Rate Adaptive Multimedia Streams: Optimization and Admission Control

Steven Weber, *Member, IEEE,* and Gustavo de Veciana, *Senior Member, IEEE*

***Abstract*—This work investigates support of rate adaptive multimedia streams on communication networks. Optimal and practical mechanisms to maximize the customer average quality of service (QoS), defined in terms of a normalized time average received rate, are established. By scaling the arrival rate and link capacity, we obtain asymptotic expressions for customer average QoS in the case of networks with single bottleneck links. The optimal adaptation policy is identified as the solution to an integer program which has an intuitive "sort by volume" interpretation for the case of single bottleneck links, where stream volume is the total number of bits associated with a stream at its maximum resolution. Our asymptotic analysis shows the optimal adaptation policy may yield performance improvements of up to 42% over baseline policies. We demonstrate that a static multi-class admission control policy can achieve the same asymptotic QoS as that of the optimal adaptation policy. This implies that dynamic adaptation may be unnecessary for large capacity networks with appropriate call admission.**

***Index Terms*—Admission control, rate adaptation, resource allocation, streaming media.**

I. Introduction

We investigate how to best allocate network bandwidth among clients receiving multimedia streams. In particular, we focus on how to allocate bandwidth when streams are available at a number of possible subscription levels, where different levels correspond to different qualities of stream resolution. We refer to streams offering multiple subscription levels as rate adaptive multimedia streams. We assume the network guarantees that each admitted client will be able to subscribe at least at its lowest subscription level, i.e., we assume a next-generation network with an admission control mechanism. However, we are interested in identifying the optimal allocation of the remaining bandwidth, i.e., after all of the minimum subscription levels have been satisfied. To pose the problem in a different light, rate adaptive streams have the flexibility to dynamically match their bandwidth requirements to changing bandwidth availability along their routes. Thus, rate adaptivity permits a stream to utilize a higher transmission rate when network congestion is low, and a lower transmission rate when network congestion is high. The optimal bandwidth allocation is therefore the optimal adaptation policy, where an adaptation policy specifies the subscription level to which the client should subscribe at a given point in time.

There are two complementary approaches to studying rate adaptive streams: the client and system perspectives. The client perspective views the network congestion experienced by a given stream as an exogenous process, and seeks to maximize that client's overall quality of service (QoS). The system perspective views congestion as the superposition of active streams, and seeks to maximize the client-average QoS by specifying an adaptation policy for all active streams. In this paper we focus on the system perspective. Some of the conclusions obtained from the system perspective are at odds with conclusions obtained from the client perspective.

A. Optimization and Number of Subscription Levels

From the client perspective, it would seem logical that a stream should be encoded with as many subscription levels as possible. This would allow clients to choose a subscription level appropriate for a wide variety of congestion levels. We show, however, that from the system perspective, the client-average QoS can be maximized under an adaptation policy which requires only two subscription levels per stream. These two subscription levels correspond to the "coarsest" and "finest" resolutions, i.e., the minimally acceptable subscription level and the maximally useful subscription level. Thus, from a system perspective, there is little benefit in offering a wide variety of subscription levels, assuming flows are not peak rate constrained by their access line. This result follows from our definition of QoS as a linear measure in the encoding rate. In particular, we define the QoS of a stream client as the fraction of the stream volume received by the client, i.e., the number of bits received by the client divided by the number of bits used to encode the media content at its maximum resolution.

B. Admission Control and Dynamic Adaptation

From the client perspective, it appears reasonable to expect that achieving a high client QoS requires the client be able to assess changes in congestion level and respond by adjusting the subscription level quickly and accurately. We show that, for large numbers of streams sharing bandwidth on large capacity links, there is in fact little need for dynamic adaptation. We present a static multi-class admission policy whereby a stream is assigned to a class (subscription level) at the time of admission based on its volume, which it then maintains throughout its duration, i.e., no dynamic adaptation. Here, the volume of a stream is the product of its time-average maximum subscription level and its duration, i.e., the number of bits associated with the stream at its maximum resolution. We show that the asymptotic QoS obtained under the optimal multi-class admission policy equals

Manuscript received March 13, 2004; revised January 4, 2005; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor M. Krunz.

S. Weber was with the Department of Electrical and Computer Engineering, University of Texas at Austin. He is now with the Department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA 19104-2875 USA (e-mail: sweber@ece.drexel.edu).

G. de Veciana is with the Department of Electrical and Computer Engineering, University of Texas at Austin, Austin TX 78712 USA (e-mail: gustavo@ece.utexas.edu).

Digital Object Identifier 10.1109/TNET.2005.860105

the asymptotic QoS obtained under the optimal dynamic adaptation policy. Intuitively, for large capacity links the ensemble of active streams is essentially constant, and the optimal adaptation policy will assign a given stream the same subscription level throughout its duration.

The rest of this paper is organized as follows. In Section II, we introduce our model for rate adaptive streams, client QoS, and the network. In Section III, we identify the optimal adaptation policy for the case where stream durations are known a priori and the case where they are unknown (but their distribution is known). We then introduce a scaling regime appropriate for large numbers of streams sharing large capacity links, which yields the asymptotic client-average QoS under the optimal adaptation policy. In Section IV, we introduce a multi-class admission control policy and show that, when correctly configured, the asymptotic client-average QoS obtained under the optimal admission control policy equals the asymptotic QoS under the optimal dynamic adaptation policy. Finally, Section V concludes with a discussion of related work.

II. The Model

Throughout this paper, we assume all random variables are continuous so that their cumulative distribution functions (CDFs) have an inverse. We also assume all random variables to have finite support on a subset of $(0, \infty)$. We will use the following convention for random variables. A random variable X will have a CDF F_X, a complementary CDF (CCDF) $\bar{F}_X$, a PDF f_X, and an inverse CDF F_X^{-1}.

A. Stream Characteristics

We model a rate adaptive stream by four parameters: stream duration, maximum subscription level, adaptivity, and the set of offered subscription levels.

Stream Duration: Stream durations are random variables, denoted by D, with a common distribution F_D, and mean $\mathbb{E}[D] = \delta$. A known stream duration is denoted by d. We assume all encodings of a given stream share the same duration, i.e., compression does not impact the stream duration. The stream duration need not necessarily equal the content duration, i.e., clients may terminate a stream prior to the completion of the content. We do assume, however, that the stream duration is independent of the client perceived QoS.

Maximum Subscription Level: The maximum subscription level is defined as the effective bandwidth of the stream when encoded at the maximum resolution deemed useful by the provider, i.e., an encoding such that a higher resolution yields a negligible increase in perceived quality.

Maximum subscription levels of streams are modeled via random variables, denoted by S, with a common distribution F_S, and mean $\mathbb{E}[S] = \sigma$. A known maximum subscription level is denoted by s. Maximum subscription levels are assumed to be independent of stream durations.

Adaptivity: Stream adaptivities are defined as the ratio between the maximum subscription level and the minimum subscription level. The minimum subscription level is the effective bandwidth of the stream when encoded at the minimum resolution deemed useful by the provider, i.e., an encoding such that a lower resolution yields a stream with unacceptable perceived quality. Adaptivities are random variables, denoted by A, with a common distribution F_A, and mean $\mathbb{E}[A] = \alpha$. The support of A is necessarily within (0,1]. A known adaptivity is denoted by a. Adaptivities are assumed to be independent of both durations and maximum subscription levels. Note that a stream with maximum subscription level S and adaptivity A has a minimum subscription level of AS. In contrast with S and D, F_A may be discrete.

Offered Subscription Levels: A stream is offered at a set of *discrete* subscription levels, denoted by $\mathcal{S} = (S_i \mid AS = S_1 < \cdots < S_K = S)$, where $K \in \mathbb{Z}$ denotes the number of subscription levels available to clients, and S_k is the effective bandwidth corresponding to subscription/encoding k. This abstraction is independent of the type of encoding used to create the subscription levels, e.g., hierarchical or simultaneous encoding. In hierarchical encoding, subscription level S_k corresponds to the sum of the first k layers, while for simultaneous encoding, S_k corresponds to the kth smallest encoding. We will focus in the sequel on the two subscription levels AS and S, the minimum and maximum subscription levels.

B. Network Model

We let $\mathcal{L}$ denote the set of links, and the vector $\mathbf{c} = (c_l, l \in \mathcal{L})$ denote the capacities of those links. We assume this capacity is shared by rate adaptive streams. In particular, we assume that the streaming traffic is given priority over the best-effort traffic on the network, so that the entire link capacity is available to the streaming traffic and the available capacity on each link is therefore assumed to be time invariant. Let $\mathcal{R}$ denote the set of routes, where a route r is composed of a set of links $\{l \in r\} = \{l \mid l \in r\}$. The vector $\boldsymbol{\lambda} = (\lambda_r, r \in \mathcal{R})$ denotes the arrival rate of new stream requests on each route. We assume all arrival processes are Poisson. The notation $\{r \ni l\} = \{r \mid l \in r\}$ denotes the set of routes incident on link l.

The random variables $\mathbf{N}(t) = (N_r(t), r \in \mathcal{R})$ denote the stationary numbers of active streams on each route at a given time t. We write $\mathbf{n}(t) = (n_r(t), r \in \mathcal{R})$ when this quantity is assumed known.

Finally, the notation (i, r) indexes stream i on route r. For any model parameter X, the notation $X_{i,r}$ refers to a parameter for stream (i, r).

C. Quality of Service

Modeling quality of service for multimedia streams is a difficult, and largely unsolved, problem. The Video Quality Experts Group recently performed a statistical analysis of nine proposed objective measures of video quality [1]. They found that none of the proposed models functioned adequately to replace subjective testing. In addition, the performance of the objective models were found to be statistically indistinguishable from one another.

Modeling quality of service for rate adaptive streams promises to be an even harder problem due to the dynamic changes in instantaneous rate. Below, we define three aspects of QoS which we feel are especially important.

Expected Normalized Time-Average Subscription Level: It seems reasonable to assume that, all other factors being equal, client perceived video quality for rate adaptive streams is increasing in the time-average subscription level, i.e.,

$(1/D)\int_0^D S(t)dt$, where the times $t = 0$ and $t = D$ respectively denote the arrival and departure time of a *typical* stream. Due to the highly heterogeneous nature of media rates (e.g., streaming audio versus streaming video), it would be inappropriate to compare two clients based on their time-average received rates alone. We therefore normalize the time-average received rate by the maximum subscription level S, to obtain the client QoS $Q = (1/D)\int_0^D (S(t)/S)dt$. Note that $Q \in [AS, 1]$. Our first system level QoS parameter is $\mathbb{E}^0[Q]$, the average QoS seen by a typical client:

$$\mathbb{E}^0[Q] = \mathbb{E}^0\left[\frac{1}{D}\int_0^D \frac{S(t)}{S}dt\right] \tag{1}$$

where $\mathbb{E}^0[\cdot]$ denotes expectation taken with respect to a customer average. We emphasize that our quality of service metric is normalized by stream volume. Our justification for this normalization is to permit fair comparison between the QoS experienced by streams whose encoding rate and/or stream duration may vary over an order of magnitude. The effect of this normalization is that our theorems on resource allocation and admission control are volume dependent.

Expected Rate of Adaptation: Rate of adaptation is defined as the number and magnitude of the changes in subscription level. It has been shown that client QoS is adversely affected by these changes [2]. We define the set of subscription level change times for a given client as $\mathcal{C} = (t \in (0, D) \mid S(t^-) \neq S(t^+))$. We then define the rate of adaptation $R = (1/D)\sum_{t\in\mathcal{C}} |S(t^-) - S(t^+)|$. Our second system level QoS parameter is $\mathbb{E}^0[R]$, the average rate of adaptation seen by a typical client:

$$\mathbb{E}^0[R] = \mathbb{E}^0\left[\frac{1}{D}\sum_{t\in\mathcal{C}} |S(t^-) - S(t^+)|\right]. \tag{2}$$

Blocking Probability: To minimize the blocking probability we restrict ourselves to policies which admit as many streams as possible while respecting the minimum rates required by already admitted streams. Thus, we allow admissions even if that admission requires one or more admitted streams to reduce their subscription levels in order to accommodate the new stream. In particular, a stream with parameters (s, a) will be admitted on route r at time t provided

$$\sum_{r'\ni l}\sum_{i=1}^{n_{r'}(t)} a_{i,r'}s_{i,r'} + as \leq c_l \quad \forall l \in r. \tag{3}$$

We define the stationary blocking probability for such a stream by

$$B(r, as) = \mathbb{P}\left(\sum_{r'\ni l}\sum_{i=1}^{N_{r'}(t)} A_{i,r'}S_{i,r'} + as > c_l \quad \forall l \in r\right) \tag{4}$$

where the probability is taken with respect to the *stationary* distribution of the network. Note that the admission policy is independent of the adaptation policy.

