ASYMPTOTIC ANALYSIS OF RATE ADAPTIVE MULTIMEDIA STREAMS

Steven Weber and Gustavo de Veciana
Department of Electrical and Computer Engineering
The University of Texas at Austin
{sweber,gustavo}@ece.utexas.edu

Abstract We investigate dynamic adaptation policies for rate adaptive multimedia streams in a network where each route traverses at most one bottleneck link. Dynamic adaptation allows clients to dynamically adapt the stream subscription level, i.e., time-average stream rate, in response to changes in available link capacity, and allows the system to maintain a lower blocking probability than is possible with non-adaptive streams. We define the quality of service for rate adaptive streams using the metrics of time-average subscription level, rate of adaptation, i.e., change in subscription level, and blocking probability. We investigate two baseline policies, namely, fair share adaptation and two rate randomized adaptation, and show that each suffers from significant implementation drawbacks. We then show that the adaptation policy which maximizes the mean subscription level overcomes these drawbacks, although streams with a duration near a critical threshold may experience unacceptably high rates of adaptation. This motivates the investigation of admission policies for rate adaptive streams where a stream is given a static subscription level at the time of admission which it maintains throughout its lifetime. We identify the asymptotically optimal admission policy for rate adaptive streams and show that it achieves an expected subscription level equal to that under the optimal dynamic adaptation policy. We also show that it maintains the asymptotic zero blocking property achievable using dynamic rate adaptation but does not incur the implementation overhead and QoS drawbacks of dynamic rate adaptation. The conclusion is that near optimal QoS can be obtained using a simple admission policy which gives the maximum subscription level to short duration streams and the minimum subscription level to long duration streams.

Keywords: multimedia streams, rate adaptation, multi-service loss networks

1. Introduction

Streaming connections allow clients to play multimedia content in real time as it is transferred over the network, and therefore streams require strict service guarantees, e.g., bandwidth, delay and loss, to guarantee satisfactory client perceived performance. Rich multimedia content may consume large amounts of network resources relative to other applications—resources that may well be available on certain routes at certain times. During congestion, however, these streams lack the ability to adjust their resource consumption in response to heavier traffic. This results in heavy loss, if service is not guaranteed, or unfairness, if service is guaranteed and non-preemptive. Multimedia data, however, is adaptive in the sense that satisfactory playback may be obtained over a large range of compression levels. This fact has motivated the investigation of rate adaptive multimedia streams which offer the client the ability to dynamically change the compression/resolution of the stream during playback in response to network congestion. The canonical service model for rate adaptive streams is hierarchical encoding, e.g., McCanne, 1996; Vishwanath and Chou, 1994, where multimedia content is simultaneously encoded into a set of subscription levels offering a range of stream resolutions with a commensurate range of required bandwidth. Clients may subscribe to as many subscription levels as their available bandwidth permits and may adapt their resolution by adding or dropping levels in response to changing network congestion.

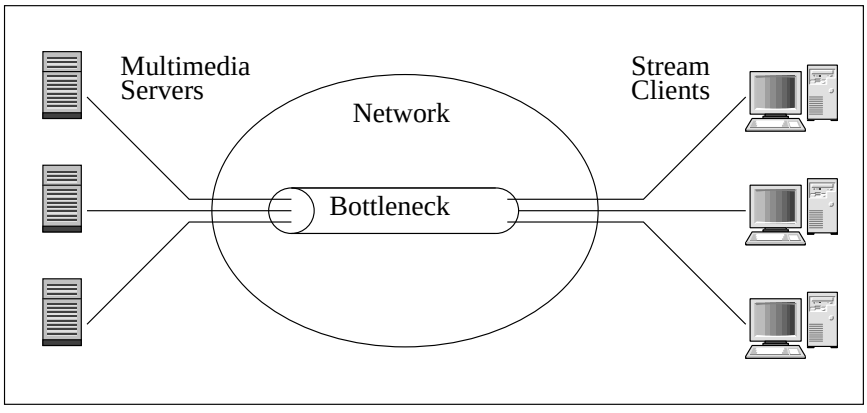

Figure 1. A network consisting of multimedia servers providing streaming content to clients, where stream routes traverse at most one bottleneck link.

The setup for the paper is shown in Figure 1 which depicts a network with stream content being transmitted from multimedia servers to

clients. We assume all active streams travel routes comprising at most one bottleneck link. The restriction to a single bottleneck link is required for purposes of obtaining closed form expressions for the mean subscription level and rate of adaptation. The policies themselves, however, are similar for general networks, see Weber and de Veciana, 2002. We assume a network protocol, e.g., DiffServ or IntServ, that guarantees a fixed amount of bandwidth is available for these streams, i.e., streams will not incur any loss.

In this paper, extending previous work in Weber and de Veciana, 2002, we investigate the critical issue of adaptation policies, i.e., when should streams adjust their subscription level, which streams should adjust, to which subscription level should they adjust, etc. Our approach is unique in that we provide a system level, as opposed to a client level, analysis of various adaptation policies in a dynamic network, i.e., where streams come and go. We investigate three aspects of quality of service associated with rate adaptive streams: the mean subscription level, the rate of adaptation, and the blocking probability. We identify three dynamic adaptation policies: fair share adaptation, two rate randomized adaptation, and an optimal adaptation policy that maximizes the time average subscription level. Although the optimal dynamic adaptation policy maximizes the mean subscription level, investigation of the rate of adaptation as a function of stream duration identifies that streams with a duration within a critical interval will experience an unacceptably high rate of adaptation. The threshold nature of the optimal dynamic adaptation policy suggests that a near optimal mean subscription level may be obtained by a duration dependent admission policy that does not dynamically adapt streams, thereby reducing the rate of adaptation to zero. That is, instead of using a single class admission policy and setting subscription levels using a given dynamic adaptation policy, we use a multi-class admission policy that sets static subscription levels for streams at the time of admission based on stream durations, which are assumed to be known a priori, and then we do not make use of dynamic adaptation. The price paid for reducing the rate of adaptation to zero is a higher blocking probability because streams may no longer adapt their subscription level to make room for new streams, but it is shown that this increase in blocking probability is negligibly small for large capacity links. The conclusion is that duration dependent admission policies obtain near optimal mean subscription levels, zero rate of adaptation, marginally higher blocking probabilities, all without the protocol and complexity overhead of dynamic rate adaptation.

Relevant work includes Saparilla and Ross, 2000; Argiriou and Georgiadis, 2002; Rejaie et al., 1999; Gorinsky and Vin, 2001; Gorinsky et al.,

2000; Kar et al., 2000; B.Vickers et al., 2000. The work in Saparilla and Ross, 2000 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. Recent work in Argiriou and Georgiadis, 2002 uses an almost identical model for QoS as ours, but pursues a very different line of analysis. Their approach doesn't seem to permit an investigation of optimal adaptation policies, which is a major focus of our work. A different tack on the problem is taken in Rejaie et al., 1999 which 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. The work in Gorinsky and Vin, 2001; Gorinsky et al., 2000 investigates many of the same issues, but with notably different results, particularly with respect to suggesting the benefit of providing additional encoding levels. The work in Kar et al., 2000 offers a system level analysis of rate adaptive streams, but in a static context, i.e., a fixed number of streams. Finally, B.Vickers et al., 2000 investigates a model where the server dynamically adjusts the number and rate of each subscription layer in response to congestion feedback.

The paper is organized as follows. Section 2 specifies the model and notation. Section 3 defines the three aspects of QoS that we deem most important for rate adaptive streams. Section 4 analyzes all three aspects of QoS for three dynamic adaptation policies. Section 5 analyzes admission policies for rate adaptive streams that identify subscription levels for streams at the time of stream initiation, but don't make use of dynamic adaptation. Section 6 concludes the paper.

2. Model and Notation

Let stream durations be independent random variables, denoted D, with a common distribution F_D, and a mean $\mathbb{E}[D] = \mu^{-1}$. We will use d to denote a known stream duration, and will write $D \sim exp(\mu))$ to denote $F_D(x) = 1 - exp(-\mu x)$. Let new stream requests arrive as a Poisson process with parameter λ. Let $\rho = \frac{\lambda}{\mu}$ denote the offered load to the link and let c be the capacity of the bottleneck link.

We abstract stream compression and encoding by considering the resulting time average mean of the compressed stream. That is, if we consider a VBR multimedia stream of duration d encoded so that the instantaneous transmissionrate is $(b(t), 0 \leq t \leq d)$ and the time average mean is

$$s = \frac{1}{d} \int_0^d b(t)dt,$$

then we speak of the stream as having a *subscription level* s. We will also speak of s as a *rate*, although this is not to be confused with the instantaneous rate $b(t)$.

Let $\bar{s}$ denote the maximally useful subscription level and $\underline{s}$ denote the minimally acceptable subscription level. The maximum subscription level corresponds to the coarsest resolution such that any finer resolution yields a negligible increase in user perception, while the minimum subscription level corresponds to the coarsest resolution deemed acceptable. Thus the interval $[\underline{s}, \bar{s}]$ defines the range of acceptable subscription levels for the stream. We define the *adaptivity* β of a stream as the ratio of its minimum and maximum subscription levels, i.e., $\beta \equiv \underline{s}/\bar{s} \in (0, 1]$.

We define the set of supported subscription levels as $\mathcal{S} \equiv \{\bar{s} = s_1 > ... > s_K = \underline{s}\}$ for $K \geq 2$. Clients may adapt their subscription level over the course of stream playback in response to changing network congestion, where the choice of the instantaneous subscription level is dictated by the enforced adaptation policy. Note that for the case of hierarchical encoding a subscription level s corresponds to subscribing to a set of layers such that the aggregate subscription is s. We abstract away the layering aspect and just consider the set of feasible subscription levels.

