Network Design For Rate Adaptive Media Streams

Steven Weber
Department of Electrical and Computer Engineering
The University of Texas at Austin
Austin, TX 78712
sweber@ece.utexas.edu

Gustavo de Veciana
Department of Electrical and Computer Engineering
The University of Texas at Austin
Austin, TX 78712
gustavo@ece.utexas.edu

Abstract— **Rate adaptive multimedia streams offer significant system and client benefits over non-adaptive streams. These benefits come at the price of increased complexity in providing adequate network support and difficulty in understanding how rate adaptation protocols affect client perceived QoS. In this paper we define quality of service in terms of the mean rate seen by the client. We identify an intuitive optimal adaptation policy that maximizes QoS. We suggest an appropriate scaling regime for rate adaptive streams and identify asymptotic QoS for large capacity networks under the optimal adaptation policy. Implementation of the optimal adaptation policy presents several obstacles that render it infeasible. We define a multi-class admission control policy that achieves asymptotically equivalent QoS to that achieved under the optimal adaptation policy, but without the need for dynamic adaptation. Our work carries implications for network designers and content providers.**

I. Introduction

Continued increasing demand for streaming multimedia data coupled with reliance on best-effort networks like the Internet has spurred interest in rate-adaptive multimedia streams. Rate adaptive multimedia streams offer the client benefit of being resilient to changing network congestion and the system benefit of permitting a large number of streams to concurrently share network resources. Multimedia streams can be adaptive because client perceived QoS is often satisfactory over a range of stream compression levels. This application adaptivity offers content providers and network designers greater flexibility than is possible for non adaptive streams. Concomitant with this flexibility are a host of questions regarding service model design and implementation feasibility.

In this paper we address several of these questions.

- *Quality of Service:* What constitutes an acceptable definition of quality of service for rate adaptive streams?
- *Optimal Adaptation:* What is the optimal adaptation policy, i.e., how do we decide to allocate network bandwidth among the various active streams so as to maximize overall QoS?
- *Analysis:* Can we characterize the expected QoS under the optimal adaptation policy and identify an appropriate scaling for large capacity networks?
- *Implementation:* How can we feasibly implement adaptation policies in a scalable distributed fashion?
- *Implications:* What are the implications of the answers to the above questions for content providers and network designers?

The rest of this section will outline the contributions of this paper for each of the above questions.

A. Quality of Service

Defining QoS for multimedia content is a particularly thorny issue. A recent publication by the Video Quality Experts Group [1] performed a statistical analysis of nine proposed objective measures of video quality. 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. We have yet to develop a satisfactory objective model of video quality.

Rate adaptive streams complicate the situation even further. Rate adaptive clients receive a video stream encoded at a time-varying compression level, i.e., they receive a high resolution stream during periods of low network congestion and a lower resolution stream during periods of higher congestion.

To maintain tractability we assume rate adaptive clients assess video quality as a function of the time-average rate of the stream and the rate at which the stream resolution changes. These QoS measures are appropriate as a first order proxy for client perceived performance. We also include blocking probability considerations in our assessment of the QoS offered by a given adaptation or admission protocol.

B. Optimal Adaptation

An adaptation policy allocates network bandwidth to active streams. We identify the adaptation policy that maximizes the expected normalized time average stream rate under two different sets of available information. First we assume all stream durations are known a priori and second we assume only the current ages of all streams are known. The former assumption is natural for the case of stored media while the latter is natural for the case of live media. Both optimal policies are "sort by volume" policies that grant the maximum feasible rate to the smallest volume streams such that there remains adequate capacity to grant the remaining streams their minimally feasible rate. For the case of stored media stream volume is the product of the stream duration times the maximum stream rate, while for the case of live media stream volume is a function of the maximum stream rate and the current age of the stream at the time of adaptation.

0-7803-7753-2/03/$17.00 (C) 2003 IEEE

C. Analysis

We identify a network scaling appropriate for large capacity networks and develop asymptotic expressions for QoS under the two optimal adaptation policies. These QoS expressions may be used to dimension large capacity links for a given QoS much as the Erlang blocking probability formula may be used for dimensioning loss networks for non-adaptive streams.

D. Implementation

Implementing the optimal adaptation policies presents severe complications which render them infeasible for all but the simplest networks. To circumvent some of these drawbacks we propose a multi-class admission policy where streams are assigned a rate upon admission to the network which they maintain throughout their duration. That is, streams do not make use of dynamic rate adaptation. There exists a simple two class admission policy that obtains an asymptotic QoS equal to that obtained under the optimal adaptation policy.

E. Implications

We assess our findings and discuss their implications. First, content providers need only offer two encodings of media content: an encoding at the lowest quality and an encoding at the highest quality. Second, network designers may profitably utilize our modes of analysis to provision capacity on networks servicing rate adaptive streams subject to blocking and QoS constraints, similar to how the Erlang blocking formula predicts the blocking probability for non-adaptive streams. In particular, bottleneck links should be provisioned to lie within the "rate adaptive scaling regime". Finally, the multi-class admission policy is feasible for networking environments with stationary offered loads.

After introducing the model in Section II, the remaining sections of the paper will investigate each of these questions in turn. In particular, Section III discusses QoS for rate adaptive multimedia content, Section IV identifies the optimal adaptation policy, Section V analyzes asymptotic QoS for large capacity networks, Section VI discusses our admission control implementation, and Section VII offers a conclusion.

II. The Model

A. Content Providers

Content providers offer multimedia content encoded at various compression levels. We abstract the complex space of possible encodings and compressions into a single parameter corresponding to the time average rate of the resulting stream, which we term the *subscription level* of the encoding. The subscription level of a stream of duration d encoded so that the instantaneous transmission rate is $(b(t), 0 \leq t \leq d)$ is

$$\frac{1}{d}\int_0^d b(t)dt. \tag{1}$$

We use the term maximum subscription level to denote the time average rate of a stream encoded at the finest resolution deemed necessary by the provider, and denote this quantity by s. We will constrain all streams to a common *adaptivity*, denoted by $\beta \in (0,1]$, with the understanding that βs is the minimum subscription level and corresponds to the coarsest resolution deemed useful by the provider. That is, the adaptivity of a stream is the ratio of the minimum subscription level over the maximum subscription level, and this ratio is assumed constant for all streams. This assumption is reasonable when considering that compression algorithms are somewhat linear, i.e., if a given compression algorithm is able to compress a 1 Mbps stream to 100 kbps, then it is likely to be able to compress a 100 kbps stream to 10 kbps. The maximum and minimum subscription levels may vary across streams, it is just the ratio that is assumed constant. To reiterate, finer resolutions than s are deemed unnecessary and coarser resolutions than βs are deemed useless. The interval $[\beta s, s]$ defines the range of reasonable subscription levels for a given stream.

The set of supported subscription levels for a given stream is denoted by $\mathcal{S} \equiv \{\beta s = s_1 < ... < s_K = s\}$ for $K \geq 2$. This definition abstracts away the underlying implementation, be it through hierarchical or simultaneous encoding. A provider utilizing hierarchical encoding would create multicast groups corresponding to each *layer* in the hierarchy with the understanding that a client subscribing to *level* s_k would subscribe to the first k layers. The aggregate bandwidth received by the client would then sum to s_k. A provider utilizing simultaneous encoding would offer K separate encodings, one at each of the offered subscription levels.