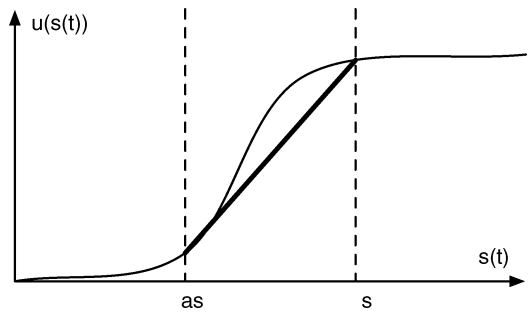

Fig. 1. A typical utility curve of client satisfaction versus encoding rate. Content providers would likely offer encodings at rates between the minimum rate as and the maximum rate s; QoS may be approximated as a linear function within this region.

Of these three QoS measures, we will focus on $\mathbb{E}^0[Q]$, the normalized time–average subscription level of a typical client. In particular, we will identify the dynamic adaptation policy which maximizes this quantity. We will, however, investigate the resulting rate of adaptation and blocking probability of this policy, and contrast these aspects of QoS with our results on optimal admission control policies in Section IV.

We emphasize that our primary QoS metric is *linear* in the encoding rate. The actual utility curve describing client satisfaction versus encoding rate is likely to have a structure like that shown in Fig. 1 [3]. Media providers might provide minimum and maximum encodings at encoding rates as and s, where as is the minimally acceptable encoding rate and s is the maximally useful encoding rate. The shape of the utility curve between as and s is not linear but can be reasonably well approximated by a linear function.

III. Optimal Adaptation

An adaptation policy assigns each active stream a feasible subscription level at each time t such that capacity constraints are satisfied on each link. In particular, a policy π, assigns subscription levels to ongoing streams $\mathbf{s}^\pi(t) = (s^\pi_{i,r}(t), i = 1, \ldots, n_r(t), r \in \mathcal{R})$ such that $s^\pi_{i,r}(t) \in \mathcal{S}_{i,r}$ and $\sum_{r\ni l}\sum_{i=1}^{n_r(t)} s^\pi_{i,r}(t) \leq c_l, \forall l \in \mathcal{L}$.

We define the optimal adaptation policy as that which maximizes the expected normalized time-average subscription level $\mathbb{E}^0[Q]$. To ensure feasibility we restrict ourselves to nonanticipatory policies, i.e., those policies which determine an allocation for a time t based on information available at that time. We identify the optimal adaptation policy under two different assumptions on available information: when stream durations are known at the time of admission (Section II-A), and when they are unknown (Section II-B). The former corresponds to the case of stored media, and the latter corresponds to the case of live media. For each case, we show the optimal adaptation policy is found by solving an integer program at each time t. For the specific case of single bottleneck routes, we obtain a closed form solution of which offers the basic insight on the nature of the optimal policy. Simulation results are presented in Section III-C. Finally, we present a linear scaling regime in Section III-D which permits closed form asymptotic expressions for $\mathbb{E}^0[Q]$ under the optimal adaptation policy.

A. Known Stream Durations

In this section, we assume stream durations are known at the time of stream admission. We define the instantaneous aggregate QoS at time t as

$$q_{\text{agg}}(t) = \sum_{r\in\mathcal{R}} \sum_{i=1}^{n_r(t)} \frac{s_{i,r}(t)}{s_{i,r}d_{i,r}} \tag{5}$$

where $\mathbf{s}(t) = (s_{i,r}(t), i = 1,\ldots,n_r(t), r \in \mathcal{R})$ is the allocation given to each stream. We write $q^{\pi}_{\text{agg}}(t)$ for the instantaneous aggregate QoS under a policy π and $\mathbf{s}^{\pi}(t)$ for the corresponding allocation. The following theorem demonstrates that maximizing $\mathbb{E}^0[Q]$ is equivalent to maximizing $q_{\text{agg}}(t)$ at each time t subject to capacity and subscription level constraints. We let π_k denote the optimal policy which maximizes $\mathbb{E}^0[Q]$ and will also identify a near-optimal policy $\tilde{\pi}_k$.

Theorem 1: The adaptation policy π_k that maximizes $\mathbb{E}^0[Q]$ when stream durations are known is the instantaneous bandwidth allocation $\mathbf{s}^{\pi_k}(t)$ at each time t resulting from the solution of the following integer programming problem:

$$\begin{aligned}
\max_{\mathbf{s}(t)} \quad & q_{\text{agg}}(t) = \sum_{r\in\mathcal{R}} \sum_{i=1}^{n_r(t)} \frac{s_{i,r}(t)}{s_{i,r}d_{i,r}} \\
\text{s.t.} \quad & \sum_{r\ni l} \sum_{i=1}^{n_r(t)} s_{i,r}(t) \le c_l \quad \forall l \in \mathcal{L}, \\
& s_{i,r}(t) \in \mathcal{S}_{i,r}, i = 1,\ldots,n_r(t) \quad \forall r \in \mathcal{R}.
\end{aligned} \tag{6}$$

There exists a feasible allocation $\mathbf{s}^{\tilde{\pi}_k}(t)$ with $s^{\tilde{\pi}_k}_{i,r}(t) \in \{a_{i,r}s_{i,r},\ s_{i,r}\}$ for all $i = 1,\ldots,n_r(t)$ and all $r \in \mathcal{R}$ such that the value of the objective under $\mathbf{s}^{\tilde{\pi}_k}(t)$ is nearly optimal. In particular,

$$\frac{q^{\pi_k}_{\text{agg}}(t) - q^{\tilde{\pi}_k}_{\text{agg}}(t)}{q^{\pi_k}_{\text{agg}}(t)} \le \frac{\kappa_k}{n(t)} \tag{7}$$

where $n(t) = \sum_{r\in\mathcal{R}} n_r(t)$, and $\kappa_k < \infty$. See the Appendix for proof.

The first part of the theorem demonstrates that each stream is weighted inversely by its volume $v_{i,r} = s_{i,r}d_{i,r}$, i.e., the product of its maximum subscription level and its duration. The intuition is that the system is able to maximize the customer average QoS by granting higher QoS to customers consuming fewer network resources. The second part of the theorem illustrates the existence of a near–optimal allocation such that all streams use either their minimum or maximum subscription level. Thus, for networks supporting large numbers of streams we may achieve a close to optimal solution by only using the minimum and maximum subscription levels. Note that the integer programming problem in Theorem 1 corresponds to a 0 – 1 multidimensional knapsack problem which is known to be NP-hard [4].

This implies, from the system perspective, that there is little need for content providers to offer intermediate subscription levels, i.e., between $a_{i,r}s_{i,r}$ and $s_{i,r}$. This conclusion is markedly different from that obtained if one considers the problem of supporting rate adaptive multimedia streams from the client perspective, which suggests streams are more resilient to congestion when they have numerous subscription levels available.

We define a bottleneck link as any link requiring adaptation, i.e., $\sum_{r\ni l}\sum_{i=1}^{n_r(t)} s_i > c_l$. We use the phrase "single bottleneck link" to denote any bottleneck link that is the unique bottleneck for all streams traversing it. For this case we can write down the allocations $\mathbf{s}^{\pi_k}(t)$, $\mathbf{s}^{\tilde{\pi}_k}(t)$ from Theorem 1 in closed form. We simplify notation by dropping the subscripts l and r, so that $n(t)$ denotes the number of streams traversing the bottleneck link, c denotes the bottleneck link capacity, and the subscript i refers to the i^{th} stream traversing the link.

Corollary 1: Consider a bottleneck link traversed by $n(t)$ active streams, labeled in order of increasing volume $v_1^{-1} > \cdots > v_n^{-1}$. The allocations $\mathbf{s}^{\pi_k}(t)$, $\mathbf{s}^{\tilde{\pi}_k}(t)$ of Theorem 1 for the case of single bottleneck links are

$$\begin{aligned}
s_i^{\pi_k}(t) = s_i^{\tilde{\pi}_k}(t) &= \begin{cases} s_i, i = 1,\ldots,\bar{n}-1 \\ a_i s_i, i = \bar{n}+1,\ldots,n(t) \end{cases} \\
s_{\bar{n}}^{\pi_k}(t) &= c - \sum_{i\neq\bar{n}} s_i^{\pi_k}(t) \\
s_{\bar{n}}^{\tilde{\pi}_k}(t) &= a_{\bar{n}} s_{\bar{n}},
\end{aligned} \tag{8}$$

where

$$\bar{n} = \max\left\{ m \mid \sum_{i=1}^{m-1} s_i + \sum_{i=m}^{n(t)} a_i s_i \le c \right\}. \tag{9}$$

See the Appendix for proof.

On single bottleneck links, the optimal adaptation policy sorts the active streams on the bottleneck link by volume, granting the full subscription level to as many streams as possible while ensuring sufficient capacity is available to allow the remaining clients to subscribe at their minimum subscription level. For large capacity links servicing large numbers of streams the difference in the objective between $s^{\pi_k}(t)$ and $s^{\tilde{\pi}_k}(t)$ will be negligible, and we may obtain a QoS comparable to the optimal by using only the minimum and maximum subscription levels for each stream.

B. Unknown Stream Durations

In this section, we assume stream durations are unknown at the time of stream admission. We denote the optimal adaptation policy under this assumption by π_u, and the approximate optimal adaptation policy by $\tilde{\pi}_u$.

Theorem 2: The adaptation policy π_u that maximizes $\mathbb{E}[Q]$ when stream durations are unknown is the instantaneous bandwidth allocation $\mathbf{s}^{\pi_u}(t)$ at each time t resulting from the solution of (6) with the quantity $1/d_{i,r}$ replaced with $\mathbb{E}[(1/D) \mid D > l_{i,r}(t)]$, where $l_{i,r}(t)$ is the current age of stream (i,r) at time t. There exists a feasible allocation $\mathbf{s}^{\tilde{\pi}_u}(t)$ with $s^{\tilde{\pi}_u}_{i,r}(t) \in \{a_{i,r}s_{i,r},\ s_{i,r}\}$ for all $i = 1,\ldots,n_r(t)$ and all $r \in \mathcal{R}$ such that the value of the objective under $\mathbf{s}^{\tilde{\pi}_u}(t)$ is nearly optimal. In particular,

$$\frac{q^{\pi_u}_{\text{agg}}(t) - q^{\tilde{\pi}_u}_{\text{agg}}(t)}{q^{\pi_u}_{\text{agg}}(t)} \le \frac{\kappa_u}{n(t)} \tag{10}$$

for $\kappa_u < \infty$. See the Appendix for proof.

If a stream is admitted at time b then its current age at time t is $l = t - b$. For the case of unknown stream durations we see that streams are weighted according to their expected inverse volume at time t, i.e., $(1/s_{i,r})\mathbb{E}[(1/D) \mid D > l_{i,r}(t)]$, as opposed to

being weighted according to their inverse volume $1/s_{i,r}d_{i,r}$ as in the case for known stream durations.

The solution to (6) when there is at most one bottleneck link per route and stream durations are unknown is to sort streams traversing a given bottleneck by "*expected*" volume. We define the expected volume $v(t)$ as $v(t)^{-1} = (1/s)\mathbb{E}[(1/D) \mid D > l(t)]$.

Corollary 2: Consider a bottleneck link traversed by $n(t)$ active streams, labeled in order of increasing expected volume $v_1(t)^{-1} > \cdots > v_n(t)^{-1}$. The allocations $\mathbf{s}^{\pi_u}(t)$, $\mathbf{s}^{\tilde{\pi}_u}(t)$ for the case of at most one bottleneck link per route and unknown stream durations are given by (8) and (9) with $(\mathbf{s}^{\pi_u}(t), \mathbf{s}^{\tilde{\pi}_u}(t))$ replacing $(\mathbf{s}^{\pi_k}(t), \mathbf{s}^{\tilde{\pi}_k}(t))$. See the Appendix for proof.

