''(1-\gamma)=&-\frac{(1+\lambda)^7(1-\lambda)}{8\lambda^4}\delta^3\\&+\frac{(1+\lambda)^8(5\lambda^2-6\lambda+5)}{16\lambda^5}\delta^4+O(\delta^5).\end{aligned} \tag{22}$$

We continue with (11). A similar approach using (17) results in

$$k=\frac{3(1-\lambda)}{2}-\frac{(1+\lambda)^3}{4\lambda}\delta+O(\delta^2). \tag{23}$$

Finally, we consider (9). Using (17), (18), and (22), the term $(\rho''(1-\gamma)/f(\gamma))$ can be represented as

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

The second term of (9) thus becomes

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

Finally, using (23), we can obtain

$$g(\gamma)=\frac{(1+\lambda)(5\lambda^2-3\lambda+2)}{2\lambda}\delta+O(\delta^2). \tag{24}$$

This completes the development of Taylor series expansions for (6)–(11). We now use these expressions to derive the Taylor series expansion for the overflow probability (5). Using (18) and (19), the first term can be expressed as

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

Using (21), the second term can be expressed as

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

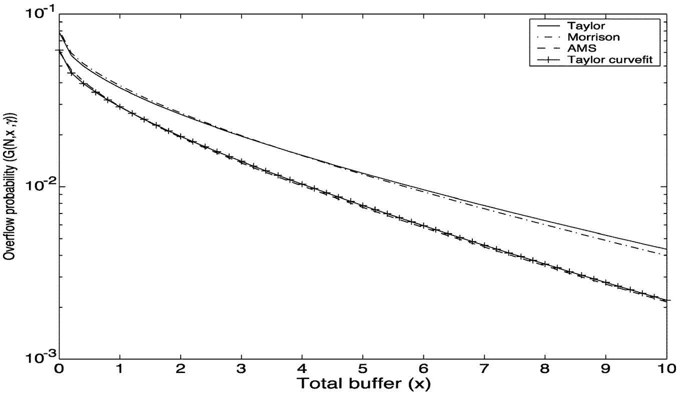

Fig. 9. Accuracy of overflow representations.

Similarly, using (18), the third term can be expressed as

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

Using (24), the fourth term can be expressed as

$$e^{-g(\gamma)x}=e^{-\frac{(1+\lambda)(5\lambda^2-3\lambda+2)}{2\lambda}\delta x+O(\delta^2x)}.$$

Finally, combining these four terms, we get (12). ■

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 Figs. 1–4. 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 (12) with Morrison's approximation for overflow probability (5) and with the exact expression (4) in Fig. 9.

The Taylor series approximation to Morrison's representation is reasonably good. A better approximation can be obtained by retaining one more term in each expansion above; however, these additional terms do not affect the principal results given in (2) and (3) and therefore we do not include them in our analysis. There is a substantial error in Morrison's approximation to the exact AMS result, however, as discussed above, more recent large deviations results in the many-sources regime typically assume that bandwidth and buffer are scaled linearly with the number of sources and cannot be used to derive our results.

In Fig. 9, we also show the overflow probability that results from a least-squares curve fit in the form of (12) to the exact AMS results. The accuracy of this result may indicate that, if such a Taylor series could be derived directly, it would produce the same form.

D. Optimal Resource Allocation

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

Theorem 2: Suppose that each unit of buffer incurs a cost p_x and each unit of bandwidth incurs a cost p_c. Assume that

$G_M(N,x,\gamma)$ is decreasing and jointly convex in x and γ. The bandwidth and buffer allocations that minimize cost subject to a maximum overflow probability of p are

$$\begin{aligned} \gamma^* &= k_1^*\sqrt{N} + O(1) \\ x^* &= k_2^*\sqrt{N} + O(1) \end{aligned} \tag{25}$$

where $m = p_c/p_x$ and k_1^* and k_2^* are the solutions to

$$c_2k_1^{*2} + c_3k_1^{*\frac{3}{2}}k_2^{*\frac{1}{2}} + c_4k_1^*k_2^* + ln\left(\frac{pk_1^*}{c_1}\right) = 0 \tag{26}$$

$$\begin{aligned} &mc_3k_1^{*\frac{5}{2}} + (mc_4 - 2c_2)k_1^{*2}k_2^{*\frac{1}{2}} \\ &- \frac{3c_3}{2}k_1^{*\frac{3}{2}}k_2^* - c_4k_1^*k_2^{*\frac{3}{2}} - k_2^{*\frac{1}{2}} = 0 \end{aligned} \tag{27}$$