Let $N(t)$ denote the number of streams that are active at time t. The maximum number of streams that can be admitted without use of adaptation is $\underline{m} \equiv \lfloor \frac{c}{\bar{s}} \rfloor$ and the maximum number of streams that can be admitted with adaptation is $\bar{m} \equiv \lfloor \frac{c}{\underline{s}} \rfloor$. We require $N(t) \leq \bar{m}$, i.e., we guarantee that all admitted streams receive sufficient bandwidth to subscribe to the minimum rate or higher.

We will make use of two distinct admission policies. In Section 4 we employ a dynamic adaptation policy and a single class full sharing admission policy. That is, admission policies have no bearing on the rate received by the stream, which is handled by the adaptation policy, and a stream is admitted at time t if $N(t) < \bar{m}$. Under this assumption the process $\{N(t)\}$ has an invariant distribution $P(N(t) = n) \equiv p_{\bar{m}}(n), 0 \leq n \leq \bar{m}$ of an $M/GI/\bar{m}/\bar{m}$ queue.

In Section 5 we employ admission policies for rate adaptive streams which are a special form of multi-class stochastic knapsacks with full sharing, i.e., an arriving stream of a given class is always admitted if there is sufficient capacity for an additional stream at the subscription level associated with that class. Thus the stream class determines the subscription level, which it maintains throughout its duration in the system, i.e., no dynamic adaptation.

Let $D_1, ..., D_{N(t)}$ be the durations of the $N(t)$ streams active at time t. An adaptation policy π identifies instantaneous subscription levels for

all active streams subject to the subscription feasibility constraint

$$S_i^\pi(t) \in \mathcal{S}, i = 1, ..., N(t),$$

and the link capacity constraint

$$\sum_{i=1}^{N(t)} S_i^\pi(t) \leq c,$$

where $S_1^\pi(t), ..., S_{N(t)}^\pi(t)$ denotes the random variables associated with the instantaneous stream subscription levels of the active streams at time t under policy π. We will concentrate on dynamic adaptation policies that always make maximum use of the available capacity. This implies that admitting streams when $N(t) \geq \underline{m}$ requires that existing streams adapt their subscription levels to accommodate the newly admitted stream, i.e., for $\underline{m} \leq N(t) \leq \bar{m}$ we assume $\sum_{i=1}^{N(t)} S_i(t) = c$.

Thus, under dynamic adaptation, clients may experience streams encoded with a time-varying subscription level. We denote the client *subscription schedule* under the dynamic adaptation policy π as the random process $(S^\pi(t), 0 \leq t \leq D)$.

3. Quality of Service

We consider three aspects of the overall client perceived performance when viewing a stream encoded with a time-varying instantaneous subscription level: the time-average mean subscription level, the rate of adaptation, and the blocking probability.

The normalized time-average mean subscription level is defined as:

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

where $Q^\pi = \beta$ corresponds to a stream that receives rate $\underline{s}$ throughout its duration, and $Q^\pi = 1$ corresponds to a stream that receives rate $\bar{s}$ throughout its duration.

The time-average mean subscription level is not a complete characterization of client perceived performance. The time-average mean does not incorporate the number of changes of subscription level nor the size of those changes. Work by Girod, 1992 demonstrates that frequent changes in image resolution have deleterious effects on overall client perceived performance. The metric also is valuable from the standpoint of implementation. Real dynamic adaptation protocols will have an upper bound on the minimum time between subscription changes. We can analyze the feasibility of a suggested protocol by analyzing its rate of

adaptation to see if it falls below the specified bound. To this end we suggest a second QoS metric, the rate of adaptation, defined as

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

wher $\mathcal{C}^\pi \equiv \{t \mid 0 < t < D, S^\pi(t^+) \neq S^\pi(t^-)\}$ is the set of times at which the client subscription level changes. Thus the rate of adaptation is the time-average rate of change of the subscription level.

The third aspect of overall client perceived performance is the probability that a client is denied service, i.e., blocked. We consider only full sharing admission policies, i.e., a client is admitted whenever there exist adequate resources to support the client. In Section 4 we consider single class admission policies, and so the blocking probability is given by the Erlang B blocking formula $B(\rho, \bar{m}) = p_{\bar{m}}(\bar{m})$. In Section 5 we consider multi-class admission policies, where the blocking probability depends on the class to which the stream is assigned. In that case we will use as our metric the overall blocking probability, i.e., if streams of class k arrive at rate λ_k and have a blocking probability B_k then the overall blocking probability is $\sum_{k=1}^{K} \frac{\lambda_k}{\lambda} B_k$.

Previous work Weber and de Veciana, 2002 identified an appropriate joint load and capacity scaling regime for rate adaptive streams, and showed that non-trivial asymptotic expressions for the mean subscription level were obtainable. We define the rate adaptive scaling regime as choosing $c(\lambda) \equiv \alpha \bar{s} \rho$, for $\alpha > 0$ the rate adaptive scaling parameter, and investigating QoS as $\lambda \to \infty$ and $c = c(\lambda)$. We define

$$q^{\alpha,\pi} \equiv \lim_{\lambda \to \infty, c = c(\lambda)} \mathbb{E}[Q^\pi], \qquad r^{\alpha,\pi} \equiv \lim_{\lambda \to \infty, c = c(\lambda)} \mathbb{E}[R^\pi],$$

as the asymptotic mean subscription level and rate of adaptation under the policy π with a rate adaptive scaling parameter α. Note that $\alpha = \frac{c}{\rho \bar{s}}$ represents the capacity divided by the desired overall workload. In Weber and de Veciana, 2002 we identify $\alpha \leq \beta$ as an overloaded regime, characterized by a high blocking probability of $1 - \frac{\alpha}{\beta}$ and minimum $q^{\alpha,\pi} = \beta$. We also identify $\alpha \geq 1$ as an under-loaded regime, characterized by zero blocking and maximum $q^{\alpha,\pi} = 1$. The regime $\beta < \alpha < 1$ corresponds to the critically loaded regime with zero blocking, but a policy dependent asymptotic value for q.

Finally, a note about utility functions. Our previous work Weber and de Veciana, 2002 uses a utility function $u(s)$ giving the utility of a given subscription level,and the definition of Q is correspondingly changed to the time average mean utility. Recent work by Kimura Kimura, 1999 on MPEG-2 encoding demonstrates that our assumption of an (implicitly defined) linear utility function may be reasonable.

4. Dynamic Adaptation Policies

Throughout this section we assume a single class admission policy with full sharing as described in Section 2. Thus the blocking probability $B(\rho, \bar{m})$ is independent of the adaptation policy. We investigate three dynamic adaptation policies: fair share ($\pi = fs$), two rate randomized adaptation ($\pi = ra$), and optimal adaptation ($\pi = \pi^*$), defined as the policy which maximizes the expected mean subscription level $\mathbb{E}[Q]$. We identify the expected mean subscription level $\mathbb{E}[Q^\pi]$, the expected rate of adaptation $\mathbb{E}[R^\pi]$ for all three policies, and, when possible, identify their asymptotic analogues $q^{\alpha,\pi}$ and $r^{\alpha,\pi}$.

Fair Share Dynamic Adaptation

Under fair share adaptation, with n active streams in the system, each stream chooses subscription level

$$s_{fs}(n) \equiv \begin{cases} \bar{s}, & 0 < n \leq \underline{m} \\ \frac{c}{n}, & \underline{m} \leq n \leq \bar{m} \end{cases} .$$

That is, we assume $\mathcal{S} = \{\underline{s}, \frac{c}{\bar{m}-1}, ..., \frac{c}{\underline{m}+1}, \bar{s}\}$, so that the number of required subscription levels grows linearly in c. The following lemma gives finite capacity and asymptotic expressions for the expected mean subscription level and expected rate of adaptation under the fair share adaptation policy.

Lemma 1 *Under the fair share ($\pi = fs$) adaptation policy we have that*

$$\mathbb{E}[Q^{fs}] = \sum_{n=0}^{\bar{m}-1} p_{\bar{m}-1}(n) \frac{s_{fs}(n+1)}{\bar{s}}, \tag{1}$$

$$q^{\alpha,fs} = \begin{cases} \beta, & \alpha \leq \beta \\ \alpha, & \beta < \alpha < 1 \\ 1, & \alpha \geq 1 \end{cases} , \tag{2}$$

$$\mathbb{E}[R^{fs}] = 2\mu \sum_{n=\underline{m}}^{\bar{m}-1} p_{\bar{m}-1}(n) s_{fs}(n+1), \tag{3}$$

$$r^{\alpha,fs} = \begin{cases} 2\mu\underline{s}, & \alpha \leq \beta \\ 2\mu\bar{s}\alpha, & \beta < \alpha < 1 \\ 0, & \alpha \geq 1 \end{cases} . \tag{4}$$

Thus both $q^{\alpha,fs}, r^{\alpha,fs}$ are linear in α in the critical regime $\beta < \alpha < 1$. Figures 2 and 3 exhibit the above equations along with simulation results for $\lambda = 40$ and 320 versus α. The plots illustrate a good match between