The corollary illustrates that although stream durations may be unknown, the near-optimal allocation still only makes use of two subscription levels: $a_i s_i$ and s_i.

C. Simulation Results

We performed simulations to investigate the performance of the optimal adaptation policies under the assumption that stream durations are known and unknown. All simulations in this paper are for a network consisting of a single bottleneck. For this and all the simulations in this paper, we used the following distributions for stream characteristics. Stream durations and maximum subscription levels were drawn from a bounded exponential distribution, i.e.,

$$F_X(x) = \begin{cases} \frac{1-e^{-\mu x}}{1-e^{-\mu M}}, & 0 \le x \le M \\ 1, & x > M. \end{cases} \tag{11}$$

We also investigated many other distributions and will comment on the results in the sequel. Stream durations were drawn from a bounded exponential distribution with a maximum duration $M_d = 6000$ seconds, and exponent $\mu_d = 1/180$, yielding a mean $\delta = \mathbb{E}[D] \approx 180$ seconds, and a variance $Var(D) = 32{,}400$. Maximum subscription levels were drawn from a bounded exponential distribution with maximum rate $M_s = 10$ MB/s, and exponent $\mu_s = 10/3$, yielding a mean $\sigma = \mathbb{E}[S] \approx 0.3$ MB/s, and a variance $Var(S) = 0.09$. Note that the values M_s and M_d are large enough that the mean and variance of S and D are nearly identical to the corresponding values for the unbounded exponential case. Note that the stream volumes have a range between 0 (MB) and $M_v = M_s * M_d = 60\,000$ MB, and a mean of $\mathbb{E}[V] = \mathbb{E}[S]\mathbb{E}[D] = 54$ MB. The variance of V can be calculated to be $Var(V) = 8748$. Stream adaptivities are uniformly distributed between 1/4 and 3/4, i.e., $A \sim Uni(1/4, 3/4)$. Note in particular that $\mathbb{E}[A] = \alpha = 1/2$. We varied the mean arrival rate, using $\lambda \in \{0.5, 5\}$ and a corresponding link capacity of $c = (3/4)\sigma\delta\lambda \in \{20.25, 202.5\}$ MB/s depending on λ. As will become clear in the sequel, this corresponds to a capacity scaling where $\gamma = 3/4$. All simulations were run for about 15 000 clients.

Figs. 2 and 3 show histograms of the client averages of Q and R, grouped by volume into bins, i.e., $\mathbb{E}[Q \mid V = v]$ and $\mathbb{E}[R \mid V = v]$ versus volume v (in MB). Several points are noteworthy. Foremost, the figures demonstrate that streams with very small or very large volume experience a very low rate of adaptation, especially when stream durations are known, while streams with intermediate volumes experience very high rates

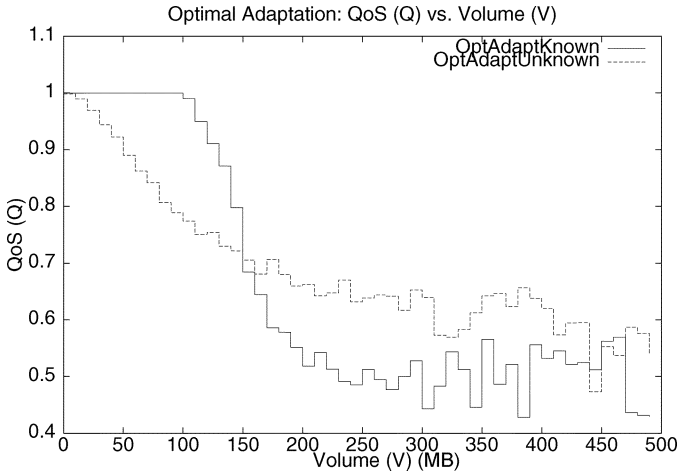

Fig. 2. Simulation histogram of the time-average normalized subscription level Q versus the stream volume $V = SD$ (in MB) under the optimal adaptation policies (known and unknown stream durations).

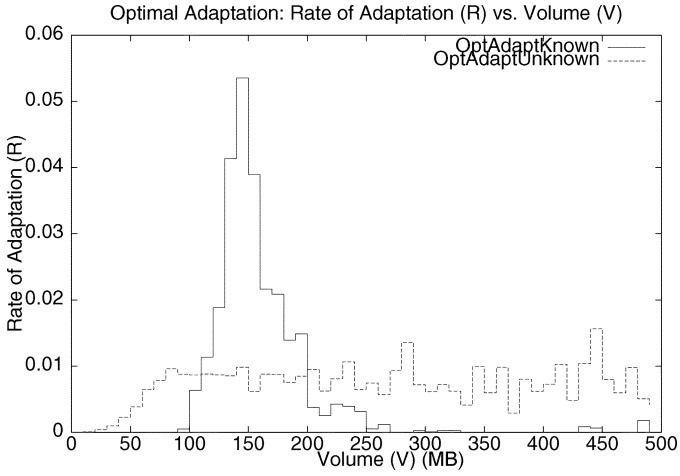

Fig. 3. Simulation histogram of the rate of adaptation R versus the stream volume $V = SD$ (in MB) under the optimal adaptation policies (known and unknown stream durations).

of adaptation. Similarly, streams with small volumes experience the maximum possible QoS, i.e., 1, those with large volumes experience the minimum average possible QoS, i.e., $\mathbb{E}[A] = 0.5$, while those with intermediate volumes have a QoS that decays in stream volume. These observations align with our intuition on how optimal adaptation policies will perform. Since we sort streams by volume into two classes, it follows naturally that those with small volumes will always be in one class, those with large volumes will always be in the other, and those with intermediate volumes will toggle between the two classes, depending on the makeup of the volumes of the active streams. Also note that the optimal policy is more efficient at sorting streams by volume when stream durations are known than when they are unknown; this follows naturally from the fact that more information is available in the former case than in the latter.

D. Asymptotic Results

We introduce a scaling regime to model the case where large numbers of streams share large capacity links. Our results in this section are limited to the case of single bottleneck links. We drop

the link and route subscripts l and r so that c is the bottleneck link capacity and λ is the arrival rate of streams on that link.

Consider a sequence of links with arrival rates and link capacities indexed by m, i.e., $\{(\lambda(m), c(m))\}_{m\in\mathbb{Z}_+}$. We linearly scale the arrival rate as $\lambda(m) = m\lambda$, and we linearly scale the bottleneck link capacity as

$$c(m) = \gamma\lambda(m)\delta\sigma. \tag{12}$$

Here, $\gamma > 0$ is a scaling parameter. Recall that δ is the mean stream duration and σ is the mean peak subscription level. Note that the average offered load (in units of bandwidth) assuming no adaptation takes place is $\lambda(m)\delta\sigma$, so that γ can be understood as the ratio of available capacity over this offered load, i.e., $\gamma = c(m)/\lambda(m)\delta\sigma$. Note that the average number of active streams (in a low–blocking regime) will be $\lambda(m)\delta$, so we can also interpret γ as the average fraction of the average maximum subscription level available to each stream sharing the link, assuming the bandwidth is distributed evenly, i.e., $c(m)/\lambda(m)\delta = \gamma\sigma$.

This scaling encompasses three distinct regimes, parameterized by γ.

- *Overloaded Regime:* $\gamma < \alpha$. Here, the bandwidth divided by the average number of active streams is less than that required to support streams at their average minimum subscription level, i.e., $\gamma\sigma < \alpha\sigma$. The asymptotic average blocking probability in this regime is $1 - (\gamma/\alpha)$. We call this the overloaded regime.
- *Rate Adaptive Regime:* $\alpha \le \gamma \le 1$. Here, the bandwidth divided by the average number of active streams lies between the average minimum subscription level and the average maximum subscription level, i.e., $\alpha\sigma < \gamma\sigma < 1\sigma$. The asymptotic average blocking probability in this regime is 0. We call this the rate adaptive scaling regime; this will be the regime of primary interest in the sequel.
- *Underloaded Regime:* $\gamma > 1$. Here the bandwidth divided by the average number of active streams strictly exceeds the average maximum subscription level, i.e., $\gamma\sigma > \sigma$. The asymptotic average blocking probability in this regime is 0. We call this the underloaded regime.

We define the asymptotic client average QoS under the optimal adaptation policy π_k, when the system is scaled with scaling parameter γ, to be

$$q^{\gamma,\pi_k} = \lim_{m\to\infty} \mathbb{E}[Q^{\pi_k,m}] \tag{13}$$

where $\mathbb{E}[Q^{\pi_k,m}]$ denotes the steady-state QoS seen by a typical stream in the m^{th} scaling of the link.

We also define some new notation for CDFs. If X is a random variable with CDF F_X, then the random variable $\hat{X}$ is defined as having a CDF $F_{\hat{X}}(x) = (1/\mathbb{E}[X])\int_0^x y dF_X(y)$. Also, for two independent random variables $X \sim F_X$ and $Y \sim F_Y$, the quantity $F_{XY}(z)$ is defined as

$$F_{XY}(z) = \int_0^\infty F_X\left(\frac{z}{y}\right) dF_Y(y) = \int_0^\infty F_Y\left(\frac{z}{x}\right) dF_X(x) \tag{14}$$

and corresponds to the CDF of the product XY. Combining these definitions implies $F_{\widehat{XY}}(z)$ can be interpreted to be

$$F_{\widehat{XY}}(z) = \frac{1}{\mathbb{E}[XY]}\int_0^z w dF_{XY}(w). \tag{15}$$

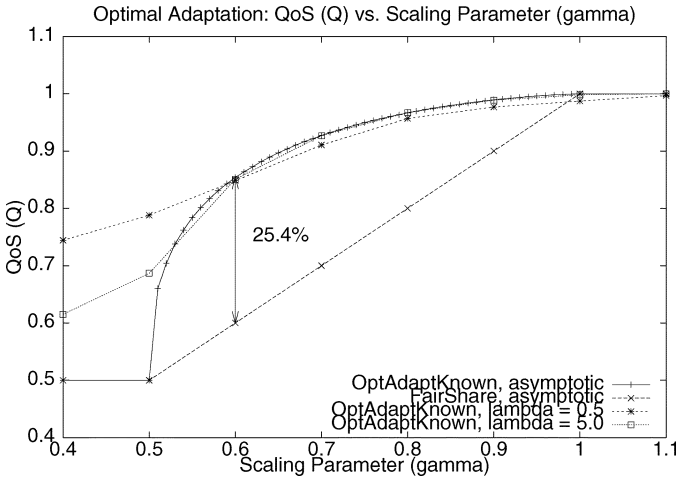

Fig. 4. Plot of the asymptotic QoS q^{π_k} versus the capacity scaling parameter γ for S and D exponentially distributed. Two plots of simulation results under the optimal adaptation policy (known stream durations) show convergence to the asymptotic QoS. In addition, the plot marked fair share denotes the asymptotic QoS under a policy where streams share bandwidth fairly. The label indicates that at $\gamma = 0.6$ the optimal policy achieves an asymptotic QoS 25.4% greater than under fair share, i.e., a 42% increase in performance.

The following theorem demonstrates the asymptotic QoS for the optimal adaptation policy for each of the three regimes.

Theorem 3: Under the optimal adaptation policy for known stream durations, π_k, the asymptotic normalized time-average subscription level for the case of single bottleneck links is

$$q^{\gamma,\pi_k} = \begin{cases} \alpha, & \gamma < \alpha \\ 1 - (1-\alpha)\bar{F}_{SD}\left(F^{-1}_{\widehat{SD}}(\xi)\right), & \alpha \le \gamma \le 1 \\ 1, & \gamma > 1 \end{cases} \tag{16}$$

where $\xi = (\gamma - \alpha)/(1 - \alpha)$. See the Appendix for proof.

