# Flow-level QoS for a Dynamic Load of Rate Adaptive Sessions Sharing a Bottleneck Link

Steven Weber
Dept. of ECE
Drexel University
Philadelphia, PA 19104
sweber@ece.drexel.edu

Gustavo de Veciana
Dept. of ECE
The University of Texas at Austin
Austin, TX 78712
gustavo@ece.utexas.edu

## ABSTRACT

We consider the flow-level quality of service (QoS) seen by a dynamic load of rate adaptive sessions sharing a bottleneck link based on fair share bandwidth allocation. This is of interest both in considering wired networks supporting rate adaptive multimedia sessions and wireless networks supporting voice with rate adaptation to realize graceful degradation during congested periods. Two QoS metrics are considered: the time-average instantaneous utility of the allocated bandwidth, and the time-average of transition penalties associated with the changes in allocation seen by a flow. We present a simple model for rate adaptation, where (heterogeneous) flows can vary their rates within (different) ranges, and present closed-form results for these perceived flow-level QoS metrics. We then prove asymptotic results for large capacity systems exhibiting the salient features of rate adaptation in a dynamic network. Finally we provide a concrete example, showing how the QoS seen by sessions with different degrees of adaptivity would vary under a natural fair bandwidth allocation policy.

## 1. INTRODUCTION

Multimedia streams are *rate adaptive* in that the corresponding data may be encoded at a variety of resolutions, and the appropriate resolution for a streaming client may vary with time, depending on the current level of network congestion. That is, clients may *dynamically adapt* their subscription level among a set of stream encodings offered by a server in response to network congestion or lack thereof. In particular, streaming clients may wish to increase their subscription level if extra capacity becomes available along their route, or may wish to decrease their subscription level at the onset of congestion. Similarly for some wireless systems it is of interest to consider supporting a higher rate encoding for voice streams, when the system is not congested, but then when congestion arises to adapt to lower rate encodings in order to enable a graceful degradation and permit a wireless system to effectively support large variations in the offered loads.

Rate adaptive streams have characteristics akin to both elastic and inelastic traffic [37]. They are like inelastic traffic in that the streams have a minimum acceptable rate for satisfactory performance, and subject to which, their sojourn in a system might be assumed independent of their actual resource allocation. They are like elastic traffic in that the bandwidth per stream may be adjusted dynamically in response to changes in the number of streams multiplexed on the link.

There are several advantages to rate adaptation for multimedia streams. One advantage is that the number of streams that can share a link simultaneously is increased above the maximum number without adaptation; this is due to the fact that adaptation reduces the rate given to active streams in order to accommodate newly arriving streams, whereas a link multiplexing inelastic traffic would be forced to block these requests. Another benefit is that, like elastic traffic, the streams may increase or decrease their bandwidth allocation in response to departures and arrivals, thereby keeping the aggregate bandwidth consumption on a link at or near the link capacity resulting in enhanced quality of service (QoS) to flows.

These advantages come at a cost: employing adaptation means streams endure a time-varying bandwidth allocation. This time varying allocation is certainly less satisfying from the customer's perspective than receiving the stream's highest rate/QoS encoding for the duration of the stream. Thus we seek to measure the *quality of service (QoS) cost of adaptation*, i.e., the extent to which a given time-varying bandwidth allocation detracts from the user's perceived performance. Quantitatively measuring QoS is a thorny issue, e.g., see [41], which establishes that most proposed QoS metrics do not adequately correlate with subjective user testing.

In this paper we will use two performance metrics that we feel adequately cover a wide range of QoS issues: the time-average utility and the time-average transition cost.

- The *time-average utility* metric assumes that the instantaneous utility of a bandwidth allocation $x$ can be captured through a concave increasing utility function $g(x)$; the time average utility is then $\frac{1}{d}\int_0^d g(x(t))dt$ where $x(t)$ denotes the bandwidth allocated to a session at time $t$ and $d$ denotes its duration.

- The *time-average transition penalty* metric assumes that a penalty in QoS resulting from a change in bandwidth allocation from $x$ to $x'$ which is captured by a function $h(x, x')$; in this paper we will for example consider $h(x, x') = |x - x'|$ where large changes are worse that small changes. The time average transition penalty is then given by $\frac{1}{d}\int_0^d h(x(t^-), x(t))M(dt)$, where $M(\cdot)$ is the point measure for the times at which changes in bandwidth allocations occur.

These metrics are "gross" in that they ignore many of the subtleties involved in the psycho-visual perception of image quality. Nevertheless, all other things being equal, higher video quality generally implies a higher associated average rate for the stream and vice versa. Similarly, given two clients receiving the same stream with the same average rate, the stream with fewer resolution changes would generally be thought to have a superior quality. The work in [10] offers a more detailed investigation of this phenomenon.

Our stream model incorporates a minimum rate for each stream to perform at a minimally acceptable visual quality. Having satisfied the minimum rate for each active stream, the remaining link bandwidth, if any, may be distributed among the competing streams to improve their QoS. There are a myriad of policies that may be employed to guide this allocation, and it is natural to consider "optimal" policies. For simplicity and tractability in this work we restrict our attention to "fair share" policies where the bandwidth is divided equitably. Our other work [42, 43, 44, 45, 46] has studied adaptation policies that maximize the QoS for a more restrictive class of metrics than that presented here.

The primary contribution of this paper is a careful analysis of a simple model for a bottleneck link subject to a dynamic load of rate adaptive sessions. We derive closed-form expressions for the above mentioned flow-level QoS metrics as a function of the key fundamental parameters, e.g., the link capacity, the stream characteristics, the arrival rate, etc., assuming a fair share bandwidth adaptation policy. By considering a large capacity regime we obtain asymptotic results for the perceived QoS which captures the salient features of rate adaptation in a dynamic system. Our analysis yields highly intuitive expressions for QoS which combined with a numerical example provide some interesting insights on the QoS when (possibly heterogeneous) rate adaptive sessions are carried, and provides some insights on convergence of the performance of finite capacity systems towards our asymptotic results.

## 1.1 Related work

McCanne's seminal receiver-driven layered multicast protocol (RLM) [28] introduced the idea of streams dynamically adapting their subscription levels in response to changing levels of available bandwidth (congestion). Several authors have analyzed the performance of dynamic rate adaptive systems, e.g., [1, 2, 5, 6, 17, 32, 34, 40] and our own work [42, 43, 44, 45, 46]. Most of these papers study systems where streams are competing for dynamically changing available resources and seek to characterize the system performance. Our own work has focused on finding *optimal* resource allocation policies that maximize customer perceived QoS. Rate distortion theory has historically provided the key insights on the tradeoff between rate and distortion, where distortion can be viewed as a proxy for QoS, see e.g., [3, 4, 16, 30, 31, 35, 39].

Capacity allocation for utility maximization has been studied extensively for the *static* case, i.e., when the set of users in the network is fixed, beginning with the seminal work of Kelly and his collaborators [19, 20, 21, 22, 23]. Other significant contributions include [8, 24, 25, 26, 27, 29, 38] but it should be emphasized that almost all of this work is restricted to resource allocation for elastic traffic. More significantly, the bulk of the above work focuses on system level performance while our effort here is aimed at characterizing the performance seen by a given stream, and how that stream's characteristics impact the service quality it receives *in a dynamic regime.*

The rest of this paper is organized as follows. Section 2 introduces the mathematical model, including the stream model, the admission policy, the fair share adaptation policy, and the formal statement of the QoS metrics. Section 3 gives the expressions for the QoS metrics for finite capacity links. Section 4 gives the corresponding many small streams asymptotic results. Section 5 presents some numerical results comparing the finite and asymptotic results and the paper concludes in 6. Finally, we mention that all proofs are relegated to the Appendix.

# 2. THE MODEL

Random variables are denoted by capital letters, e.g., $X$, scalars are lowercase letters, e.g., $x$. We will consider two stream models: *i) homogeneous* traffic where all streams have the same minimum rate requirement, and *ii) heterogeneous* traffic where streams have individual rate requirements. The *link state* is the set of minimum rate requirements for all active streams. For the homogeneous case the state is represented by the number of active streams, usually denoted by $n$ (when known) or $N$ (when unknown). For the heterogeneous case we assume the minimum rates are drawn from some continuous distribution so that no two active streams have the same exact minimum rate requirement. We can then denote the link state by a set, denoted by $\mathbf{X}$ when the state is treated as a collection of random variables, and denoted by $\mathbf{x}$ when the state is assumed known. Due to arrivals and departures the link state set has a time-varying number of elements.

Let $n(\mathbf{x})$ denote the number of components of the vector $\mathbf{x}$. Similarly, define $a(\mathbf{x}) = x_1 + \cdots + x_{n(\mathbf{x})}$ as the sum of the components of $\mathbf{x}$. In the context of link state, $a(\mathbf{x})$ corresponds to the *aggregate minimum bandwidth requirement* of the streams on the link.

## 2.1 Rate adaptive streams

Streaming media usually demonstrates burstiness across multiple time scales, see e.g., [9, 15], the burstiness due to both the inherent time-varying bit rate required to encode the media information as well as artifacts of the encoding process. For tractability we employ a much simpler CBR model: streams are modeled by a pair of positive real-valued random variables $(X, D)$, where $X \sim F_X$ is the minimum bandwidth requested by the stream for satisfactory playback quality and the stream durations $D$ are exponentially distributed with mean $\mathbb{E}[D] = \frac{1}{\mu}$. The CBR minimum stream rate $X$ can be thought of as the effective bandwidth associated with multiplexing the stream on a link [7, 18]. The distribution $F_X$ captures the relative frequency of different "size" media streams on a link, e.g., streaming radio, VoIP traffic, streaming video, etc. Let $\mathcal{S} = (s_{min}, s_{max})$ denote the support set of $X$.

We assume all streams have a common *adaptivity* $\alpha \in (0,1]$ defined as the ratio of the minimum and maximum bandwidth. In particular, the maximum bandwidth required by a stream with minimum bandwidth requirement $X$ is $\frac{X}{\alpha}$. These minimum and maximum rates can be thought of either as bounds set by the content provider (making available a range of media encodings at rates between $X$ and $\frac{X}{\alpha}$) or inherent to the media content (rates corresponding to effective minimum and maximum quality levels). Note that small $\alpha$ means the stream is highly "compressible" or elastic, while $\alpha$ near one means the stream is basically inelastic. A natural extension of our model would be to allow the adaptivity parameter $\alpha$ to vary across streams according to a specified distribution. We have elected not to pursue this generalization in the interest of keeping our model, and thus our results, as simple as possible. Moreover, it is is natural to model compression algorithms as "scale invariant", meaning that the algorithm is capable of compressing a small stream by the same factor as a large stream. Under this assumption the fixed $\alpha < 1$ represents the compression factor of the algorithm, and there is a tacit assumption that all media employs the same compression algorithm. We recognize this assumption ignores many of the second-order effects on media compression, but our aim is to focus on tractable models for client performance instead of accurate models for compression algorithms.

The random variables $X$ and $D$ are assumed independent and we denote their means by $\mathbb{E}[X] = \sigma$ and $\mathbb{E}[D] = \frac{1}{\mu}$. We denote expectation taken with respect to $(X, D)$ as $\mathbb{E}^0[f(X,D)] \equiv \int_{\mathcal{S}\times\mathbb{R}^+} f(x,d)dF_X(x)dF_D(d)$ for bounded measurable functions $f$ of the stream state.