Stream durations are independent random variables, denoted by D, with common distribution F_D, and mean $\mathbb{E}[D] \equiv \mu^{-1}$. A known stream duration is denoted by d. All encodings of a given stream share the same duration, i.e., the compression level 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 levels are independent random variable, denoted by S, with a common distribution F_S and a mean $\mathbb{E}[S] \equiv \sigma$. We assume D and S are independent. The assumption that all streams share a common adaptivity requires that the minimum subscription level equal βS with probability 1 for all streams.

We define the product SD as the volume of a stream. The volume of a stream is therefore the product of its maximum subscription level and its duration, and corresponds to the total number of bits required to encode the stream at its highest useful resolution.

Throughout the paper we assume all random variables to have support on $[0, \infty)$, on which their CDF is continuous and increasing so as to guarantee the existence of an inverse and a density. We introduce some notation for CDF's. If F_X is a CDF for a random variable X then $\bar{F}_X$ is the CCDF for X. If X is a random variable with CDF F_X, then the random variable $\hat{X}$ is defined as having a CDF $F_{\hat{X}}(x) \equiv \frac{1}{\mathbb{E}[X]}\int_0^x y dF_X(y)$. Also, for two independent random variables

$X \sim F_X$ and $Y \sim F_Y$, the quantity $F_{XY}(z)$ is defined as

$$F_{XY}(z) \equiv \int_0^\infty F_X(\frac{z}{y})dF_Y(y) \equiv \int_0^\infty F_Y(\frac{z}{x})dF_X(x), \quad (2)$$

and corresponds to the probability $\mathbb{P}(XY \leq z)$. Combining these definitions implies $F_{\widehat{XY}}(z)$ be interpreted to mean

$$F_{\widehat{XY}}(z) = \frac{1}{\mathbb{E}[XY]} \int_0^z w dF_{XY}(w). \quad (3)$$

B. Network

Let $\mathcal{L}$ denote the set of links comprising the network. We assume that the capacity available on each link for multimedia content, denoted by $c_l, l \in \mathcal{L}$, is fixed. This assumption is valid in several realistic scenarios where non-streaming content is granted a lower priority under streaming traffic in deference to the more demanding service constraints of the latter. Such a protocol is envisioned to varying degrees in the proposed DiffServ and IntServ network architectures. This assumption is, roughly speaking, somewhat reflective of the Internet considering that most streaming providers use congestion–insensitive UDP for transport while most elastic traffic uses the congestion–sensitive TCP protocol.

Let $\mathcal{R}$ denote the set of routes, i.e., client/server pairs, on the network. The notation $(l \in r)$ denotes the set of links comprising route r and $(r \ni l)$ denotes the set of routes incident on link l. Let λ denote the arrival rate of new stream requests, either in the context of arrivals on a given route or on a given link. As is natural in such models, we will assume new stream requests form a Poisson process. Define the offered load as $\rho \equiv \frac{\lambda}{\mu}$, again either in the context of a route or link.

The random variables $\mathbf{N}(t) \equiv (N_r(t), r \in \mathcal{R})$ denote the number of active streams on each route at a given time t. We write $\mathbf{n}(t) \equiv (n_r(t), r \in \mathcal{R})$ when this quantity is assumed known at time t. The notation (i, r) indexes stream i on route r. For any model parameter X, the notation $X_{i,r}$ refers to stream (i, r) and implies the value $X_{i,r}$ will in general be stream dependent.

C. Admission

Rate adaptive streams have characteristics reminiscent of both elastic and inelastic traffic. They have the capability of adapting their subscription levels in response to congestion, i.e., they share network bandwidth much as elastic flows do. This adaptivity, however, is finite because streams have minimally acceptable subscription levels below which their service quality is unacceptable. In this sense they are similar to inelastic flows. Minimum resource requirements naturally lead to admission control policies. We utilize a non-preemptive full sharing admission policy, i.e., a stream is blocked only if there is insufficient bandwidth along its route to support it when all previously admitted streams incident on the route are at their respective minimum subscription levels. That is, a new stream with minimum rate βs on route r' is admitted provided

$$\sum_{r \ni l} \sum_{i=1}^{n_r} \beta s_{i,r} + \beta s \leq c_l, \ \forall l \in r'. \quad (4)$$

The admission control decision does not depend on the instantaneous subscription levels of the active streams at the time of admission, but only the minimum subscription levels associated with the active streams. Recall that stream durations are independent random variables that are also independent of the received QoS. In terms of the number of active/admitted streams, this network functions like a loss network [2] where the sizes of the arriving calls are the corresponding minimum subscription levels.

At first blush such an admission scheme would appear infeasible due to the immense amount of state information required for an admission decision, i.e., knowledge of the minimum rates associated with all active streams incident on any link comprising the route. If, however, stream packets corresponding to the minimum subscription level are aggregated in high priority queues throughout the network, then admission decisions may be made via end to end probing, as proposed in [3].

D. Adaptation

We are interested in investigating dynamic adaptation where a client changes its subscription level throughout stream playback in response to congestion changes along its route. Define an adaptation policy π as assigning an instantaneous subscription level $S^\pi(t)$ to each active stream at each time t. That is, the instantaneous subscription levels $(S^\pi_{i,r}(t), i = 1, ..., N_r(t), r \in \mathcal{R})$ of all streams active at time t are random variables which depend on the adaptation policy π.

III. Quality of Service

As mentioned in the introduction, it has proved difficult to develop accurate objective models of video quality that correspond well with results obtained from subjective tests [1]. This difficulty is exacerbated for the case of adaptive video in that the client receives a stream with a time-varying compression rate corresponding to changes in the client subscription level. We make two natural assumptions about client perceived performance for rate adaptive streams:

- stream resolution, and thus the client perceived quality, is roughly proportional to the time average bandwidth used by the video stream;
- changes in stream subscription level have an adverse effect on client perceived quality due to the distraction caused by the changing video resolution.

The first assumption is gross in that it ignores many of the intricacies inherent to most modern digital compression algorithms such as motion compensation, psychovisual considerations, the means of compression, etc. All other things being equal, however, higher video quality generally implies a higher associated average rate for the stream and vice versa. So too with the second assumption: given two clients receiving the same stream with the same average rate, the stream with fewer resolution changes will generally be thought to be of superior quality over one with more. The work in [4] offers a more detailed investigation of this phenomenon.

The more contentious issue is how to model the functional dependence of client perceived performance on these two quantities. The fact that the proposed models evaluated in [1] failed to outperform a simple linear model such as SNR suggests a linear dependence of video quality on the mean stream rate may be adequate. Indeed, the subjective evaluation of the BT.500 video quality assessment standard found in [5] suggests that client perceived performance is linear in the mean rate over the operating regime of interest. There is naturally a law of diminishing returns as the average rate of the stream increases beyond $\bar{s}$, modeled by a utility function which is eventually concave. There is also naturally a convex neighborhood around 0 corresponding to negligible video quality for stream encodings with average rates below $\underline{s}$, see [6] for a more complete discussion. Within the interval $[\underline{s}, \bar{s}]$, however, a linear model may indeed be adequate.

Assuming the preceding is valid, there is still the issue of ascribing the relative importance of these two measures in determining client perceived performance. For example, how do we compare a high bandwidth stream with frequent subscription level changes with a low bandwidth stream with few subscription level changes? Lacking any better insight, we define two quality of service measures for rate adaptive multimedia streams: the normalized time-average subscription level and the rate of adaptation, i.e., the rate at which the subscription level changes.