The above expression for the asymptotic QoS depends on four quantities: the scaling parameter γ, the average adaptivity α, the distribution of the maximum subscription levels F_S, and the distribution of stream durations F_D. These equations can be thought of as asymptotic analogues to Erlang's blocking probability equations for loss networks. Erlang's equation gives the QoS for a loss model, i.e., the blocking probability $E(\rho, c)$ as a function of the offered load ρ and the number of circuits on the link c, is

$$E(\rho, c) = \frac{\frac{\rho^c}{c!}}{\sum_{k=0}^{c}\frac{\rho^k}{k!}}. \tag{17}$$

The equation permits link designers to provision the number of circuits to achieve a target blocking probability (QoS) for the estimated load. The above expression gives the asymptotic QoS for rate adaptive streams, i.e., the asymptotic normalized average subscription level, as a function of the scaling parameter and stream distributions. The overloaded and underloaded scaling regimes yield the trivial QoS bounds of α and 1, respectively.

Fig. 4 exhibits the asymptotic QoS for known stream durations, q^{γ,π_k}, versus the scaling parameter γ. Also shown are two plots of simulation results illustrating the convergence to the asymptotic QoS. Recall the scaling regimes have transitions at $\gamma = \alpha = 0.5$ and $\gamma = 1$. The two simulation results

use $\lambda = 0.5$ and $\lambda = 5.0$, respectively, and a link capacity $c = \gamma\lambda\sigma\delta$. The figure illustrates that the simulation results agree with the computed asymptotic QoS for $\gamma > \alpha$. The simulations show a slower convergence to the computed asymptotic values for the overloaded regime $\gamma \leq \alpha = 0.5$. For small m in this regime, blocking of streams with large minimum subscription levels AS will permit admitted streams to temporarily increase their subscription levels until a stream with smaller AS is admitted to use that capacity. In the asymptotic regime, however, the aggregate minimum subscription level is always at capacity and admitted streams receive their minimum subscription level throughout their duration.

In our previous work [5], we analyzed sub-optimal adaptation policies. One such policy was the fair-share policy, denoted π_{fs}, where streams share the available bandwidth equally, subject to capacity and subscription level constraints. We also investigated performance under a a two-rate randomized adaptation policy, denoted π_{ra}, which selects a random subset of the streams to receive their maximum subscription level and grants the remaining streams their minimum subscription level. One can show that the asymptotic QoS under these adaptation policies is given by

$$q^{\gamma,\pi_{fs}} = q^{\gamma,\pi_{ra}} = \begin{cases} \alpha, & \gamma \leq \alpha \\ \gamma, & \alpha < \gamma < 1 \\ 1, & \gamma \geq 1. \end{cases} \tag{18}$$

Thus, the asymptotic QoS under these two sub-optimal policies achieves a linear performance improvement in γ, also plotted in Fig. 3. Note that $q^{\gamma,\pi_{fs}} = q^{\gamma,\pi_{ra}}$ are independent of the stream duration and maximum subscription level distributions. The distance between $q^{\gamma,\pi_{fs}}$ and q^{γ,π_k} indicates the optimal adaptation policy achieves an asymptotic QoS up to 25% greater than these baseline policies, which translates to a relative improvement of $(0.25/0.6)100 = 42\%$.

We investigated several other probability distributions for F_S and F_D, several of which are plotted in Fig. 5. In addition to the bounded exponential distribution, we also studied the uniform distribution and the bounded Pareto distribution. To facilitate comparison we kept the means of all the distributions the same, namely, $\mathbb{E}[S] = 0.3$ MB/s and $\mathbb{E}[D] = 180$ seconds. The uniform distributions we used are $S \sim Uni(1/10, 5/10)$ and $D \sim Uni(60, 300)$. The bounded Pareto distributions we used are $S \sim Par(1.382, 1/10, 10)$ and $D \sim Par(1.382, 60, 6000)$, where $Par(\alpha_p, m, M)$ means a Pareto distribution over (m, M) with an exponent of α_p. Table I shows the distributions, the variance of the corresponding volume $V = SD$, and five summary statistics over the rate adaptive regime ($\alpha = 1/2 \leq \gamma \leq 1$). The statistics we considered are

$$z_a = \int_\alpha^1 q^{\gamma,\pi_k} d\gamma$$
$$z_b = \int_\alpha^1 (q^{\gamma,\pi_k} - \gamma) d\gamma$$
$$z_c = \int_\alpha^1 \frac{q^{\gamma,\pi_k} - \gamma}{\gamma} d\gamma$$
$$z_d = \max_{\alpha \leq \gamma \leq 1} \{q^{\gamma,\pi_k} - \gamma\}$$
$$z_e = \max_{\alpha \leq \gamma \leq 1} \left\{ \frac{q^{\gamma,\pi_k} - \gamma}{\gamma} \right\}.$$

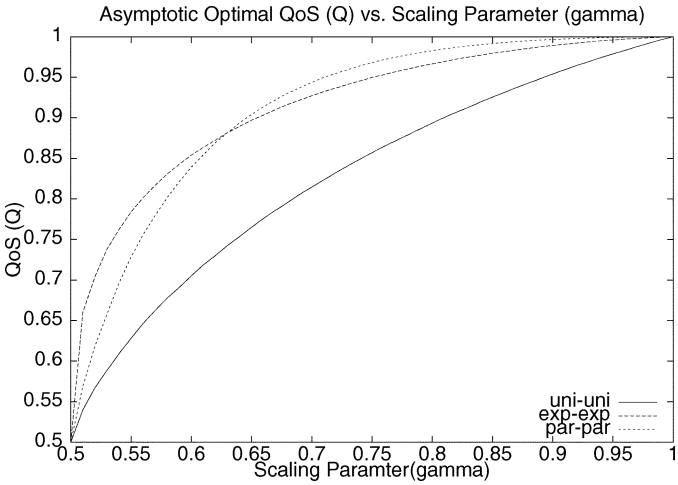

Fig. 5. Plot of the asymptotic QoS q^{γ,π_k} versus the capacity scaling parameter γ for several pairs of distributions for S and D. The pairs in Table I not plotted here are bounded above and below by one of the plots shown.

TABLE I
VARIOUS COMBINATIONS OF DISTRIBUTIONS FOR THE MAXIMUM SUBSCRIPTION LEVEL AND STREAM DURATION, THE CORRESPONDING VARIANCE, AND THE STATISTICS $z_a, \ldots, z_e$

F_S	F_D	$Var(V)$	z_a	z_b	z_c	z_d	z_e
Uni	Uni	928	0.824	0.077	0.110	0.116	0.178
Uni	Exp	3780	0.877	0.131	0.192	0.199	0.328
Exp	Uni	3780	0.877	0.131	0.192	0.199	0.328
Exp	Exp	8748	0.907	0.163	0.241	0.254	0.433
Uni	Par	10599	0.874	0.129	0.183	0.196	0.293
Par	Uni	10599	0.874	0.129	0.183	0.196	0.293
Exp	Par	20626	0.905	0.161	0.236	0.246	0.403
Par	Exp	20626	0.905	0.161	0.236	0.246	0.403
Par	Par	44600	0.904	0.160	0.231	0.254	0.402

These correspond to the average QoS, the average increase in QoS, the average improvement in QoS, the maximum increase in QoS, and the maximum improvement in QoS.

To our surprise there is not a direct correlation between the variance $Var(V)$ and any of the five summary statistics considered above.

IV. OPTIMAL ADMISSION CONTROL

Figs. 1 and 2 illustrate a serious drawback to the optimal adaptation policy: streams with intermediate volumes suffer an unacceptably high rate of adaptation. The figures also exhibit that streams with small or large volumes experience little to no adaptation. Intuitively, most streams experience zero adaptation when there are large numbers of streams sharing large capacity links because the ensemble of active streams is fairly constant, and therefore the optimal adaptation policy results in a given stream either consistently being granted its full subscription level, or consistently being granted its minimum subscription level. This insight suggests that dynamic adaptation may not be necessary for large capacity links. In particular, it suggests a static admission control policy whereby small volume streams maintain a maximum subscription level while large volume streams maintain a minimum subscription level throughout their duration.

In this section, we investigate admission and adaptation policies where, upon admission, streams are assigned a permanent subscription level based on their volume. We will in fact identify an asymptotically optimal multi-class admission control policy and show that the asymptotic QoS obtained under this policy equals that obtained under the optimal dynamic adaptation policy. The implication is that there is little benefit to dynamic adaptation when streams share large capacity links. For this section we restrict our attention to the case where stream durations are known at the time of admission, i.e., live media streams with unknown durations are not considered.

We parameterize a set of admission/adaptation policies as follows. Upon admission each stream is assigned to an adaptation class based on its volume. Streams admitted to class 1 receive their full subscription level and streams admitted to class 2 receive their minimum subscription level. We will show that two classes suffice to obtain an asymptotic QoS equal to that obtained under the dynamic adaptation policy. Note that the class of admission policies can be thought of as a subset of the class of adaptation policies, where adaptation decisions are only made at the time of a stream's admission into the network. Thus, showing that an admission policy achieves an asymptotic QoS equal to that under the optimal adaptation policy implies that admission policy is optimal.

Let $\mathbf{v} = (v_r, r \in \mathcal{R})$ denote a set of volume thresholds such that a stream with volume v admitted on route r is assigned its maximum subscription level if $v \leq v_r$ and is assigned its minimum subscription level otherwise. Note that, by definition, $s_{i,r} \in \mathcal{S}_{i,r}$ and $a_{i,r}s_{i,r} \in \mathcal{S}_{i,r}$, so that these allocations are feasible. We let $k \in (1,2)$ index the two classes on each route and we number the active streams in each class so that (i, k, r) refers to stream i in class k on route r. Define $\beta_1 = 1$ and $\beta_2 = 0$. We can then say the bandwidth assigned to an arbitrary stream admitted to class k is $\beta_k s_{i,k,r} + (1-\beta_k)a_{i,k,r}s_{i,k,r}$. With this notation we can describe the admission rule for arriving streams. A stream on route r with parameters a and s is admitted provided

$$\sum_{r \ni l} \sum_{i=1}^{n_{k,r}(t)} \beta_k s_{i,k,r} + (1-\beta_k)a_{i,k,r}s_{i,k,r} + \beta_k s + (1-\beta_k)as \leq c_l \quad \forall l \in r' \tag{19}$$

where $\mathbf{n}(t) = (n_{k,r}(t), k \in (1,2), r \in \mathcal{R})$ is the number of active streams in each class on each route.

We extend our definition of the scaling regime presented in Section III-D to an arbitrary network. We consider a sequence of networks indexed by m, where the arrival rate on route r in the m^{th} network is $\lambda_r(m) = m\lambda_r$. We define $\nu_l(m) = \sum_{r \ni l} \lambda_r(m)$ as the aggregate arrival rate on link l in the m^{th} network. Finally, we let the link capacities in the m^{th} network be given by $c_l(m) = \gamma_l \nu_l(m) \delta\sigma$. The condition $\gamma_l > \alpha$ for each $l \in \mathcal{L}$ is necessary for a low blocking regime.

The arrival rate of class k streams on route r in the m^{th} network is

$$\begin{aligned} \lambda_{1,r}(m) &= \lambda_r(m)\mathbb{P}(V \leq v_r) \\ \lambda_{2,r}(m) &= \lambda_r(m)\mathbb{P}(V > v_r). \end{aligned} \tag{20}$$

We consider multi-class admission policies that achieve an asymptotic zero blocking probability by requiring the asymptotic utilization be 1 on each link $l \in \mathcal{L}$, i.e.,

$$\lim_{m\to\infty} \frac{1}{c_l(m)} \sum_{r \ni l} \Big(\lambda_{1,r}(m)\mathbb{E}[V \mid V \leq v_r] + \alpha\lambda_{2,r}(m)\mathbb{E}[V \mid V > v_r] \Big) \leq 1 \quad \forall l \in \mathcal{L}. \tag{21}$$

It is shown in [6] that blocking is asymptotically zero for this case, although convergence is $O(1/\sqrt{c})$. Our objective is to maximize the asymptotic customer average normalized subscription level which, under the assumed asymptotic zero blocking regime, is given by

$$\lim_{m\to\infty} \frac{\sum_{r\in\mathcal{R}} \lambda_{1,r}(m) + \alpha\lambda_{2,r}(m)}{\sum_{r\in\mathcal{R}} \lambda_r(m)}. \tag{22}$$

Thus, we need to identify the optimal set of volume thresholds $\mathbf{v}$ that maximizes the asymptotic normalized subscription level (22) subject to the asymptotic utilization being bounded by 1 on each link (21).

