The Dissertation Committee for Steven Patrick Weber

certifies that this is the approved version of the following dissertation:

Supporting Rate Adaptive Multimedia Streams on the Internet

Committee:

Gustavo de Veciana, Supervisor

Aristotle Arapostathis

Ross Baldick

Scott Nettles

Harrick Vin

Supporting Rate Adaptive Multimedia Streams on the Internet

by

Steven Patrick Weber, BS, MS

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May 2003

To my parents, and to Erin.

Acknowledgments

My sincere thanks to my adviser, Professor Gustavo de Veciana, for all of his effort on my behalf over our five years of working together. I would also like to thank Professors Arapostathis, Baldick, Nettles, and Vin for sitting on my dissertation committee. Finally, my deepest gratitude to my parents, Thomas and Dianne Weber, and to Erin, for their continual love and support.

Steven Patrick Weber

The University of Texas at Austin
May 2003

Supporting Rate Adaptive Multimedia Streams on the Internet

Publication No. ____________________

Steven Patrick Weber, Ph.D.
The University of Texas at Austin, 2003

Supervisor: Gustavo de Veciana

This thesis investigates the feasibility of using rate adaptation, i.e., selective service degradation, as a mechanism for achieving various system level Quality of Service (QoS) targets on communication networks. In particular, we investigate how to optimally support rate adaptive multimedia streams on best-effort networks like the Internet.

Optimal and practical mechanisms to maximize the client average QoS, defined in terms of a normalized time average received rate, are established. By scaling the arrival rate and link capacity, we obtain closed form expressions for asymptotic client average QoS. The optimal adaptation policy is identified as the solution to an integer programming problem which has an intuitive “sort by volume” interpretation. Our asymptotic analysis shows the optimal adaptation policy may yield performance improvements of up to 42% over baseline policies.

We demonstrate that a static multi–class admission control policy can achieve the same asymptotic QoS as that of the optimal adaptation policy. This implies that dynamic adaptation may be unnecessary for large capacity networks with appropriate call admission.

The multi–class admission policy, however, requires the stream load characteristics be both stationary and known a priori. To address this drawback we investigate a class of distributed algorithms where the frequency of rate adaptations depends on the stream "volume." We show that these algorithms are able to achieve a QoS comparable to that achieved under the optimal adaptation policy, but without requiring knowledge of system wide parameters. Our simulations indicate our algorithm may yield performance improvements of up to 28% over baseline algorithms.

Finally, we investigate using optimal adaptation in a networking environment supporting multiple service classes with distinct QoS guarantees.

Our results confirm that rate adaptation, i.e., selective service degradation, is a viable means of achieving several different types of system level quality of service targets.

Table of Contents

List of Figures

Chapter 1

Introduction

This chapter aims to motivate the importance of investigating the problem of how to design networks to support rate adaptive multimedia streams. Section 1.1 describes why controlling streaming media traffic is important, and Section 1.2 introduces what streaming media is and how it is significant from a business, cultural, and networking standpoint. Section 1.3 describes rate-adaptive streaming media, Section 1.4 describes some of the related work found in the literature, and Section 1.5 summarizes the contributions of this thesis.

1.1 Importance of controlling streaming media traffic

One of the fastest growing applications on the Internet today is streaming multimedia. The cultural, business, and networking implications of this technology are enormous. We can envision a time in the near future when streaming media will be as commonplace as email and the web. What's preventing this revolution? The current resource bottleneck is access line speed; it is still the case that the majority of home Internet users use dial-up, but the growing popularity of broadband access promises that in the near future we can expect to see a majority of users with high speed access.

But high access line speeds are not a panacea. That is, we cannot expect that giving all users broadband speeds will eliminate network congestion. Just as nature abhors a vacuum, so too it appears that Internet application developers abhor unused bandwidth. We can expect stream resolutions to increase, i.e., higher rate applications, and longer content to be made available for streaming. In the future we might look back at today's current standard of 128 kbps small window media streams as unwatchable and set our minimally acceptable bandwidth quality "threshold" at a megabit per second or higher.

We will therefore have to do congestion control on links supporting potentially hundreds or thousands of simultaneous streams. How should this be done? Economics is likely to play a part; there are multitude of recent papers available on Internet congestion pricing where price serves as an incentive to discriminate between high and low demand users. But even with the economics, *we are still faced with the problem of allocating scarce bandwidth among competing streams with minimum Quality of Service requirements.*

This thesis addresses this fundamental issue. In particular, I will address four topics related to handling congestion control for streaming multimedia:

- optimal dynamic bandwidth allocation among competing streams so as to maximize QoS;
- optimal static bandwidth allocation among large numbers of streams on large capacity networks using admission control;
- decentralized bandwidth allocation algorithms on networks with non-stationary traffic and/or unknown system parameters;
- multiple service classes matched to distinct traffic classes with heterogeneous QoS requirements.