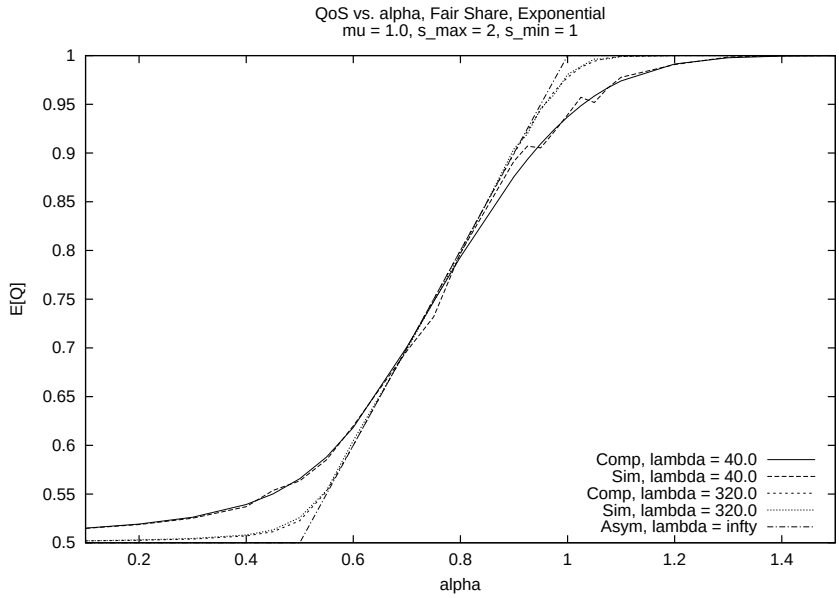

Figure 2. Simulation and computation results for the fair share dynamic adaptation policy: $\mathbb{E}[Q^{fs}]$ *and* $q^{\alpha,fs}$ *vs.* α *for* $\lambda = 40, 320$, $\mu^{-1} = 1$, $\underline{s} = 1, \bar{s} = 2$, *and* $c(\lambda) = \alpha\bar{s}\rho$. *The overloaded regime is* $\alpha \leq \beta = 0.5$, *the critically loaded regime is* $0.5 < \alpha < 1$, *and the under-loaded regime is* $\alpha \geq 1$.

computational and simulation results, as well as the convergence to the asymptotic values.

Although $r^{\alpha,fs}$ is finite, the asymptotic expected number of adaptations is infinite. Straightforward manipulations show that the expected number of subscription level changes under the fair share policy is $2\lambda\mathbb{P}(\underline{m} - 1 \leq N < \bar{m} - 1)$ which goes to infinity as λ gets large. This result is easily understood: in a loss network the number of active streams changes at rate $2\lambda(1 - B(\rho, \bar{m}))$. The difference here is that no change in rate is required by a change in the number of streams when $n < \underline{m}$.

Two Rate Randomized Dynamic Adaptation

To realize fair share adaptation content servers must provide a large set of subscription levels. The idea behind two rate randomized dynamic adaptation is that instead of adapting all streams by a small amount we can do equally wellon average by adapting a small set of streams by a larger amount. Under two rate randomized dynamic adaptation, when

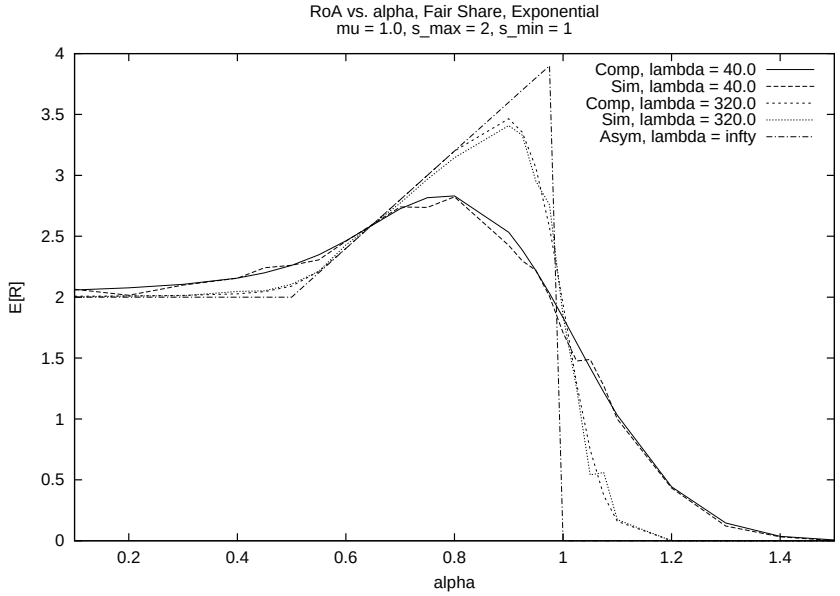

Figure 3. Simulation and computation results for the fair share dynamic adaptation policy: $\mathbb{E}[R^{fs}]$ and $r^{\alpha,fs}$ vs. α for varying λ. Same simulation parameters as Figure 2.

there are n active streams, we allocate a rate $\bar{s}$ to $\bar{n}(n)$ of the streams, chosen at random, and a rate $\underline{s}$ to the remaining $\underline{n}(n)$ streams. The functions $(\underline{n}(n), \bar{n}(n))$ are defined as

$$(\underline{n}(n), \bar{n}(n)) \equiv \begin{cases} (0, n), & 0 \leq n \leq \underline{m} \\ \left(\left\lceil \frac{n\bar{s}-c}{\bar{s}-\underline{s}} \right\rceil, \left\lfloor \frac{c-n\underline{s}}{\bar{s}-\underline{s}} \right\rfloor \right), & \underline{m} < n \leq \bar{m} \end{cases} .$$

The quantity $\bar{n}(n)$ is the maximum number of streams that can be supported at subscription level $\bar{s}$ while leaving sufficient capacity for the remaining $\underline{n}(n)$ streams to maintain a subscription level $\underline{s}$. Note that $\mathcal{S} = \{\bar{s}, \underline{s}\}$, which is significantly smaller than the set of required supported subscription levels under fair share. The following lemma gives finite capacity and asymptotic expressions for the expected mean subscription level and expected rate of adaptation under the two rate randomized adaptation policy.

Lemma 2 *Under the two rate randomized adaptation policy ($\pi = ra$) we have that*

$$\mathbb{E}[Q^{ra}] = \sum_{n=0}^{\bar{m}-1} p_{\bar{m}-1}(n)\left[\frac{\underline{n}(n+1)}{n+1}\beta + \frac{\bar{n}(n+1)}{n+1}\right], \quad (5)$$

$$q^{\alpha,ra} = q^{\alpha,fs}, \quad (6)$$

$$\mathbb{E}[R^{ra}] = 2(\bar{s}-\underline{s})\mu \sum_{n=\underline{m}}^{\bar{m}-1} p_{\bar{m}-1}(n)\frac{\bar{n}(n)\underline{n}(n+1)+\bar{n}(n+1)\underline{n}(n)}{n+1}, \quad (7)$$

$$r^{\alpha,ra} = \begin{cases} \infty, & \alpha \le \beta \\ \infty, & \beta < \alpha < 1 \\ 0, & \alpha \ge 1 \end{cases} . \quad (8)$$

Thus two rate randomized adaptation achieves an asymptotic mean subscription level equal to that under the fair share policy, but suffers from an infinite asymptotic infinite rate of adaptation for $\alpha < 1$.

Figures 4 and 5 exhibit the above equations along with simulation results for $\lambda = 40$ and 320 versus α. The plots illustrate a good match between computational and simulation results, as well as the convergence to the asymptotic values.

It can be shown that $\mathbb{E}[Q^{fs}] - \mathbb{E}[Q^{ra}] \le \frac{\bar{s}-\underline{s}}{\underline{m}}$, so that for large capacity links the expressions are nearly equal, and the difference goes to zero as the link capacity increases, i.e., $q^{\alpha,ra} = q^{\alpha,fs}$ Weber and de Veciana, 2002. The rate of adaptation, however, is drastically different for randomized adaptation than for fair share. The problem is that our formulation of the two rate randomized policy randomly selects a new set of the appropriate size to be adapted each time $N(t)$ changes and $N(t) \ge \underline{m}$. We have also investigated a randomized adaptation policy that keeps state information on stream subscription levels, and changes rate for as few streams as required by changes in $\bar{n}$ and $\underline{n}$. Under this policy we find equivalent expressions for $\mathbb{E}[Q^{ra}]$ but the values for $\mathbb{E}[R^{ra}]$ are on par with $\mathbb{E}[R^{fs}]$. The drawback to the fair share policy is that it requires a large number of supported subscription levels and requires a large number of adaptations, although the overall rate of adaptation is reasonably small. The two rate randomized adaptation policy either suffers from unacceptably high rate of adaptation or requires link state be kept to keep the rate of adaptation reasonably low.

The drawbacks to the fair share and two rate randomized adaptation policies are serious: the unacceptably high number of adaptations required by fair share and the unacceptably high rate of adaptation required by randomized adaptation render these policies infeasible for large

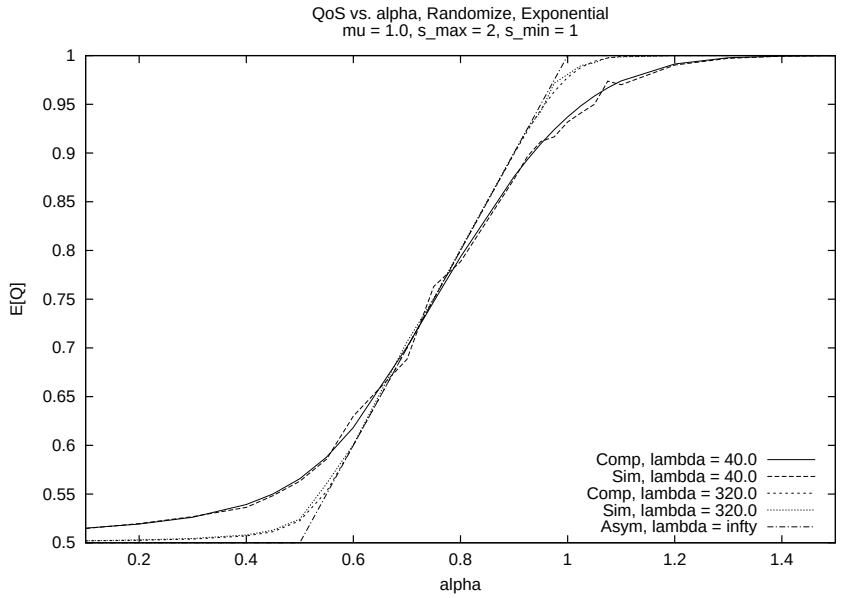

Figure 4. Simulation and Computation of $\mathbb{E}[Q^{ra}], q^{\alpha,ra}, \mathbb{E}[R^{ra}]$ *vs.* α *for varying* λ*. Same simulation parameters as Figure 2.*

capacity links. These drawbacks motivate investigation of the optimal adaptation policy, which we explore next.