The following result proves that two adaptation classes are sufficient to obtain the asymptotic QoS obtained under dynamic adaptation.

Theorem 4: The asymptotically optimal two-class admission policy, π_a, that achieves asymptotic zero blocking, has a volume threshold

$$v_r^{\pi_a} = \frac{\mathbb{E}[V]}{\sum_{l\in r} z_l^*} \tag{23}$$

where $z^* = (z_l^*, l \in \mathcal{L})$ is a vector of optimal Lagrange multipliers on the constraints (21). See the Appendix for proof.

The format of (23) suggests the intuitive understanding that the optimal route threshold is inversely proportional to the route cost $\sum_{l\in r} z_l^*$, using the interpretation of the Lagrange multiplier as the marginal cost of the congestion level of the link.

For the case of single bottleneck links we are able to obtain a closed form solution for the asymptotic QoS under the asymptotically optimal admission policy, π_a.

Theorem 5: The asymptotically optimal two-class admission policy, π_a, that achieves asymptotic zero blocking for the special case of single bottleneck links has a volume threshold

$$v^{\pi_a} = \begin{cases} 0, & \gamma \leq \alpha \\ F_{SD}^{-1}(\xi), & \alpha < \gamma < 1 \\ \infty, & \gamma \geq 1 \end{cases} \tag{24}$$

where $\xi = (\gamma - \alpha)/(1-\alpha)$. The asymptotic normalized subscription level under this policy is

$$q^{\gamma,\pi_a} = \begin{cases} \alpha, & \gamma \leq \alpha \\ 1 - (1-\alpha)\bar{F}_{SD}(v^{\pi_a}), & \alpha < \gamma \leq 1 \\ 1, & \gamma \geq 1. \end{cases} \tag{25}$$

See the Appendix for proof.

Notice that the asymptotic normalized subscription level under the optimal admission policy given by (25) is identical to the asymptotic QoS under the optimal dynamic adaptation policy given by (16). Thus, there is no need to perform dynamic adaptation when streams share large capacity links if intelligent admission control is performed. One caveat to this result is that

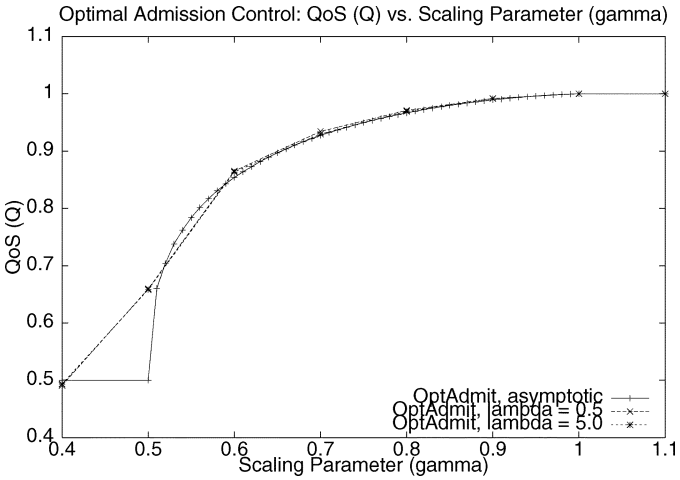

Fig. 6. Plot of the asymptotic QoS q versus the capacity scaling parameter γ. Two plots of simulation results under the optimal admission policy (known stream durations) show convergence to the asymptotic QoS.

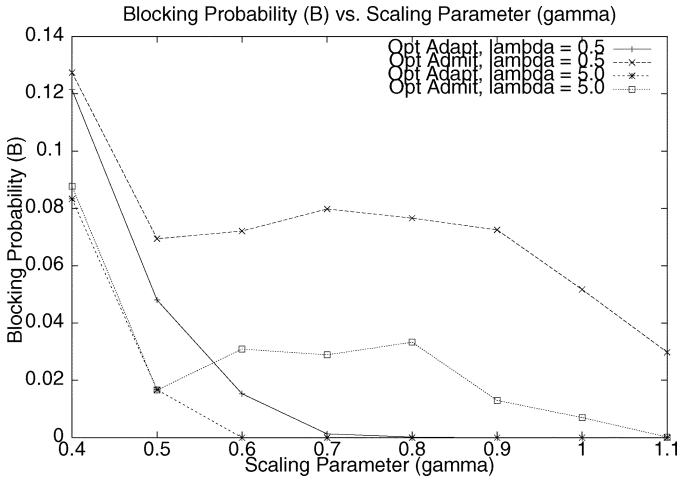

Fig. 7. Plot of the blocking probability b versus the capacity scaling parameter γ under the optimal adaptation policy and under the optimal admission policy. Two simulations were run for each case ($\lambda = (0.5, 5.0)$).

the blocking probability under the optimal dynamic adaptation policy goes to zero asymptotically fast, while the blocking probability under the optimal admission policy goes to zero as $O(1/\sqrt{c})$. In addition, the multi-class admission control implementation requires accurate assessment of the system parameters and a stationary load, while optimal adaptation does not. For this reason, optimal adaptation may outperform optimal admission control for networks servicing nonstationary workloads. Finally, optimal admission control relies upon stream durations being known at the time of admission. Optimal admission control is therefore not viable for live media.

We investigate the optimal admission control policy in Figs. 6 and 7. Fig. 6 shows a plot of the asymptotic QoS $q^{\pi_a} = q^{\pi_k}$ under the optimal admission policy and two simulation plots (with λ at 0.5 and 5.0, respectively) illustrating the convergence to the asymptotic QoS. Fig. 7 shows a plot of the fraction of streams blocked under the optimal adaptation policy and under the optimal admission control policy. Again, two simulations were used for each case. The optimal asymptotic threshold v^{π_a} was used for all simulations of optimal admission control. Fig. 7 illustrates the higher blocking probability incurred by the optimal admission policy in the regime $\alpha \leq \gamma \leq 1$. This might be expected since the optimal adaptation policy permits adaptation to admit as many streams as possible, while the optimal admission control policy does not. Both approaches yield acceptably low blocking for $\gamma > 1$, and both incur high blocking probability approaching $1 - \gamma/\alpha$ for $\gamma < \alpha$.

V. Related and Future Work

The "client" versus "system" views can be used to classify related work in the area of supporting rate adaptive multimedia streams. Representative papers investigating the client perspective include [7] and [8]. The work in [7] investigates optimal policies for streams to dynamically adapt the fraction of their available bandwidth given to base and enhancement layers. In [8], the authors propose a TCP-friendly congestion control scheme for rate adaptive video which makes smart use of buffering to absorb short time scale congestion.

Papers investigating the system perspective include [9]–[13]. The work presented in [9], [10], and [11] uses an almost identical model for QoS to ours, but neither investigates optimal adaptation, which is central to our effort. In [12], the authors offer a system level analysis of rate adaptive streams, but in a static context, i.e., a fixed number of streams. Also related is [13], which investigates a model where the server dynamically adjusts the number and rate of each subscription layer in response to congestion feedback. We feel such server adaptive models are of less interest than client adaptive models because the former does not generalize well to multicast scenarios. Our other work in this field [5], [14] also takes a system perspective.

Another body of work addresses distributed algorithms for rate adaptive multimedia streams. Representative papers include [15]–[17]. The work in [15] analyzes the performance of a rate adaptation algorithm where clients probe the network to determine congestion and then adjust their subscription level accordingly. In [16], the authors contrast a server which adapts the compression level of the stream to match with client requirements versus a server which provides a fixed set of encodings. Finally, [17] proposes a distributed algorithm for layered media with emphasis on efficient use in a multicast scenario.

A different approach to the problem of admission control is taken in [18] which identifies competitively optimal admission policies; it might be interesting to extend this work to the rate-adaptive case.

The field of media quality assessment has developed several metrics for media quality versus encoding rate [19]–[25]. These "distortion measures", e.g., sum of squared differences (SSD), mean squared error (MSE), peak signal-to-noise ratio (PSNR), are quantifiable means of assessing quality, but their correlation to human subjective evaluation is tenuous due to the complexities in the human psychovisual system [21]. These metrics are in general nonlinear functions of the encoding rate, but linear approximations (see Fig. 1) to these functions would seem reasonable within a range of interest.

Our future work on this topic is currently focused on pricing models for multiple classes of streams with different QoS guarantees. We are also interested in extending our current distributed algorithm implementations to actual networks for testing.

VI. Conclusion

The primary contributions of this work are: 1) identification of optimal adaptation policies as two-rate policies; 2) proof that static admission control achieves asymptotic optimal QoS for large capacity links multiplexing large numbers of streams; and 3) distributed algorithms that might feasibly be implemented on real networks and achieve near-optimal QoS.

We emphasize several points. First, the near-optimality of two-rate policies is a consequence of our linear QoS measure. Second, the optimality of volume discrimination is a consequence of our QoS measure being normalized by stream volume. Third, although stream volume discrimination may be unfair to large volume streams during congestion, we emphasize that it is precisely these streams that are most responsible for the congestion itself. We take the approach that such streams are perhaps therefore justified in bearing the brunt of the "adaptation load", and that to do otherwise is to risk incentive incompatibility with reducing network congestion.

Appendix

Proof of Theorem 1

Let (i, r) denote the i^{th} stream admitted into the system on route r. Define the random instantaneous QoS of stream (i, r) at a stationary time t as

$$Q_{i,r}(t) = \begin{cases} \frac{S_{i,r}(t)}{S_{i,r}D_{i,r}}, & B_{i,r} \le t \le B_{i,r} + D_{i,r} \\ 0, & \text{otherwise} \end{cases}$$

for $B_{i,r}$ the arrival time of stream (i, r). Now define the random instantaneous aggregate QoS of the network at a stationary time t as

$$Q_{\text{agg}}(t) = \sum_{r\in\mathcal{R}} \sum_{i=1}^{N_r(t)} Q_{i,r}(t).$$

Next, define the time average instantaneous aggregate QoS of the network as

$$Q_{\text{agg}} = \lim_{t\to\infty} \frac{1}{t} \int_0^t Q_{\text{agg}}(s)ds = \mathbb{E}[Q_{\text{agg}}(t)] \text{ a.s.}$$

where the second equality follows by ergodicity, and $\mathbb{E}[\cdot]$ denotes expectation with respect to the stationary distribution. Let the customer average QoS on route r be defined as

$$Q_r^0 = \lim_{n_r\to\infty} \frac{1}{n_r} \sum_{i=1}^{n_r} \int_{-\infty}^{\infty} Q_{i,r}(s)ds = \mathbb{E}^0[Q_r] \text{ a.s.}$$

where we again have the second equality by ergodicity, and $\mathbb{E}^o[\cdot]$ denotes expectation with respect to a customer average. Note that 1) the admission policy is independent of the adaptation policy, and 2) in terms of admission, the system is a stochastic knapsack with continuous sizes [6]. Thus, there is a blocking probability on each route, p_r, and the rate at which streams are admitted into the system on route r is $\lambda_r^a = \lambda_r(1 - p_r)$. Let the expected QoS of a typical stream be

$$\mathbb{E}^0[Q] = \sum_{r\in\mathcal{R}} \frac{\lambda_r^a}{\lambda^a} \mathbb{E}^0[Q_r]$$

where $\lambda^a = \sum_{r\in\mathcal{R}} \lambda_r^a$. This can be thought of as choosing a random customer by conditioning on the probability of choosing a customer from a given route. Now, straightforward application of Brumelle's Theorem [26] yields

$$\sum_{r\in\mathcal{R}} \lambda_r^a \mathbb{E}^0[Q_r] = \mathbb{E}[Q_{\text{agg}}(t)]$$

which gives

$$\lambda^a \mathbb{E}^0[Q] = \mathbb{E}[Q_{\text{agg}}(t)].$$

Thus, maximizing the expected QoS of a typical stream is equivalent to maximizing the expected instantaneous aggregate QoS at a stationary time t. Brumelle's Theorem can be understood as a generalization of Little's Law.