In general, this thesis is an investigation of the claim that *rate adaptation, i.e., selective service degradation, is a general mechanism capable of achieving a wide variety of system level quality of service targets.*

1.2 Streaming media

1.2.1 File transfers and streaming media

The conventional means of media transfer over best-effort networks is to encode the media object into a file and to then transfer the file over the network from the content provider to the client's local disk. Upon completing the file transfer, the client is then able to play back the media object locally from the disk. The drawbacks to this approach are *i*) the client must wait for the file transfer to complete before commencing playback, and *ii*) it is more difficult for content providers to control the client's usage of the media object once it is stored locally on the client's disk.

The principle underlying the operation of streaming media is the simple observation that media playback may be performed concurrently with media transfer over the network provided the network is able to offer sufficient bandwidth on the route connecting the content provider and the client. Sufficient bandwidth here simply means the instantaneous rate of the encoding of the media object. Client side streaming media players typically employ a playback buffer to protect against network jitter, loss, and fluctuations in the available capacity. In addition, content providers may utilize smoothing algorithms which minimize the variability in the instantaneous transmission rate of the encoded stream. Streaming media is attractive because it overcomes both of the difficulties raised with respect to file transfers. Namely, streaming clients need not wait for file transfers to complete before commencing playback since these happen concurrently with streaming media, and client side streaming media players may be designed so that the client is unable to save

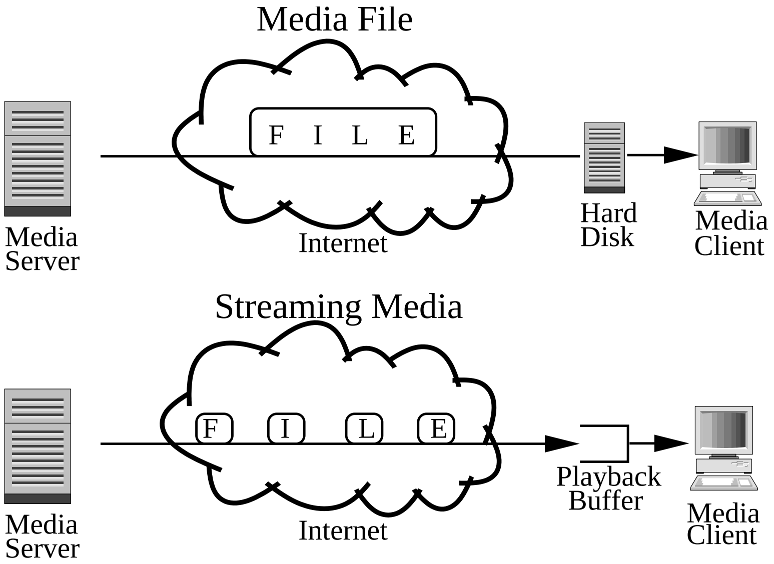

Figure 1.1: Illustration of file transfers and streaming media.

the media object to disk.

The contrast between file transfers and streaming media is illustrated in Figure 1.1. With file transfers (top), the entire media object must be transferred over the network and stored on the client's disk before the client may commence playback. With streaming media (bottom), transfer and playback are simultaneous in that the encoding is sent to a local playback buffer and then immediately to the client side streaming media player.

This simple innovation in media transfer has had profound implications in business, culture, and on how data networks are used and designed.

1.2.2 Business impact

Streaming media has become a major industry both in terms of media servers and content providers. Evidence of this fact is provided in the following statistics.

- The Cahner's In-Stat Group, a specialist in digital communications business research, estimates the streaming media industry will be worth $5 billion by

2005 [15].

- The Gartner group, a specialist in IT business research, predicts that 80% of the top 2000 businesses worldwide will deploy video applications to employee desktops by 2006 [5].
- Enterprise video streaming traffic will grow by an average of 90% each year [22].

These statistics underscore the point that the business potential of streaming media is enormous.

1.2.3 Cultural impact

Aside from the business impact, streaming media is also having a tremendous impact on our culture. Namely, streaming media is expected to become as common a mode of accessing media content as television and radio. The following statistics support this claim.

- 103 million Americans have used streaming media [26].
- 90% of web sites expected to offer some type of streaming content within the next two years [19].
- 56% of Internet users access streaming video and audio at work [21].
- A person with broadband home access spends as much time accessing online media as watching television or listening to radio [25].