Optimal Dynamic Adaptation

We use the term optimal dynamic adaptation policy to denote the policy that maximizes $\mathbb{E}[Q^{\pi}]$ over all feasible policies. The following theorem, from Weber and de Veciana, 2002, identifies this policy as granting preference to short duration streams.

Theorem 1 *Order the streams active at time t by increasing duration so that* $D_1 < ... < D_{N(t)}$*. The dynamic adaptation policy* π^* *that maximizes* $\mathbb{E}[Q^{\pi}]$ *is*

$$S_i^{\pi^*}(t) = \begin{cases} \bar{s}, & i = 1, ..., \bar{n}(N(t)) \\ \underline{s}, & i = \bar{n}(N(t)) + 1, ..., N(t) \end{cases} .$$

The intuition behind this result is simple: we maximize the long term client average mean subscription level by giving priority to short duration streams since those streams consume fewer resources.

The optimal rate of an admitted stream depends on the number of other streams in the system as well as on their durations. Let D denote

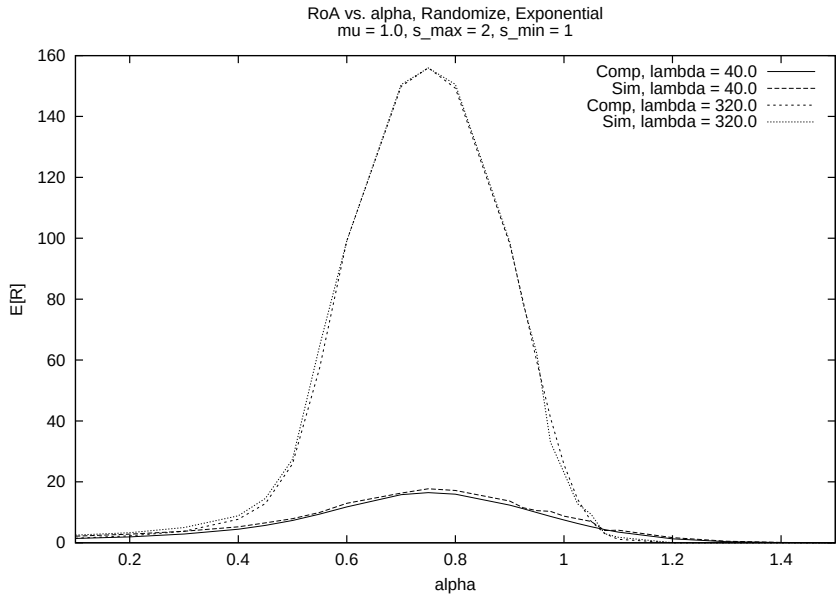

Figure 5. Simulation and Computation of $\mathbb{E}[Q^{ra}], q^{\alpha,ra}, \mathbb{E}[R^{ra}]$ *vs.* α *for varying* λ*. Same simulation parameters as Figure 2.*

the duration of an arbitrary active stream, let N denote the number of other active streams, and let $D_1, ..., D_N$ be their durations. Define

$$Y_{N,D} \equiv \sum_{i=1}^{N} 1(D_i \leq D)$$

as the number of streams with shorter durations than the considered stream of duration D. The distribution of stream durations when viewed at an arbitrarily chosen time is not $F_D(d)$ because we are more likely to see longer duration streams than short duration streams at an arbitrary time. The distribution is $F_U(u) \equiv \mu \int_0^u z dF_D(z)$ for the $M/GI/\infty$ queue Walrand, 1988, which should approximate the distribution for the $M/GI/\bar{m}/\bar{m}$ queue when the blocking probability is low. In the sequel we consider approximations that are valid whenever $B(\rho, \bar{m})$ is acceptably small. In this regime, $Y_{N,D} \sim Bin(N, F_U(D))$ since $D_1, ..., D_N$ are i.i.d. with distribution $F_U(D)$. The following theorem gives approximate expressions for the expected mean subscription level and expected rate of adaptation under the optimal adaptation policy.

Theorem 2 *Under the optimal adaptation policy and in a low blocking regime*

$$\mathbb{E}[Q^{\pi^*}] \approx 1-(1-\beta)\sum_{n=1}^{\bar{m}-1} p_{\bar{m}-1}(n)\int_0^\infty \mathbb{P}(Y_{n,d}\geq \bar{n}(n+1))dF_D(d) \quad (9)$$

$$\mathbb{E}[R^{\pi^*}] \approx 2\lambda(\bar{s}-\underline{s})\sum_{n=\underline{m}-1}^{\bar{m}-2}\Big(p_{\bar{m}-1}(n) \times \int_0^\infty \mathbb{P}\Big(Y_{n,d}<\bar{n}(n+1), Y_{n+1,d}\geq \bar{n}(n+2)\Big)dF_D(d)\Big). \quad (10)$$

Under the optimal adaptation policy

$$q^{\alpha,\pi^*} = \begin{cases} \beta, & \alpha\leq\beta \\ 1-(1-\beta)\bar{F}_D(F_U^{-1}(\frac{\alpha-\beta}{1-\beta})), & \beta<\alpha<1 \\ 1, & \alpha\geq 1 \end{cases}. \quad (11)$$

A simple expression for r^{α,π^*}, $\beta<\alpha<1$ appears difficult to obtain, although it may be shown that $r^{\alpha,\pi^*}=\infty$ for $\alpha\leq\beta$, and $r^{\alpha,\pi^*}=0$ for $\alpha\geq 1$.

Figures 6 and 7 exhibit the above equations along with simulation results for $\lambda=40$ and 320 versus α for the case of exponentially distributed stream durations. The plots show a good match between computational and simulation results for the finite capacity case when the low blocking assumption is valid $\alpha\geq 0.5$, as well as the convergence to the asymptotic values. The region $\alpha\leq\beta$ illustrates the divergence between computed and simulated results due to the low blocking assumption being violated in this regime. Comparing the plot of $\mathbb{E}[Q^{\pi^*}]$ with $\mathbb{E}[Q^{fs}]$ in Figure 2 and $\mathbb{E}[Q^{ra}]$ in Figure 4, we see an increase in mean subscription level under the optimal policy of as much as 20% in the critical regime. The plot of $\mathbb{E}[R^{\pi^*}]$ shows that the rate of adaptation decreases in α. The intuition, made clear in the following discussion on client perceived performance, is that the optimal adaptation policy effectively creates a duration threshold and streams near that threshold experience high adaptation. For α near β that threshold is very short, so that short streams experience frequent adaptation, yielding a very high rate of adaptation, while for α near 1 the threshold is very long, so that only the very longest streams experience adaptation, which, when divided by their long stream duration, gives them a small rate of adaptation.

Client perceived performance measures may be obtained by considering the QoS metrics conditioned on a particular client stream duration d. The following lemma gives expressions for these quantities for both

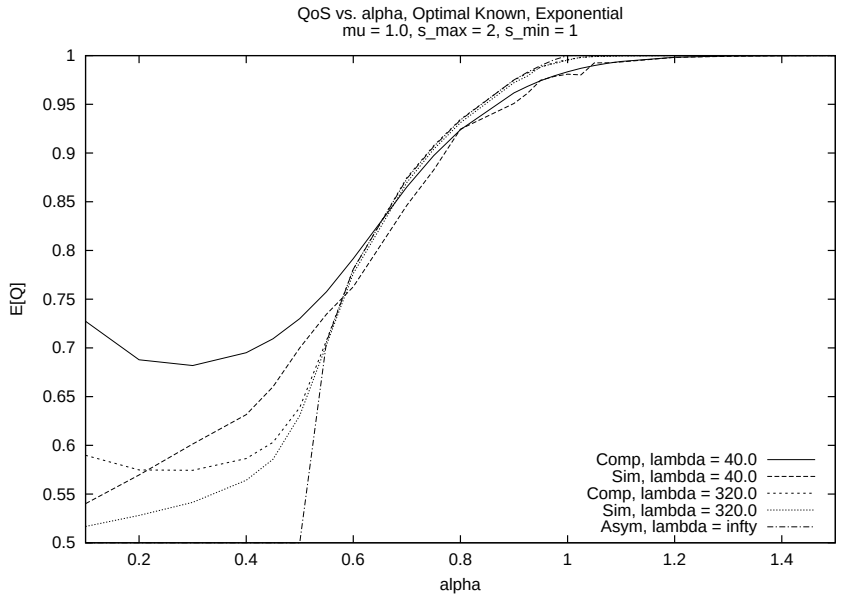

Figure 6. Simulation and Computation of $\mathbb{E}[Q^{\pi^*}]$ *and* q^{α,π^*} *vs.* α *for varying* λ*,* $\beta = 0.5$*,* $D \sim exp(1)$*.*

finite capacity and asymptotic cases. We use the following notation

$$q_d^{\alpha,\pi^*} \equiv \lim_{\lambda\to\infty, c=c(\lambda)} \mathbb{E}[Q^{\pi^*} \mid D = d], \quad r_d^{\alpha,\pi^*} \equiv \lim_{\lambda\to\infty, c=c(\lambda)} \mathbb{E}[R^{\pi^*} \mid D = d]$$

to denote the asymptotic mean subscription level and rate of adaptation for a stream with duration d.