We consider the case where content providers dynamically adapt the encoding of the media stream to raise or lower quality in response to network congestion (or lack thereof). For example, streams might increase their subscription level (choose a higher resolution encoding) during periods of low congestion, or decrease their subscription level to mitigate loss during periods of high congestion. These decisions are abstracted into our notion of *adaptation policy*, described in the sequel. Let $\mathcal{A}_x$ denote the set of possible media encodings for a given media object with minimum bandwidth requirement $x \in \mathcal{S}$ made available by the content provider. We make the following assumption about media encoding availability.

Assumption 1. **Source adaptation.** *The source is able to (dynamically and instantaneously) adjust the encoding of the stream to match any rate, between the minimum and maximum rate, i.e.,* $\mathcal{A}_x = [x, \frac{x}{\alpha}]$.

## 2.2 Admission policies

We consider a single link of capacity $c$. Single link models are often adequate models for network scenarios where backbone bandwidth is plentiful and streams are constrained at either the source or destination. Streams arrive requesting service forming a Poisson process with rate $\lambda$. Let $\rho = \frac{\lambda}{\mu}$ so that $\rho$ is the offered load in terms of number of streams. We assume a *full-sharing admission policy*, meaning streams are always admitted if there are resources available to do so. In particular, suppose the state of the link is $\mathbf{x} = (x_1, \ldots, x_n)$ where the $x_i$'s are the minimum bandwidth requirements of the $n$ active streams. Define $y = a(\mathbf{x})$ as the *aggregate minimum load*; an arriving stream with minimum bandwidth $x$ is admitted provided $y + x \le c$, and is blocked otherwise.

The system dynamics specified here are those of a stochastic knapsack with continuous sizes, see [33] §2.8. In particular, let $\mathcal{X} = \bigcup_{n\in\mathbb{Z}} \mathcal{S}^n \subset \mathbb{R}_+^\infty$ denote the *link state space*; recall the link state is the vector of minimum bandwidth requirements for the active streams. It follows that the $\mathcal{X}$ valued random process $\{\mathbf{X}(t)\}$ is a (homogeneous, stationary, ergodic) Markov process with transition kernel $Q$ specified as

$$\begin{array}{rcl} Q(\mathbf{x}, \mathbf{x} \cup \{x\}) = & \lambda \mathbb{I}(a(\mathbf{x}) + x \le c), & x \in \mathcal{S} \\ Q(\mathbf{x}, \mathbf{x} \setminus \{x_i\}) = & \mu, & i = 1, \ldots, n(\mathbf{x}) \\ Q(\mathbf{x}, B) = & 0, & \text{else,} \end{array}$$

for all Borel $B \subset \mathcal{X}$. This Markov process is a pure jump process with bounded rates, whose sample paths are in the space $\mathcal{D}$, the set of right-continuous functions with left-limits. Thus, given the occurrence of a state transition at time $t$ (arrival or departure), $\mathbf{X}(t^-)$ is the state of the system immediately prior to the transition, and $\mathbf{X}(t)$ is the state immediately following the transition.

Let $M$ be the random point process consisting of the points $\ldots, T_{-1}, T_0, T_1, \ldots$ where the $T_i$'s correspond to the points where $\mathbf{X}(t) \neq \mathbf{X}(t^-)$. We can equivalently view $M$ as a random measure induced by $\{\mathbf{X}(t)\}$ defined as $M(A) = \sum_n \mathbb{I}(T_n \in A)$ for all Borel sets $A \subset \mathbb{R}$. Intuitively, $M(A)$ is the number of transition times of $\{\mathbf{X}(t)\}$ occurring in each Borel set $A$.

We denote the invariant distribution of $\{\mathbf{X}(t)\}$ as $\mathbf{p} = \{p(B)\}$ for all Borel $B \subset \mathcal{X}$. We will have cause to consider the invariant distribution conditioned on a particular stream being admitted in the system. In particular, define $\mathcal{X}_x \subset \mathcal{X}$ as the set of link states containing the stream state $x$, i.e., $\mathcal{X}_x = \{\mathbf{x} \in \mathcal{X} \mid x \in \mathbf{x}\}$. The invariant distribution on $\mathcal{X}_x$ is denoted $\mathbf{q}_x = \{q_x(B)\}$, for all Borel $B \subset \mathcal{X}_x$. It is easily shown that the system dynamics are time-reversible so the distribution $\mathbf{q}_x$ is found from that of $\mathbf{p}$ by a truncation argument, i.e.,

$$q_x(B) = \frac{p(B)}{p(\mathcal{X}_x)}, \; B \subset \mathcal{X}_x, \; x \in \mathcal{S}. \quad (1)$$

We define expectation with respect to these distributions as follows. Let $f : \mathcal{X} \to \mathbb{R}$ be a bounded measurable function of the link state. Then $\mathbb{E}^{\mathbf{p}}[f(\mathbf{X})]$ denotes expectation of $f(\cdot)$ taken with respect to distribution $\mathbf{p}$, and $\mathbb{E}^{\mathbf{q}_x}[f(\mathbf{X})]$, denotes expectation of $f(\cdot)$ assuming a stream with state $x \in \mathcal{S}$ has been admitted to the link, i.e., conditioned on $x \in \mathbf{X}$. Finally, let $\mathbb{P}(x \in \mathbf{X}) = p(\mathcal{X}_x)$ denote the probability the system contains at least one stream with rate $x$ when $\mathbf{X} \sim \mathbf{p}$.

## 2.3 Adaptation policies

Having specified the stream admission process it remains to discuss how the bandwidth $c$ is allocated among the admitted streams. Let $Y = a(\mathbf{X})$ denote the (random) aggregate minimum load when the link is in steady state, and let $\{Y(t)\}$ denote the corresponding (homogeneous, stationary, ergodic) stochastic process. The system dynamics ensure $Y(t) \le c$ a.s. for all $t$. An adaptation policy $\pi$ defines how to allocate the *residual bandwidth* $c - Y(t)$ at each time $t$ among the active streams. Let $N = n(\mathbf{X})$ denote the (random) number of active streams when the link is in steady state, and let $\{N(t)\}$ denote the corresponding stochastic process.

Suppose the state vector is $\mathbf{x} = (x_1, \ldots, x_n)$ at some time $t$, i.e., there are $n$ active streams with minimum rates $x_i$. An adaptation policy $\pi$ is represented by a function $\mathbf{s}^\pi(\mathbf{x}) = (s^\pi_{x_1}(\mathbf{x}), \ldots, s^\pi_{x_n}(\mathbf{x}))$ that assigns a subscription level to each active stream. In particular, $s^\pi_x(\mathbf{x})$ is the allocation under policy $\pi$ to a stream with state (minimum subscription level $x$) when the link state is $\mathbf{x}$. We define $\Pi$ as the set of policies that are feasible and work-conserving.

Definition 1. **Feasibility** *Feasible adaptation policies satisfy the stream and capacity constraints.*

- **stream constraint:** *each stream's allocation* $s^\pi_{x_i}(\mathbf{x}) \in \mathcal{A}_{x_i}$ *for each* $i = 1, \ldots, n(\mathbf{x})$.
- **capacity constraint:** *the aggregate allocation obeys the link capacity constraint:* $\sum_{i=1}^{n(\mathbf{x})} s^\pi_{x_i}(\mathbf{x}) \leq c$.

Note that $\frac{Y}{\alpha}$ is the *aggregate maximum load*, i.e., the maximum bandwidth that can be used by the active streams given their respective maximum bandwidth requirements.

Definition 2. **Work-conserving** *Work-conserving adaptation policies satisfy:*

$$\frac{y(t)}{\alpha} > c \quad \Rightarrow \quad \sum s^\pi_{x_i}(\mathbf{x}(t)) = c,$$
$$\frac{y(t)}{\alpha} \leq c \quad \Rightarrow \quad s^\pi_{x_i}(\mathbf{x}(t)) = \frac{x_i}{\alpha}, i = 1, \ldots, n(\mathbf{x}).$$

The first equation simply requires that we utilize all available capacity when the maximum load exceeds the link capacity, and the second equation requires each stream receive its maximum possible rate when the maximum load is under the link capacity.

### 2.3.1 Fair share adaptation policy

We consider the fair share adaptation policy (denoted $\pi_f \in \Pi$) under three different link scenarios. Let $c$ denote the link capacity.

- **Scenario 1: no min/max rates.** Arriving streams have no minimum or maximum bandwidth requirements, i.e., the minimum rate is 0, the maximum rate is (the link capacity) $c$, the set of available encoding rates is $\mathcal{A} = (0, c]$. The fair share allocation when there are $n$ active streams is $s^{\pi_f}(n) = \frac{c}{n}$ for $n = 1, 2, \ldots$.
- **Scenario 2: homogeneous min/max rates.** Arriving streams have common minimum and maximum bandwidth requirements, i.e., the minimum rate is $\sigma$, the maximum rate is $\frac{\sigma}{\alpha} \leq c$, the set of available encoding rates is $\mathcal{A} = [\sigma, \frac{\sigma}{\alpha}]$. The fair share allocation when there are $n$ active streams is $s^{\pi_f}(n) = \min\{\frac{c}{n}, \frac{\sigma}{\alpha}\}$ for $n = 1, \ldots, \bar{n}$, and $\bar{n} = \lfloor \frac{c}{\sigma} \rfloor$.
- **Scenario 3: heterogeneous min/max rates.** Arriving streams have varying minimum and maximum bandwidth requirements, i.e., the minimum rate is $x$ (or $X \sim F_X$ for a random stream), the maximum rate is $\frac{x}{\alpha}$, the set of available encoding rates is $\mathcal{A}_x = [x, \frac{x}{\alpha}]$. The fair share allocation for a stream with minimum bandwidth requirement $x$ when the aggregate minimum load is $x \leq y \leq c$ is $s^{\pi_f}_x(y) = \frac{x}{\alpha} \min\{\frac{c\alpha}{y}, 1\}$. In words, the fair share allocation is the maximum rate $\frac{x}{\alpha}$ when the aggregate maximum load is less than capacity, $\frac{y}{\alpha} < c$, and is proportional to the fraction of the aggregate minimum load $\frac{x}{y}c$ when the aggregate maximum load exceeds capacity.

We emphasize that the fair-share allocation for a stream $(x, d)$ is independent of the stream duration $d$.

## 2.4 QoS metrics

Note that a stream with minimum bandwidth $x$ with duration $d$ admitted to a link operating under an adaptation policy $\pi$ is assigned a (time-varying) subscription level $(s^\pi_x(\mathbf{X}(t)), 0 \leq t \leq d)$ throughout its tenure in the system.

The question considered in this paper is the following: *what is the effect of the fair share adaptation policy on the quality of service seen by the stream?* Answering this question requires a definition of quality of service for rate-adaptive multimedia streams, which is a thorny issue, e.g., [1, 2, 5, 6, 11, 12, 17, 32, 34, 40].

We define two general QoS measures that we feel encompass a wide range of QoS issues. These two measures are the *time-average utility of the instantaneous bandwidth allocation* and the *time-average transition cost of change in bandwidth allocation.* The time average utility metric assumes the existence of an instantaneous utility function $g(s)$ that maps assigned subscription levels $s$ to a user satisfaction level $g(s)$. The metric captures the time-average instantaneous utility over the duration of the stream. The time average transition cost assumes the existence of a transition cost function $h(s, s')$ that captures the visual/aural disruption induced by shifting the stream resolution from $s$ to $s'$. The metric captures the time-average transition cost over the duration of the stream. The formal definitions are below.