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

Proof: We start with the constraint $G_M(N,x,\gamma) = p$, with $G_M(N,x,\gamma)$ given by (12). Taking logarithms on both sides and rearranging, we obtain

$$\begin{aligned} &c_2N\delta^2 + c_3\sqrt{\delta^3Nx} + c_4\delta x \\ &\quad + O(N\delta^3) + O\left(\delta^{\frac{5}{2}}N^{\frac{1}{2}}x^{\frac{1}{2}}\right) + O(\delta^2 x) \\ &= -ln\left(\frac{p}{\frac{c_1}{(\sqrt{N}\delta)} + O\left(\frac{1}{\sqrt{N}}\right)}\right). \end{aligned} \tag{28}$$

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

$$-\frac{\frac{\partial G_M}{\partial x}}{\frac{\partial G_M}{\partial c}} = -\frac{p_x}{p_c} = -\frac{1}{m}. \tag{29}$$

Differentiating (12) and substituting into (29) gives

$$\begin{aligned} &mc_4\delta + \frac{mc_3}{2}\sqrt{\frac{\delta^3N}{x}} + O\left(\delta^{\frac{5}{2}}N^{\frac{1}{2}}x^{-\frac{1}{2}}\right) + O(\delta^2) \\ &= \frac{1}{N\delta} + 2c_2\delta + \frac{3c_3}{2}\sqrt{\frac{x\delta}{N}} + c_4\frac{x}{N} \\ &\quad + O(\delta^2) + O\left(\delta^{\frac{3}{2}}N^{-\frac{1}{2}}x^{\frac{1}{2}}\right) + O\left(\frac{\delta x}{N}\right). \end{aligned} \tag{30}$$

Now suppose that $\delta = k_1/\sqrt{N}$ and $x = k_2\sqrt{N}$ for some $k_1 = O(1)$ and $k_2 = O(1)$. By substitution into (28), we obtain

$$c_2k_1^2 + c_3k_1^{\frac{3}{2}}k_2^{\frac{1}{2}} + c_4k_1k_2 + ln\left(\frac{pk_1}{c_1}\right) = O\left(\frac{1}{\sqrt{N}}\right). \tag{31}$$

Similarly, by substitution into (30), we obtain

$$\begin{aligned} &mc_3k_1^{\frac{5}{2}} + (mc_4 - 2c_2)k_1^2k_2^{\frac{1}{2}} - \frac{3c_3}{2}k_1^{\frac{3}{2}}k_2 \\ &\qquad - c_4k_1k_2^{\frac{3}{2}} - k_2^{\frac{1}{2}} = O\left(\frac{1}{\sqrt{N}}\right). \end{aligned} \tag{32}$$

Let k_1^* and k_2^* be the solutions to (26) and (27), and let k_1 and k_2 be the solutions to (31) and (32). Then, $k_1 = k_1^* + \Delta k_1$ and $k_2 = k_2^* + \Delta k_2$. It can be shown that $\Delta k_1 = O(1/\sqrt{N})$ and $\Delta k_2 = O(1/\sqrt{N})$.