Lemma 3 *Under the optimal adaptation policy and in a low blocking regime*

$$\mathbb{E}[Q^{\pi^*} \mid D = d] \approx 1 - (1-\beta) \sum_{n=1}^{\bar{m}-1} p_{\bar{m}-1}(n) \mathbb{P}(Y_{n,d} \geq \bar{n}(n+1)), \quad (12)$$

$$\mathbb{E}[R^{\pi^*} \mid D = d] \approx 2\lambda(\bar{s} - \underline{s}) \sum_{n=\underline{m}-1}^{\bar{m}-2} \Big(p_{\bar{m}-1}(n) \times \mathbb{P}\Big(Y_{n,d} < \bar{n}(n+1), Y_{n+1,d} \geq \bar{n}(n+2)\Big)\Big). \quad (13)$$

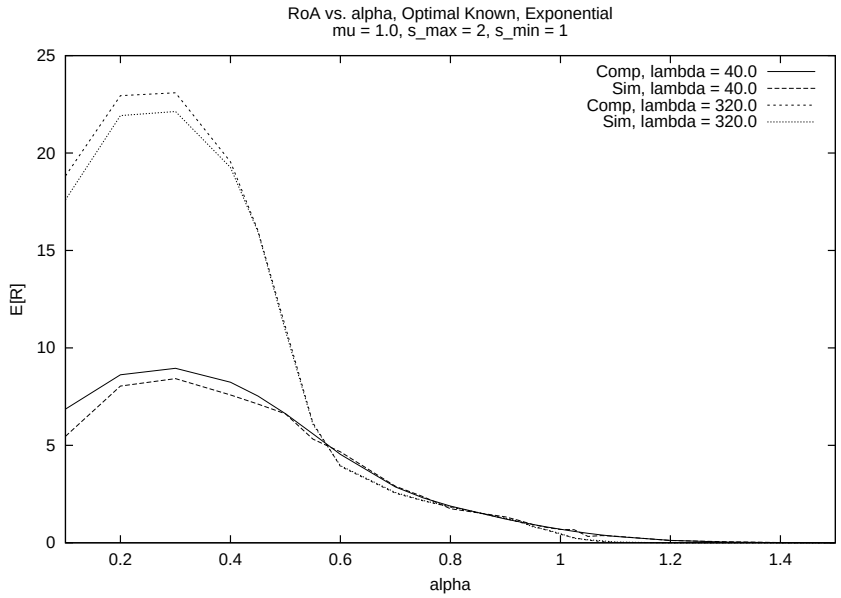

Figure 7. Simulation and Computation of $\mathbb{E}[R^{\pi^*}]$*, vs.* α *for varying* λ*,* $\beta = 0.5$*,* $D \sim exp(1)$*.*

Under the optimal adaptation policy

$$q_d^{\alpha,\pi^*} = \begin{cases} \beta, & \alpha \leq \beta, \\ \beta, & \beta < \alpha < 1, d \geq F_U^{-1}(\frac{\alpha-\beta}{1-\beta}) \\ 1, & \beta < \alpha < 1, d < F_U^{-1}(\frac{\alpha-\beta}{1-\beta}) \\ 1, & \alpha \geq 1 \end{cases} . \tag{14}$$

A simple expression for r_d^{α,π^*} appears difficult to obtain, although $r_d^{\alpha,\pi^*} = 0$ for $|d - F_U^{-1}(\frac{\alpha-\beta}{1-\beta})| > \epsilon$, for some unspecified $\epsilon > 0$.

Figures 8 and 9 exhibit the above equations along with simulation results for $\lambda = 320$ versus d for $\alpha = 0.75$ and $\beta = 0.5$. The plots illustrate a good match between computational and simulation results, as well as the convergence to the asymptotic values. Several points are worth mentioning. First, it is easily seen from these plots that the optimal dynamic adaptation policy grants a constant subscription level of $\bar{s}$ for streams with durations significantly shorter than the threshold $F_U^{-1}(\frac{\alpha-\beta}{1-\beta})$, and a constant subscription level of $\underline{s}$ for streams with durations significantly longer than the threshold. Streams with durations in the vicinity of the threshold experience a mean subscription level in $(\beta, 1)$, and a relatively high rate of adaptation.

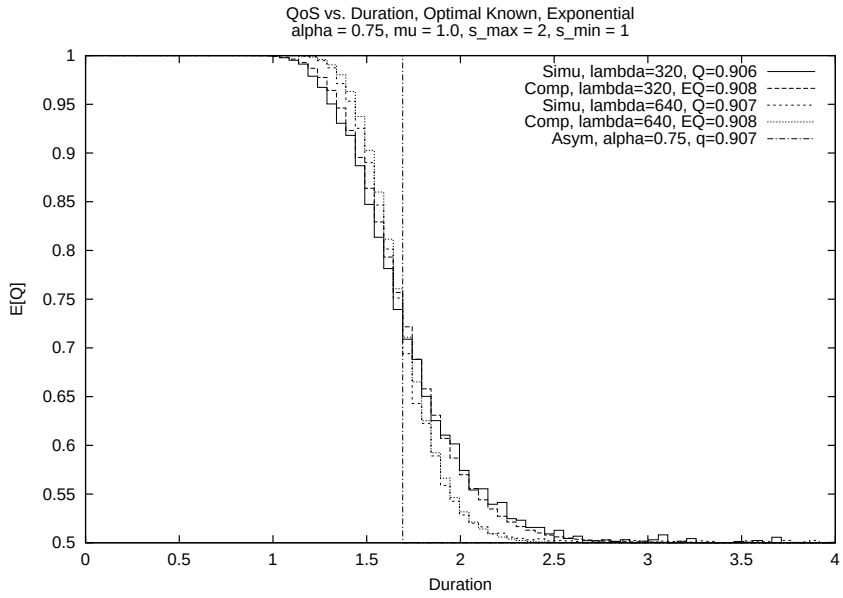

Figure 8. Simulation and Computation of $\mathbb{E}[Q^{\pi^*} \mid D = d]$ *and* q_d^{α,π^*} *vs.* d *for* $\alpha = 0.75$ *and* $\lambda = 320, 640$, $\beta = 0.5$, $D \sim exp(1)$.

The main idea that is gleaned from these figures is that the optimal dynamic adaptation policy only utilizes dynamic adaptation for a small number of streams with duration near the threshold. These streams are the ones that are toggled between $\bar{s}$ and $\underline{s}$ as streams depart and arrive respectively. These observations suggest that near optimal mean subscription levels may be obtainable by an admission policy where streams are granted a fixed subscription level depending on their duration at the time of stream initiation. A fixed subscription level means $\mathbb{E}[R^{\pi}] = 0$ but that the blocking probability B would be higher because streams no longer can adapt their subscription levels to accommodate newly admitted streams. We investigate admission policies for rate adaptive streams in the next section.

5. Admission Policies For Rate Adaptive Streams

We define an admission policy for rate adaptive streams as an admission policy that assigns a subscription level to a stream which that stream maintains throughout its lifetime in the system, i.e., the stream is not dynamically adapted. The previous section demonstrates that

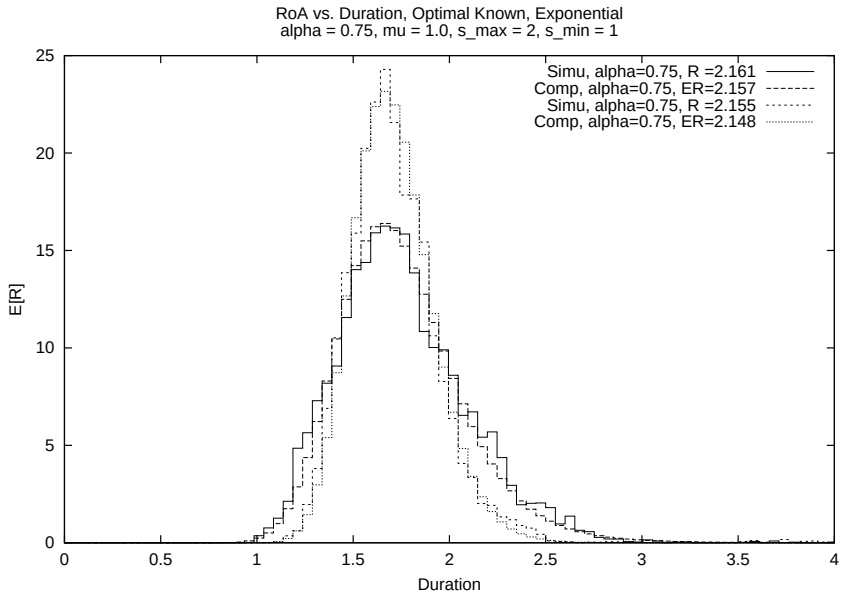

Figure 9. Simulation and Computation of $\mathbb{E}[R^{\pi^*} \mid D = d]$ *vs.* d *for* $\alpha = 0.75$ *and* $\lambda = 320, 640$*,* $\beta = 0.5$*,* $D \sim exp(1)$*.*

on large capacity links the optimal dynamic adaptation policy effectively creates a duration threshold and gives short duration streams the maximum subscription level and long duration streams the minimum subscription level. This suggests designing an admission policy for rate adaptive streams whereby streams are admitted at a duration dependent subscription level which they maintain throughout their lifetime. In this section we will show that such an admission policy can obtain an expected subscription level that is asymptotically equal to that under the optimal dynamic adaptation policy, and has the added benefit of not requiring dynamic adaptation. We make a slight abuse of notation and say $\pi = aa$ to refer to an admission policy for rate adaptive streams where dynamic adaptation is not employed.