Definition 3. *Consider a stream with minimum subscription level $x$ and duration $d$, and let $g_x : [x, \frac{x}{\alpha}] \to \mathbb{R}^+$ be a bounded continuous measurable function where $g_x(s)$ is the* utility *of a bandwidth allocation $s$ to a stream with available rates $\mathcal{A}_x \subset [x, \frac{x}{\alpha}]$. The time-average utility of the instantaneous bandwidth allocation for this stream under policy $\pi$ is the random variable*

$$U^\pi_{x,d} \equiv \frac{1}{d} \int_0^d g_x\big(s^\pi_x(\mathbf{X}(t))\big) dt. \qquad (2)$$

Definition 4. *Consider a stream with minimum subscription level $x$ and duration $d$, and let $h_x : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+$ be a bounded measurable function where $h_x(s, s')$ is the* transition cost *associated with the change in bandwidth allocation from $s$ to $s'$ for a stream with minimum bandwidth requirement $x$. Since the bandwidth allocation is a function of the link state, we may write $h_x\Big(s^\pi_x\big(\mathbf{X}(t^-)\big), s^\pi_x\big(\mathbf{X}(t)\big)\Big)$. Assume $h_x(s, s) = 0$ and that $h_x(s, s') = h_x(s', s)$ for all $s, s'$. The time-average transition cost of change in bandwidth allocation for this stream under policy $\pi$ is the random variable*

$$R^\pi_{x,d} \equiv \frac{1}{d} \int_0^d h_x\Big(s^\pi_x\big(\mathbf{X}(t^-)\big), s^\pi_x\big(\mathbf{X}(t)\big)\Big) M(dt). \qquad (3)$$

As stated in the introduction, formulating objective measures of QoS that match experimental subjective testing of customer perceived quality is difficult, but it is reasonable to suppose that these two metrics capture in broad strokes the quality of a stream.

At various times we will consider a particular transition cost function where the transition cost is the magnitude of the change in allocation.

Assumption 2. **Transition cost assumption.** *Assume the transition cost function obeys*

$$h_x\big(s_x^\pi(\mathbf{x}), s_x^\pi(\mathbf{x}')\big) = \big|s_x^\pi(\mathbf{x}) - s_x^\pi(\mathbf{x}')\big|, \ \forall \mathbf{x}, \mathbf{x}' \in \mathcal{X}. \qquad (4)$$

We are interested in both stream-specific and stream-average expectations of these QoS metrics. Stream average quantities yield the QoS for a typical *admitted* stream. Let a stream's minimum rate, $x$, denote its "type." Calculating the admitted stream average QoS in terms of the expected QoS for each stream type requires the distribution on admitted stream types. Let $R_x = \{\mathbf{x} \in \mathcal{X} : a(\mathbf{x}) + x \leq c\}$ be the subset of link states where an arriving stream of size $x$ is admitted. Then, by PASTA (Poisson arrivals see time averages), the link state at a typical arrival time is in its steady state distribution, $\mathbf{X} \sim \mathbf{p}$, and as such the admission probability for a stream of type $x$ is $p(R_x)$. Then the distribution of *admitted* stream types is

$$\tilde{X} \sim F_{\tilde{X}}(x) \equiv \frac{\int_{s_{\min}}^x p(R_u) dF_X(u)}{\int_{\mathcal{S}} p(R_u) dF_X(u)}, \quad x \in \mathcal{S}. \qquad (5)$$

The admitted stream average is now defined as expectation with respect to the distribution $F_{\tilde{X}}$, and will be denoted by $\tilde{\mathbb{E}}[\cdot]$. We will seek expressions for the following four quantities.

$\mathbb{E}^{\mathbf{q}_x}\big[U_{x,d}^\pi\big]$: the expected value of $U_{x,d}$ for a stream of type $(x,d)$ under policy $\pi$ conditioned on that stream being admitted.

$\mathbb{E}^{\mathbf{q}_x}\big[R_{x,d}^\pi\big]$: the expected value of $R_{x,d}$ for a stream of type $(x,d)$ under policy $\pi$ conditioned on that stream being admitted.

$\tilde{\mathbb{E}}[U^\pi] \equiv \int_{\mathcal{S}\times\mathbb{R}+} \mathbb{E}^{\mathbf{q}_x}[U_{x,d}^\pi] dF_{\tilde{X}}(x) dF_D(d)$: the admitted stream average value of $U$.

$\tilde{\mathbb{E}}[R^\pi] \equiv \int_{\mathcal{S}\times\mathbb{R}+} \mathbb{E}^{\mathbf{q}_x}[R_{x,d}^\pi] dF_{\tilde{X}}(x) dF_D(d)$: the admitted stream average value of $R$.

# 3. ANALYTICAL RESULTS

## 3.1 General results

Lemma 1. *For all policies $\pi \in \Pi$ the QoS metrics obey:*

$$\begin{aligned}
\mathbb{E}^{\mathbf{q}_x}\big[U_{x,d}^\pi\big] &= \mathbb{E}^{\mathbf{q}_x}\big[g_x\big(s_x^\pi(\mathbf{X})\big)\big], & (6)\\
\mathbb{E}^{\mathbf{q}_x}\big[R_{x,d}^\pi\big] &= \mathbb{E}^{\mathbf{q}_x}\Big[\int_{\mathcal{X}_x} Q(\mathbf{X}, d\mathbf{x}) h_x\Big(s_x^\pi(\mathbf{X}), s_x^\pi(\mathbf{x})\Big)\Big], & (7)\\
\tilde{\mathbb{E}}[U^\pi] &= \int_{\mathcal{S}} \mathbb{E}^{\mathbf{q}_x}[g_x(s_x^\pi(\mathbf{X}))] dF_{\tilde{X}}(x), & (8)\\
\tilde{\mathbb{E}}[R^\pi] &= \int_{\mathcal{S}} \mathbb{E}^{\mathbf{q}_x}\Big[\int_{\mathcal{X}_x} Q(\mathbf{X}, d\mathbf{x}) h_x(s_x^\pi(\mathbf{X}), s_x^\pi(\mathbf{x}))\Big] dF_{\tilde{X}}(x). & (9)
\end{aligned}$$

*See the appendix for proof.*

The lemma allows us to calculate values for $U$ and $R$ under various policies $\pi \in \Pi$ provided we specify the distribution on the state $\mathbf{p}$.

In the following subsections we develop expressions for stream-specific and stream-average QoS metrics under the fair share adaptation policy for each of the three scenarios described in Subsection 2.3.1.

## 3.2 Scenario 1: Unconstrained rates

In this scenario we suppose the arriving streams have no minimum or maximum rates. This means $i$) there is no admission control, $ii$) the queue dynamics are those of the $M/GI/\infty$ queue with offered load $\rho$, $iii$) the relevant state information is just the number of active streams $N(t)$. The steady-state distribution $\mathbf{p}$ is Poisson with parameter $\rho$, i.e.,

$$p(n) = \mathbb{P}(N = n) = e^{-\rho}\frac{\rho^n}{n!}, \ n = 0, 1, 2, \ldots, \qquad (10)$$

and the conditional distribution $\mathbf{q}$ is the steady state distribution of $N(t)$ conditioned on at least one stream being present in the system, i.e.,

$$q(n) = \mathbb{P}(N = n \mid N > 0) = \frac{p(n)}{1 - p(0)}, \ n = 1, 2, 3, \ldots \qquad (11)$$

Corollary 1. *Under the fair share adaptation policy, and assuming arriving streams have no minimum or maximum rates, the QoS metrics obey*

$$\begin{aligned}
\mathbb{E}^{\mathbf{q}}\big[U^{\pi_f}\big] &= \mathbb{E}^{\mathbf{q}}\big[g\big(\frac{c}{N}\big)\big], & (12)\\
\mathbb{E}^{\mathbf{q}}\big[R^{\pi_f}\big] &= 2\lambda\mathbb{E}^{\mathbf{q}}\Big[h\Big(\frac{c}{N}, \frac{c}{N+1}\Big)\Big] & (13)
\end{aligned}$$

*See the appendix for proof.*

The expressions for stream average QoS in the lemma admit an intuitive interpretation. The time-average utility of the instantaneous bandwidth allocation is simply the utility of the expected bandwidth per stream when the number of streams is taken according to its steady-state distribution, conditioned on the system being non-empty.

The time-average transition cost of change in bandwidth allocation is simply twice the rate of stream arrivals times the expected transition cost of an arrival when the number of streams is taken according to its steady-state distribution, conditioned on the system being non-empty. The factor 2 comes from the fact that departures also cause fluctuations, and the rate of departures is also $\lambda$.

## 3.3 Scenario 2: Homogeneous min/max rates

As with the first scenario, all streams again have the same relevant characteristics, i.e., $\mathcal{S} = \{\sigma\}$ (durations may still vary across streams, but this doesn't affect the QoS under the fair-share policy). The relevant link state information is again the number of active streams, $N(t)$. In contrast to the first model, the inclusion of minimum subscription levels $x$ puts an upper bound $\bar{n} = \lfloor \frac{c}{\sigma} \rfloor$ on the number of streams that may simultaneously share the link. Similarly, define $\underline{n} = \lfloor \frac{\alpha c}{\sigma} \rfloor$ as the maximum number of streams that may simultaneously share the link and each receive their maximum subscription level $\frac{\sigma}{\alpha}$.

The link state space is $\mathcal{X} = \{0, \ldots, \bar{n}\}$. The system dynamics are those of the $M/GI/\bar{n}/\bar{n}$ queue and the steady-state distribution $\mathbf{p}$ is a truncated Poisson with parameter $\rho$, i.e.,

$$p(n) = \mathbb{P}(N = n) = \frac{\frac{\rho^n}{n!}}{\sum_{i=0}^{\bar{n}} \frac{\rho^i}{i!}}, \ n = 0, \ldots, \bar{n}. \qquad (14)$$

Define the conditioned distribution $\mathbf{q}$ as the steady state distribution of $N(t)$ conditioned on at least one stream being present in the system, i.e.,

$$q(n) = \mathbb{P}(N = n \mid N > 0) = \frac{p(n)}{1 - p(0)},\ n = 1, \dots, \bar{n}. \quad (15)$$

Finally, let $b = E(\rho, \bar{n})$ denote the blocking probability, where $E(\rho, \bar{n})$ is the Erlang-b formula.