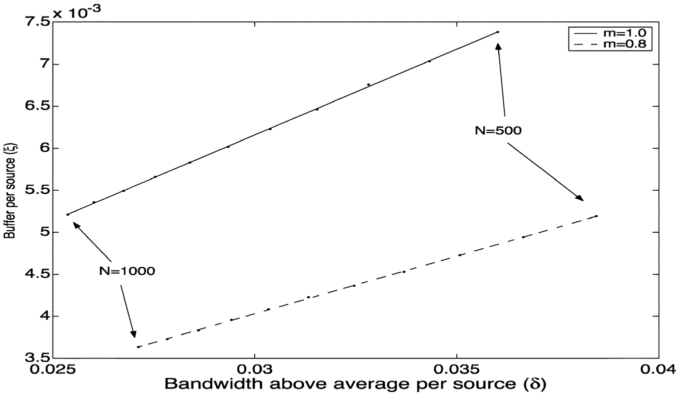

Fig. 10. Optimal buffer versus optimal bandwidth.

It follows that

$$\begin{aligned} \delta^* &= \frac{k_1^* + O\left(\frac{1}{\sqrt{N}}\right)}{\sqrt{N}} = \frac{k_1^*}{\sqrt{N}} + O\left(\frac{1}{N}\right) \\ x^* &= \left(k_2^* + O\left(\frac{1}{\sqrt{N}}\right)\right)\sqrt{N} = k_2^*\sqrt{N} + O(1). \end{aligned}$$

The theorem follows. ■

These results imply that, as the number of sources increase, the minimum cost solution (under fixed per-unit bandwidth and buffer costs) is to not add bandwidth and buffer 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.

We believe that the dependence of the optimal bandwidth and buffer on N are properties of the form of the overflow probability with respect to the number of sources, bandwidth and buffer—not on the Taylor series expansion (12). The terms we chose to include in the Taylor series are exactly those that affect the forms for the optimal bandwidth and buffer given in (2) and (3). Inclusion of any additional terms do not change the first-order dependence on N.

The result is predicated upon the assumption that $G_M(N,x,\gamma)$ is decreasing and jointly convex in x and γ. While this has not been proven analytically for the expression in (12), similar results have been uniformly empirically found to hold in the literature, including the numerical cases investigated in this paper. In addition, a similar result has been proven for overflow probabilities given by large deviations results in the many-source regime [27].

For numerical illustration and validation, the optimal bandwidth (above average) and buffer allocations per source were shown in Figs. 1 and 2, as a function of the number of sources N (for $m = 1$). Also shown are least-squares curvefits to a $1/\sqrt{N}$ form, as predicted by (25). The fits are extremely good, over a wide range of N.

In Fig. 10, we plot the optimal bandwidth (above average) and buffer allocations per source versus each other. As illustrated in Fig. 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.

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 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 bandwidth and buffer allocation is calculated. Correspondingly, denote $\hat{x}$ and $\hat{c}$ as the minimum cost allocation of buffer and bandwidth and such that $G_M(\hat{N}, \hat{x}, \hat{c}/N) = p$.

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 FB resource-allocation policy allocates a buffer of $x' = \hat{x}$ and a bandwidth of c', where c' is the value that satisfies $G_M(N, x', c'/N) = p$. The 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 $G_M(N, x', c'/N) = p$.

The cost of the optimal policy is $C^* = p_x x^* + p_c c^*$, where x^* and c^* are the optimal bandwidth and buffer 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 given in the following theorem.

Theorem 3: Consider either the FB or IB policy given above, with $\hat{N}$ as the nominal number of sources upon which the initial bandwidth and buffer 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

$$\Delta C \sim \left(\frac{\Delta N}{\hat{N}}\right)^2 \sqrt{\hat{N}}. \tag{33}$$

Proof: For the FB policy, we have

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

provided that $(\Delta N/\hat{N}) \ll 1$.

For the IB policy, we have

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

provided that $(\Delta N/\hat{N}) \ll 1$.

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

$$\Delta C(N) \approx \left.\frac{\partial^2 c}{\partial x^2}\right|_{x^*} \frac{(x' - x^*)^2}{2}. \tag{34}$$

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 (28). Dropping the $O()$ terms, and substituting $y = \sqrt{x}$, we can restate this as

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

where

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

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

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

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)

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

Consequently,

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

and thus

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

since $(\partial c/\partial x) = O(1)$.