These statistics suggest that users with broadband access, at work or at home, make use of streaming media. Thus we can expect that the popularity of streaming media will increase as more and more modem users make the switch to broadband access.

1.2.4 Network impact

The increasing volume of streaming traffic on the Internet will have a potentially negative effect on other network traffic. This point is underscored by the following statistics.

- Service providers expect 24% of network traffic to be streaming content by 2006 [4].
- 60-80% of streaming media traffic is transmitted via UDP and is therefore not responsive to network congestion [18].

Content providers and client side streaming media players utilize the RTSP protocol [29] which transmits streaming content over UDP. The RTSP protocol was designed to use UDP precisely because streaming content requires a minimum transmission rate for satisfactory operation and therefore stream quality will be reduced if the transmission rate is reduced in the face of network congestion, as is the case with TCP.

Transmitting streaming content over UDP is fine when the volume of streaming traffic is small relative to the volume of TCP traffic because the majority of traffic is congestion sensitive. But, as the first statistic predicts, we may well see streaming traffic grow to comprise a significant fraction of overall traffic in the near future. A 1999 New York Times article entitled "Multimedia transmissions drive net toward gridlock" points out that network goodput may drop significantly when a significant fraction of traffic is insensitive to congestion.

Researchers have long been aware of this problem. Sally Floyd and Kevin Fall describe the danger of congestion collapse that arises when UDP traffic comprises a significant fraction of network traffic [6]. Figure 1.2 is a simple example adapted from their paper. The figure shows two links, one with large capacity C and another with small capacity $c < C$. This network is servicing two flows, the TCP link traverses

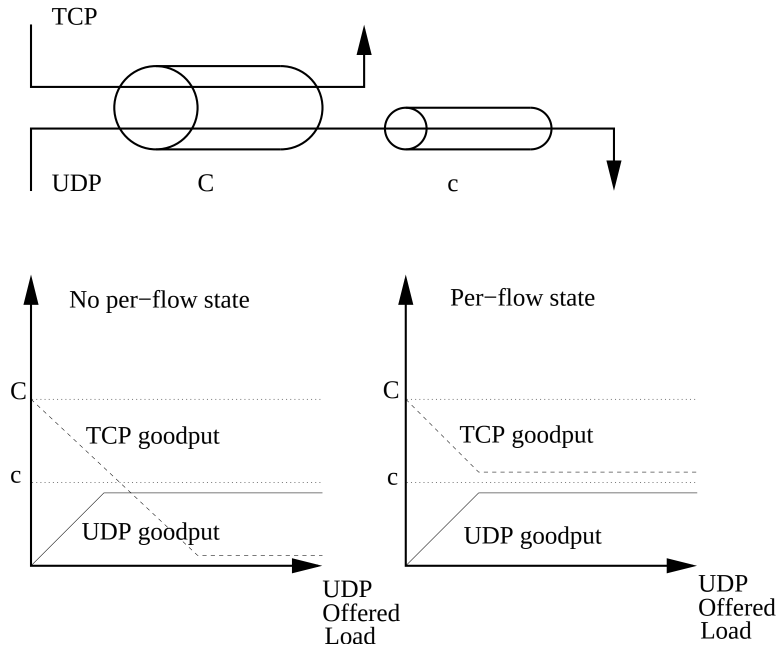

Figure 1.2: Illustration of congestion collapse caused by unresponsive UDP traffic.

only the large link C and the UDP link traverses both links. The plot on the left shows the resulting goodputs of the two streams as the intensity of the UDP flow is increased when no per-flow state is used, e.g., FIFO queueing. Because the flows are not protected from one another, increasing the intensity of the UDP flow above its feasible goodput of c results in wasting the bandwidth $C - c$ with UDP packets that will be dropped on the second link. TCP responds to the congestion on the link by continually reducing its transmission rate to near zero. Contrast this with the plot on the right where per-flow state is employed to protect the flows from one another. Here increasing the intensity of the UDP has no effect on the TCP flow above the UDP flow's reservation level at c, permitting the TCP flow to make use of the remaining capacity of $C - c$ on the first link.

This simple example illustrates how streaming traffic can have an adverse effect on TCP traffic when UDP traffic is not controlled. The example points to the more general problem of designing networks to support applications with heterogeneous QoS requirements. The proposed IntServ [12] and DiffServ [11] network architectures offer solutions to this problem through the use of network state, on a per-flow and per-traffic class basis respectively.