We restrict ourselves to nonanticipatory policies, i.e., those which only make use of information available at time t. To this end, define the filtration $\{\sigma(t), t \in \mathbb{R}\}$ to represent what is known at time t, which in this case includes the adaptivities, arrival times, durations, and maximum subscription levels of all active streams, i.e.,

$$\sigma(t) = \sigma(\{(a_{i,r}, b_{i,r}, d_{i,r}, s_{i,r}) \mid b_{i,r} \le t\})$$

where $b_{i,r}$ is the time of arrival of stream (i, r). To find the optimal adaptation policy, we will seek to maximize

$$\mathbb{E}[Q_{\text{agg}}(t) \mid \sigma(t)] = \sum_{r\in\mathcal{R}} \sum_{i=1}^{n_r(t)} \frac{s_{i,r}(t)}{s_{i,r}d_{i,r}}$$

over all feasible $\mathbf{s}(t) = (s_{i,r}(t), i = 1, \ldots, n_r(t), r \in \mathcal{R})$, where we can assume the random variables $N_r(t)$ and $D_{i,r}$ are known because they are in $\sigma(t)$. Feasible $\mathbf{s}(t)$ requires $s_{i,r}(t) \in \mathcal{S}_{i,r}$ and that the link capacity constraints be obeyed. This yields (6).

We next prove the existence of an allocation $\mathbf{s}^{\tilde{\pi}_k}(t)$ with $s_{i,r}^{\tilde{\pi}_k}(t) \in \{a_{i,r}s_{i,r},\ s_{i,r}\}$ that is nearly optimal, with a bound given by (7). We denote the value of the objective under an allocation $\mathbf{s}(t)$ as

$$G(\mathbf{s}(t)) = \sum_{r\in\mathcal{R}} \sum_{i=1}^{n_r(t)} \frac{s_{i,r}(t)}{s_{i,r}d_{i,r}}$$

and denote the load on each link under an allocation $\mathbf{s}(t)$ as

$$g_l(\mathbf{s}(t)) = \sum_{r\ni l} \sum_{i=1}^{n_r(t)} s_{i,r}(t), \quad l \in \mathcal{L}.$$

The capacity constraints will be written

$$\mathbf{g}(\mathbf{s}(t)) \le \mathbf{c} \ \Rightarrow\ g_l(\mathbf{s}(t)) \le c_l \quad \forall l \in \mathcal{L}.$$

We will also use the following notation, where the right-hand sides are understood to hold for all $i = 1, \ldots n_r(t), r \in \mathcal{R}$:

$$\mathbf{s}(t) \in \{\mathbf{as}, \mathbf{s}\} \Rightarrow s_{i,r}(t) \in \{a_{i,r}s_{i,r}, s_{i,r}\}$$
$$\mathbf{s}(t) \in \mathcal{S} \Rightarrow s_{i,r}(t) \in \mathcal{S}_{i,r}$$
$$\mathbf{s}(t) \in [\mathbf{as}, \mathbf{s}] \Rightarrow a_{i,r}s_{i,r} \le s_{i,r}(t) \le s_{i,r}.$$

Consider the problems P^x, P^{π_k}, P^y:

$$P^x : \max_{\mathbf{s}(t)} \Big\{ G(\mathbf{s}(t)) \Big| \mathbf{g}(\mathbf{s}(t)) \le \mathbf{c},\ \mathbf{s}(t) \in \{\mathbf{as}, \mathbf{s}\} \Big\}$$
$$P^{\pi_k} : \max_{\mathbf{s}(t)} \Big\{ G(\mathbf{s}(t)) \Big| \mathbf{g}(\mathbf{s}(t)) \le \mathbf{c},\ \mathbf{s}(t) \in \mathcal{S} \Big\}$$
$$P^y : \max_{\mathbf{s}(t)} \Big\{ G(\mathbf{s}(t)) \Big| \mathbf{g}(\mathbf{s}(t)) \le \mathbf{c},\ \mathbf{s}(t) \in [\mathbf{as}, \mathbf{s}] \Big\}.$$

Let us denote a solution of P^x, P^{π_k}, P^y by $\mathbf{s}^x(t), \mathbf{s}^{\pi_k}(t), \mathbf{s}^y(t)$. Note that P^y is a relaxation of P^{π_k}, and that P^{π_k} is a relaxation of P^x, implying $G(\mathbf{s}^x(t)) \leq G(\mathbf{s}^{\pi_k}(t)) \leq G(\mathbf{s}^y(t))$.

We next show there exists a solution of P^y which assigns at most all but one stream per route either its minimum or maximum subscription level by showing the value of the objective function is not decreased by changing the allocation to one that does satisfy that property. Suppose $\mathbf{s}^y(t)$ is a solution to P^y and let $(i,\ r)$ and $(j,\ r)$ be two streams on route r receiving an intermediate allocation, i.e., $a_{i,r}s_{i,r} < s^y_{i,r}(t) < s_{i,r}$ and $a_{j,r}s_{j,r} < s^y_{j,r}(t) < s_{j,r}$. Suppose that we assume $s_{i,r}d_{i,r} \leq s_{j,r}d_{j,r}$ and define $w_1 = s_{j,r}d_{j,r} - s_{i,r}d_{i,r} \geq 0$. Define $w_2 = \min\{s_{i,r} - s^y_{i,r}(t), s^y_{j,r}(t) - a_{j,r}s_{j,r}\}$. Now consider the allocation $\mathbf{s}'(t)$ where $\mathbf{s}'(t) = \mathbf{s}^y(t)$ aside from $s'_{i,r}(t) = s^y_{i,r}(t) + w_2$ and $s'_{j,r}(t) = s^y_{j,r}(t) - w_2$. Note that $\mathbf{s}'(t)$ is feasible and that either $s'_{i,r}(t) = s_{i,r}$ or $s'_{j,r}(t) = a_{j,r}s_{j,r}$ so that the allocation $\mathbf{s}'(t)$ decreases the number of streams on route r with intermediate rates by one. We can show the value of the objective function under $\mathbf{s}'(t)$ exceeds that under $\mathbf{s}^y(t)$ as follows:

$$\begin{aligned} G(\mathbf{s}'(t)) - G(\mathbf{s}^y(t)) &= \frac{s'_{i,r}(t) - s^y_{i,r}(t)}{s_{i,r}d_{i,r}} + \frac{s'_{j,r}(t) - s^y_{j,r}(t)}{s_{j,r}d_{j,r}} \\ &= \frac{w_2}{s_{i,r}d_{i,r}} - \frac{w_2}{s_{j,r}d_{j,r}} \\ &= \frac{w_1 w_2}{s_{i,r}d_{i,r}s_{j,r}d_{j,r}} \geq 0. \end{aligned}$$

We can continue to shift the allocations in this manner until at most one stream has an intermediate rate on each route.

Let $\mathbf{s}^y(t)$ therefore denote a solution to P^y with at most one stream receiving an intermediate rate per route. Define the allocation $\mathbf{s}^{\tilde{\pi}_k}(t)$ as equaling $\mathbf{s}^y(t)$ but with the allocation for the streams receiving intermediate rates set to their respective minimum. Similarly, define the allocation $\mathbf{s}^z(t)$ as equaling $\mathbf{s}^y(t)$ but with the allocation for the stream receiving intermediate rates set to their respective maximum. Clearly, $G(\mathbf{s}^{\tilde{\pi}_k}(t)) \leq G(\mathbf{s}^y(t)) \leq G(\mathbf{s}^z(t))$. Moreover, $G(\mathbf{s}^{\tilde{\pi}_k}(t)) \leq G(\mathbf{s}^x)$ since $\mathbf{s}^{\tilde{\pi}_k}(t)$ is an allocation satisfying the constraints of P^x and $\mathbf{s}^x$ is a solution to P^x. Combining these observations yields

$$G(\mathbf{s}^{\tilde{\pi}_k}(t)) \leq G(\mathbf{s}^x(t)) \leq G(\mathbf{s}^{\pi_k}(t)) \leq G(\mathbf{s}^y(t)) \leq G(\mathbf{s}^z(t)).$$

We may therefore obtain a bound in the difference in the value of the objective under π_k versus $\tilde{\pi}_k$ as

$$\begin{aligned} G(\mathbf{s}^{\pi_k}(t)) - G(\mathbf{s}^{\tilde{\pi}_k}(t)) &\leq G(\mathbf{s}^z(t)) - G(\mathbf{s}^{\tilde{\pi}_k}(t)) \\ &= \sum_{r\in\mathcal{R}} \frac{s_{i,r} - a_{i,r}s_{i,r}}{s_{i,r}d_{i,r}} \\ &= \sum_{r\in\mathcal{R}} \frac{1 - a_{i,r}}{d_{i,r}} \end{aligned}$$

where $(i,\ r)$ denotes the stream receiving an intermediate allocation under $\mathbf{s}^y(t)$.

We can then bound the relative difference in the value of the objective under π_k versus $\tilde{\pi}_k$ as

$$\begin{aligned} \frac{G(\mathbf{s}^{\pi_k}(t)) - G(\mathbf{s}^{\tilde{\pi}_k}(t))}{G(\mathbf{s}^{\pi_k}(t))} &\leq \frac{\sum_{r\in\mathcal{R}} \frac{1-a_{i,r}}{d_{i,r}}}{\sum_{r\in\mathcal{R}} \sum_{j=1}^{n_r(t)} \frac{s^{\pi_k}_{j,r}(t)}{s_{j,r}d_{j,r}}} \\ &\leq \frac{\sum_{r\in\mathcal{R}} \frac{1-a_{i,r}}{d_{i,r}}}{\sum_{r\in\mathcal{R}} \sum_{j=1}^{n_r(t)} \frac{a_{j,r}}{s_{j,r}d_{j,r}}} \\ &\leq \frac{\frac{1}{\underline{d}} \sum_{r\in\mathcal{R}} 1 - a_{i,r}}{\frac{1}{\bar{d}} \sum_{r\in\mathcal{R}} \sum_{j=1}^{n_r(t)} \frac{a_{j,r}}{s_{j,r}}} \\ &\leq \frac{\bar{s}\bar{d}}{\underline{d}} \frac{\sum_{r\in\mathcal{R}} 1 - a_{i,r}}{\sum_{r\in\mathcal{R}} \sum_{j=1}^{n_r(t)} a_{j,r}} \\ &\leq \frac{(1-\underline{a})\bar{s}\bar{d}|\mathcal{R}|}{\underline{a}\underline{d}} \frac{1}{n(t)} \end{aligned}$$

where $\underline{d}$ and $\bar{d}$ are the minimum and maximum possible durations, $\bar{s}$ is the maximum full subscription level, and $\underline{a}$ is the minimum possible adaptivity. Finally, $n(t) = \sum_{r\in\mathcal{R}} n_r(t)$ is the total number of active streams on the network at time t. Recall all random variables are assumed to have finite support, and are bounded away from 0, so $\kappa_k = (1-\underline{a})\bar{s}\bar{d}|\mathcal{R}|/\underline{a}\underline{d} < \infty$. Note this bound is very loose. Thus, for networks servicing large numbers of streams the bound goes to 0. ■

Proof of Corollary 1

The integer program (6) for the case of a single bottleneck is

$$\max_{\mathbf{s}(t)} \left\{ \sum_{i=1}^{n(t)} \frac{s_i(t)}{s_i d_i} \Big| \sum_{i=1}^{n(t)} s_i(t) \leq c, s_i(t) \in \mathcal{S}_i \right\}.$$

We use integer relaxation to transform the discrete constraint $s_i(t) \in \mathcal{S}_i$ to a continuous box constraint of the form $a_i s_i \leq s_i(t) \leq s_i$, then use the change of variables $x_i(t) = (s_i(t) - a_i s_i)/(1-a_i)$ to obtain

$$\begin{aligned} \max_{\mathbf{x}(t)} \quad & \sum_{i=1}^{n(t)} \frac{(1-a_i)x_i(t)}{s_i d_i} \\ \text{s.t.} \quad & \sum_{i=1}^{n(t)} (1-a_i)x_i(t) \leq c', \\ & 0 \leq x_i(t) \leq 1. \end{aligned}$$

where $c' = c - \sum_{i=1}^{n(t)} a_i s_i$. This is a standard knapsack relaxation problem where the weights are the $(1-a_i)s_i$, the values are $(1-a_i)/s_i d_i$, and the size of the knapsack is c'. We fill the knapsack sorted in order of decreasing value per unit weight, i.e., starting with the smallest $s_i d_i$. ■