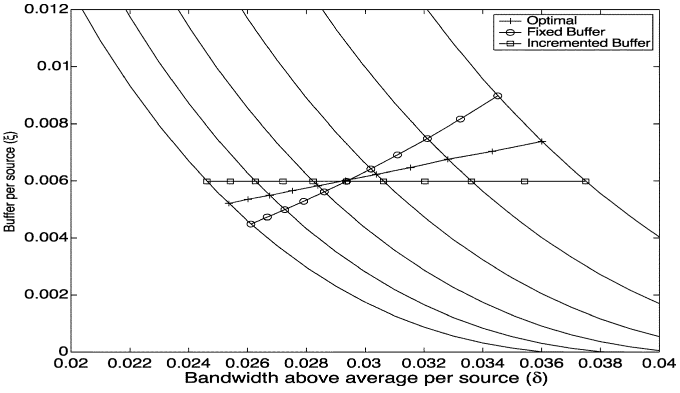

Fig. 11. Resource allocations for alternative policies.

Substituting this back into the Taylor series (34) and using $N \sim \hat{N}$, we obtain

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

Finally, using $x' - x^* \sim (\Delta N/\sqrt{\hat{N}})$, we have

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

■

For numerical illustration, in Fig. 11, we plot the bandwidth (above average) and buffer allocations per source for the optimal policy, for the FB policy and for the IB policy.

The FB policy constitutes a curve through the set of contours, and the IB policy constitutes a horizontal line. We have set the nominal number of sources upon which the initial bandwidth and buffer 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 FB policy changes *only the bandwidth* so that the new allocation is on the new buffer-versus-bandwidth contour. The incremented buffer policy varies the buffer *linearly* and sets the bandwidth so that the new allocation is also on the new contour.

An examination of Fig. 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. Similar reasoning regarding the total buffer concludes that the optimal policy will increase the total buffer allocation with N. Thus, for our set of parameters, the optimal policy lies strictly between the FB and IB policies.

The analysis for the cost comparison explains Figs. 3 and 4, which show the cost differences between the optimal policy, FB, and IB. As in Fig. 11, N is varied about the nominal value of $\hat{N} = 750$. All three policies are generated directly using the exact AMS results for overflow probability. The Taylor series analysis above (33) predicts that the resulting cost savings should be quadratic in ΔN for a fixed $\hat{N}$ (for small values of ΔN). Least-squares quadratic curvefits are shown on the figures, and we find the cost differences agree well with this form. The asymmetry can be attributed to the presence of a third-order term, which was neglected in the analysis.

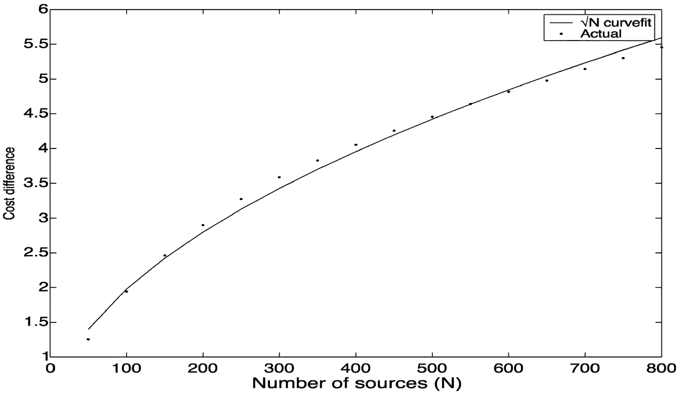

Fig. 12. Cost difference between optimal policy and FB policy for a fixed $(N - \hat{N}/N)$.

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

The Taylor series analysis (33) 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. A plot of the cost difference between the optimal policy and IB is almost identical.