These network architecture proposals are based on the assumption that individual streams can be adequately served by assigning a static bandwidth reservation to each stream, where the bandwidth reservation may be the mean rate of the stream, the peak rate of the stream, or some effective bandwidth falling between the two. As we will discuss in the next section, the fact that streaming content is rate adaptive implies media streams are not characterized by single rate, but rather should be characterized by a set of rates corresponding to the set of offered encodings of the media content.

1.3 Rate adaptive streaming media

Sophisticated media encoding techniques exist which permit access to streaming media at an arbitrary number of distinct stream resolutions. This section will introduce hierarchical encoding as the primary example of multi-resolution encoding and will then explain how multi-resolution encodings of media content are beneficial in *i*) servicing clients with heterogeneous access line speeds and, *ii*) allowing streams a mechanism for dynamic congestion control.

1.3.1 Hierarchical encoding

Hierarchical encoding [13] is a multi-resolution encoding, meaning that media content may be displayed at various resolutions, depending on the needs of the user. Hierarchical encoding consists of a base-layer and one or more enhancement layers. The base layer comprises a coarse-grained encoding of the media content and each enhancement layer offers a successively finer encoding. The enhancement layer

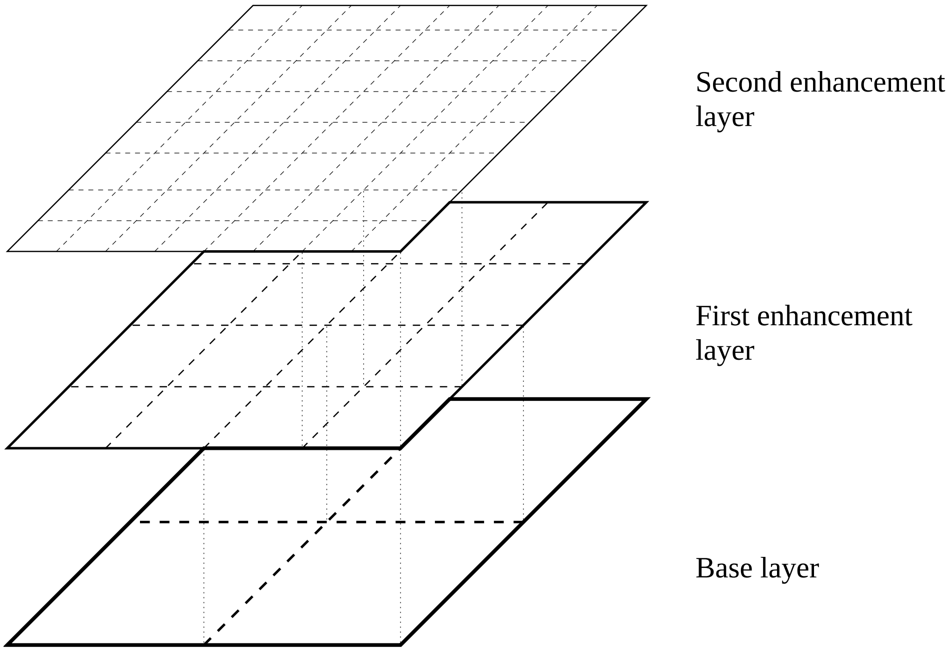

Figure 1.3: Illustration of hierarchical encoding

encodings are efficient in that they encode only the information required to make successive refinements on the information encoded in the underlying layers, i.e., they store the differential between the desired resolution and the resolution obtained under all of the previous layers.

This process is illustrated in Figure 1.3 for the particular case of image encoding. The base layer encoding uses a coarse resolution of 2 × 2. Each square of the base layer is then subdivided and the information for each subdivision is stored as the difference from the corresponding square of the base layer. A similar process yields the second enhancement layer.

An enhancement layer is only useful if the base layer and all of the underlying enhancement layers are also available. Clients utilizing different numbers of layers are said to have different subscription levels. This terminology comes from the Receiver-driven layered multicast (RLM) algorithm [17] where each layer is assigned a multicast address and clients subscribe to as many multicast groups as permitted by their available bandwidth. The minimum subscription level corresponds to subscribing only to the multicast group assigned to carry the base layer encoding, while the maximum subscription level corresponds to subscribing to all of the multicast groups associated with the various layers of the encoding.

1.3.2 Static rate adaptation

A perfect example of different streaming clients having heterogeneous QoS requirements is the case where clients with different access line speeds desire the same media content. Consider the situation in Figure 1.4 where a modem client and a broadband client are accessing media content from a server which has been hierarchically encoded into a base layer and a single enhancement layer. The modem client subscribes to the base layer alone because of its access line constraint while the broadband client is able to subscribe to both layers. If monolithic encoding were used then either the modem client would be unable to access the content (if the encoding included all of the media content) or the broadband client would receive unsatisfactory service (if the encoding included only the content included in the base layer). We term this type of differential service static rate adaptation to denote the fact that heterogeneous clients are able to adapt their subscription levels to match their available rate. The rate adaptation is static in that the clients will maintain their subscription level throughout the duration of the stream since their access line rates are fixed.