COROLLARY 2. *Under the fair share adaptation policy, and assuming arriving streams have homogeneous minimum and maximum bandwidth requirements $\sigma$ and $\frac{\sigma}{\alpha}$, the QoS metrics obey*

$$\begin{aligned} \mathbb{E}^{\mathbf{q}}\big[U^{\pi_f}\big] &= \mathbb{E}^{\mathbf{q}}\big[g\big(\min\{\frac{c}{N}, \frac{\sigma}{\alpha}\}\big)\big], & (16) \\ \mathbb{E}^{\mathbf{q}}\big[R^{\pi_f}\big] &= 2\lambda q(\underline{n})h\Big(\frac{\sigma}{\alpha}, \frac{c}{\underline{n}+1}\Big) + \\ & \quad 2\lambda \sum_{n=\underline{n}+1}^{\bar{n}-1} q(n)h\Big(\frac{c}{n}, \frac{c}{n+1}\Big). & (17) \end{aligned}$$

*See the appendix for proof.*

### 3.4 Scenario 3: Heterogeneous min/max rates

As opposed to the first two models, now streams have heterogeneous minimum subscription levels $x_i \in \mathcal{S}$. For this scenario we make use of the transition cost Assumption 2. Recall that $\{Y(t)\}$ is the aggregate minimum load process, where $Y(t) = a(\mathbf{X}(t))$. The steady state aggregate minimum load is $Y = a(\mathbf{X})$. The distribution of $Y$ is given by [33] (Eqn. 2.22):

$$F_Y^c(y) = \mathbb{P}(Y \le y) = \frac{1 + \sum_{l=1}^{\infty} \frac{\rho^l}{l!}\sigma_l(y)}{1 + \sum_{l=1}^{\infty} \frac{\rho^l}{l!}\sigma_l(c)}, 0 \le y \le c, \quad (18)$$

where $\rho = \frac{\lambda}{\mu}$ and

$$\sigma_l(y) = \int_{\mathcal{S}^l} \mathbb{I}(x_1 + \dots + x_l \le y) \prod_{i=1}^{l} dF_X(x_i). \quad (19)$$

Note that the blocking probability for an arriving stream with minimum bandwidth requirement $x$ is denoted $b(x) \equiv 1 - F_Y^c(c - x)$. Define $F_{Y|x}^c(y)$ for $0 \le y \le c$ as the distribution of $Y$ conditioned on $x \in \mathbf{X}$. It is straightforward to see that

$$F_{Y|x}^c(y) = \frac{F_Y^c(y) - F_Y^c(x)}{1 - F_Y^c(x)}, \quad x \le y \le c. \quad (20)$$

COROLLARY 3. *Under the fair share adaptation policy and the transition cost assumption, and assuming arriving streams have heterogeneous minimum rates $X \sim F_X$ and maximum rates $\frac{X}{\alpha}$, the QoS metrics obey*

$$\begin{aligned} \mathbb{E}^{\mathbf{q}_x}\big[U_x^{\pi_f}\big] &= \mathbb{E}^{F_{Y|x}^c}\Big[g_x\Big(\frac{x}{\alpha}\big(\frac{c\alpha}{Y} \wedge 1\big)\Big)\Big], & (21) \\ \mathbb{E}^{\mathbf{q}_x}\big[R_x^{\pi_f}\big] &= 2\lambda \mathbb{E}^{F_{Y|x}^c}\Big[\int_0^c \mathbb{I}(Y + x' \le c) \times & (22) \\ & \quad h_x\Big(\frac{x}{\alpha}\big(\frac{c\alpha}{Y} \wedge 1\big), \frac{x}{\alpha}\big(\frac{c\alpha}{Y + x'} \wedge 1\big)\Big) dF_X(x')\Big], \\ \tilde{\mathbb{E}}[U^\pi] &= \int_{\mathcal{S}} \mathbb{E}^{F_{Y|x}^c}\Big[g_x\Big(\frac{x}{\alpha}\big(\frac{c\alpha}{Y} \wedge 1\big)\Big)\Big] dF_{\tilde{X}}(x), & (23) \\ \tilde{\mathbb{E}}[R^\pi] &= 2\lambda \int_{\mathcal{S}} \mathbb{E}^{F_{Y|x}^c}\Big[\int_0^c \mathbb{I}(Y + x' \le c) \times & (24) \\ & h_x\Big(\frac{x}{\alpha}\big(\frac{c\alpha}{Y} \wedge 1\big), \frac{x}{\alpha}\big(\frac{c\alpha}{Y + x'} \wedge 1\big)\Big)\Big] dF_X(x') dF_{\tilde{X}}(x). \end{aligned}$$

*See the appendix for proof.*

The expected time-average utility of the instantaneous bandwidth allocation for a stream of type $x$ is the expected value over all link load levels (conditioned on $x$) of the instantaneous utility of the fair share allocation corresponding to each load level. The expected time average transition cost for a stream of type $x$ is the product of the average rate of changes ($2\lambda$) times the average transition cost. The average transition cost is the expected value over all link load levels (conditioned on $x$) of the instantaneous transition cost associated with the change in fair share allocation corresponding to a stream $x$ and a link load $Y$ seeing all possible link load changes $x'$. The admitted stream average quantities are obtained from the corresponding expected QoS for each stream type $x$ by taking an expectation with respect to the admitted type distribution $F_{\tilde{X}}$.

## 4. ASYMPTOTICS AND SCALINGS

The results in the previous section apply for finite capacity links and are given in terms of expectations due to the inherent dependence of the performance metrics on the stochastic system state. For large capacity links multiplexing a large number of streams this system state undergoes a law of large numbers effect and many functions of the state converge to constants yielding simplified expressions for asymptotic performance. The many small users regime is obtained by letting the link capacity $c$ and the stream arrival rate $\lambda$ both go linearly to infinity. More formally, we introduce a linear capacity scaling consisting of a sequence of links, indexed by $m = 1, 2, \dots$, where the $m^{th}$ link has arrival rate $\lambda(m) = m\lambda$ and link capacity $c(m) = \big(\lambda(m)\frac{1}{\mu}\big)\big(\gamma\frac{\sigma}{\alpha}\big)$ for some $\gamma > 0$. Define $\rho(m) = \lambda(m)\frac{1}{\mu}$. We identify three distinct scaling regimes, parameterized by $\gamma$.

- *Overloaded Regime:* $\gamma < \alpha$. Here, the bandwidth divided by the average number of active streams is less than that required to support streams at their average minimum subscription level, i.e., $\gamma\frac{\sigma}{\alpha} < \sigma$. The asymptotic average blocking probability in this regime is $1 - \frac{\gamma}{\alpha}$. We call this the overloaded regime.

- *Rate Adaptive Regime:* $\alpha \le \gamma \le 1$. Here, the bandwidth divided by the average number of active streams lies between the average minimum subscription level and the average maximum subscription level, i.e., $\sigma < \gamma\frac{\sigma}{\alpha} < \frac{\sigma}{\alpha}$. The asymptotic average blocking probability in this regime is 0. We call this the rate adaptive scaling regime; this will be the regime of primary interest in the sequel.

- *Underloaded Regime:* $\gamma > 1$. Here the bandwidth divided by the average number of active streams strictly exceeds the average maximum subscription level, i.e., $\gamma\frac{\sigma}{\alpha} > \frac{\sigma}{\alpha}$. The asymptotic average blocking probability in this regime is 0. We call this the underloaded regime.

Let $U_x^{m,\pi}, R_x^{m,\pi}$ be the values of $U, R$ for a stream with minimum bandwidth requirement $x$ in the $m^{th}$ system under policy $\pi$. We define the asymptotic QoS under policy $\pi$ for a stream with minimum bandwidth requirement $x$ and with scaling parameter $\gamma$ as

$$\begin{aligned} u_x^{\gamma,\pi} &= \lim_{m\to\infty} U_x^{m,\pi} = \lim_{m\to\infty} g_x\big(s_x^{m,\pi}(\mathbf{X}(m))\big), \\ r_x^{\gamma,\pi} &= \lim_{m\to\infty} R_x^{m,\pi} = \lim_{m\to\infty} \int_{\mathcal{X}_x} Q(\mathbf{X}(m), d\mathbf{x}) h_x^{m,\pi}(\mathbf{X}(m), \mathbf{x}). \end{aligned}$$

We break our notational convention and let $u_X^{\gamma,\pi}, r_X^{\gamma,\pi}$ denote *random variables*, where $u_X^{\gamma,\pi}$ is the asymptotic utility for a random stream $X$ drawn from $F_X$, and similarly for $r_X^{\gamma,\pi}$. We define the stream average asymptotic QoS under policy $\pi$ and with scaling parameter $\gamma$ as

$$\begin{aligned} u^{\gamma,\pi} &= \tilde{\mathbb{E}}^{\gamma}\big[u_X^{\gamma,\pi}\big] = \int_{\mathcal{S}} u_x^{\gamma,\pi} dF_{\tilde{X}}^{\gamma}(x), \\ r^{\gamma,\pi} &= \tilde{\mathbb{E}}^{\gamma}\big[r_X^{\gamma,\pi}\big] = \int_{\mathcal{S}} u_x^{\gamma,\pi} dF_{\tilde{X}}^{\gamma}(x), \end{aligned}$$

where

$$F_{\tilde{X}}^{\gamma}(x) = \lim_{m\to\infty} F_{\tilde{X}}^{(m)}(x)$$

is the asymptotic distribution on admitted stream minimum rates.

Lemma 2. *Under the fair share adaptation policy and the transition cost assumption, and assuming arriving streams have heterogeneous minimum rates $X \sim F_X$ and maximum rates $\frac{X}{\alpha}$, the asymptotic QoS metrics obey*

$$\begin{aligned} u_x^{\gamma,\pi_f} &= g_x\Big(\big((\gamma \vee \alpha) \wedge 1\big)\frac{x}{\alpha}\Big) && (25) \\ r^{\gamma,\pi_f} &= 2\frac{x}{\alpha}\mu\gamma,\ \alpha < \gamma < 1. && (26) \end{aligned}$$

*See the appendix for proof.*

## 5. NUMERICAL AND SIMULATION RESULTS

All numerical results are computed using Mathematica. We have also written a simulator in Perl. All simulation results are given with 90% confidence intervals, although in some cases the intervals too small to be perceptible. In all cases our simulation results show very strong agreement with the numerical results.

### 5.1 Utility and transition costs for a given stream

Consider a link of capacity $c$ with arrival rate $\lambda$, mean stream duration $\mathbb{E}[D] = \frac{1}{\mu}$, offered load $\rho = \frac{\lambda}{\mu}$, and adaptivity $\alpha$. Suppose in particular that the distribution on minimum rate requests of arriving streams is uniform over $[0, c]$, i.e., $F_X = \mathrm{Uni}(0, c)$. From Corollary 3, all we need to calculate is the distribution for the aggregate minimum load $Y$, i.e., $F_Y^c$. As shown in [33], p. 65,

$$F_Y^c(y) = \frac{I_0\big(2\sqrt{\frac{\rho}{c}y}\big)}{I_o\big(2\sqrt{\rho}\big)},\ 0 \le y \le c, \tag{27}$$

where $I_n(z)$ denotes a modified Bessel function of the first kind of order $n$. As shown in [13] p. 206, we can express such functions in terms of hypergeometric series:

$$F\left(\begin{matrix}1\\ n,1\end{matrix}\,\middle|\, z\right) = I_{n-1}(z)\big(2\sqrt{z}\big)\frac{(n-1)!}{z^{\frac{n-1}{2}}}, \tag{28}$$

where

$$F\left(\begin{matrix}1\\ n,1\end{matrix}\,\middle|\, z\right) \equiv \sum_{k=0}^{\infty} \frac{(n-1)!}{(n-1+k)!}\frac{z^k}{k!}. \tag{29}$$

It is straightforward to show that

$$\frac{d}{dz}F\left(\begin{matrix}1\\ n,1\end{matrix}\,\middle|\, z\right) = \frac{1}{n}F\left(\begin{matrix}1\\ n+1,1\end{matrix}\,\middle|\, z\right). \tag{30}$$