In particular, define the random variable

$$Q^\pi \equiv \frac{1}{D}\int_0^D \frac{S^\pi(t)}{S} dt \in [\beta, 1] \tag{5}$$

as the normalized time average subscription level seen by a randomly selected client, where the random variable D is the stream duration and the random process $(S^\pi(t), 0 \le t \le D)$ is the subscription schedule the client sees in a network operated under rate adaptation policy π. Note that $Q^\pi = \beta$ corresponds to a client receiving subscription level βS throughout its duration, and $Q^\pi = 1$ corresponds to a client receiving subscription level S throughout its duration.

Similarly, let the random variable

$$R^\pi \equiv \frac{1}{D}\sum_{t \in \mathcal{C}^\pi} |S^\pi(t^+) - S^\pi(t^-)|, \tag{6}$$

denote the rate of adaptation seen by a randomly selected client. These same two metrics have been used in several recent papers on rate adaptive streams, e.g., [7], [8].

Finally, the blocking probability is also an important aspect of system level quality of service. The blocking probability $B_{r'}(\underline{s})$ of a client with minimum subscription level βs on route r' is

$$1 - B_{r'}(\underline{s}) \equiv \mathbb{P}\Big(\sum_{r \ni l}\sum_{i=1}^{N_r} \beta S_{i,r} + \beta s \le c_l,\ \forall l \in r'\Big), \tag{7}$$

where the probability is taken with respect to the stationary distribution of the active streams. This equation simply states that the probability of acceptance is the probability that the minimum bandwidth commitment on each link is less than the link capacity. Note that the blocking probability is independent of the adaptation policy π because the admission policy is independent of π. We will focus on adaptation and admission protocols achieving asymptotic zero blocking whenever possible.

Of these three metrics we will grant precedence to the mean subscription level and will use it as our objective in searching for the optimal adaptation policy.

IV. Optimal Adaptation Policy

An adaptation policy is an allocation of network capacity to the set of active streams, subject to the constraint that the aggregate allocation to all streams incident on a given link not exceed the link capacity, i.e., $\sum_{r \ni l}\sum_{i=1}^{n_r} s_{i,r}(t) \le c_l\ \forall l \in \mathcal{L}$, and that each stream's allocation be feasible, i.e., $s_{i,r}(t) \in \mathcal{S}_{i,r}, i = 1, ..., n_r\ \forall r \in \mathcal{R}$.

Our objective is to maximize the expected normalized time average subscription level, i.e., $\mathbb{E}^\pi[Q]$, the client average performance. We will identify the optimal adaptation policy under two different sets of available information: first when stream durations are known and second when they are unknown but share a common distribution. The former corresponds to the case of stored media and the latter corresponds to the case of live media. We also assume the minimum and maximum subscription levels, i.e., βS and S, are known.

A. Known Stream Durations

In this subsection we consider the case of stored media, where the durations of all active streams are known.

Theorem 1: The adaptation policy π_k that maximizes $\mathbb{E}^\pi[Q]$ when stream durations are known is the instantaneous bandwidth allocation at each time t resulting from the solution of the following integer linear program:

$$\begin{array}{ll} \max\limits_{\mathbf{s}(t)} & \displaystyle\sum_{r \in \mathcal{R}}\sum_{i=1}^{n_r(t)} \frac{s_{i,r}(t)}{s_{i,r} d_{i,r}} \\ s.t. & \displaystyle\sum_{r \ni l}\sum_{i=1}^{n_r(t)} s_{i,r}(t) \le c_l\ \forall l \in \mathcal{L}, \\ & s_{i,r}(t) \in \mathcal{S}_{i,r}, i = 1, ..., n_r(t), \forall r \in \mathcal{R}. \end{array} \tag{8}$$

where $\mathbf{s}(t) = (s_{i,r}(t), i = 1, ..., n_r(t), r \in \mathcal{R})$ is the allocation given to each active stream. Proof: see appendix.

The theorem demonstrates two important aspects of optimal adaptation. First, optimal adaptation is "static between events", i.e., the optimal allocation only changes upon a stream departure or arrival. This follows from the fact that the parameters of (8) only change upon a departure or arrival. Second, the optimal adaptation policy is instantaneous since the optimal allocation is independent of any past adaptation experienced by the active streams.

The special case of a network with a single bottleneck link per route provides some insight into the form of the solution to the program. The solution to (8) when there is at most one bottleneck link per route is to sort streams traversing a given bottleneck by volume $v_{i,r} = s_{i,r}d_{i,r}$, granting the maximum subscription $s_{i,r}(t) = s_{i,r}$ to the small volume streams and the minimum subscription $s_{i,r}(t) = \beta s_{i,r}$ to the large volume streams.

Corollary 1: Consider a bottleneck link traversed by n active streams, labeled in order of increasing volume $v_1^{-1} > ... > v_n^{-1}$. The solution to (8) for the case of at most one bottleneck link per route is

$$s_i^{\pi_k}(t) = \begin{cases} s_i, i = 1, ..., \bar{n} - 1 \\ s_{\bar{n}}^*, i = \bar{n} \\ \beta s_i, i = \bar{n} + 1, ..., n \end{cases} \tag{9}$$

where

$$\bar{n} = \max\Big\{m \mid \sum_{i=1}^{m-1} s_i + \sum_{i=m}^{n} \beta s_i \leq c_l\Big\} \tag{10}$$

and

$$s_{\bar{n}}^* = \max\{s \in \mathcal{S}_{\bar{n}} \mid s \leq c_l - \sum_{\substack{i=1 \\ i \neq \bar{n}}}^{n} s_i^{\pi_k}(t)\}. \tag{11}$$

Proof: see appendix.

The intuition behind these results is clear: the client average QoS is maximized by granting precedence to streams using fewer network resources. The value $s_{\bar{n}}^*$ states stream $\bar{n}$ should use any residual bandwidth remaining after all other streams have received their allocation. For large capacity links servicing large numbers of streams this increase will be negligible.

The corollary is noteworthy in several points. First, the optimal adaptation policy uses only two of the offered subscription levels for each stream, i.e., s and βs. This implies content providers need only provide two encodings for each stream, provided clients are not access-line limited. If clients are access-line limited, i.e., if a_i is the access line for client i and $a_i < s_i$, then the solution in Corollary 1 requires changing s_i to $\min\{s_i, a_i\}$ in (9,10). In this case the optimal allocation may indeed make use of more than two subscription levels.

Second, the "sort by volume" policy offers clients and content providers the correct incentives to make efficient use of shared capacity. During network congestion, using a policy where small volume streams achieve a higher QoS than large volume streams encourages the use and provision of smaller volume streams.

B. Unknown Stream Durations

In this subsection we consider the case of live media where the durations of all active streams are unknown. We maintain the assumption that the minimum and maximum subscription levels, i.e., βS and S, are known.

Theorem 2: The adaptation policy π_u that maximizes $\mathbb{E}^{\pi}[Q]$ when stream durations are unknown is the instantaneous bandwidth allocation at each time t resulting from the solution of (8) with the quantity $\frac{1}{d_{i,r}}$ replaced with $\mathbb{E}[\frac{1}{D} \mid D > l_{i,r}(t)]$, where $l_{i,r}(t)$ is the current age of stream (i,r) at time t. Proof: see appendix.

If a stream is admitted at time a then its current age at time t is $l = t - a$. The allocation properties of being "static between events" and instantaneous apply to this case as well.

The solution to (8) 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 $v(t)^{-1} = \frac{1}{s}\mathbb{E}[\frac{1}{D} \mid D > l(t)]$.