1.3.3 Dynamic rate adaptation

Earlier we argued that the increasing popularity of broadband access will gradually eliminate access line constraints, thus static rate adaptation will only be valuable if media content resolutions continue to match client access line speeds. But hierar-

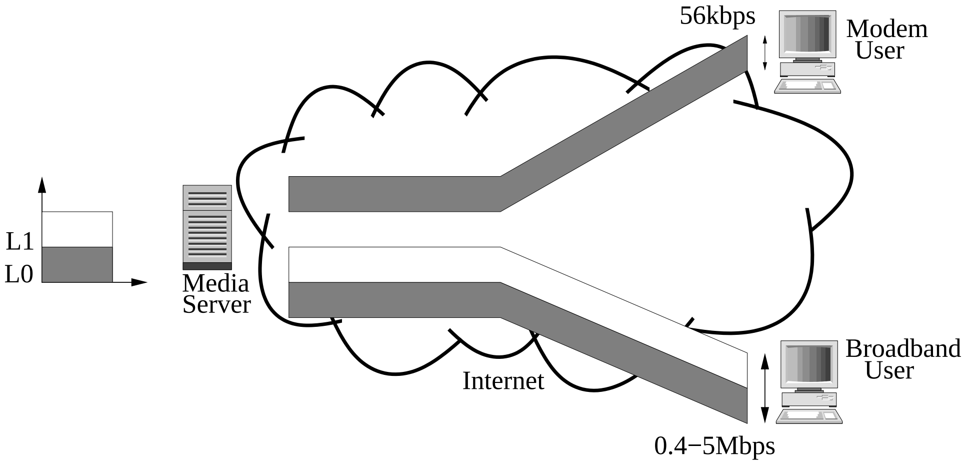

Figure 1.4: Illustration of static rate adaptation.

chical encoding of media content and rate adaptation will still be valuable even if static rate adaptation is no longer necessary. In particular, rate adaptation permits streams to utilize dynamic changes in subscription level as a means of reacting to network congestion and avoiding excessive packet loss.

Consider the situation depicted in Figure 1.5 where the top panel shows the instantaneous load of a link over time and the link capacity. The bottom panel shows a streaming client's subscription level changing as the link congestion changes. At first, when link congestion is moderate, we assume the bandwidth available to the stream is sufficient for the client to subscribe to the base layer and first enhancement layer of the stream. Then, when the congestion level drops appreciably, the client may well be able to subscribe to both enhancement layers. Finally, when heavy congestion arises on the link, the client may need to unsubscribe from both enhancement layers in order to avoid heavy packet loss. Thus, clients can dynamically adjust their subscription levels in response to changing levels of network congestion.

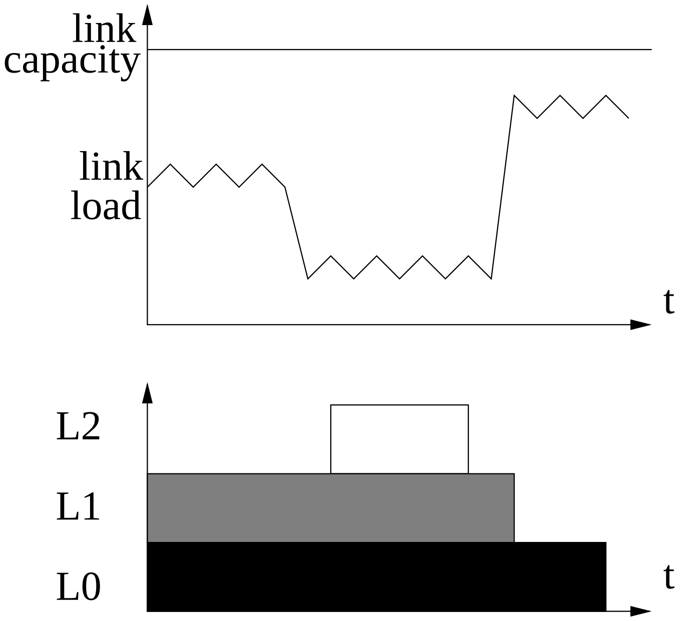

Figure 1.5: Illustration of dynamic rate adaptation.