Using these identities, we can write the CDF and PDF for $Y$ as

$$\begin{aligned} F_Y^c(y) &= \frac{F\left(\begin{matrix}1\\ 1,1\end{matrix}\,\middle|\, \frac{\rho}{c}y\right)}{F\left(\begin{matrix}1\\ 1,1\end{matrix}\,\middle|\, \rho\right)},\ 0 \le y \le c, && (31) \\ f_Y^c(y) &= \begin{cases} \dfrac{1}{F\left(\begin{matrix}1\\ 1,1\end{matrix}\,\middle|\, \rho\right)}, & y = 0 \\ \dfrac{\rho}{c}\dfrac{F\left(\begin{matrix}1\\ 2,1\end{matrix}\,\middle|\, \frac{\rho}{c}y\right)}{F\left(\begin{matrix}1\\ 1,1\end{matrix}\,\middle|\, \rho\right)}, & 0 < y \le c. \end{cases} && (32) \end{aligned}$$

It is then straightforward to compute the CDF for $Y$ conditioned on $Y \ge x$ is:

$$\begin{aligned} F_{Y|x}^c(y) &= \frac{1}{F\left(\begin{matrix}1\\ 1,1\end{matrix}\,\middle|\, \rho\right)} + \Bigg(1 - \frac{1}{F\left(\begin{matrix}1\\ 1,1\end{matrix}\,\middle|\, \rho\right)}\Bigg) \times \\ &\quad \frac{F\left(\begin{matrix}1\\ 1,1\end{matrix}\,\middle|\, \frac{\rho}{c}y\right) - F\left(\begin{matrix}1\\ 1,1\end{matrix}\,\middle|\, \frac{\rho}{c}x\right)}{F\left(\begin{matrix}1\\ 1,1\end{matrix}\,\middle|\, \rho\right) - F\left(\begin{matrix}1\\ 1,1\end{matrix}\,\middle|\, \frac{\rho}{c}x\right)}. \end{aligned} \tag{33}$$

The PDF for $Y$ conditioned on $Y \ge x$ is:

$$f_{Y|x}^c(y) = \begin{cases} \dfrac{1}{F\left(\begin{matrix}1\\ 1,1\end{matrix}\,\middle|\, \rho\right)}, \\ \dfrac{\rho}{c}\dfrac{F\left(\begin{matrix}1\\ 2,1\end{matrix}\,\middle|\, \frac{\rho}{c}y\right)}{F\left(\begin{matrix}1\\ 1,1\end{matrix}\,\middle|\, \rho\right)}\dfrac{F\left(\begin{matrix}1\\ 1,1\end{matrix}\,\middle|\, \rho\right)-1}{F\left(\begin{matrix}1\\ 1,1\end{matrix}\,\middle|\, \rho\right)-F\left(\begin{matrix}1\\ 1,1\end{matrix}\,\middle|\, \frac{\rho}{c}x\right)}, \end{cases} \tag{34}$$

where the top expression holds for $y = x$ and the bottom expression holds for $x < y \le c$.

Figure 1 shows numerical plots of the expected utility under the fair-share policy $\mathbb{E}^{F_{Y|x}^c}[U_x^{\pi_f}]$ (21) versus the conditioned minimum stream rate $x$ for varying loads $\rho = 0.1, 1, 10, 100$ on a link of capacity $c = 100$. The streams have an adaptivity $\alpha = \frac{1}{2}$ and a linear utility function $g_x(s) = s$, hence the utility of a rate allocation is the allocation itself. The lines $x$ and $\frac{x}{\alpha} = 2x$ represent the minimum and maximum rate allocations. As expected the average rate is decreasing in $\rho$ and increasing in $x$. At $x = \alpha c = 50$ the allocation for small $\rho$ (here, $\rho = 0.1$) changes slope since the allocation for $x < 50$ is roughly $\frac{x}{\alpha} = 2x$, while for $x > 50$ the allocation is roughly $c = 100$ due to the capacity constraint kicking in. For heavy loads (here, $\rho = 100$) the allocation is roughly $x$ for all $x$.

Figure 2 shows numerical plots of the expected transition cost $\mathbb{E}^{\mathbf{q}_x}[R_x^{\pi_f}]$ (22) versus the conditioned stream rate $x$ for $\rho = 0.1, 1, 10, 100$, and the same parameters as in Figure 1. The transition cost function $h$ is taken as the magnitude of the allocation change (4). We see that the $2\lambda$ coefficient dominates the transition cost for $x$ small, but as the conditioned stream rate approaches the link capacity $x \to c$, high blocking sets in, fewer streams are admitted, and hence the rate of allocation changes goes to zero. The transition penalty for $\rho = 10$ and $\rho = 100$ are fairly close. This is due to the fact that the system is overloaded for both these offered loads, and increasing $\rho$ from 10 to 100 has little effect on the admitted streams.

All simulation results for Figures 1 and 2 are computed by generating $1,000,000$ streams and storing for each admitted stream the triple $(x, u, r)$, denoting the minimum rate, the utility, and the transition cost. The $(x, u)$ and $(x, r)$ pairs are then binned into a histogram with 100 bins each of size 1 from $x = 0$ to $x = c = 100$. 90% Confidence intervals for each bin value are obtained from the $u$ and $r$ values in each bin. This procedure is then repeated for each value of $\rho$.

## 5.2 Utility and transition costs for a typical stream

Figure 3 shows numerical plots of the expected utility $\mathbb{E}^{\mathbf{q}}[U^{\pi_f}]$ (16) under a logarithmic utility function $g(s) = \log s$ versus the scaling parameter $\gamma$ for $m = 20, 40, 60$. Recall $m$ is the arrival rate and link capacity scaling index. Also shown is the corresponding asymptotic utility $u^{\gamma,\pi_f}$ (25). The figure illustrates the convergence to the asymptotic utility under the linear scaling. The jagged behavior for small $m$ and small $\gamma$ comes from the fact that small increases in $\gamma$ increase the average utility, but once the extra capacity is sufficient to allow admission of an additional stream then the average allocation decreases. The utility is decreasing in $m$ for $\gamma$ small because the time between a departure and the subsequent next admission is decreasing in $m$; the active streams receive a higher allocation during these intervals, which vanishes as $m$ gets large.

Figure 4 shows numerical plots of the expected transition cost $\mathbb{E}^{\mathbf{q}}[R^{\pi_f}]$ (17) under a transition cost $h(s, s') = |s - s'|$ versus the scaling parameter $\gamma$ for $m = 20, 40, 60, 80, 100$. Also shown is the corresponding asymptotic transition cost $r^{\gamma,\pi_f}$ (26). The figure illustrates the slower convergence (relative to Figure 3) to the asymptotic transition cost under the linear scaling.

All simulation results for Figures 3 and 4 are found by running 2000 streams at each $\gamma$ level between $\gamma = 0$ and $\gamma = 1.2$ in steps of 0.005 for a link scaling of $m = 40$. 95% confidence intervals are obtained from the 2000 or so $u$ and $r$ values for each stream at each value of $\gamma$.

# 6. CONCLUSION

Fair share is a reasonable policy for bandwidth allocation among competing rate adaptive streams. However, the analysis here identifies several drawbacks to fair share adaptation. First, it is evident that the transition costs associated with fair share can be prohibitively high. In practice the aural and visual disruption caused by such frequent changes in the encoding rate would not justify the benefit obtained by having an occasionally higher rate. Second, it is in practice not feasible to adjust the encoding rates as a continuous parameter. This is due to the fact that large media servers will not likely possess the computational power to dynamically adjust the encoding rate of each active stream. Instead it is more reasonable to suppose that each media object will be available at several discrete encoding levels. Finally, the fair share policy is not optimal. As discussed in [46], the optimal allocation policy that maximizes the stream-average normalized received rate is a *volume-discrimination* policy, meaning that small volume streams receive superior service at the expense of large volume streams. Stream volume here denotes the maximum size in bytes of the stream, and is computed as the product of the bit rate at the maximum encoding level times the stream duration.

In this paper we investigate a model for a bottleneck link subject to a dynamic load of rate adaptive flows. We develop closed-form expressions for QoS metrics seen by such flows under the fair share adaptation policy, for finite capacity links and asymptotic results for large capacity systems. The results permit one to evaluate the sensitivity of flow-level QoS for wired/wireless systems given the flows' characteristics and system load. This enables a rough cost benefit analysis for designing such systems, e.g., deciding the range of encodings worth supporting for an expected offered load. A key contribution of this paper is the focus on flow-level QoS in a dynamic system and specifically, evaluating the dependence on a flow's particular characteristics in a heterogeneous system.

Several important extensions of these results are possible. The most notable extension is to develop corresponding results under other adaptation policies besides fair share. This is the focus of ongoing work. Also we expect that the results above hold for general stream duration distributions, i.e., not just the exponential distribution.

# 7. REFERENCES

[1] N. Argiriou and L. Georgiadis. Channel sharing by rate-adaptive streaming applications. In *IEEE Conference on Computer Communications (Infocom)*, 2002.

[2] A. Bain and P. Key. Modeling the performance of in-call probing for multi-level adaptive applications. Technical Report Technical Report MSR-TR-2002-06, Microsoft Research, October 2001.

[3] J.-J. Chen and D. Lin. Optimal bit allocation for coding of video signals over ATM networks. *IEEE Journal on Selected Areas in Communications*, 15(6):1002–1015, August 1997.

[4] P.-Y. Cheng, J. Li, and J. Kuo. Rate control for an embedded wavelet video coder. *IEEE Transactions on Circuits and Systems for Video Technology*, 7(4):696–702, August 1997.

[5] C.-T. Chou and K. G. Shin. Analysis of adaptive bandwidth allocation in wireless networks with multilevel degradable quality of service. *IEEE Transactions on Mobile Computing*, 3(1):5–17, January–February 2004.

[6] P. Chou, A. Mohr, A. Wang, and S. Mehrotra. Error control for receiver-driven layered multicast of audio and video. *IEEE Transactions on Multimedia*, 3(1), March 2001.

[7] G. de Veciana and J. Walrand. Effective bandwidths: call admission, traffic policing, and filtering for atm networks. *Queueing Systems*, 20:37–59, 1995.

[8] S. Deb and R. Srikant. Global stability of congestion controllers for the Internet. *IEEE Transactions on Automatic Control*, 48(6):1055–1060, 2003.

[9] M. W. Garrett and W. Willinger. Analysis, modeling and generation of self-similar VBR video traffic. In *Proceedings of ACM SIGCOMM*, pages 269–280, 1994.

[10] B. Girod. Psychovisual aspects of image communications. *Signal Processing*, 28:239–251, 1992.

[11] S. Gorinsky, K. Ramakrishnan, and H. Vin.

Addressing heterogeneity and scalability in layered multicast congestion control. Technical Report TR2000–31, Department of Computer Sciences, The University of Texas at Austin, November 2000.

[12] S. Gorinsky and H. Vin. The utility of feedback in layered multicast congestion control. In *Network and Operating Systems Support for Digital Audio and Video (NOSSDAV)*, June 2001.

[13] R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics, $2^{nd}$ edition*. Addison–Wesley, 1994.

[14] G. Grimmett and D. Stirzaker. *Probability and random processes*. Oxford, 1992.

[15] M. Grossglauser and S. Keshav. RCBR: A simple and efficient service for multiple time-scale traffic. *IEEE/ACM Transactions on Networking*, 5(6):741–755, December 1997.

[16] Z. He, J. Cai, and C. W. Chen. Joint source channel rate-distortion analysis for adaptive mode selection and rate control in wireless video coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 12(6):511–523, June 2002.