Recall that $\mathcal{S} \equiv \{\bar{s} = s_1, ..., s_K = \underline{s}\}$ is the set of supported subscription levels, and that without loss of generality we assume $s_k > s_{k+1}$. Let $\mathcal{D} \equiv \{d_1, ..., d_{K-1}\}$ denote a set of duration thresholds where $d_k \geq d_{k-1} \geq 0$ for $k = 1, ..., K-1$. A stream of duration d is assigned a static subscription level given by

$$S^{aa}(d) = s_{k^*},\ k^* = \max\{0 < k \leq K \mid d_{k-1} \leq d < d_k\}.$$

We assume $d_0 = 0$ and $d_K = \infty$ in the above definition. The stream is admitted at time t provided

$$\sum_{k=1}^{K} s_k N_k(t) + S^{aa}(d) \leq c$$

where we redefine $N(t) \equiv (N_1(t), ..., N_K(t))$ as the number of active streams at each subscription level at time t.

This system is a K class stochastic knapsack with a full sharing admission policy Ross, 1995. The parameters of the stochastic knapsack are the capacity c, the class size s_k, and the class load $\rho_k = \frac{\lambda_k}{\mu_k}$ where λ_k is the class arrival rate and μ_k^{-1} is the mean class duration. We calculate the arrival rate, mean class duration, and class load as

$$\begin{aligned}
\lambda_k &= \lambda(F_D(d_k) - F_D(d_{k-1})), \\
\mu_k^{-1} &= \mathbb{E}[D \mid d_{k-1} \leq d < d_k] = \mu^{-1} \frac{F_U(d_k) - F_U(d_{k-1})}{F_D(d_k) - F_D(d_{k-1})}, \\
\rho_k &= \rho(F_U(d_k) - F_U(d_{k-1})),
\end{aligned}$$

for $k = 1, ..., K$. Recall $F_U(u) \equiv \mu u \int_0^u z f_D(z) dz$ is the distribution of stream durations when the system is viewed at an arbitrary time.

The blocking probabilities are class dependent and we write $B_k(\mathcal{D}) \equiv \mathbb{P}(\sum_{l=1}^{K} s_l N_l(t) + s_k > c)$ for $k = 1, ..., K$. We define the overall blocking probability as $B(\mathcal{D}) \equiv \sum_{k=1}^{K} \frac{\lambda_k}{\lambda} B_k(\mathcal{D})$. We can bound this as $B(\rho, \bar{m}) \leq B(\mathcal{D}) \leq B(\rho, \underline{m})$ since $B(\rho, \bar{m})$ corresponds to the single class system where all streams are admitted at rate $\underline{s}$ and $B(\rho, \underline{m})$ corresponds to the single class system where all streams are admitted at rate $\bar{s}$.

The expected subscription level under the admission policy for rate adaptive streams is defined as

$$\mathbb{E}[Q^{aa}] \equiv \sum_{k=1}^{K} \frac{\lambda_k (1 - B_k(\mathcal{D})) s_k}{\lambda (1 - B(\mathcal{D})) \bar{s}},$$

which can be thought of as a normalized revenue function, i.e., an admitted stream of class k earns revenue s_k, and so $\mathbb{E}[Q^{aa}]$ is the normalized rate at which revenue is earned. Similarly, we define the asymptotic expected subscription level under the admission policy for rate adaptive streams as

$$q^{\alpha,aa} \equiv \lim_{\lambda \to \infty, c = c(\lambda)} \mathbb{E}[Q^{aa}].$$

The following theorem identifies the asymptotically optimal admission policy for rate adaptive streams that maximizes the asymptotic expected

subscription level subject to maintaining an asymptotic blocking probability of zero.

Theorem 3 *The two class admission policy for rate adaptive streams with duration threshold*

$$d^* = \begin{cases} 0, & \alpha \leq \beta \\ F_U^{-1}(\frac{\alpha-\beta}{1-\beta}), & \beta < \alpha < 1 \\ \infty, & \alpha > 1 \end{cases} \tag{15}$$

maximizes the asymptotic expected subscription level $q^{\alpha,aa}$ over all K class stochastic knapsacks that achieve an asymptotic blocking probability of 0 for $\alpha > \beta$. Moreover, this admission policy $\pi = aa^*$ achieves an asymptotic expected subscription level equal to that under the optimal dynamic adaptation policy, i.e., $q^{\alpha,aa^*} = q^{\alpha,\pi^*}$.

Theorem 3 gives the asymptotically optimal admission policy for rate adaptive streams and, more importantly, tells us that there is no loss in asymptotic mean subscription level incurred by not using dynamic adaptation. Thus admission policies for rate adaptive streams obtain the equivalent mean subscription level as dynamic adaptation, have the benefit of a rate of adaptation of 0, and maintain the low blocking characteristic of dynamic adaptation. The cost paid is that the blocking probability using dynamic adaptation goes to 0 exponentially fast for $\alpha > \beta$ while the blocking probability using the optimal admission policy for rate adaptive streams goes to 0 like $O(\frac{1}{\sqrt{c}})$ for $\beta < \alpha \leq 1$, and exponentially for $\alpha > 1$.

Figures 10 and 11 provide a comparison between the optimal dynamic adaptation policy and the optimal admission policy for rate adaptive streams. Figure 10 provides computational and simulation results for the expected subscription level under the two class admission policy which maximizes the expected subscription level subject to an overall blocking probability of $B^* = 1\%$ for $\lambda = 40,320$. That is, the duration d^* solves

$$\max_d \{\mathbb{E}[Q^{aa}] \mid B(d) \leq B^*\}$$

making the slight abuse of notation by writing d for $\mathcal{D} = \{d\}$. A plot of $q^{\alpha,aa^*} = q^{\alpha,\pi^*}$ is also provided. The plot illustrates the convergence to the asymptotic expected subscription level.

Figure 11 gives the overall blocking probability for a two class admission policy where the duration threshold is chosen so that the expected subscription level equals that under the optimal dynamic adaptation-policy, i.e., d^* is chosen as the unique d such that $\mathbb{E}[Q^{aa}] = \mathbb{E}[Q^{\pi^*}]$ for $\lambda = 40,320$. Plots of the computed and simulated overall blocking

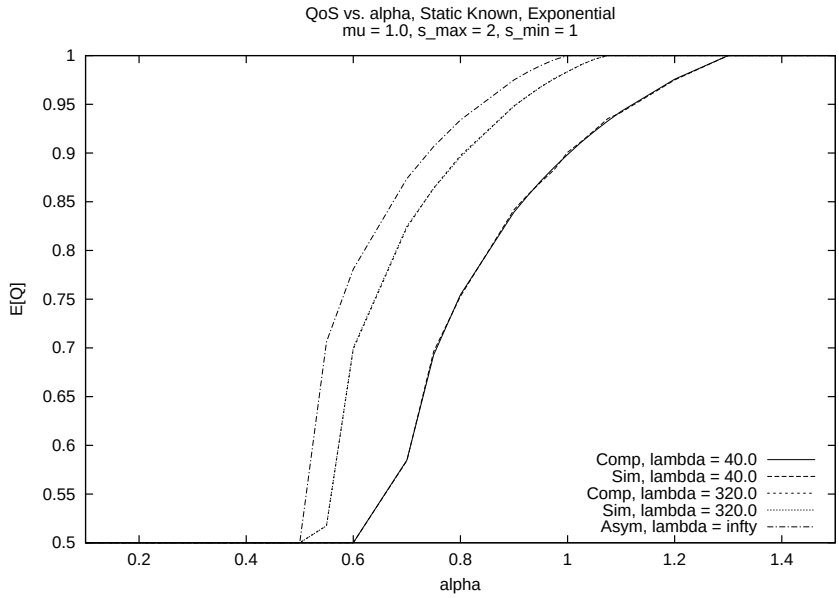

Figure 10. Simulation and computation of $\mathbb{E}[Q^{aa^*}]$ *and* q^{α,aa^*} *vs.* α *for* $B^* = 1\%$*,* $\lambda = 40, 320$*,* $\beta = 0.5$*, and* $D \sim exp(1)$*. Not all lines are visible because the computed and simulation values being nearly equal.*

probability are given, along with a plot of the asymptotic blocking probability, i.e., $1 - \frac{\alpha}{\beta}$ for $\alpha \leq \beta$ and 0 for $\alpha > \beta$. The panel illustrates the convergence to the asymptotic blocking probability.