1.3.4 A taxonomy of network applications

We have shown how rate adaptive streams have minimum QoS constraints, namely, clients must be able to successfully receive the base layer encoding of media content. We have also shown how rate adaptive streams are flexible in that they operate satisfactorily over a range of transmission rates corresponding to the average rates of the minimum and maximum subscription levels.

In this light rate adaptive streams can be viewed as a hybrid between traditional elastic and inelastic network applications. Elastic applications are network applications which require no particular minimum bandwidth for satisfactory operation, but their performance increases as their available bandwidth increases (subject to a law of diminishing returns). Typical elastic applications include WWW traffic and email. Elastic applications are ideally suited for best-effort networks like the Internet which offer no guaranteed service rate.

Inelastic applications, on the other hand, are network applications which do require a minimum bandwidth for satisfactory operation, but their performance does not significantly increase when they are given additional bandwidth above the minimum. Typical inelastic applications include phone, radio, and monolithic stream encodings. Inelastic applications are best served on loss networks, where the network performs admission control to ensure that admitted streams receive their required minimum rates.

These concepts are illustrated in Figure 1.6. The top row shows representative utility functions where the argument is the instantaneous rate given to the application. The bottom row shows abstractions of loss networks, best-effort networks, and rate-adaptive networks. Inelastic applications have a utility function reflecting the minimum rate constraint and show zero marginal benefit to rates above that minimum. Loss networks ensure this rate constraint is satisfied by occasionally blocking arriving calls during periods of congestion. The relevant QoS parameter for a loss network is therefore the blocking probability. Elastic applications have concave increasing utility functions, where the decreasing marginal utility corresponds to the law of diminishing returns for increased bandwidth. Best effort networks don't employ admission control since aggregate utility is maximized by sharing network resources equitably among the active flows. The relevant QoS parameter for a best-effort network is therefore the average bandwidth a typical flow receives. Rate adaptive applications reflect the minimum service rate requirement with a utility function that has a convex increasing neighborhood around zero, and reflect the law of diminishing returns with a utility function that has decreasing marginal utility for large rates. Rate adaptive networks, i.e., networks servicing rate adaptive streams, must perform admission control to ensure the minimum rate constraint of each admitted stream is met, but must also allocate excess bandwidth among the competing streams when congestion is low. Thus the relevant QoS pa-

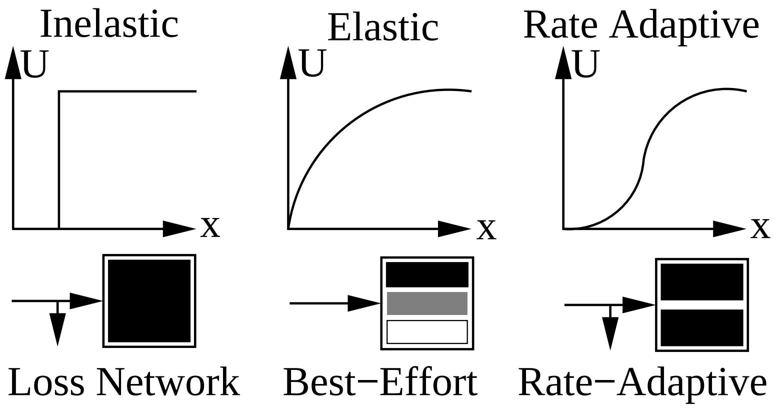

Figure 1.6: Illustration of network application utility functions and architectures.

rameters for a rate adaptive network is the blocking probability *and* the average bandwidth a typical stream receives.

1.4 Related work

We have seen that adequately supporting rate adaptive multimedia streams over best-effort networks requires two components: *i*) an admission control mechanism to ensure admitted streams are able to maintain a transmission rate sufficient to adequately receive the minimum subscription level, and *ii*) a bandwidth allocation mechanism to allocate residual bandwidth (capacity remaining after base layer subscriptions are handled) among competing streams. This residual bandwidth will be used by the active streaming clients to subscribe to one or more of the enhancement layers offered by the content provider.

There is an extensive body of literature on streaming media, we will focus on reviewing a subset of that work which focuses on resource allocation. It is convenient to dichotomize the literature into what we call the client and system perspective.

1.4.1 Client perspective

The client perspective on adequately supporting rate adaptive streams views network congestion as an exogenous process, and studies how to maximize the Quality of Service of an individual stream facing that congestion.

Saparilla and Ross's work, entitled "Optimal streaming of layered video" [28] models the bandwidth available to a stream as a stochastic process and assumes the media content has been encoded as a base layer and a single enhancement layer. They identify how to dynamically optimize the allocation of the available bandwidth among the two layers so as to minimize the impact of client starvation, i.e., an empty client playback buffer. They find the optimal policy is dependent upon the length of the video and the relative importance of the base and enhancement layers.