[17] K. Kar, S. Sarkar, and L. Tassiulas. Optimization based rate control for multirate multicast sessions. In *IEEE Conference on Computer Communications (Infocom)*, pages 123–132, 2001.

[18] F. Kelly. Notes on effective bandwidths. In F. Kelly, S. Zachary, and I. Zeidins, editors, *Stochastic Networks: Theory and Applications*, pages 141–168. Oxford University Press, 1996.

[19] F. Kelly. Charging and rate control for elastic traffic. *European Transactions on Communications*, 8:33–37, 1997.

[20] F. Kelly. Models for a self-managed internet. *Philosophical transactions of the Royal Society*, A358:2335–2348, 2000.

[21] F. Kelly. Fairness and stability of end-to-end congestion control. *European Journal of Control*, 9:149–165, 2003.

[22] F. Kelly, P. B. Key, and S. Zachary. Distributed admission control. *IEEE Journal on Selected Areas in Communications*, 18:2617–2628, 2000.

[23] F. Kelly, A. Maulloo, and D. Tan. Rate control in communication networks: shadow prices, proportional fairness, and stability. *Journal of the Operational Research Society*, 49:237–252, 1998.

[24] R. La and V. Anantharam. Utility based rate control in the internet for elastic traffic. *IEEE/ACM Transactions on Networking*, 10(2):272–286, 2002.

[25] T.-J. Lee, G. de Veciana, and T. Konstantopoulos. Stability and performance analysis of networks supporting elastic services. *IEEE/ACM Transactions on Networking*, 9(1):2–14, 1999.

[26] S. Low and D. E. Lapsley. Optimization flow control, I: basic algorithm and convergence. *IEEE/ACM Transactions on Networking*, pages 861–875, 1999.

[27] L. Massoulie. Stability of distributed congestion control with heterogeneous feedback delays. *IEEE Transactions on Automatic Control*, 47:895–902, 2002.

[28] S. McCanne. *Scalable compression and transmission of Internet multicast video*. PhD thesis, University of California at Berkeley, December 1996.

[29] J. Mo and J. Walrand. Fair end-to-end window-based congestion control. *IEEE/ACM Transactions on Networking*, 8(5):556–567, 2000.

[30] A. Ortega and K. Ramchandran. Rate-distortion methods for image and video compression. *IEEE Signal Processing Magazine*, November 1998.

[31] K. Ramchandran, A. Ortega, and M. Vetterli. Bit allocation for dependent quantization with applications to multiresolution and MPEG video coders. *IEEE Transactions on Image Processing*, 3:533–545, September 1994.

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

[33] K. Ross. *Multiservice loss models for broadband telecommunication networks*. Springer, 1995.

[34] D. Saparilla and K. Ross. Optimal streaming of layered video. In *IEEE Conference on Computer Communications (Infocom)*, March 2000.

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

[36] R. Serfozo. *Introduction to stochastic networks*. Springer, 1999.

[37] S. Shenker. Fundamental design issues for the future internet. *IEEE Journal on Selected Areas in Communication*, 13(7), September 1995.

[38] R. Srikant. *The mathematics of Internet congestion control*. Birkhauser, 2004.

[39] G. Sullivan and T. Wiegand. Rate-distortion optimization for video compression. *IEEE Signal Processing Magazine*, November 1998.

[40] B. Vickers, C. Albuquerque, and T. Suda. Source-adaptive multi-layered multicast algorithms for real-time video distribution. Technical Report Technical Report ICS-TR 99-45, University of California, Irvine, June 1999.

[41] Video Quality Experts Group. Current results and future directions. In *Proc. SPIE Visual Communications and Image Processing*, volume 4067, pages 742–753, 2000.

[42] S. Weber. *Supporting rate adaptive multimedia streams on the Internet*. PhD thesis, The University of Texas at Austin, Austin, TX, May 2003.

[43] S. Weber and G. de Veciana. Asymptotic analysis of rate adaptive multimedia streams. In G. Anandalingam and S. Raghavan, editors, *Telecommunications network design and management*, Operations research / computer science interfaces series, chapter 9, pages 167–192. Kluwer academic press, Boston, MA, 2003.

[44] S. Weber and G. de Veciana. Network design for rate adaptive multimedia streams. In *Proceedings of IEEE INFOCOM*, 2003.

[45] S. Weber and G. de Veciana. Multiple service classes for rate adaptive streams. In G. Anandalingam and S. Raghavan, editors, *Telecommunications planning: innovations, pricing, and revenue management*, Operations research / computer science interfaces

series. Kluwer academic press, Boston, MA, 2005.

[46] S. Weber and G. de Veciana. Rate adaptive multimedia streams: optimization and admisssion control. *IEEE/ACM Transactions on Networking*, December 2005.

# APPENDIX

**Proof of Lemma 1.** *Consider (6). In steady state:*

$$\mathbb{E}^{\mathbf{q}_x}\big[U^{\pi}_{x,d}\big] = \mathbb{E}^{\mathbf{q}_x}\Big[\frac{1}{d}\int_0^d g_x\big(s^{\pi}_x(\mathbf{X}(t))\big)dt\Big] = \mathbb{E}^{\mathbf{q}_x}\big[g_x\big(s^{\pi}_x(\mathbf{X})\big)\big]. \tag{35}$$

*Consider next (7). We restate Lévy's formula for stationary Markov processes, i.e., ([36], p. 103),*

$$\mathbb{E}\Big[\frac{1}{b-a}\int_{(a,b]} h(\mathbf{X}(t^-),\mathbf{X}(t))M(dt)\Big] = \mathbb{E}\Big[\int_{\mathcal{X}} Q(\mathbf{X},d\mathbf{y})h(\mathbf{X},\mathbf{y})\Big], \tag{36}$$

*where $M(\cdot)$ is the counting measure induced by the link state changes in $\{\mathbf{X}(t)\}$. The equation follows. The remaining two equations follow from the first two by noting the expected QoS for each stream type $x$ is independent of the stream duration $d$.* ■

**Proof of Corollary 1.** *Streams vary only by their duration in this model, and the fair share allocation is independent of stream duration, so the expected QoS is constant for all streams. Applying Lemma 1 yields immediately that $\mathbb{E}[U^{\pi_f}] = \mathbb{E}^{\mathbf{p}}[g(\frac{c}{N}) \mid N > 0]$.*

*Consider next $\mathbb{E}^{\mathbf{q}}[R^{\pi_f}]$. Lemma 1 yields:*

$$\begin{aligned}\mathbb{E}^{\mathbf{q}}\big[R^{\pi_f}\big] &= \mathbb{E}^{\mathbf{q}}\Big[\sum_{n'\neq N} Q(N,n')h(s^{\pi_f}(N),s^{\pi_f}(n'))\Big]\\ &= \sum_{n=1}^{\infty} q(n)\sum_{n'=1,n'\neq n}^{\infty} Q(n,n')h(s^{\pi_f}(n),s^{\pi_f}(n')).\end{aligned}$$

*For the $M/G/\infty$ queue the transition rates satisfy*

$$Q(n,n') = \lambda\mathbb{I}(n'=n+1) + n\mu\mathbb{I}(n'=n-1). \tag{37}$$

*Substituting $Q(n,n')$ yields:*

$$\begin{aligned}\mathbb{E}^{\mathbf{q}}[R^{\pi_f}] &= \sum_{n=1}^{\infty}\lambda q(n)h\big(s^{\pi_f}(n),s^{\pi_f}(n+1)\big) +\\ &\quad \sum_{n=2}^{\infty} n\mu q(n)h\big(s^{\pi_f}(n),s^{\pi_f}(n-1)\big)\\ &= 2\lambda\sum_{n=1}^{\infty} q(n)h\big(s^{\pi_f}(n),s^{\pi_f}(n+1)\big)\\ &= 2\lambda\sum_{n=1}^{\infty} q(n)h\big(\frac{c}{n},\frac{c}{n+1}\big)\end{aligned}$$

*where in the second equality we invoke detailed balance and the fact that $h(s,s') = h(s',s)$.* ■

**Proof of Corollary 2.** *Streams vary only by their duration in this model, and the fair share allocation is independent of stream duration, so the expected QoS is constant for all streams. Applying Lemma 1 yields immediately that $\mathbb{E}^{\mathbf{q}}[U^{\pi_f}] = \mathbb{E}^{\mathbf{p}}[g\big(\min\{\frac{c}{N},\frac{\sigma}{\alpha}\}\big) \mid N > 0]$.*

*Consider next $\mathbb{E}^{\mathbf{q}}\big[R^{\pi_f}\big]$. Lemma 1 yields:*

$$\begin{aligned}\mathbb{E}^{\mathbf{q}}\big[R^{\pi_f}\big] &= \mathbb{E}^{\mathbf{q}}\Big[\sum_{n'\neq N} Q(N,n')h\big(s^{\pi_f}(N),s^{\pi_f}(n')\big)\Big]\\ &= \sum_{n=1}^{\bar{n}} q(n)\sum_{n'=1,n'\neq n}^{\bar{n}} Q(n,n')h\big(s^{\pi_f}(n),s^{\pi_f}(n')\big).\end{aligned}$$

*For the $M/G/\bar{n}/\bar{n}$ queue the transition rates satisfy*

$$\begin{aligned}Q(n,n+1) &= \lambda, && n = 0,\ldots,\bar{n}-1\\ Q(n,n-1) &= n\mu, && n = 1,\ldots,\bar{n}\\ Q(n,n') &= 0, && \textit{else}.\end{aligned}$$

*Applying these transition rates yields:*

$$\begin{aligned}\mathbb{E}^{\mathbf{q}}[R^{\pi_f}] &= \sum_{n=1}^{\bar{n}-1} q(n)\lambda h\big(s^{\pi_f}(n),s^{\pi_f}(n+1)\big) +\\ &\quad \sum_{n=2}^{\bar{n}} q(n)n\mu h\big(s^{\pi_f}(n),s^{\pi_f}(n+1)\big)\\ &= 2\lambda\sum_{n=1}^{\bar{n}-1} q(n)h\big(s^{\pi_f}(n),s^{\pi_f}(n+1)\big),\end{aligned}$$

*where the second equality follows from detailed balance equations. Note the lower limit of summation is $n = 1$ (because one stream is assumed present), and the upper limit is $\bar{n}-1$ (because arrivals are blocked in state $\bar{n}$). The fair share allocation is*

$$s^{\pi_f}(n) = \begin{cases} \frac{\sigma}{\alpha}, & 1 \le n \le \underline{n}\\ \frac{c}{n}, & \underline{n} < n \le \bar{n}\end{cases}. \tag{38}$$

*The transition cost is $0$ when $n < \underline{n}$ since all streams receive their maximum rate:*

$$h(s^{\pi_f}(n),s^{\pi_f}(n+1)) = \begin{cases} 0, & 1 \le n < \underline{n}\\ h(\frac{\sigma}{\alpha},\frac{c}{n+1}), & n = \underline{n}\\ h(\frac{c}{n},\frac{c}{n+1}), & \underline{n} < n < \bar{n}\end{cases}. \tag{39}$$

*Substituting the transition costs yields the desired expression.* ■

**Proof of Corollary 3.** *Consider first $\mathbb{E}^{\mathbf{q}_x}\big[U^{\pi_f}_x\big]$. Applying Lemma 1:*

$$\mathbb{E}^{\mathbf{q}_x}\big[U^{\pi_f}_x\big] = \mathbb{E}^{F^c_{Y|x}}\Big[g_x\Big(\frac{x}{\alpha}\big(\frac{c\alpha}{Y}\wedge 1\big)\Big)\Big]. \tag{40}$$