Finally, we offer a brief comment on implementation. The optimal adaptive admission policy requires that the bottleneck link identify its duration threshold (15), which depends on $F_D, \rho, c, \underline{s}, \bar{s}$. The distribution F_D could be analyzed and hard-coded, or could be estimated empirically by keeping track of stream durations. The load ρ could be estimated empirically by monitoring arriving service request times and service durations. Indeed, such a measurement would make the model more robust to the observed non-stationarities present in Internet traffic. The parameters $\underline{s}$ and $\bar{s}$ could be estimated by monitoring stream rate associated with long duration and short duration streams. The algorithm could easily be made to be distributed—stream service requests would traverse the route from client to server and at each node pick up the duration threshold for that link along with an admission decision. Upon returning to the client, provided all link admission decisions are positive, the client subscribes to the stream content provider at the maximum rate if

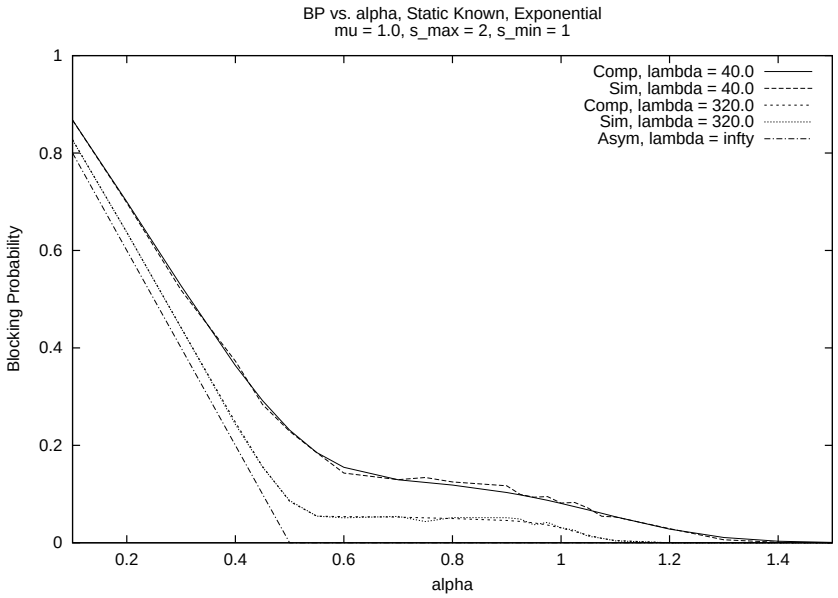

Figure 11. Simulation and computation of $B(d^*)$ (see text) and asymptotic blocking probability vs α for $\lambda = 40, 320$, $\beta = 0.5$, and $D \sim exp(1)$.

the stream duration is smaller than the maximum duration threshold, and subscribes at the minimum rate otherwise.

6. Conclusion

The rationale behind dynamic adaptation of rate adaptive streams is that by dynamically adjusting the subscription level we obtain the system benefit of a low blocking probability and the client benefit of making full use of all available capacity. The price paid for these gains is that clients incur a lowered perceived QoS both in terms of the mean subscription level and the rate of adaptation compared with non-adaptive streams admitted at their maximum subscription level $\bar{s}$. Our investigation of adaptive admission shows that, for large capacity links, a simple duration threshold two class admission policy obtains the asymptotic optimal mean subscription level, asymptotic zero blocking probability, and zero rate of adaptation. Thus, for large capacity links, the protocol, resource, and implementation overhead required by dynamic adaptation is not justifiable. Moreover, optimal QoS is obtainable using only two subscription levels—this has the significant implication that multimedia content providers gain little benefit from providing more than two en-

codings. Of course, this is assuming clients aren't access line limited, in which case additional encodings will be of use to such clients. Future work in this area will analyze the case of heterogeneous minimum and maximum subscription levels.

Appendix

Proof of Lemma 1 The proof of (1) and (2) is found in Weber and de Veciana, 2002.

Proof of (3). Conditioned on a stream being admitted, that stream sees the system with capacity $\bar{m}-1$, i.e., the process $N(t)$ denoting the number of other streams has an invariant distribution $p_{\bar{m}-1}(n), 0 \leq n \leq \bar{m}-1$, where n is the number of other streams in the system. By PASTA the conditioned stream sees the other streams in steady state. In steady state, when $N(t)=n$, the number of other streams changes due to arrivals and departures at rate $\lambda + n\mu$. The definition of the fair share rate implies that not all changes in the number of streams will change the fair share rate. When $n \in \{0, ..., \underline{m}-2\}$ neither an arrival or departure causes a change in rate. When $n \in \{\underline{m}-1, ..., \bar{m}-2\}$ an arrival causes a change in rate from $\frac{c}{n+1}$ to $\frac{c}{n+2}$. When $n \in \{\underline{m}, ..., \bar{m}-1\}$ a departure causes a change in rate from $\frac{c}{n+1}$ to $\frac{c}{n}$. Admissions occur at rate λ for $n < \bar{m}-1$ and departures occur at rate $n\mu$ for $n > 0$. Putting these observations together we obtain

$$\mathbb{E}[R^{fs}] = \sum_{n=\underline{m}-1}^{\bar{m}-2} \lambda p_{\bar{m}-1}(n)(\frac{c}{n+1} - \frac{c}{n+2}) + \sum_{n=\underline{m}}^{\bar{m}-1} n\mu p_{\bar{m}-1}(n)(\frac{c}{n} - \frac{c}{n+1}).$$

Using detailed balance equations for $\pi_{\bar{m}-1}(n)$ and relabeling indices yields the result.

Proof of (4). We may write (3) as

$$\mathbb{E}[R^{fs}] = 2\mu\bar{s}\alpha\mathbb{E}\Big[\frac{1}{N/\rho}\Big|\underline{m} < N \leq \bar{m}\Big]\mathbb{P}\Big(\underline{m} < N \leq \bar{m}\Big).$$

Define the random process $\{N^{\alpha,\rho}(t)\}$ for the number of streams on the link at time t when the load is ρ and the capacity is $c(\lambda) = \alpha\bar{s}\rho$. The distribution of $N^{\alpha,\rho}(t)$ is that of an $M/GI/\bar{m}/\bar{m}$ queue with load ρ and capacity $\bar{m} = \frac{\alpha}{\beta}\rho$. By the Law of Large Numbers

$$\lim_{\lambda\to\infty} \mathbb{E}[\frac{N^{\alpha,\rho}}{\rho}] = \begin{cases} \frac{\alpha}{\beta}, & \alpha \leq \beta \\ 1, & \beta < \alpha < 1 \\ 1, & \alpha > 1 \end{cases}$$

and

$$\lim_{\lambda\to\infty} \mathbb{P}\Big(\alpha < \frac{N^{\alpha,\rho}}{\rho} \leq \frac{\alpha}{\beta}\Big) = \begin{cases} 1, & \alpha \leq \beta \\ 1, & \beta < \alpha < 1 \\ 0, & \alpha \geq 1 \end{cases}.$$

The asymptotic rate of adaptation is

$$\begin{aligned} r^{\alpha,fs} &= \lim_{\lambda\to\infty, c=c(\lambda)} \mathbb{E}[R^{fs}] \\ &= \lim_{\lambda\to\infty} 2\mu\bar{s}\alpha\mathbb{E}\Big[\frac{1}{N^{\alpha,\rho}/\rho}\Big|\alpha < \frac{N^{\alpha,\rho}}{\rho} \leq \frac{\alpha}{\beta}\Big]\mathbb{P}\Big(\alpha < \frac{N^{\alpha,\rho}}{\rho} \leq \frac{\alpha}{\beta}\Big). \end{aligned}$$

Thus when $\alpha \leq \beta$ we have $r^{\alpha,fs} = (2\mu\bar{s}\alpha)(\frac{\beta}{\alpha})(1) = 2\mu\bar{s}\beta = 2\mu\underline{s}$. When $\beta < \alpha < 1$ we have $r^{\alpha,fs} = (2\mu\bar{s}\alpha)(1)(1) = 2\mu\bar{s}\alpha$. Finally, when $\alpha \geq 1$ we have $r^{\alpha,fs}(2\mu\bar{s}\alpha)(1)(0) = 0$.

■

Proof of Lemma 2

The proof of (5) and (6) is found in Weber and de Veciana, 2002.

Proof of (7). The preliminary remarks in the proof of Lemma 1 apply here as well.

All changes in rate are of size $\bar{s} - \underline{s}$, but the number of streams that change rate depends on the number of streams and the new random selection of streams to be adapted. Let t denote a jump time, i.e., a stream arrival or departure, $S(t)$ the rate allocated to the stream we've conditioned on being present, and $S(t^-)$ the rate allocated to that stream immediately prior to the jump time. The probability that the conditioned stream changes rate is

$$\mathbb{P}(S^{ra}(t) \neq S^{ra}(t^-)) = \mathbb{P}(S^{ra}(t) = \bar{s})\mathbb{P}(S^{ra}(t^-) = \underline{s}) + \mathbb{P}(S^{ra}(t) = \underline{s})\mathbb{P}(S^{ra}(t^-) = \bar{s})$$

When $n \in \{0, ..., \underline{m} - 2\}$ no streams change rate for either an arrival or departure. When $n \in \{\underline{m}-1, ..., \bar{m}-2\}$ an arrival causes a stream to change rate with probability

$$g(n+1) \equiv \frac{\bar{n}(n+2)\underline{n}(n+1) + \underline{n}(n+2)\bar{n}(n+1)}{(n+1)(n+2)}.$$

When $n \in \{\underline{m}, ..., \bar{m}-1\}$ a departure causes a stream to change rate with probability $g(n)$. This gives

$$\mathbb{E}[R^{ra}] = \sum_{n=\underline{m}=1}^{\bar{m}-2} \lambda p_{\bar{m}-1}(n)g(n+1) + \sum_{n=\underline{m}}^{\bar{m}-1} n\mu p_{\bar{m}-1}(n)g(n).$$

Detailed balance equations yield the result.