Rejaie, Handley, and Estrin's work, entitled "Quality adaptation for congestion controlled video playback over the Internet" [23] studies the combined use of subscription level adaptation for long-term coarse-grain adaptation with a TCP-friendly congestion control mechanism for short-term fine-grain adaptation. They find the key issue to be the allocation of buffer resources among the various encoding layers.

Chou et. al. focus on error control in their work "Error control for receiver-driven layered multicast of audio and video" [3]. They propose a layered encoding consisting of multiple source layers each of which has a corresponding parity layer. Clients subscribe to both parity and source layers and use the parity layer for forward error correction (FEC). They augment this error correction with a pseudo-automatic repeat request (ARQ) mechanism.

Vickers, Albuquerque, and Suda take a different approach on system design in their paper "Source adaptive multi-layered multicast algorithms for real-time video distribution" [31]. Source adaptive encoding refers to a scenario where the source uses congestion feedback from the network to dynamically adjust the number of encoded layers and the rate of each layer so as to be of maximum benefit to the subscribing clients. We will argue in the sequel that such sophisticated encoding techniques are not necessary on large capacity links, and in fact two static encodings, a base layer and a single enhancement layer, are sufficient to maximize QoS.

Gorinsky and Vin analyze the benefit of source adaptivity in their work "The Utility of Feedback in Layered Multicast Congestion Control" [10]. They provide a layered multicast congestion control protocol in their work "Addressing Heterogeneity and Scalability in Layered Multicast Congestion Control" [9]. Their work identifies that source adaptivity is beneficial when clients have heterogeneous rates at which they can receive streaming traffic. Our work, on the other hand, focuses on the case when clients are not limited by access line constraints. That is, we view stream adaptivity as a congestion control mechanism.

1.4.2 System perspective

In contrast to the client perspective, the system perspective recognizes that network congestion is not really an exogenous process, but is instead the superposition of the traffic generated by the individual clients, each of which can be controlled. The question from the system perspective is one of resource allocation, i.e., how to optimally allocate network bandwidth among competing streams so as to make efficient use of the bandwidth without allowing congestion to arise.

Argiriou and Georgiadis propose fair channel sharing as a reasonable network bandwidth allocation in their work "Channel sharing by rate adaptive streaming applications" [1]. They define QoS measures similar to those we will propose in the sequel, but they do not identify bandwidth allocations which maximize those QoS measures. They are able to obtain closed form expressions for the performance of their fair channel sharing allocation using the method of Laplace transforms.

Bain and Key's work, entitled "Modeling the performance of in-call probing for multi-level adaptive applications" [2] obtains analytic models to quantify the benefit in QoS as a function of the rate at which the client is permitted to "probe" the network to test for sufficient bandwidth to increase its subscription level. They find that small amounts of in-call probing yields significant increases in QoS.

Kar, Sarkar, and Tassiulas investigate optimal bandwidth allocations for multicast rate adaptive streams in their paper "Optimization based rate control for multirate multicast sessions" [14]. Their optimization criterion is the aggregate instantaneous utility, i.e., they optimize over a static network where the set of active connections is fixed. Our work defines QoS as a time-average over the duration of the stream and thus can be said to optimize over dynamic networks where the set of active connections changes.

1.5 Summary of findings

The contributions of this thesis are fourfold: we identify optimal adaptation policies, asymptotically optimal admission control policies, near-optimal distributed algorithms, and optimal adaptation and admission for networks with multiple service classes.

The contribution of this thesis is an analysis of optimal bandwidth allocation among competing streams on dynamic networks from the system perspective. In addition, we show that on large capacity networks the optimal bandwidth allocation is fixed for each stream, and hence dynamic adaptation policies, which are difficult to implement, may be replaced with simple static admission control policies. We propose a class of distributed algorithms which are able to regain most of the increases in client Quality of Service seen under the optimal allocations, but without the need for centralized control or knowledge of system parameters. Finally, we introduce a network model for multiple service classes, where traffic characteristics and quality of service guarantees vary across classes.

1.5.1 Optimization

From the client perspective it would seem logical that a stream should be encoded with as many subscription levels as possible. This would allow clients to choose a subscription level appropriate for a wide variety of congestion levels. We show, however, that from the system perspective, the client average QoS can be maximized under an adaptation policy which requires only two subscription levels per stream. These two subscription levels correspond to the "coarsest" and "finest" resolutions, i.e., the minimally acceptable subscription level and the maximally useful subscription level. Thus, from a system perspective, there is little benefit in offering a wide variety of subscription levels, assuming flows are not peak rate constrained by their access line.