*Consider next $\mathbb{E}^{\mathbf{q}_x}\big[R^{\pi_f}_x\big]$. By the transition cost assumption,*

$$\begin{aligned}h_x\big(s^{\pi_f}_x(\mathbf{x}),s^{\pi_f}_x(\mathbf{x}\cup\{y\})\big) &= s^{\pi_f}_x(\mathbf{x}) - s^{\pi_f}_x(\mathbf{x}\cup\{y\}), \forall y > 0\\ h_x\big(s^{\pi_f}_x(\mathbf{x}),s^{\pi_f}_x(\mathbf{x}\setminus\{x_i\})\big) &= s^{\pi_f}_x(\mathbf{x}\setminus\{x_i\}) - s^{\pi_f}_x(\mathbf{x}),\end{aligned}$$

*for each $i = 1,\ldots,n(\mathbf{x})$.*

*Let $\mathbf{x} = (x_1,\ldots,x_n)$ denote a generic link state. The transition rates satisfy*

$$\begin{aligned}Q(\mathbf{x},\mathbf{x}\cup\{x\}) = &\quad \lambda dF_X(x)\mathbb{I}(x_1 + x_n + x \le c), && x > 0\\ Q(\mathbf{x},\mathbf{x}\setminus\{x_i\}) = &\quad \mu, && i = 1,\ldots,n\\ Q(\mathbf{x},\mathbf{x}') = &\quad 0, && \textit{else}\end{aligned}$$

*Applying the transition cost function and transition rates* *to the equation in Lemma 1, we obtain*

$$\begin{aligned}\mathbb{E}^{\mathbf{q}_x}\big[R_x^{\pi_f}\big] \;=\;& \mathbb{E}^{\mathbf{q}_x}\Big[\lambda\int_0^c \mathbb{I}(a(\mathbf{X})+x'\le c)\times\\ &\big(s_x^{\pi_f}(\mathbf{X})-s_x^{\pi_f}(\mathbf{X}\cup\{x'\})\big)dF_X(x')+\\ &\mu\sum_{i=1}^{n(\mathbf{X})}\big(s_x^{\pi_f}(\mathbf{X}\setminus\{X_i\})-s_x^{\pi_f}(\mathbf{X})\big)\Big]\end{aligned}$$

*It follows from [33] p.62 that*

$$\begin{aligned}&\mathbb{E}^{\mathbf{q}_x}\Big[\lambda\int_0^c \mathbb{I}(v(\mathbf{X})+x'\le c)\times\\ &\big(s_x^{\pi_f}(\mathbf{X})-s_x^{\pi_f}(\mathbf{X}\cup\{x'\})\big)dF_X(x')\Big]\\ =\;&\mathbb{E}^{\mathbf{q}_x}\Big[\mu\sum_{i=1}^{n(\mathbf{X})}\big(s_x^{\pi_f}(\mathbf{X}\setminus\{X_i\})-s_x^{\pi_f}(\mathbf{X})\big)\Big].\end{aligned}$$

*The equation follows. The expressions for $\tilde{\mathbb{E}}[U^\pi]$ and $\tilde{\mathbb{E}}[R^\pi]$ follow from simple conditioning arguments.* ■

**Proof of Lemma 2.** *The proof will make use of the following standard facts on the convergence of random variables. Let $X_n \xrightarrow{\mathcal{D}} X$ denotes convergence in distribution and $X_n \xrightarrow{P} X$ denotes convergence in probability. These two modes of convergence are equivalent for convergence to a constant.*

- **Fact 1.** *If $X_n \xrightarrow{\mathcal{D}} X$ and $f$ is continuous then $f(X_n) \xrightarrow{\mathcal{D}} f(X)$ ([14] p.283);*
- **Fact 2.** *$X_n \xrightarrow{\mathcal{D}} X$ is equivalent to $\mathbb{E}[g(X_n)] \to \mathbb{E}[g(X)]$ for all bounded continuous functions $g$ ([14] p.283);*
- **Fact 3.** *$X_n \to X$ (in any mode) and $c_n \to c$ then $c_nX_n \to cX$ (in that mode) ([14] p. 285);*
- **Fact 4. (Slutsky)** *If $X_n \xrightarrow{\mathcal{D}} X$ and $Y_n \xrightarrow{P} c$ then $f(X_n, Y_n) \xrightarrow{\mathcal{D}} f(X,c)$ for all continuous functions $f$.*

*Consider $u^{\gamma,\pi_f}$. We first show*

$$W(m) = \frac{Y(m)}{\sigma\rho(m)\big(1-b(m)\big)} \xrightarrow{\mathcal{D}} 1, \qquad (41)$$

*where $b(m) = \mathbb{E}^0[b^m(X)]$ is the expected blocking probability on link $m$. As before, one can show that $\lim_{m\to\infty} b(m) = \frac{\gamma}{\alpha}$ for $\gamma \le \alpha$ and $0$ otherwise. By Chebychev's inequality, for all $\epsilon > 0$:*

$$\mathbb{P}\big(\big|W(m)-1\big| > \epsilon\big) \le \frac{Var\big(Y(m)\big)}{\Big(\sigma\rho(m)\big(1-b(m)\big)\epsilon\Big)^2}. \qquad (42)$$

*Recall that $Y(m) = \sum_{i=1}^{N(m)} X_i$. The distribution on $N$ is given by ([33], p. 61)*

$$\mathbb{P}(N=n) = \frac{\frac{\rho^n}{n!}\sigma_n(c)}{1+\sum_{l=1}^{\infty}\frac{\rho^l}{l!}\sigma_l(c)},\; n = 0,1,\ldots, \qquad (43)$$

*and $\sigma_l(c)$ is given by (19). Moreover, because of the capacity constraint, the admitted $X_i$'s comprising the load are not independent and are not drawn from $F_X$ due to selective blocking. Thus the variance of $Y(m)$ is difficult to compute directly and must be bounded.*

*Define $\hat{Y}(m)$ as $\hat{Y}(m) = \sum_{i=1}^{\hat{N}(m)} \hat{X}_i$, $\hat{N}(m) \sim Poisson(\rho(m))$ and the $\hat{X}_i$ are iid with $\hat{X} \sim F_X$ and are independent of $\hat{N}(m)$. Note that $Y(m)$ has the distribution of $\hat{Y}(m)$ conditioned on $\hat{Y}(m) \le c(m)$, i.e.,*

$$\mathbb{P}(Y(m)\le y) = \mathbb{P}(\hat{Y}(m)\le y\big|\hat{Y}(m)\le c(m)), \qquad (44)$$

*for $0 \le y \le c(m)$. We now demonstrate that*

$$Var\big(Y(m)\big) \le \frac{\rho(m)\mathbb{E}[\hat{X}^2]}{1-b(m)}. \qquad (45)$$

*Defining $\hat{Z}(m) = \mathbb{I}(\hat{Y}(m) \le c(m))$ we bound the variance of $\hat{Y}(m)$ by conditioning on $\hat{Z}(m)$:*

$$\begin{aligned}Var(\hat{Y}(m)) &= \mathbb{E}\big[Var(\hat{Y}(m)\big|\hat{Z}(m))\big] + Var\big(\mathbb{E}[\hat{Y}(m)\big|\hat{Z}(m)]\big)\\ &\ge \mathbb{E}\big[Var(\hat{Y}(m)\big|\hat{Z}(m))\big]\\ &= Var(\hat{Y}(m)\big|\hat{Z}(m)=0)\mathbb{P}(\hat{Z}(m)=0)\\ &\quad +Var(\hat{Y}(m)\big|\hat{Z}(m)=1)\mathbb{P}(\hat{Z}(m)=1)\\ &\ge Var(\hat{Y}(m)\big|\hat{Z}(m)=1)\mathbb{P}(\hat{Z}(m)=1)\\ &= Var(Y(m))\mathbb{P}(\hat{Y}(m)\le c(m)).\end{aligned}$$

*This gives an upper bound on the variance of $Y(m)$:*

$$Var(Y(m)) \le \frac{Var(\hat{Y}(m))}{\mathbb{P}(\hat{Y}(m)\le c(m))}. \qquad (46)$$

*The variance of $\hat{Y}(m)$ is easily found to be*

$$\begin{aligned}Var\big(\hat{Y}(m)\big) &= \mathbb{E}[\hat{N}(m)]Var(\hat{X}) + \mathbb{E}\big[\hat{X}\big]^2 Var\big(\hat{N}(m)\big)\\ &= \rho(m)\mathbb{E}\big[\hat{X}^2\big]. \qquad (47)\end{aligned}$$

*Suppose $\gamma > \alpha$, then Markov's inequality yields:*

$$\begin{aligned}\mathbb{P}(\hat{Y}(m)\le c(m)) &\ge 1-\frac{\mathbb{E}[\hat{Y}(m)]}{c(m)}\\ &= 1-\frac{\sigma\rho(m)}{\sigma\rho(m)\gamma/\alpha} = 1-\frac{\alpha}{\gamma}. \qquad (48)\end{aligned}$$

*Substituting (47) and (48) into (46) yields:*

$$Var(Y(m)) \le \frac{\rho(m)\mathbb{E}[\hat{X}^2]}{1-\frac{\alpha}{\gamma}}. \qquad (49)$$

*Substitution of (49) into (42):*

$$\begin{aligned}\mathbb{P}\big(\big|W(m)-1\big|>\epsilon\big) &\le \frac{\frac{\rho(m)\mathbb{E}[\hat{X}^2]}{1-\frac{\alpha}{\gamma}}}{\Big(\sigma\rho(m)\big(1-b(m)\big)\epsilon\Big)^2}\\ &= \frac{\gamma}{\gamma-\alpha}\frac{\mathbb{E}[\hat{X}^2]}{\mathbb{E}[\hat{X}]^2}\frac{1}{\rho(m)(1-b(m))^2\epsilon^2} \to 0.\end{aligned}$$

*For $\gamma \le \alpha$ the Markov inequality bound is trivial; in this case an application of the Chebychev bound suffices:*

$$\mathbb{P}(\hat{Y}(m)\le c(m)) \ge 1-\frac{Var(\hat{Y}(m))}{(c(m)-\mathbb{E}[\hat{Y}(m)])^2}. \qquad (50)$$

*Applying $\mathbb{E}[\hat{Y}(m)] = \sigma\rho(m)$, $Var(\hat{Y}(m)) = \rho(m)\mathbb{E}[\hat{X}^2]$, and $c(m) = \sigma\rho(m)\gamma/\alpha$, substituting into (46) and then into (42) yields an expression that goes to 0 in $m$.*

*The fair share allocation for a stream with minimum bandwidth requirement $x$ on link $m$ is*

$$Z_x(m) = \frac{x}{\alpha}\Big(\frac{\alpha c(m)}{Y(m)}\wedge 1\Big) = \frac{x}{\alpha}\Big(\frac{\gamma}{W(m)\big(1-b(m)\big)}\wedge 1\Big), \quad (51)$$

*which is easily shown to converge to* $\big((\gamma \vee \alpha) \wedge 1\big)\frac{x}{\alpha}$.