Proof of Theorem 2

The approach used to prove Theorem 1 applies here as well. The difference is that the filtration $\sigma(t)$ doesn't include the durations of the active streams. We can recover the current ages $l_{i,r}$ of the active streams from the arrival times $b_{i,r}$ as $\{l_{i,r} = t - b_{i,r} \mid b_{i,r} \leq t\}$. This yields

$$\mathbb{E}[Q_{\text{agg}}(t) \mid \sigma(t)] = \sum_{r\in\mathcal{R}} \sum_{i=1}^{n_r(t)} \frac{s_{i,r}(t)}{s_{i,r}} \mathbb{E}\left[\frac{1}{D_{i,r}} \mid D_{i,r} > l_{i,r} \right].$$

The same considerations on feasible $\mathbf{s}(t)$ apply here yielding the same equation as (6), with $1/d_{i,r}$ replaced by $\mathbb{E}[(1/D) \mid D > l_{i,r}]$. Obtaining the bound on (10) is similar to the proof for the bound on (7), yielding

$$\kappa_u = \frac{\bar{s}\bar{d}\mathbb{E}[\frac{1}{D}](1-\underline{a})|\mathcal{R}|}{\underline{a}} < \infty. \qquad \blacksquare$$

Proof of Corollary 2

The proof follows directly from the proofs of Theorem 2 and Corollary 1. ∎

Proof of Theorem 3

Proof of (16). Let Q^{m,π_k} denote the QoS of a typical stream in the m^{th} scaling of the link capacity under the optimal adaptation policy π_k. Similarly, let $S^{m,\pi_k}(t)$ denote the instantaneous allocation to a typical stream at some time t after that stream's admission:

$$q^{\gamma,\pi_k} = \lim_{m\to\infty} \mathbb{E}^0[Q^{m,\pi_k}] = \lim_{m\to\infty} \mathbb{E}^0\left[\frac{1}{D}\int_0^D \frac{S^{m,\pi_k}(t)}{S}dt\right].$$

We can condition on $S = s$ and $D = d$ to obtain

$$q^{\gamma,\pi_k} = \lim_{m\to\infty}\int_0^\infty\int_0^\infty \mathbb{E}^0\left[\frac{1}{D}\int_0^D \frac{S^{m,\pi_k}(t)}{S}dt \mid D=d, S=s\right] dF_D(d)dF_S(s).$$

Note that since the optimal adaptation policy does not depend on the time t since the stream's admission into the system, we can claim

$$\mathbb{E}^0\left[\frac{1}{D}\int_0^D \frac{S^{m,\pi_k}(t)}{S}dt \mid D=d, S=s\right] = \mathbb{E}^0\left[\frac{S^{m,\pi_k}(t)}{S} \mid D=d, S=s\right]$$

for the t in the right-hand side understood to be a typical time. This allows

$$q^{\gamma,\pi_k} = \lim_{m\to\infty}\int_0^\infty\int_0^\infty \mathbb{E}^0\left[\frac{S^{m,\pi_k}(t)}{S} \mid D=d, S=s\right] dF_D(d)dF_S(s).$$

Next, note that under the optimal adaptation policy $S(t)/S$ is either 1 or A depending on whether or not the stream is adapted at time t. Also, note that the whether or not the stream is adapted is independent of A. We write $p(m,t,s,d)$ for the probability that a typical stream with parameters $S = s$ and $D = d$ is adapted at a typical time t in the m^{th} link.

$$\begin{aligned}
&\mathbb{E}^0\left[\frac{S^{m,\pi_k}(t)}{S} \mid D=d, S=s\right] \\
&= \int_0^1 \mathbb{E}^0\left[\frac{S^{m,\pi_k}(t)}{S} \mid D=d, S=s, A=a\right] dF_A(a) \\
&= \int_0^1 \Big[ap(m,t,s,d)] + 1(1-p(m,t,s,d))\Big] dF_A(a) \\
&= \int_0^1 \Big[1-(1-a)p(m,t,s,d)\Big] dF_A(a) \\
&= 1-(1-\alpha)p(m,t,s,d).
\end{aligned}$$

Dominated convergence allows us to move the limit inside the integrals:

$$q^{\gamma,\pi_k} = 1-(1-\alpha)\int_0^\infty\int_0^\infty \lim_{m\to\infty} p(m,t,s,d)dF_D(d)dF_S(s).$$

We focus now on $\lim_{m\to\infty} p(m,t,s,d)$. Let $N(m,t)$ denote the number of other active streams, besides the stream with volume sd, in the m^{th} system at a typical time t. The event that a stream with volume sd is adapted at a typical time t is equivalent to the event

$$\sum_{i=1}^{N(m,t)} S_i\mathbb{I}(S_i\hat{D}_i \le sd) + s + \sum_{i=1}^{N(m,t)} A_iS_i\mathbb{I}(S_i\hat{D}_i > sd) \ge c(m)$$

where we write $\hat{D}$ to denote that the durations of the $N(m,t)$ other streams active at time t have stretched distributions [26]. Thus,

$$\begin{aligned}
&p(m,t,s,d) \\
&= \mathbb{P}\left(\sum_{i=1}^{N(m,t)} S_i\mathbb{I}(S_i\hat{D}_i \le sd) + s + \sum_{i=1}^{N(m,t)} A_iS_i\mathbb{I}(S_i\hat{D}_i > sd) \ge c(m)\right) \\
&= \mathbb{P}\left(\frac{1}{m\sigma\lambda\delta}\sum_{i=1}^{N(m,t)} S_i\mathbb{I}(S_i\hat{D}_i \le sd) + \frac{s}{m\sigma\lambda\delta} + \frac{1}{m\sigma\lambda\delta}\sum_{i=1}^{N(m,t)} A_iS_i\mathbb{I}(S_i\hat{D}_i > sd) \ge \gamma\right).
\end{aligned}$$

We now define the random variable

$$Z(m,t,s,d) = \frac{1}{m\sigma\lambda\delta}\left(\sum_{i=1}^{N(m,t)} S_i\mathbb{I}(S_i\hat{D}_i \le sd) + \sum_{i=1}^{N(m,t)} A_iS_i\mathbb{I}(S_i\hat{D}_i > sd)\right)$$

so that

$$\lim_{m\to\infty} p(m,t,s,d) = \lim_{m\to\infty} \mathbb{P}\left(Z(m,t,s,d) \ge \gamma - \frac{s}{m\sigma\lambda\delta}\right).$$

We next find the mean and variance of $Z(m,t,s,d)$.

$$\begin{aligned}
\mathbb{E}[Z(m,t,s,d)] &= \frac{1}{m\sigma\lambda\delta}\mathbb{E}\left[\sum_{i=1}^{N(m,t)} S_i\mathbb{I}(S_i\hat{D}_i \le sd)\right] \\
&\quad + \frac{1}{m\sigma\lambda\delta}\mathbb{E}\left[\sum_{i=1}^{N(m,t)} A_iS_i\mathbb{I}(S_i\hat{D}_i > sd)\right].
\end{aligned}$$

By Wald's identity

$$\mathbb{E}\left[\sum_{i=1}^{N(m,t)} S_i\mathbb{I}(S_i\hat{D}_i \le sd)\right] = \mathbb{E}[N(m,t)]\mathbb{E}[S\mathbb{I}(S\hat{D} \le sd)].$$

Recall $N(m,t) \sim \text{Poisson}(m\lambda\delta)$, so that $\mathbb{E}[N(m,t)] = m\lambda\delta$. Also,

$$\mathbb{E}[S\mathbb{I}(S\hat{D} \le sd)] = \int_0^\infty\int_0^\infty x\mathbb{I}(xy \le sd)dF_{\hat{D}}(y)dF_S(x)$$

$$= \int_0^\infty x \left[\int_0^{sd/x} dF_{\hat{D}}(y) \right] dF_S(x)$$
$$= \int_0^\infty x \left[\int_0^{sd/x} \frac{1}{\mathbb{E}[D]} y dF_D(y) \right] dF_S(x).$$

Now introduce the change of variables $z = xy$:

$$\mathbb{E}[S\mathbb{I}(S\hat{D} \le sd)] = \frac{1}{\mathbb{E}[D]} \int_0^\infty \int_0^{sd} z dF_D\left(\frac{z}{x}\right) \frac{1}{x} dF_S(x)$$
$$= \frac{1}{\mathbb{E}[D]} \int_0^{sd} \left[\int_0^\infty \frac{z}{x} f_D\left(\frac{z}{x}\right) f_S(x) dx \right] dz$$
$$= \frac{1}{\mathbb{E}[D]} \int_0^{sd} z \left[f_{SD}(z) \right] dz$$
$$= \frac{\mathbb{E}[SD]}{\mathbb{E}[D]} \int_0^{sd} \frac{z}{\mathbb{E}[SD]} dF_{SD}(z)$$
$$= \sigma F_{\widehat{SD}}(sd).$$

A similar argument shows that $\mathbb{E}[AS\mathbb{I}(S\hat{D} > sd)] = \alpha\sigma\bar{F}_{\widehat{SD}}(sd)$. We combine the above results to obtain

$$\mathbb{E}[Z(m,t,s,d)] = F_{\widehat{SD}}(sd) + \alpha\bar{F}_{\widehat{SD}}(sd).$$

We next bound the variance of $Z(m,t,s,d)$. We can write

$$Z(m,t,s,d) = \frac{1}{m\sigma\lambda\delta} \sum_{i=1}^{N(m,t)} W_i$$

for $W_i = S_i(1-(1-A_i)\mathbb{I}(S_i\hat{D}_i \ge sd))$ and thereby obtain

$$Var(Z(m,t,s,d)) = \frac{1}{(m\sigma\lambda\delta)^2} \left[\mathbb{E}[N(m,t)] Var(W) + \mathbb{E}[W]^2 Var(N(m,t)) \right].$$

Recalling that $\mathbb{E}[N(m,t)] = Var(N(m,t)) = m\lambda\delta$, we obtain

$$Var(Z(m,t,s,d)) = \frac{1}{m\sigma^2\lambda\delta}\mathbb{E}[W^2] \le \frac{\mathbb{E}[S^2]}{m\sigma^2\lambda\delta}.$$

We consider three cases: 1) $\mathbb{E}[Z(m,t,s,d)] < \gamma$; 2) $\mathbb{E}[Z(m,t,s,d)] = \gamma$; and 3) $\mathbb{E}[Z(m,t,s,d)] > \gamma$. Consider the first case. Define $\epsilon(m) = \gamma - (s/m\sigma\lambda\delta) - \mathbb{E}[Z(m,t,s,d)]$. Note that $\mathbb{E}[Z(m,t,s,d)] < \gamma$ implies there exists an m' such that $\epsilon > 0$ for all $m > m'$. A little thought shows

$$\mathbb{P}\left(Z(m,t,s,d) \ge \gamma - \frac{s}{m\sigma\lambda\delta}\right) \le \mathbb{P}(|Z(m,t,s,d) - \mathbb{E}[Z(m,t,s,d)]| > \epsilon(m))$$

for all $m > m'$. Chebychev's inequality yields

$$\mathbb{P}(|Z(m,t,s,d) - \mathbb{E}[Z(m,t,s,d)]| > \epsilon(m)) \le \frac{Var(Z(m,t,s,d))}{\epsilon(m)^2} \quad \forall m > m'.$$

Noting that $\lim_{m\to\infty}\epsilon(m)$ is a constant and that $\lim_{m\to\infty} Var(Z(m,t,s,d)) = 0$ implies

$$\lim_{m\to\infty} \mathbb{P}\left(Z(m,t,s,d) \ge \gamma - \frac{s}{m\sigma\lambda\delta}\right) = 0$$

when $\mathbb{E}[Z(m,t,s,d)] < \gamma$. A similar analysis for the third case yields

$$\lim_{m\to\infty} \mathbb{P}\left(Z(m,t,s,d) \ge \gamma - \frac{s}{m\sigma\lambda\delta}\right) = 1$$

when $\mathbb{E}[Z(m,t,s,d)] > \gamma$. Finally, the set of pairs (s,d) such that $\mathbb{E}[Z(m,t,s,d)] = \gamma$ has measure zero. Thus, we conclude

$$\lim_{m\to\infty} p(m,t,s,d) = \lim_{m\to\infty} \mathbb{P}\left(Z(m,t,s,d) \ge \gamma - \frac{s}{m\sigma\lambda\delta}\right) = \mathbb{I}(\mathbb{E}[Z(m,t,s,d)] > \gamma).$$