Corollary 2: Consider a bottleneck link traversed by n active streams, labeled in order of increasing expected volume $v_1(t)^{-1} > ... > v_n(t)^{-1}$. The solution to (8) for the case of at most one bottleneck link per route and unknown stream durations is again given by (9) through (11). Proof: see appendix.

Not knowing stream durations a priori does not change the fact that the optimal solution still uses only two subscription levels. Also, the issue of incentives applies here as well. For the case of live streams, newly initiated streams will have lower expected durations and will therefore have better chances of being allocated a high subscription level. As the live stream gets longer, however, it becomes increasingly likely that the stream will be allocated a low subscription level.

V. Asymptotic Analysis

Our results in this section are limited to networks where each route traverses at most one bottleneck link. In practice this is often the case, especially for clients with shared access links, e.g., cable modem users, or at network peering points.

We propose a network scaling, which we call "Rate Adaptive Scaling", consisting of a linear scaling of bottleneck link capacity in the link arrival rate. Consider a sequence of links, indexed by m where the arrival process for the m^{th} link is Poisson with rate parameter $m\lambda$ and the link capacity for the m^{th} link is

$$c(m) \equiv m\alpha\sigma\mu^{-1}\lambda = m\alpha\sigma\rho. \tag{12}$$

Define the asymptotic normalized time-average subscription level under adaptation policy π and under the rate adaptive scaling with scaling parameter α to be

$$q^{\alpha,\pi} \equiv \lim_{m\to\infty} \mathbb{E}^{\pi}[Q]. \tag{13}$$

The average number of active streams in a low blocking regime is $m\rho = m\lambda\mu^{-1}$, i.e., the product of the arrival rate of new stream requests, $m\lambda$, times the average stream duration, μ^{-1}. The maximum offered load, in units of bandwidth, is $m\sigma\rho$, which is the product of the average number of active streams, $m\rho$, times the average maximum subscription level σ. The minimum offered load, in units of bandwidth, is $\beta m\sigma\rho$, which is the product of the average number of active streams, $m\rho$, times the average minimum subscription level $\beta\sigma$. The scaling parameter $\alpha = \frac{c(m)}{m\sigma\rho}$ is the ratio of available capacity, $c(m)$, over the maximum offered load, $m\sigma\rho$.

There are three natural scaling regimes, parameterized by α, that describe the average bandwidth available to a stream. The average bandwidth available to a stream is $\frac{c}{\rho}$, i.e., the bottleneck link capacity, c, divided by the average number of active streams, ρ. The scaling parameter α may be thought of as the fraction of the maximum subscription level available on the link, i.e., $\frac{c}{\rho} = \alpha\sigma$. When $\alpha > 1$, the average available

TABLE I

SCALING REGIMES, BLOCKING, AND ASYMPTOTIC QOS.

Regime	α	Blocking	QoS
Overloaded	$\alpha \leq \beta$	$1 - \frac{\alpha}{\beta}$	$q^{\alpha,\pi} = \beta$
Rate Adaptive	$\beta < \alpha < 1$	0	Theorem 3
Underloaded	$1 \leq \alpha$	0	$q^{\alpha,\pi} = 1$

bandwidth per stream $\frac{c}{\rho} > \sigma$ exceeds the average maximum subscription level σ. When $\alpha < \beta$, the average available bandwidth per stream $\frac{c}{\rho} < \beta\sigma$ is less than the average minimum subscription level $\beta\sigma$. The regime $\beta < \alpha < 1$ is characterized by an average available bandwidth per stream between the average minimum subscription level and the average maximum subscription level, i.e., $\beta\sigma < \frac{c}{\rho} < \sigma$. The characteristics of these three scaling regimes are summarized in Table I.

Consider the overloaded regime, parameterized by $\alpha \leq \beta$, where the average available bandwidth per stream is less than the average minimum subscription level. In an overloaded network, admitted streams will almost always receive their minimum subscription level, and so the normalized time average subscription level will be β for most streams. Because the system is almost always full, the Erlang blocking probability is given by

$$E(m\rho, \frac{c(m)}{\beta\sigma}) = \frac{m\rho - \frac{c(m)}{\beta\sigma}}{m\rho} = 1 - \frac{\alpha}{\beta}, \qquad (14)$$

where $\frac{c(m)}{\beta\sigma}$ is the average maximum number of streams admissible on the bottleneck link.

Consider the underloaded regime, parameterized by $\alpha \geq 1$, where the average available bandwidth per stream exceeds the average maximum subscription level. In an underloaded network, admitted streams will almost always receive their maximum subscription level, and so the normalized time average subscription level will be 1 for most streams. It follows trivially that the blocking probability will be negligible in this regime.

Finally consider the rate adaptive regime, parameterized by $\beta < \alpha < 1$, where the average available bandwidth per stream lies between the average minimum and maximum subscription levels. Because of this, admitted streams will receive a normalized time average subscription level somewhere between β and 1. This regime will also have zero blocking because

$$\lim_{m\to\infty} E(m\rho, \frac{c(m)}{\beta\sigma}) = \lim_{m\to\infty} E(m\rho, \frac{\alpha}{\beta} m\rho) = 0. \qquad (15)$$

The following theorem gives the asymptotic normalized time-average subscription level under the optimal adaptation policies.

Theorem 3: Under the optimal adaptation policy for known stream durations π_k, the asymptotic normalized time-average subscription level for networks with at most one bottleneck link per route is

$$q^{\alpha,\pi_k} = 1 - (1-\beta)\bar{F}_{SD}(F_{\hat{SD}}^{-1}(\xi)), \text{ for } \beta < \alpha < 1, \qquad (16)$$

where $\xi \equiv \frac{\alpha-\beta}{1-\beta}$.

Under the optimal adaptation policy for unknown stream durations π_u, the asymptotic normalized time-average subscription level for networks with at most one bottleneck link per route is

$$q^{\alpha,\pi_u} = 1 - (1-\beta)\int_0^\infty \int_0^\infty \frac{1}{d}\int_0^d \mathbb{I}(\frac{s\mu}{\gamma(l)} > F_{\hat{S}}^{-1}(\xi))\, dl\, dF_D(d)\, dF_S(s). \qquad (17)$$

for $\beta < \alpha < 1$, where $\xi \equiv \frac{\alpha-\beta}{1-\beta}$ and $\gamma(l) \equiv \mathbb{E}[\frac{1}{D} \mid D > l]$.
Proof: see appendix.

The equations in the theorem show that asymptotic QoS under the optimal adaptation policy is a function of the scaling parameter α, the stream adaptivity parameter β, the stream duration distribution F_D, and the maximum average rate distribution F_S. These equations can be seen as the rate adaptive analogue to the Erlang blocking probability formula for non-adaptive streams.

We claim the rate adaptive scaling regime to be the appropriate way to scale capacity for rate adaptive streams because it makes efficient use of capacity, has a low blocking probability, and exploits the adaptive capability of these streams. To investigate the significance of optimal adaptation within the rate adaptive scaling regime, we contrast the asymptotic normalized subscription level under the optimal adaptation policy with that achieved under two baseline policies which we term the fair share and two rate randomized adaptation policies. These policies are defined for the case when stream subscription levels are homogeneous, i.e., all streams have a common maximum subscription level s and a common minimum subscription level βs. The fair share adaptation policy grants all active streams an equal share of the bottleneck link capacity, bounded below by βs and bounded above by s. The two rate randomized policy grants a random set of streams s and the remaining streams βs. We show in [9] that these two baseline policies provide an asymptotic QoS of $q^\alpha = \alpha$ for $\beta < \alpha < 1$.