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. Network Model

We again consider a single queue fed by N sources. As above, we denote the aggregate bandwidth by c, the aggregate buffer by x, and the maximum acceptable overflow probability by p. In contrast to the assumption in previous sections that the sources are i.i.d. ON/OFF fluid flows, we now consider any estimate of overflow probability $G(N, x, c/N)$ that is a function of N, x, and c. For a single source, given an allocated buffer of x, we denote the bandwidth required to obtain a loss probability of p by

$$eb(x) \equiv c\Big| \left[G\left(1, x, \frac{c}{N}\right) = p\right]$$

and call this quantity the *effective bandwidth* of one source. We restrict our analysis to overflow estimates G that result in an effective bandwidth that is a decreasing convex function of x. Furthermore, we assume that the bandwidth required to obtain a loss probability of p for N multiplexed flows, called the effective bandwidth of the multiplexed stream, is given by N times the effective bandwidth of one source as follows:

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

The convexity property is satisfied by many effective bandwidth derivations in the literature [21], [27]. 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

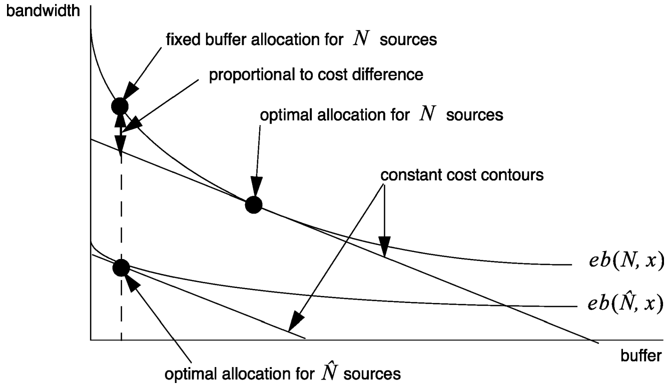
bandwidth
fixed buffer allocation for N sources
proportional to cost difference
optimal allocation for N sources
constant cost contours
$eb(N, x)$
$eb(\hat{N}, x)$
buffer
optimal allocation for $\hat{N}$ sources

Fig. 13. Illustration of cost difference.

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 gives 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).

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

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

and the resulting optimal cost by

$$C^*(N) = p_x x^*(N) + p_c N eb\left(x^*(N)\right).$$

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

$$\left.\frac{\partial N eb(x)}{\partial x}\right|_{x^*(N)} = N \left.\frac{\partial eb(x)}{\partial x}\right|_{x^*(N)} = -\frac{p_x}{p_c} = -\frac{1}{m}$$

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

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

Our principal result in this section is given in the following theorem.

Theorem 4: 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

$$N \left.\frac{\partial eb(x)}{\partial x}\right|_{x^*(N)} = (N+1) \left.\frac{\partial eb(x)}{\partial x}\right|_{x^*(N+1)} = -\frac{1}{m}$$

and therefore that

$$\frac{\left.\frac{\partial eb(x)}{\partial x}\right|_{x^*(N+1)}}{\left.\frac{\partial eb(x)}{\partial x}\right|_{x^*(N)}} = \frac{N}{N+1} < 1.$$

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 2, which states that, for ON/OFF sources, the optimal buffer allocation is proportional to $\sqrt{N}$. Theorem 4 considers a wider class of flows, but is weaker than Theorem 2 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 an FB policy. The cost of the FB policy, using $\hat{N}$ as the nominal number of sources upon which the initial bandwidth and buffer allocation is calculated, is

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

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

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

Our principal result in this section is given in the following theorem.

Theorem 5: $\Delta C_{\text{FB}}(N, \hat{N})$ is increasing and convex in $|N - \hat{N}|$, when $x^*(N) > 0$.

Proof: Substituting expressions for $C_{\text{FB}}(N)$ and $C^*(N)$ from above, we have

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

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

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

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 Fig. 13.

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

$$\begin{aligned} &\Delta C_{\text{FB}}(N+1, \hat{N}) - \Delta C_{\text{FB}}(N, \hat{N}) \\ &\quad = p_c \int_{x^*(N)}^{x^*(N+1)} \left[-(N+1)\frac{\partial eb(x)}{\partial x} - \frac{1}{m}\right] dx \\ &\quad + p_c \int_{x^*(\hat{N})}^{x^*(N)} \left[-\frac{\partial eb(x)}{\partial x}\right] dx \end{aligned} \quad (35)$$

and as

$$\begin{aligned}\Delta C_{\mathrm{FB}}&(N+2,\hat{N})-\Delta C_{\mathrm{FB}}(N+1,\hat{N})\\ &=p_c\int_{x^*(N+1)}^{x^*(N+2)}\left[-(N+2)\frac{\partial eb(x)}{\partial x}-\frac{1}{m}\right]dx\\ &\quad+p_c\int_{x^*(\hat{N})}^{x^*(N+1)}\left[-\frac{\partial eb(x)}{\partial x}\right]dx. \qquad (36)\end{aligned}$$