Note that $\mathbb{I}(\mathbb{E}[Z(m,t,s,d)] > \gamma)$ is equivalent to $\mathbb{I}(sd > F^{-1}_{\widehat{SD}}(\xi))$ for $\xi = (\gamma-\alpha)/(1-\alpha)$. Substituting this into the integral yields

$$q^{\gamma,\pi_k} = 1 - (1-\alpha)\int_0^\infty \int_0^\infty \mathbb{I}(sd > F^{-1}_{\widehat{SD}}(\xi)) dF_D(d) dF_S(s)$$

which easily simplifies to the equation given in the Theorem. ■

Proof of Theorem 4

It is not difficult to show that, for any random variable V

$$\mathbb{E}[V \mid V \le v] = \mathbb{E}[V]\frac{F_{\hat{V}}(v)}{F_V(v)}$$
$$\mathbb{E}[V \mid V \ge v] = \mathbb{E}[V]\frac{\bar{F}_{\hat{V}}(v)}{\bar{F}_V(v)}.$$

Using the above, straightforward manipulations of the objective and the constraints allows us to write the problem as

$$\max_{\mathbf{v}\ge\mathbf{0}} \left\{ \sum_{r\in\mathcal{R}} \lambda_r F_V(v_r) \mid \sum_{r\ni l} \lambda_{rl} F_{\hat{V}}(v_r) \le \xi_l \quad \forall l \in \mathcal{L} \right\}$$

where $\lambda_{rl} = (\lambda_r / \sum_{s\ni l}\lambda_s) \in [0,1]$ and $\xi_l = (\gamma_l - \alpha)/(1-\alpha) \in [0,1]$. We use the change of variables $y_r = F_{\hat{V}}(v_r) \in [0,1]$ so that the problem becomes

$$\max_{\mathbf{0}\le\mathbf{y}\le\mathbf{1}} \left\{ \sum_{r\in\mathcal{R}} \lambda_r F_V(F^{-1}_{\hat{V}}(y_r)) \mid \sum_{r\ni l} \lambda_{rl} y_r \le \xi_l \quad \forall l \in \mathcal{L} \right\}.$$

Note that the constraints are linear so that the feasible region is a convex set. It is easily shown that

$$\frac{dF_V(F^{-1}_{\hat{V}}(y))}{dy} = \frac{\mathbb{E}[V]}{F^{-1}_{\hat{V}}(y)} > 0$$
$$\frac{d^2F_V(F^{-1}_{\hat{V}}(y))}{dy^2} = -\frac{\mathbb{E}[V]^2}{F^{-1}_{\hat{V}}(y)^2 f_V(F^{-1}_{\hat{V}}(y))} < 0,$$

which means the objective function is an increasing concave function. We can therefore use Lagrangian methods to identify the unique maximum. The Lagrangian is

$$L(\mathbf{y},\mathbf{z}) = \sum_{r\in\mathcal{R}} \lambda_r F_V(F^{-1}_{\hat{V}}(y_r)) + \sum_{l\in\mathcal{L}} z_l \left(\sum_{r\ni l} \lambda_{rl} y_r - \xi_l \right).$$

Taking derivatives with respect to v_r and simplifying yields

$$\frac{\partial L(\mathbf{y},\mathbf{z})}{\partial y_r} = \lambda_r \left(\frac{\mathbb{E}[V]}{F^{-1}_{\hat{V}}(y_r)} - \sum_{l\in r} z_l \right).$$

Optimality requires $(\partial L(\mathbf{y},\mathbf{z})/\partial y_r) = 0$, $\forall r \in \mathcal{R}$, which means $F^{-1}_{\hat{V}}(y_r^*) = (\mathbb{E}[V]/\sum_{l\in r} z_l)$. Using $v_r = F^{-1}_{\hat{V}}(y_r)$ yields the result. ■

Proof of Theorem 5

For the case of single bottleneck links, the constraint becomes

$$v^{\pi_a} = F_{\hat{V}}^{-1}\left(\frac{\gamma-\alpha}{1-\alpha}\right).$$

When $\gamma \leq \alpha$, asymptotic zero blocking is impossible, but is minimized by admitting all streams at their minimum subscription level, i.e., $v^* = 0$. When $\gamma \geq 1$, we obtain asymptotic zero blocking by admitting all streams at their maximum subscription level, i.e., $v^* = \infty$.

We next find the asymptotic QoS under the optimal admission policy. Let Q^{m,π_a} denote the QoS of a typical stream in the m^{th} scaling under the asymptotically optimal admission policy π_a. Then,

$$\begin{aligned} q^{\gamma,\pi_a} &= \lim_{m\to\infty} \mathbb{E}^0[Q^{m,\pi_a}] \\ &= \lim_{m\to\infty} \int_0^\infty \mathbb{E}^0[Q^{m,\pi_a} \mid V = v] dF_V(v). \end{aligned}$$

Note that, under π_a, $\mathbb{E}^0[Q^{m,\pi_a} \mid V = v]$ equals A if $v > v^{\pi_a}$ and 1 otherwise. We condition on A to get

$$\begin{aligned} \mathbb{E}^0[Q^{m,\pi_a} \mid V{=}v] &= \int_0^1 \mathbb{E}^0[Q^{m,\pi_a} \mid V{=}v, A{=}a] dF_A(a) \\ &= \int_0^1 \mathbb{I}(v{\leq}v^{\pi_a}) + a\mathbb{I}(v{>}v^{\pi_a}) dF_A(a) \\ &= 1-(1-\alpha)\mathbb{I}(v > v^{\pi_a}). \end{aligned}$$

This allows

$$\begin{aligned} q^{\gamma,\pi_a} &= 1-(1-\alpha)\int_0^\infty \mathbb{I}(v > v^{\pi_a}) dF_V(v) \\ &= 1-(1-\alpha)\bar{F}_V(v^{\pi_a}). \end{aligned}$$

■

REFERENCES

[1] Video Quality Experts Group (VQEG), "Current results and future directions," in *Proc. SPIE Visual Communications and Image Processing*, vol. 4067, 2000, pp. 742–753.

[2] B. Girod, "Psychovisual aspects of image communication," *Signal Processing*, vol. 28, no. 3, pp. 239–251, Sep. 1992.

[3] S. Shenker, "Fundamental design issues for the future internet," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 7, pp. 1176–1188, Sep. 1995.

[4] D. Bertsimas and J. N. Tsitsiklis, *Introduction to Linear Optimization*. Belmont, MA: Athena Scientific, 1997.

[5] S. Weber and G. de Veciana, "Asymptotic analysis of rate adaptive multimedia streams," in *Telecommunications Network Design and Management*, G. Anandalingam and S. Raghaven, Eds. Boston, MA: Kluwer Academic, 2003.

[6] K. Ross, *Multiservice Loss Models for Broadband Telecommunication Networks*. London, U.K.: Springer-Verlag, 1995.

[7] D. Saparilla and K. Ross, "Optimal streaming of layered video," in *Proc. IEEE INFOCOM*, 2000, pp. 737–746.

[8] R. Rejaie, M. Handley, and D. Estrin, "Quality adaptation for congestion controlled video playback over the internet," in *Proc. ACM SIGCOMM*, 1999, pp. 189–200.

[9] N. Argiriou and L. Georgiadis, "Channel sharing by rate adaptive streaming applications," in *Proc. IEEE INFOCOM*, 2002, pp. 753–762.

[10] C.-T. Chou and K. Shin, "Analysis of combined adaptive bandwidth allocation and admission control in wireless networks," in *Proc. IEEE INFOCOM*, 2002, pp. 676–684.

[11] C.-T. Chou and K. G. Shin, "Analysis of adaptive bandwidth allocation in wireless networks with multilevel degradable quality of service," *IEEE Trans. Mobile Comput.*, vol. 3, no. 1, pp. 5–17, Jan./Mar. 2004.

[12] K. Kar, S. Sarkar, and L. Tassiulas, "Optimization based rate control for multirate multicast sessions," Inst. Syst. Research, Univ. Maryland, College Park, MD, Tech. Rep., 2000.

[13] B. Vickers, C. Albuquerque, and T. Suda, "Source-adaptive multi-layered multicast algorithms for real-time video distribution," *IEEE/ACM Trans. Netw.*, vol. 8, no. 6, pp. 720–733, Dec. 2000.

[14] S. Weber and G. de Veciana, "Network design for rate adaptive multimedia streams," in *Proc. IEEE INFOCOM*, 2003, pp. 2122–2132.

[15] A. Bain and P. Key, "Modeling the performance of in-call probing for multi-level adaptive applications," Microsoft Research, Tech. Rep. MSR-TR-2002–06, 2001.

[16] S. Gorinsky and H. Vin, "The utility of feedback in layered multicast congestion control," in *Proc. NOSSDAV*, 2001, pp. 93–102.

[17] S. Gorinsky, K. K. Ramakrishnan, and H. Vin, "Addressing heterogeneity and scalability in layered multicast congestion control ," Dept. Comput. Sci., Univ. Texas, Austin, TX, Tech. Rep., 2000.

[18] S. Plotkin, "Competitive routing of virtual circuits in ATM networks," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 6, pp. 1128–1136, Aug. 1995.

[19] G. Schuster and A. K. Katsaggelos, *Rate-Distortion Based Video Compression; Optimal Video Frame Compression and Object Boundary Encoding*. Boston, MA: Kluwer Academic, 1997.

[20] A. Ortega and K. Ramchandran, "Rate-distortion methods for image and video compression," *IEEE Signal Process. Mag.*, vol. 15, no. 6, pp. 23–50, Nov. 1998.

[21] G. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *IEEE Signal Process. Mag.*, vol. 15, no. 6, pp. 74–90, Nov. 1998.

[22] K. Ramchandran, A. Ortega, and M. Vetterli, "Bit allocation for dependent quantization with applications to multiresolution and MPEG video coders," *IEEE Trans. Image Processing*, vol. 3, no. 5, pp. 533–545, Sep. 1994.

[23] P.-Y. Cheng, J. Li, and J. Kuo, "Rate control for an embedded wavelet video coder," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, no. 4, pp. 696–702, Aug. 1997.

[24] J.-J. Chen and D. Lin, "Optimal bit allocation for coding of video signals over ATM networks," *IEEE J. Sel. Areas Commun.*, vol. 15, no. 6, pp. 1002–1015, Aug. 1997.

[25] Z. He, J. Cai, and C. W. Chen, "Joint source channel rate-distortion analysis for adaptive mode selection and rate control in wireless video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 6, pp. 511–523, Jun. 2002.

[26] J. Walrand, *An Introduction to Queueing Networks*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

Steven Weber (M'03) received the B.S. degree in 1996 from Marquette University, Milwaukee, WI, and the M.S. and Ph.D. degrees from the University of Texas at Austin in 1999 and 2003, respectively.

In 2003, he joined the Department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA, where he is currently an Assistant Professor. His research interests are centered around mathematical modeling of computer and communication networks.

Gustavo de Veciana (S'88–M'94–SM'01) received the B.S., M.S, and Ph.D. degrees in electrical engineering from the University of California at Berkeley in 1987, 1990, and 1993, respectively.

He is currently a Professor at the Department of Electrical and Computer Engineering at the University of Texas at Austin. His research focuses on the design, analysis, and control of telecommunication networks. His current interests include measurement, modeling, and performance evaluation, wireless and sensor networks, and architectures and algorithms to design reliable computing and network systems.

Dr. de Veciana has been an editor for the IEEE/ACM TRANSACTIONS ON NETWORKING. He was the recipient of a General Motors Foundation Centennial Fellowship in Electrical Engineering and a 1996 National Science Foundation CAREER Award, co-recipient of the IEEE William McCalla Best ICCAD Paper Award for 2000, and co-recipient of the Best Paper in ACM *Transactions on Design Automation of Electronic Systems*, Jan 2002–2004.