Figure 1 plots q^{α,π_k} vs. α for various β when stream durations are exponentially distributed. It is straightforward to show that the asymptotic normalized subscription level under an exponential stream duration distribution is

$$q^{\alpha,\pi_k} = 1 - (1-\beta)\exp\Big(W(\frac{\xi - 1}{e}) + 1\Big), \qquad (18)$$

where $W(x)$ is Lambert's W function. Also included on the plot is the straight line $q^\alpha = \alpha$ corresponding to the baseline policies. Figure 1 shows the benefit of optimal adaptation over baseline adaptation is higher for more adaptive streams, i.e., smaller β. Second, the benefit of optimal adaptation over baseline adaptation is greatest when α is well within the rate adaptive scaling regime. Finally, the benefit of optimal adaptation over baseline adaptation is higher for stream duration distributions with large variance. Indeed, when all stream durations are constant, say $D = d$, we have that $q^{\alpha,\pi_k} = \alpha$.

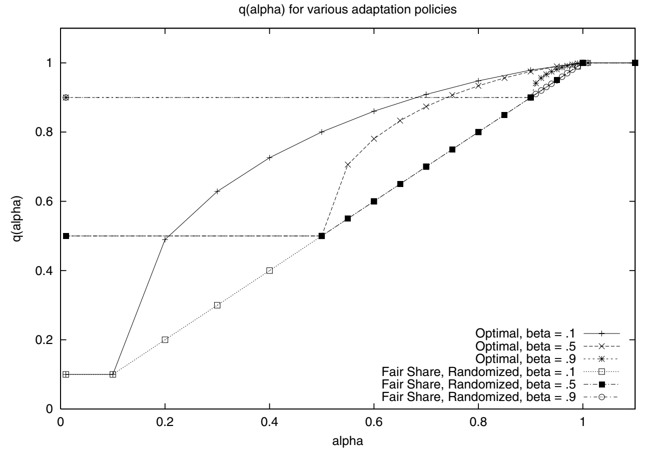

Fig. 1. q^α vs. α under the optimal adaptation policy with known stream durations and the fair share/randomized policies for $\beta = .1, .5, .9$. Stream duration distribution is exponential.

A more thorough investigation comparing simulation results with computed theoretical values may be found in [9].

Our work in [9] also points to a serious drawback with optimal adaptation. Streams with a volume near some critical threshold will undergo an unacceptably high rate of adaptation R^π. The intuition is that because the optimal adaptation policy "sorts" active streams by volume, small volume streams continuously receive their maximum subscription level S, large volume streams continuously receive their minimum subscription level βS, but those streams with intermediate volumes receive a subscription level that toggles between βS and S at a rate that increases in the arrival rate λ. The problem of a high rate of adaptation is not peculiar to the optimal adaptation policy; in [9] we demonstrate that the baseline policies we investigate suffer from this problem as well.

A more severe critique is that implementation of optimal adaptation would be infeasible due to the fact that the policy would require centralized network control which would be instantly aware of all stream arrivals and departures and would require that active streams, or the network, instantaneously adapt to changes in their allotted subscription level.

VI. Optimal Admission Control

The previous section identifies several problems with the optimal adaptation policies. In this section we do away with dynamic adaptation and instead consider a multi-class admission policy where an arriving stream request is assigned a service class based on its stream volume.

We will show that the asymptotically optimal multi-class admission policy achieves an asymptotic QoS equal to that obtained under the optimal dynamic adaptation policy. That is, because the traffic mix is roughly constant for a large capacity link, we would expect a stream's QoS under the optimal policy to depend more on *its own* volume than on the traffic it encounters on the network, which will be somewhat static in makeup. Roughly speaking, a small volume stream may expect to receive its maximum subscription level throughout its lifetime, and a large volume stream may expect to receive its minimum subscription level throughout its lifetime. This argument grants some intuition to our result in this section that the optimal multi-class admission control policy obtains an asymptotic QoS equaling that obtained under the optimal dynamic adaptation policy.

We consider the case of $K \geq 2$ classes where streams are assigned a class based on their stream volume $V = SD$. The volume thresholds are denoted $\mathbf{v} = (v_1, ..., v_{K-1})$ where $v_k \leq v_{k+1}$, $v_0 = 0$ and $v_K = \infty$. Thus a stream with volume v is assigned to class k^* if $v_{k^*-1} \leq v < v_{k^*}$. We define a vector of adaptation ratios $\vec{\beta} = (1 = \beta_1, ..., \beta_K = \beta)$ where $\beta_k > \beta_{k+1}$. The interpretation is that a stream with maximum subscription level s assigned to class k' is permitted a subscription level $\beta_{k'} s$, or the nearest feasible subscription level in $\mathcal{S}$. The stream is admitted provided

$$\sum_{k=1}^{K} \beta_k \sum_{i=1}^{n_k(t)} s_{i,k} + \beta_{k'} s \leq c \tag{19}$$

where $s_{i,k}$ is the maximum subscription level of the ith stream in class k, and $\mathbf{n}(t) = (n_k(t), k = 1, ..., K)$ is the number of active streams in each class. All streams admitted to each class k have the same normalized subscription level $\frac{\beta_k S}{S} = \beta_k$.

We will again make use of the rate adaptive scaling regime, where the arrival rate for the m^{th} link is $\lambda(m) = m\lambda$ and the capacity of the m^{th} link is $c(m) = m\alpha\sigma\rho$. The arrival rate of class k streams in the m^{th} link is

$$\lambda_k(m) = m\lambda \mathbb{P}(v_{k-1} \leq V < v_k). \tag{20}$$

The optimal adaptation policy achieves asymptotic zero blocking provided $\alpha \geq \beta$. To maintain that property for our multi-class admission policy we impose the constraint that the asymptotic utilization be 1, i.e.,

$$\lim_{m\to\infty} \frac{\sum_{k=1}^{K} \lambda_k(m) \mathbb{E}[V \mid v_{k-1} \leq V \leq v_k]\beta_k}{c(m)} = 1. \tag{21}$$

It is shown in [2] that blocking is asymptotically zero for this case, although convergence is $O(\frac{1}{\sqrt{c}})$. Our objective is to maximize the asymptotic normalized subscription level which, under the assumed asymptotic zero blocking regime, is given by

$$\lim_{m\to\infty} \sum_{k=1}^{K} \frac{\lambda_k(m)}{\lambda(m)} \beta_k. \tag{22}$$

The optimization is to identify the optimal $\mathbf{v}^*$ that solves

$$\begin{aligned} \max_{\mathbf{v}} \quad & \lim_{m\to\infty} \sum_{k=1}^{K} \frac{\lambda_k(m)}{\lambda(m)} \beta_k \qquad (23) \\ s.t. \quad & \lim_{m\to\infty} \sum_{k=1}^{K} \frac{\lambda_k(m) \mathbb{E}[V \mid v_{k-1} \leq V \leq v_k]\beta_k}{c(m)} = 1. \end{aligned}$$

The following theorem identifies the asymptotically optimal multi-class admission policy that achieves asymptotic zero blocking. The theorem gives an expression for the asymptotic normalized subscription level under that admission policy equaling that obtained for the asymptotic normalized subscription level under the optimal dynamic adaptation policy.