In (35), the first integral is an integral of a positive quantity (since $eb(x)$ is decreasing and convex and $(\partial(N+1)eb(x)/\partial x)|_{x^*(N+1)} = -1/m$) over a positive range (from Theorem 4). 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 $\Delta C_{\mathrm{FB}}(N,\hat{N})$ 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 [(35) from (36)] and collecting terms, we obtain

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

The first integral is a positive quantity, since it is the same as the first integral in (35), with N changed to $N+1$. The second integral is also a positive quantity, since $eb(x)$ is decreasing and convex and $(\partial Neb(x)/\partial x)|_{x^*(N)} = -1/m$. The sum is therefore positive, and it follows that $\Delta C_{\mathrm{FB}}(N,\hat{N})$ is convex in $|N-\hat{N}|$ when $x^*(N) > 0$. ■

This theorem can be compared to Theorem 3, which states that, for ON/OFF sources, the equivalent expression for the cost difference is proportional to the square of $|N-\hat{N}|$. Theorem 5 considers a wider class of flows but is weaker than Theorem 3 in that it only guarantees that the cost difference is 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.

References

[1] F. Kelly, "Effective bandwidths at multi-service queues," *Queueing Syst.*, vol. 9, pp. 5–16, 1991.

[2] G. Kesidis, J. Walrand, and C.-S. Chang, "Effective bandwidths for multiclass Markov fluids and other ATM sources," *IEEE/ACM Trans. Networking*, vol. 1, pp. 424–428, Aug. 1993.

[3] A. Elwalid and D. Mitra, "Effective bandwidth of general Markovian traffic sources and admission control of high-speed networks," *IEEE/ACM Trans. Networking*, vol. 1, pp. 329–344, June 1993.

[4] C.-S. Chang and J. Thomas, "Effective bandwidth in high-speed digital networks," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 1091–1100, Aug. 1995.

[5] A. W. Berger and W. Whitt, "Effective bandwidths with priorities," *IEEE/ACM Trans. Networking*, vol. 6, pp. 447–460, Aug. 1998.

[6] N. Likhanov and R. R. Mazumdar, "Cell loss asymptotics in buffers fed by heterogeneous long-tailed sources," in *Proc. Conf. Computer Communications (IEEE Infocom)*, 2000, pp. 173–180.

[7] B. Zwart, S. Borst, and M. Mandjes, "Exact queueing asymptotics for multiple heavy-tailed on-off flows," in *Proc. Conf. Computer Communications (IEEE Infocom)*, 2001, pp. 279–288.

[8] A. Weiss, "A new technique for analyzing large traffic systems," *Advances Appl. Prob.*, vol. 18, pp. 506–532, 1986.

[9] G. Choudhary, D. M. Lucantoni, and W. Whitt, "On the effectiveness of effective bandwidths for admission control in ATM networks," in *Proc. 14th Int. Teletraffic Congress*, J. Labetoulle and J. W. Roberts, Eds., Amsterdam, The Netherlands, 1994, pp. 411–420.

[10] A. Simonian and J. Guibert, "Large deviations approximation for fluid queues fed by a large number of on/off sources," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 1017–1027, Aug. 1995.

[11] D. N. Tse, R. G. Gallager, and J. Tsitsiklis, "Statistical multiplexing of multiple time-scale markov streams," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 1028–1038, Aug. 1995.

[12] D. Botvich and N. Duffield, "Large deviations, the shape of the loss curve, and economies of scale in large multiplexers," *Queueing Syst.*, vol. 20, pp. 293–320, 1995.