*Consider next* $r^{\gamma,\pi_f}$. *Define the sequence* $Z_x(m)$

$$\begin{aligned}
&= 2\lambda(m)\int_0^\infty \mathbb{I}\big(Y(m)+x' \le c(m)\big) \times \\
&\quad h_x\Big(\frac{x}{\alpha}\big(\frac{c(m)\alpha}{Y(m)} \wedge 1\big), \frac{x}{\alpha}\big(\frac{c(m)\alpha}{Y(m)+x'} \wedge 1\big)\Big) dF_X(x') \\
&= 2\frac{x}{\alpha}\lambda(m)\Big[\int_0^\infty \Big(\frac{c(m)\alpha}{Y(m)+x'} - 1\Big) \times \\
&\quad \mathbb{I}\Big(\alpha c(m) - x' < Y(m) \le \alpha c(m) \wedge \big(c(m) - x'\big)\Big) dF_X(x') \\
&+ \int_0^\infty \mathbb{I}\Big(\alpha c(m) < Y(m) \le c(m) - x'\Big) \times \\
&\quad \frac{c(m)\alpha x'}{Y(m)\big(Y(m+x')\big)} dF_X(x')\Big] \\
&= 2\frac{x}{\alpha}\mu\Big[\int_0^\infty \rho(m)\Big(\frac{\gamma}{W(m)\big(1-b(m)\big) + \frac{x'}{\sigma\rho(m)}} - 1\Big) \times \\
&\quad \mathbb{I}\Big(\frac{\gamma - \frac{x'}{\sigma\rho(m)}}{1-b(m)} \le W(m) < \frac{\gamma \wedge \big(\frac{\gamma}{\alpha} - \frac{x'}{\sigma\rho(m)}\big)}{1-b(m)}\Big)\Big] dF_X(x') \\
&+ \int_0^\infty \frac{\gamma\frac{x'}{\sigma}\mathbb{I}\Big(\frac{\gamma}{1-b(m)} \le W(m) < \frac{\frac{\gamma}{\alpha} - \frac{x'}{\sigma\rho(m)}}{1-b(m)}\Big)}{W(m)\big(1-b(m)\big)\Big(W(m)\big(1-b(m)\big) + \frac{x'}{\sigma\rho(m)}\Big)} dF_X(x').
\end{aligned}$$

*Taking limits for* $W(m), b(m), \rho(m)$ *we see the indicator function for the first term is never satisfied, while the indicator function for the second term holds for* $\alpha < \gamma < 1$. *The limit for* $\gamma$ *in this range is easily seen to be* $Z_x(m) \xrightarrow{\mathcal{D}} 2\frac{x}{\alpha}\mu\gamma$. ■

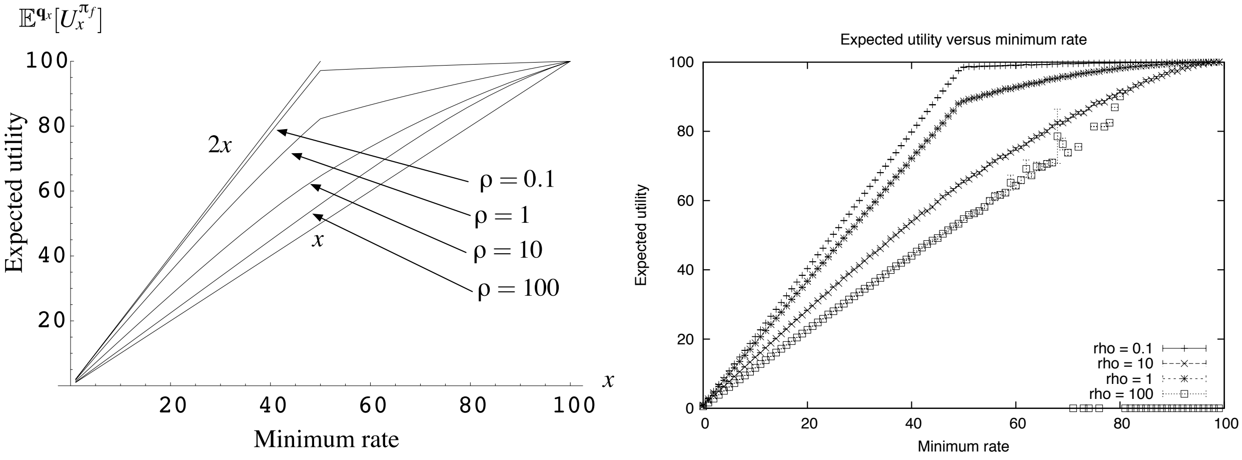

**Figure 1: Expected utility (rate allocation) $\mathbb{E}^{\mathbf{q}_x}[U_x^{\pi_f}]$ for a stream with minimum rate $x$ versus $x$ for varying offered loads $\rho = \{0.1, 1, 10, 100\}$ on a link of capacity $c = 100$. The minimum and maximum allocations are shown as the lines $y = x$ and $y = \frac{x}{\alpha} = 2x$ respectively. The right hand plot shows simulation results.**

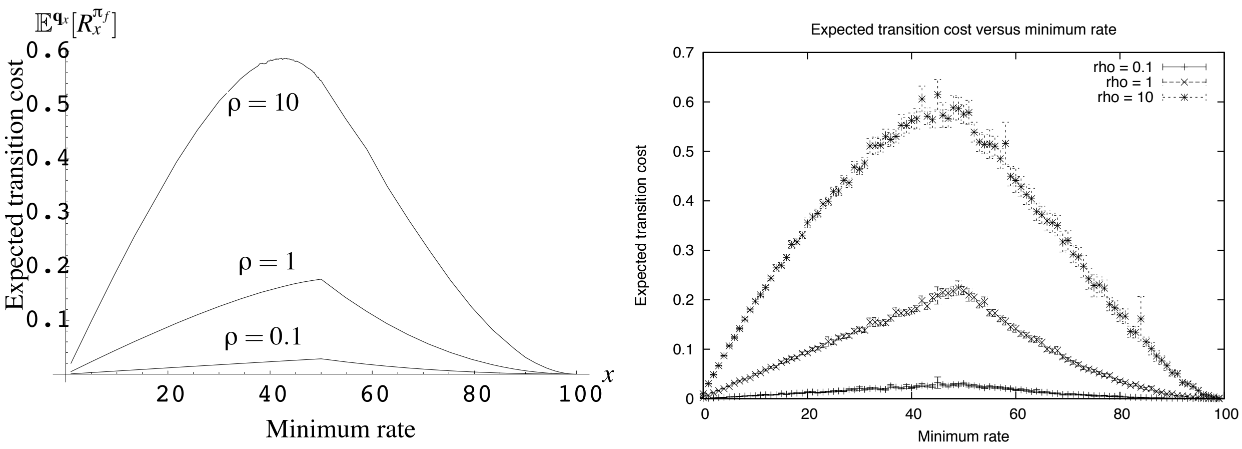

**Figure 2: Expected transition costs $\mathbb{E}^{\mathbf{q}_x}[R_x^{\pi_f}]$ for a stream with minimum rate $x$ for versus $x$ for varying offered loads $\rho = \{0.1, 1, 10, 100\}$ on a link of capacity $c = 100$. The right hand plot shows simulation results.**

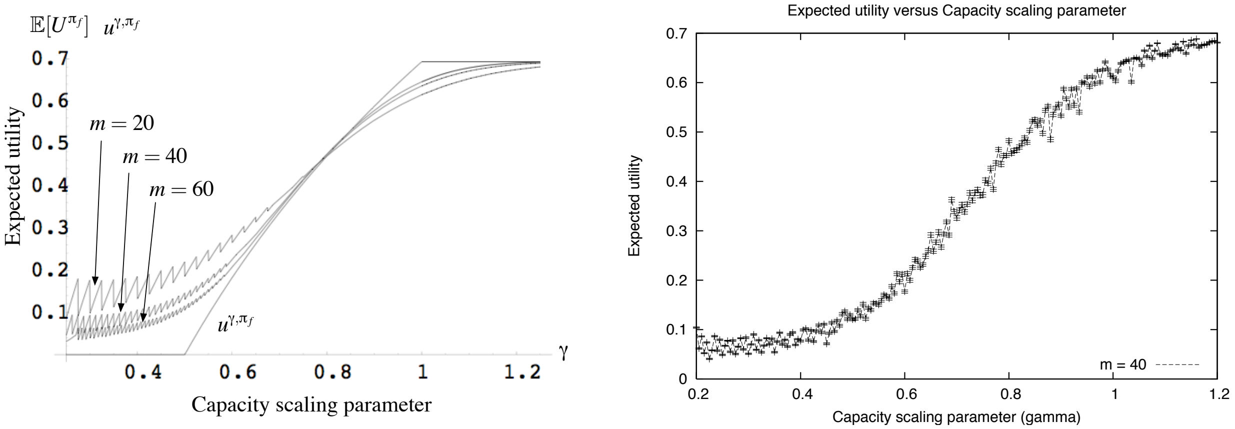

**Figure 3: The expected utility for a typical admitted stream $\mathbb{E}^{\mathbf{q}}[U^{\pi_f}]$ versus the scaling parameter $\gamma$ for $m = \{20, 40, 60\}$. Also shown is the asymptotic utility $u^{\gamma,\pi_f}$. The right hand plot shows simulation results for $m = 40$.**

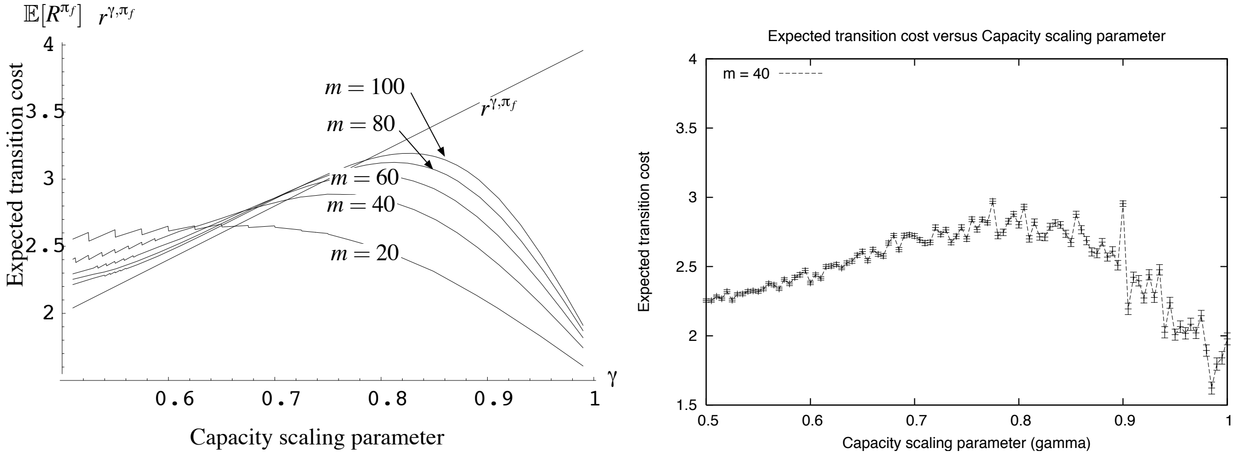

**Figure 4: The expected transition cost for a typical admitted stream $\mathbb{E}^{\mathbf{q}}[R^{\pi_f}]$ versus the scaling parameter $\gamma$ for $m = \{20, 40, 60, 80, 100\}$. Also shown is the asymptotic utility $r^{\gamma,\pi_f}$. The right hand plot shows simulation results for $m = 40$.**