Theorem 4: The asymptotically optimal multi-class admission policy that achieves asymptotic zero blocking for networks with at most one bottleneck link per route is a two-class policy with a volume threshold

$$v^* = \begin{cases} 0, & \alpha \le \beta \\ F^{-1}_{\widehat{SD}}(\frac{\alpha-\beta}{1-\beta}), & \beta < \alpha < 1 \\ \infty, & \alpha > 1 \end{cases} . \tag{24}$$

The asymptotic normalized subscription level under this policy is

$$q^{\alpha,\pi_a} = \begin{cases} \beta, & \alpha \le \beta \\ 1-(1-\beta)\bar{F}_{SD}(F^{-1}_{\widehat{SD}}(\frac{\alpha-\beta}{1-\beta})), & \beta < \alpha \le 1 \\ 1, & \alpha > 1 \end{cases} \tag{25}$$

where π_a denotes the optimal admission policy. Proof: see appendix.

Thus $\mathbf{v}^* = (v^*, ..., v^*)$ and so, as for optimal adaptation, the only required subscription levels are βS and S. The optimal threshold depends upon the rate adaptive scaling parameter α, the stream adaptivity β, the duration distribution F_D, and the maximum subscription level distribution F_S.

Optimal admission control achieves the same normalized subscription level as optimal adaptation, achieves a superior rate of adaptation, and maintains asymptotically zero blocking. The caveat here is that the blocking probability under optimal adaptation goes to zero exponentially fast, while the blocking probability under optimal admission control only goes to zero as $O(\frac{1}{\sqrt{c}})$. In addition, the multi-class admission control implementation requires accurate assessment of the system parameters, while optimal adaptation does not. For this reason optimal adaptation may outperform optimal admission control for networks servicing non-stationary 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.

VII. Conclusion and Related Work

We have provided a system level analysis of performance and design issues surrounding rate adaptive networks. Our primary contributions include the following.

- Intermediate subscription levels between βs and s may be superfluous under certain circumstances.
- Optimal adaptation involves discriminating against streams based on stream volume, offers significant performance improvement over baseline adaptation policies, but its implementation may be infeasible.
- Analysis of QoS under the rate adaptive scaling yields useful expressions that can be used to help dimension networks and identify bottlenecks.
- Multi-class admission control achieves the QoS benefits of optimal adaptation but requires accurate knowledge of system parameters.

Related work includes [10], [7], [11], [12], [13], [14], [15], [8]. [10] investigates optimal policies to dynamically adapt the fraction of the available bandwidth given to a base and enhancement layer. Their work differs from ours in that it takes is a client-centric view while ours is a system-centric view. Both [7] and [8] use an almost identical model for QoS as ours, but neither investigates optimal adaptation, which is central to our effort. [11] proposes a TCP-friendly congestion control scheme for rate adaptive video which makes smart use of buffering to absorb short time scale congestion. This paper also takes a client-centric view. [12], [13] investigates many of the same issues, but focuses on services for clients with heterogeneous access line rates. They focus on aligning offered subscription levels to the bandwidth available to clients in this environment and therefore come to different conclusions regarding the benefit of providing additional encoding levels. [14] offers a system level analysis of rate adaptive streams, but in a static context, i.e., a fixed number of streams. [15] 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 current challenge in this area is to develop loss-reactive adaptation mechanisms that achieve near optimal performance as optimal adaptation. McCanne's RLM [16] adaptation mechanism utilizes sustained packet loss as a signal to streams that their subscription level is too high for the bandwidth available along their route. We propose a similar scheme but where the stream's sensitivity to loss depends on the stream volume and the mean stream volume. Streams with a volume significantly larger than the mean would have a high sensitivity to loss, i.e., fewer packet losses are required to trigger an adaptation, while streams with a volume significantly smaller than the mean would have a lower sensitivity to loss. Such a mechanism would allow for a higher client average normalized subscription level than does RLM because it comes closer to implementing the optimal adaptation policy.

References

[1] Video Quality Experts Group, "Current Results and Future Directions," in *Proc. SPIE Visual Communications and Image Processing*, 2000, vol. 4067, pp. 742–753.

[2] Keith Ross, *Multiservice Loss Models for Broadband Telecommunication Networks*, Springer-Verlag, London, 1995.

[3] F.P. Kelly, P.B. Key, and S. Zachary, "Distributed Admission Control," *IEEE Journal on Selected Areas in Communications*, vol. 18, pp. 2617–2628, 2000.

[4] B. Girod, "Psychovisual Aspects of Image Communications," *Signal Processing*, vol. 28, pp. 239–251, 1992.

[5] Jun–ichi Kimura et. al., "Perceived Quality and Bandwidth Characterization of Layered MPEG-2 Video Encoding," in *Proceedings of the SPIE International Symposium on Voice, Video, and Data Communications*, 1999.

[6] Scott Shenker, "Fundamental Design Issues for the Future Internet," *IEEE Journal on Selected Areas in Communications*, vol. 13, 1995.

[7] N. Argiriou and L. Georgiadis, "Channel Sharing By Rate Adaptive Streaming Applications," in *Proceedings of IEEE Infocom*, 2002.

[8] Chun-Ting Chou and Kang Shin, "Analysis of Combined Adaptive Bandwidth Allocation and Admission Control in Wireless Networks," in *Proceedings of IEEE Infocom*, 2002.

[9] Steven Weber and Gustavo de Veciana, "Asymptotic Analysis of Rate Adaptive Multimedia Streams," in *Telecommunications Network Design and Management*, ed. R. Anandalingam and S. Raghaven, Kluwer Academic Press, p167—192, 2002.

[10] Despina Saparilla and Keith Ross, "Optimal Streaming of Layered Video," in *Proceedings of Infocom*, 2000.

[11] Reza Rejaie, Mark Handley, and Deborah Estrin, "Quality Adaptation for Congestion Controlled Video Playback Over the Internet," in *SIGCOMM*, 1999, pp. 189–200.

[12] Sergey Gorinsky and Harrick Vin, "The Utility of Feedback in Layered Multicast Congestion Control," in *Proceedings of NOSSDAV*, 2001.

[13] Sergey Gorinsky, K. K. Ramakrishnan, and Harrick Vin, "Addressing Heterogeneity and Scalability in Layered Multicast Congestion Control," Tech. Rep., Department of Computer Sciences, The University of Texas at Austin, 2000.

[14] Koushik Kar, Saswati Sarkar, and Leandros Tassiulas, "Optimization Based Rate Control For Multirate Multicast Sessions," Tech. Rep., Institute of Systems Research and University of Maryland, 2000.

[15] B. Vickers, C. Alburquerque, and T. Suda, "Source-Adaptive Multi-Layered Multicast Algorithms for Real-Time Video Distribution," *IEEE/ACM Transactions on Networking*, December 2000.

[16] Steve McCanne, *Scalable Compression and Transmission of Internet Multicast Video*, Ph.D. thesis, University of California at Berkeley, 1996.

[17] Karl Sigman, "Notes on Little's Law and its Generalizations," http://www.ieor.columbia.edu/∼sigman.

[18] Jean Walrand, *An Introduction to Queueing Networks*, Prentice-Hall, New Jersey, 1988.

APPENDIX

Proof of Theorem 1. Define the instantaneous QoS of stream (i,r) at time t as $Q_{i,r}(t) = \frac{S_{i,r}(t)}{S_{i,r}D_{i,r}}$ and define the instantaneous aggregate QoS of the network at time t as

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

Straightforward application of Brumelle's Result [17] for stationary ergodic processes yields

$$\mathbb{E}^{\pi}[Q] \propto \mathbb{E}^{\pi}[Q_{agg}(t)].$$

Brumelle's Result can be understood as a generalization of Little's Law. Thus maximizing $\mathbb{E}^{\pi}[Q]$ is equivalent to maximizing $\mathbb{E}^{\pi}[Q_{agg}(t)]$ at some stationary time t.