[13] A. Weiss, "An introduction to large deviations for communication networks," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 938–952, Aug. 1995.

[14] D. Wischik, "Sample path large deviations for queues with many inputs," *Ann. Appl. Probabil.*, vol. 11, pp. 379–404, 2001.

[15] A. Elwalid, D. Heyman, T. Lakshman, D. Mitra, and A. Weiss, "Fundamental bounds and approximations for atm multiplexors with applications to video teleconferencing," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 1004–1016, Aug. 1995.

[16] M. Montgomery and G. DeVeciana, "On the relevance of time scales in performance oriented traffic characterizations," in *Proc. Infocom*, 1996, pp. 513–520.

[17] D. Mitra, J. A. Morrison, and K. Ramakrishnan, "ATM network design and optimization: a multirate loss network framework," *IEEE/ACM Trans. Networking*, vol. 4, pp. 531–543, Aug. 1996.

[18] H. G. Perros and K. M. Elsayed, "Call admission control schemes: a review," *IEEE Commun. Mag.*, vol. 34, pp. 82–91, Nov. 1996.

[19] S. Low and P. Varaiya, "A new approach to service provisioning in ATM networks," *IEEE/ACM Trans. Networking*, vol. 1, pp. 547–553, Oct. 1993.

[20] A. Elwalid, D. Mitra, and R. H. Wentworth, "A new approach for allocating buffers and bandwidth to heterogeneous, regulated traffic in an ATM node," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 1115–1127, Aug. 1995.

[21] K. Kumaran and M. Mandjes, "The buffer-bandwidth trade-off curve is convex," *Queueing Syst.*, vol. 38, pp. 471–483, 2001.

[22] H. Jiang and S. Jordan, "The role of price in the connection establishment process," *Eur. Trans. Telecommun.*, vol. 6, no. 4, pp. 421–429, July–Aug. 1995.

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

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

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

[26] N. B. Shroff and M. Schwartz, "Improved loss calculations at an ATM multiplexor," *IEEE/ACM Trans. Networking*, vol. 6, pp. 411–421, Aug. 1998.

[27] K. Kumaran, M. Mandjes, and A. Stolyar, "Convexity properties of loss and overflow functions," *Oper. Res. Lett.*, vol. 31, pp. 95–100, 2003.

Scott Jordan (S'86–M'90) received the B.S./A.B., M.S., and Ph.D. degrees from the University of California, Berkeley, in 1985, 1987, and 1990, respectively.

From 1990 until 1999, he served as a faculty member with Northwestern University, Evanston, IL. Since 1999, he has served as a faculty member with the University of California, Irvine. His research interests currently include pricing and differentiated services in the Internet and resource allocation in wireless multimedia networks.

Kalpana Jogi received the B.S. degree in computer science from the Birla Institute of Technology and Science, India, in 1999, and the M.S. degree in electrical and computer engineering from the University of California, Irvine, in 2000.

Since 2000, she has been working in Silicon Valley as a software professional and has been gaining expertise in the areas of networking and storage.

Chunlin Shi received the B.S. degree from Shanghai Jiaotong University, China, and the M.S. degree from the University of California (UC Irvine), Irvine.

She was a Research Assistant with the Network Performance Group, UC Irvine, where her research interests included dynamic resource allocation and pricing model in Internet. She is currently a System Engineer with Maxspeed Corporation, Palo Alto,CA.

Ikhlaq Sidhu (S'93–M'95) received the B.S. degree in electrical engineering from the University of Illinois at Urbana-Champaign in 1988, and the M.S. and Ph.D. degrees from Northwestern University, Evanston, IL, in 1993 and 1995, respectively.

He has previously served as a Hardware Engineer with Hewlett Packard, Director of Advanced Technologies for U.S. Robotics, Vice President of Internet Communications for 3Com Corporation, and CTO for Cambia Networks. He is currently a Visiting Associate Professor with the University of Illinois at Urbana-Champaign and directs the Technology Entrepreneur Center for its College of Engineering. His research interests span business and technology issues in networking and image processing.