Proof of (8). We may rewrite (7) as

$$\begin{aligned}\mathbb{E}[R^{ra}] &= 2(\bar{s}-\underline{s})\mu\mathbb{E}\Big[\frac{\bar{n}(N)\underline{n}(N+1) + \bar{n}(N+1)\underline{n}(N)}{N+1}\Big|\underline{m} \leq N \leq \bar{m}-1\Big] \times \\ &\quad \mathbb{P}\Big(\underline{m} \leq N \leq \bar{m}-1\Big)\end{aligned}$$

The same developments found in the proof of Lemma 1 regarding $\mathbb{P}(\underline{m} \leq N \leq \bar{m}-1)$ apply. Under the rate adaptive scaling this becomes

$$\begin{aligned}r^{\alpha,ra} &= \lim_{\lambda\to\infty, c=c(\lambda)} \mathbb{E}[R^{ra}] \\ &= \lim_{\lambda\to\infty} 4(\bar{s}-\underline{s})\mathbb{E}\Big[\frac{\bar{n}(N^{\alpha,\rho})\underline{n}(N^{\alpha,\rho})}{N^{\alpha,\rho}}\Big|\alpha < \frac{N^{\alpha,\rho}}{\rho} \leq \frac{\alpha}{\beta}\Big]\mathbb{P}\Big(\alpha < N^{\alpha,\rho} \leq \frac{\alpha}{\beta}\Big).\end{aligned}$$

For $\alpha \leq 1$ $\mathbb{P}(\alpha < N^{\alpha,\rho} \leq \frac{\alpha}{\beta}) = 1$ and so $r^{\alpha,ra} = \infty$ since

$$\lim_{\lambda\to\infty} \mathbb{E}\Big[\frac{\bar{n}(N^{\alpha,\rho})\underline{n}(N^{\alpha,\rho})}{N^{\alpha,\rho}}\Big] = \infty.$$

For $\alpha > 1$, however,

$$\mathbb{P}\Big(\alpha < N^{\alpha,\rho} \leq \frac{\alpha}{\beta}\Big)$$

goes to 0 exponentially in λ while

$$\mathbb{E}\Big[\frac{\bar{n}(N^{\alpha,\rho})\underline{n}(N^{\alpha,\rho})}{N^{\alpha,\rho}}\Big|\alpha < \frac{N^{\alpha,\rho}}{\rho} \leq \frac{\alpha}{\beta}\Big]$$

grows linearly in λ, forcing $r^{\alpha,ra}$ to 0.

■

Proof of Theorem 1 See Weber and de Veciana, 2002.

Proof of Theorem 2

Proof of (9). Consider an arbitrary time t.

$$\begin{aligned}\mathbb{E}[Q^{\pi^*}] &= \mathbb{E}[\frac{S^{\pi^*}(t)}{\bar{s}}] \\ &= 1\mathbb{P}(Y_{N,D} < \bar{n}(N+1)) + \beta\mathbb{P}(Y_{N,D} \geq \bar{n}(N+1)) \\ &= 1 - (1-\beta)\mathbb{P}(Y_{N,D} \geq \bar{n}(N+1)).\end{aligned}$$

The first equality follows by ergodicity. Simple conditioning yields the equation.

Proof of (10). Consider again an arbitrary time t. Similar to the proofs of Lemmas 1 and 2, we break down the analysis for different values of $N(t)$. For $n \in \{0, ..., \underline{m}-2\}$ neither an arrival nor departure causes a change in rate. For $n \in \{\underline{m}-1, ..., \bar{m}-2\}$, an arrival causes a change in rate from $\bar{s}$ to $\underline{s}$ if $Y_{n,D} < \bar{n}(n+1)$ and $Y_{n+1,D} \geq \bar{n}(n+2)$. For $n \in \{\underline{m}, ..., \bar{m}-1\}$, a departure causes a change in rate from $\underline{s}$ to $\bar{s}$ if $Y_{n,D} \geq \bar{n}(n+1)$ and $Y_{n-1,D} < \bar{n}(n)$. Putting these observations together we obtain

$$\begin{aligned}\mathbb{E}[R^{\pi^*}] &= (\bar{s}-\underline{s}) \times \\ &\Big(\sum_{n=\underline{m}-1}^{\bar{m}-2} \lambda p_{\bar{m}-1}(n) \int_0^\infty \mathbb{P}(Y_{n,d} < \bar{n}(n+1), Y_{n+1,d} \geq \bar{n}(n+2)) dF_D(d) \\ &+ \sum_{n=\underline{m}}^{\bar{m}-1} n\mu p_{\bar{m}-1}(n) \int_0^\infty \mathbb{P}(Y_{n,d} \geq \bar{n}(n+1), Y_{n-1,d} < \bar{n}(n)) dF_D(d) \Big).\end{aligned}$$

Detailed balance equations yield the result.

Proof of (11) may be found in Weber and de Veciana, 2002.

■

Proof of Lemma 3 The proofs of (12), (13), and (14) are exactly the same as the proofs of (9), (10) and (11) respectively, but with $D = d$.

■

Proof of Theorem 3 The critical insight behind the proof is that the optimal asymptotic expected subscription level subject to the asymptotic zero blocking constraint is met when asymptotic load equals asymptotic capacity, i.e.,

$$\lim_{\lambda\to\infty, c=c(\lambda)} \frac{\sum_{k=1}^K \rho_k s_k}{c} = \frac{\sum_{k=1}^K (F_U(d_k) - F_U(d_{k-1}))s_k}{\alpha\bar{s}} = 1.$$

It is shown in Ross, 1995 that blocking is zero for this case, although the convergence is $O(\frac{1}{\sqrt{c}})$. We may then write the optimization problem as

$$\max_{\mathcal{D}} \Big\{ \sum_{k=1}^K (F_D(d_k) - F_D(d_{k-1}))s_k \Big| \sum_{k=1}^K (F_U(d_k) - F_U(d_{k-1}))s_k = \alpha\bar{s} \Big\}$$

The Lagrangian is

$$L(\mathcal{D}, z) = \sum_{k=1}^K (F_D(d_k) - F_D(d_{k-1}))s_k - z\Big(\sum_{k=1}^K (F_U(d_k) - F_U(d_{k-1}))s_k - \alpha\bar{s} \Big)$$

Taking derivatives yields

$$\frac{\partial L(\mathcal{D},z)}{\partial d_k} = f_D(d_k)s_k - f_D(d_k)s_{k+1} - z(f_U(d_k)s_k - f_U(d_k)s_{k+1}).$$

Use of the fact that $f_U(d) = \mu d f_D(d)$ allows

$$\frac{\partial L(\mathcal{D},z)}{\partial d_k} = (s_k - s_{k+1})f_D(d_k)(1 - z\mu d_k)$$

Optimality requires $\frac{\partial L(\mathcal{D},z)}{\partial d_k} = 0$ for $k = 1, ..., K-1$; inspection shows this is only true for $d_k = \frac{1}{z\mu}$,i.e., $d_k = d^* \; \forall k$. This implies the optimal threshold policy uses only two classes, i.e., $\bar{s}$ and $\underline{s}$, and so the inclusion of additional subscription levels, i.e., $K > 2$, is unnecessary.

For a two class system the blocking constraint simplifies to

$$F_U(d)\bar{s} + \bar{F}_U(d)\underline{s} = \alpha\bar{s}.$$

Solving this for d yields (for $\beta < \alpha \leq 1$)

$$d^* = F_U^{-1}(\frac{\alpha - \beta}{1 - \beta}).$$

When $\alpha \leq \beta$ asymptotic zero blocking is impossible since the system is overloaded. We minimize blocking, however, by admitting all streams at $\underline{s}$, i.e., $d^* = 0$. When $\alpha > 1$ we obtain asymptotic zero blocking by admitting all streams at $\bar{s}$, i.e., $d^* = \infty$. Combining these notions gives the result.

The asymptotic expected subscription level under the adaptive admission policy with duration threshold d^* is

$$q^{\alpha,aa^*} = F_D(d^*) + \beta\bar{F}_D(d^*).$$

Rearranging gives the result.

■

References

Argiriou, N. and Georgiadis, L. (2002). Channel sharing by rate adaptive streaming applications. In *Proceedings of Infocom.*

B.Vickers, Alburquerque, C., and Suda, T. (2000). Source-adaptive multi-layered multicast algorithms for real-time video distribution. *IEEE/ACM Transactions on Networking.*

Girod, B. (1992). Psychovisual aspects of image communications. *Signal Processing*, 28:239–251.

Gorinsky, S., Ramakrishnan, K. K., and Vin, H. (2000). Addressing heterogeneity and scalability in layered multicast congestion control. Technical report, Department of Computer Sciences, The University of Texas at Austin.

Gorinsky, S. and Vin, H. (2001). The utility of feedback in layered multicast congestion control. In *Proceedings of NOSSDAV.*

Kar, K., Sarkar, S., and Tassiulas, L. (2000). Optimization based rate control for multirate multicast sessions. Technical report, Institute of Systems Research and University of Maryland.

Kimura, J. (1999). Perceived quality and bandwidth characterization of layered MPEG-2 video encoding. In *Proceedings of the SPIE International Symposium on Voice, Video, and Data Communications*.

McCanne, S. (1996). *Scalable Compression and Transmission of Internet Multicast Video*. PhD thesis, University of California at Berkeley.

Rejaie, R., Handley, M., and Estrin, D. (1999). Quality adaptation for congestion controlled video playback over the internet. In *SIGCOMM*, pages 189–200.

Ross, K. (1995). *Multiservice Loss Models for Broadband Telecommunication Networks*. Springer-Verlag, London.

Saparilla, D. and Ross, K. (2000). Optimal streaming of layered video. In *Proceedings of Infocom*.

Vishwanath, M. and Chou, P. (1994). An efficient algorithm for hierarchical compression of video. In *Proceedings of the IEEE International Conference on Image Processing*.

Walrand, J. (1988). *An Introduction to Queueing Networks*. Prentice-Hall, New Jersey.

Weber, S. and de Veciana, G. (2002). Network design for rate adaptive multimedia streams. In Submission.