We restrict ourselves to non-anticipatory 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 durations and maximum subscription levels of all active streams, i.e.,

$$\sigma(t) = \sigma(\{(A_{i,r}, D_{i,r}, S_{i,r}) \mid A_{i,r} \le t\})$$

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

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

over all feasible $\mathbf{s}(t) = (s_{i,r}(t), i = 1, ..., n_r(t), r \in \mathcal{R})$, 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. The theorem follows. ■

Proof of Corollary 1. The linear program (8) for the case of a single bottleneck is

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

We use integer relaxation to transform the discrete constraint $s_i(t) \in \mathcal{S}_i$ to a continuous box constraint of the form $\beta s_i \le s_i(t) \le s_i$. Using the change of variables $x_i(t) = \frac{s_i(t)}{s_i}$ yields

$$\max_{\mathbf{x}(t)} \Big\{ \sum_{i=1}^{n} \frac{x_i(t)}{d_i} \Big| \sum_{i=1}^{n} s_i x_i(t) \le c, \beta \le x_i(t) \le 1 \Big\}.$$

Changing variables again via $y_i(t) = \frac{x_i(t)-\beta}{1-\beta}$ and $c' = \frac{c-\sum \beta s_i}{1-\beta}$ yields

$$\max_{\mathbf{y}(t)} \Big\{ \sum_{i=1}^{n} \frac{y_i(t)}{d_i} \Big| \sum_{i=1}^{n} s_i y_i(t) \le c', 0 \le y_i(t) \le 1 \Big\}.$$

This is a standard knapsack problem where the weights are the s_i, the values are $\frac{1}{d_i}$, 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 $\sigma(t)$ now is restricted to the arrival times $\{A_{i,r} \le t\}$, the durations of departed streams $\{D_{i,r} \mid A_{i,r} + D_{i,r} \le t\}$, the maximum subscription levels of all active streams $\{S_{i,r} \mid A_{i,r} \le t\}$, and the current ages of the active streams $\{L_{i,r} = t - A_{i,r} \mid A_{i,r} \le t\}$. This yields

$$\mathbb{E}^{\pi}[Q(t) \mid \sigma(t)] = \sum_{r\in\mathcal{R}} \sum_{i=1}^{n_r(t)} \frac{s_{i,r}(t)}{s_{i,r}} \mathbb{E}[\frac{1}{D} \mid D > l_{i,r}].$$

The same considerations on feasible $\mathbf{s}(t)$ apply here yielding the same equation as (8), with $\frac{1}{d_{i,r}}$ replaced by $E[\frac{1}{D} \mid D > l_{i,r}]$. ■

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

Proof of Theorem 3. Proof of (16). By Brumelle's Result (see Theorem 1), $\mathbb{E}[Q] = \mathbb{E}[\frac{S(t)}{S}]$ at a typical time t. Note that under the optimal adaptation policy $\frac{S(t)}{S}$ is either 1 or β depending on whether or not the stream is adapted at time t. We write $\{A(m,t)\}$ for the event that the stream is adapted in the m^{th} link at time t under π_k, and $\{A^c(m,t)\}$ for the event that the stream is not adapted.

$$\begin{aligned} q^{\alpha,\pi_k} &= \lim_{m\to\infty} \mathbb{E}[\frac{S(t)}{S}] \\ &= \lim_{m\to\infty} 1\mathbb{P}(\{A^c(m,t)\}) + \beta\mathbb{P}(\{A(m,t)\}) \\ &= 1 - (1-\beta) \lim_{m\to\infty} \mathbb{P}(\{A(m,t)\}) \end{aligned}$$

We next condition on S and D, and, by Dominated Convergence, move the limit inside the integrals.

$$\begin{aligned} q^{\alpha,\pi_k} &= 1 - (1-\beta) \int_0^\infty \int_0^\infty \\ &\quad [\lim_{m\to\infty} p(m,t,s,d)] dF_D(d) dF_S(s) \end{aligned}$$

where $p(m,t,s,d) = \mathbb{P}(\{A(m,t)\} \mid S = s, D = d)$. 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

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

where we write $\hat{D}$ to denote that the durations of the $N(m,t)$ other streams active at time t have stretched distributions [18]. Thus $p(m,t,s,d)$

$$\begin{aligned}
&= \mathbb{P}(\sum_{i=1}^{N(m,t)} S_i\mathbb{I}(S_i\hat{D}_i \le sd) + \\
&\qquad s + \beta \sum_{i=1}^{N(m,t)} S_i\mathbb{I}(S_i\hat{D}_i > sd) \ge c(m)) \\
&= \mathbb{P}(\frac{1}{m\sigma\rho} \sum_{i=1}^{N(m,t)} S_i\mathbb{I}(S_i\hat{D}_i \le sd) + \\
&\qquad \frac{s}{m\sigma\rho} + \frac{\beta}{m\sigma\rho} \sum_{i=1}^{N(m,t)} S_i\mathbb{I}(S_i\hat{D}_i > sd) \ge \alpha).
\end{aligned}$$

We now define the random variable $Z(m,t,s,d)$

$$= \frac{1}{m\sigma\rho} \sum_{i=1}^{N(m,t)} S_i\mathbb{I}(S_i\hat{D}_i \le sd) + \frac{\beta}{m\sigma\rho} \sum_{i=1}^{N(m,t)} S_i\mathbb{I}(S_i\hat{D}_i > sd)$$

so that

$$\lim_{m\to\infty} p(m,t,s,d) = \lim_{m\to\infty} \mathbb{P}(Z(m,t,s,d) \ge \alpha - \frac{s}{m\sigma\rho}).$$

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

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

By Wald's identity,

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

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

$$\begin{aligned}
\mathbb{E}[S\mathbb{I}(S\hat{D} \le sd)] &= \int_0^\infty \int_0^\infty x\mathbb{I}(xy \le sd)dF_{\hat{D}}(y)dF_S(x) \\
&= \int_0^\infty x\Big[\int_0^{\frac{sd}{x}} dF_{\hat{D}}(y)\Big]dF_S(x) \\
&= \int_0^\infty x\Big[\int_0^{\frac{sd}{x}} \frac{1}{\mathbb{E}[D]} y dF_D(y)\Big]dF_S(x).
\end{aligned}$$

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

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

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

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

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

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

for $W_i = S_i(1-(1-\beta)\mathbb{I}(S_i\hat{D}_i \ge sd))$. and thereby obtain $Var(Z(m,t,s,d)) =$

$$\frac{1}{(m\sigma\rho)^2}\Big[\mathbb{E}[N(m,t)]Var(W) + \mathbb{E}[W]^2 Var(N(m,t))\Big].$$

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

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

We consider three cases: *i)* $\mathbb{E}[Z(m,t,s,d)] < \alpha$, *ii)* $\mathbb{E}[Z(m,t,s,d)] = \alpha$, *iii)* $\mathbb{E}[Z(m,t,s,d)] > \alpha$. Consider the first case. Define $\epsilon(m) = \alpha - \frac{s}{m\sigma\rho} - \mathbb{E}[Z(m,t,s,d)]$. Note that $\mathbb{E}[Z(m,t,s,d)] < \alpha$ implies there exists an m' such that $\epsilon > 0$ for all $m > m'$. A little thought shows

$$\begin{aligned}
&\mathbb{P}(Z(m,t,s,d) \ge \alpha - \frac{s}{m\sigma\rho}) \le \\
&\qquad \mathbb{P}(|Z(m,t,s,d) - \mathbb{E}[Z(m,t,s,d)]| > \epsilon(m))
\end{aligned}$$

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

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

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

$$\lim_{m\to\infty} \mathbb{P}(Z(m,t,s,d) \ge \alpha - \frac{s}{m\sigma\rho}) = 0$$

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

$$\lim_{m\to\infty} \mathbb{P}(Z(m,t,s,d) \ge \alpha - \frac{s}{m\sigma\rho}) = 1$$

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

$$\begin{aligned}\lim_{m\to\infty} p(m,t,s,d) &= \lim_{m\to\infty} \mathbb{P}(Z(m,t,s,d) \geq \alpha - \frac{s}{m\sigma\rho}) \\ &= \mathbb{I}(\mathbb{E}[Z(m,t,s,d)] > \alpha).\end{aligned}$$

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

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

which easily simplifies to the equation given in the Theorem.

Proof of (17). The proof of q^{α,π_u} is similar to that of q^{α,π_k}. Because we sort by *expected* volume, we must condition on the age of the stream at the typical time t. It may be shown that, conditioned on a stream of duration D being in the system, the age of that stream at a typical time $t \in [0,D]$ is uniform over $[0,D]$ [18]. Proceeding as before, but also conditioning on the age of the stream yields

$$\begin{aligned}q^{\alpha,\pi_u} = 1 - (1-\beta)\int_0^\infty \int_0^\infty \frac{1}{d}\int_0^d \\ \Big[\lim_{m\to\infty} p(m,t,s,l)\Big]\, dl\ dF_D(d)\ dF_S(s).\end{aligned}$$

where $p(m,t,s,l)$ is the probability a stream with maximum subscription level s and age l is adapted in the m^{th} link at a typical time t under policy π_u. We may write

$$p(m,t,s,l) = \mathbb{P}(Z(m,t,s,l) > \alpha - \frac{s}{m\sigma\rho})$$

where

$$\begin{aligned}Z(m,t,s,l) = \frac{1}{m\sigma\rho}\sum_{i=1}^{N(m,t,)} S_i \mathbb{I}(V_i(t)^{-1} > v(t)^{-1}) \\ + \ \frac{\beta}{m\sigma\rho}\sum_{i=1}^{N(m,t,)} S_i \mathbb{I}(V_i(t)^{-1} < v(t)^{-1}).\end{aligned}$$

Here $V_i(t)^{-1} = \frac{1}{S_i}\mathbb{E}[\frac{1}{\hat{D}_i} \mid \hat{D}_i > L_i(t)]$ is the expected volume of stream i at time t, with $L_i(t)$ as the age of stream i at time t, and $v(t)^{-1} = \frac{1}{s}\mathbb{E}[\frac{1}{D} \mid D > l]$ is the expected volume of the conditioned stream. Using the fact that the joint density of the stream age and duration is $f_{L\hat{D}}(l,d) = \frac{1}{\mathbb{E}[D]} f_D(d)$ for all $0 \leq l \leq d$ ([18]) allows

$$\begin{aligned}\mathbb{E}[\frac{1}{\hat{D}_i} \mid \hat{D}_i > L_i(t)] &= \int_0^\infty \int_0^x \frac{1}{x} f_{L\hat{D}}(y,x) dy dx \\ &= \int_0^\infty \frac{1}{\mathbb{E}[D]} f_D(x) dx = \mu.\end{aligned}$$

Writing $\gamma(l) = \mathbb{E}[\frac{1}{D} \mid D > l]$, we come to

$$\begin{aligned}Z(m,t,s,l) = \frac{1}{m\sigma\rho}\sum_{i=1}^{N(m,t)} S_i \mathbb{I}(\frac{\mu}{S_i} > \frac{\gamma(l)}{s}) \\ + \ \frac{\beta}{m\sigma\rho}\sum_{i=1}^{N(m,t)} S_i \mathbb{I}(\frac{\mu}{S_i} < \frac{\gamma(l)}{s}).\end{aligned}$$

We now follow a similar program to the proof for the case of known stream durations, omitted here due to space constraints. ■

Proof of Theorem 4. It is not difficult to show that, for any random variable V,

$$\mathbb{E}[V \mid v_{k-1} \leq V < v_k] = \mathbb{E}[V]\frac{F_{\hat{V}}(v_k) - F_{\hat{V}}(v_{k-1})}{F_V(v_k) - F_V(v_{k-1})}.$$

Using the definition of $\lambda_k(m)$ we may write the optimization problem as

$$\begin{aligned}\max_{\mathbf{v}} \quad & \sum_{k=1}^{K}(F_V(v_k) - F_V(v_{k-1}))\beta_k \\ s.t. \quad & \sum_{k=1}^{K}(F_{\hat{V}}(v_k) - F_{\hat{V}}(v_{k-1}))\beta_k = \frac{\alpha\sigma\rho}{\lambda\mathbb{E}[V]} = \alpha.\end{aligned}$$

The Lagrangian is

$$\begin{aligned}L(\mathbf{v},z) &= \sum_{k=1}^{K}(F_V(v_k) - F_V(v_{k-1}))\beta_k \\ &- \ z\Big(\sum_{k=1}^{K}(F_{\hat{V}}(v_k) - F_{\hat{V}}(v_{k-1}))\beta_k - \alpha\Big).\end{aligned}$$

Taking derivatives and simplifying yields

$$\frac{\partial L(\mathbf{v},z)}{\partial v_k} = (f_V(v_k) - z f_{\hat{V}}(v_k))(\beta_k - \beta_{k+1}).$$

Using the fact that $f_{\hat{V}}(v) = \frac{1}{\mathbb{E}[V]} v f_V(v)$ yields

$$\frac{\partial L(\mathbf{v},z)}{\partial v_k} = f_V(v_k)(1 - \frac{z v_k}{\mathbb{E}[V]})(\beta_k - \beta_{k+1}).$$

Optimality requires $\frac{\partial L(\mathbf{v},z)}{\partial v_k} = 0, \forall k$, which means $v_k^* = \frac{\mathbb{E}[V]}{z}$ $\forall k$, i.e., all optimal thresholds are equal. This implies only two service classes are required.

To find the optimal threshold we consider the asymptotic zero blocking constraint for a two class admission policy:

$$\lambda_1 \mathbb{E}[V \mid V \leq v^*] + \lambda_2 \mathbb{E}[V \mid V > v^*]\beta = \alpha\sigma\rho.$$

This simplifies to

$$F_{\hat{V}}(v^*) + \bar{F}_{\hat{V}}(v^*)\beta = \alpha.$$

which yields the equation for v^* for $\beta < \alpha < 1$ since $V = SD$. When $\alpha \leq \beta$ asymptotic zero blocking is impossible, but is minimized by admitting all streams at their minimum subscription level βS, i.e., $v^* = 0$. When $\alpha \geq 1$ we obtain asymptotic zero blocking by admitting all streams at S, i.e., $v^* = \infty$.

The asymptotic normalized subscription level under the optimal admission policy is 1 for all streams with volume $V \leq v^*$ and β for all streams with volume $V > v^*$, yielding an overall asymptotic normalized subscription level $F_V(v^*) + \beta\bar{F}_V(v^*)$. Substituting the value for v^* and rearranging yields the result. ■