In addition we demonstrate that the optimal adaptation policy discriminates among active streams according to stream volume, where stream volume is the maximum number of bits corresponding to a stream, i.e., the product of a stream's maximum subscription level and its duration.

1.5.2 Admission Control

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

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

1.5.3 Distributed Algorithms

The popular RLM algorithm proposed in [17] is a distributed algorithm wherein a rate adaptive client lowers its subscription level upon receiving congestion “notifications” from the network, and also periodically performs “join” experiments to test if congestion levels are low enough to support an increase in the client’s subscription level. The client perspective suggests streams should act aggressively to make use of unclaimed bandwidth, regardless of their own resource consumption. We show that client–average QoS is increased above that achieved by conventional RLM schemes when clients respond to congestion in a manner that is volume dependent. In particular, we show through simulation results that allowing small volume streams to be more aggressive than large volume streams yields an overall increase in client average QoS. Moreover, these volume dependent algorithms achieve a client-average QoS that is nearly that achieved under the optimal adaptation and admission policies.

1.5.4 Multiple Service Classes

We will demonstrate that the optimal adaptation policy obtains a high client average QoS by discriminating against large volume streams in order to grant superior service to larger numbers of small volume streams. This type of volume discrimination may be unsatisfactory to clients of large volume streams with a high subjective QoS requirement. Multiple service classes, appropriately priced, offer the opportunity to provide discriminatory service quality independent of stream volume. We identify the optimal adaptation policy for a link offering multiple service classes and identify the asymptotic QoS under this policy.

1.5.5 Outline

We will introduce our model for rate adaptive streams and give a definition of Quality of Service for rate adaptive streams in Chapter 2. In Chapter 3 we investigate optimal adaptation policies. In Chapter 4 we identify asymptotically optimal admission control policies. In Chapter 5 we study a class of volume-dependent distributed algorithms for rate adaptive streams. In Chapter 6 we discuss how to generalize many of our results to the case of multiple service classes. Our conclusion and summary is made in Chapter 7.

Chapter 2

Modeling Rate Adaptive Multimedia Streams

We begin by defining some notational conventions in Section 2.1. Section 2.2 describes our model for rate adaptive multimedia streams. Our network model is described in Section 2.3. Section 2.4 introduces three aspects of Quality of Service (QoS) relevant to rate adaptive multimedia streams and Section 2.5 discusses some possible criticisms of the chosen metrics.

2.1 Notation

We let $\mathbb{R}$ denote the set of real numbers, $\mathbb{Z}$ denote the integers. The set of non-negative reals and non-negative integers will be denoted $\mathbb{R}^+$ and $\mathbb{Z}^+$ respectively.

Bold lowercase letters will be used to denote a vector, say $\mathbf{x}$, and bold uppercase letters will denote a vector of random variables, say $\mathbf{X}$. Sets will be written using script letters, say $\mathcal{X}$.

Expectations will be denoted by $\mathbb{E}[\cdot]$ and probabilities will be denoted by $\mathbb{P}(\cdot)$. Random variables will be denoted by capital letters, e.g., X. The cumulative distribution function (CDF) for a random variable X will be denoted F_X. The complementary cumulative distribution function (CCDF) will be denoted $\bar{F}_X = 1 - F_X$. All random variables will be real-valued and positive, i.e., $X : \Omega \to \mathbb{R}^+$, where $(\Omega, \mathcal{F}, \mathbb{P})$ is the underlying probability triple. Unless otherwise specified, random variables are assumed to have a continuous increasing CDF which guarantees the existence of a probability density function (PDF), denoted f_X, and an inverse, denoted F_X^{-1}.

2.2 Model for rate adaptive multimedia streams

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

2.2.1 Stream Duration

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

2.2.2 Maximum Subscription Level

The maximum subscription level is defined as the time–average instantaneous rate of the stream when encoded at the maximum resolution deemed useful by the provider, i.e., an encoding such that a higher resolution yields a negligible increase in perceived quality. Maximum subscription levels of streams are modeled via random variables, denoted by S, with a common distribution F_S, and mean $\mathbb{E}[S] = \sigma$. A known maximum subscription level is denoted by s. Maximum subscription levels are also assumed to take values in some finite interval $[\underline{s}, \bar{s}] \subset \mathbb{R}^+$.

Maximum subscription levels are assumed to be independent of stream durations. This assumption is primarily made to simplify the analysis. One might argue that maximum subscription levels and stream durations are likely to be correlated in reality. Note, however, that there are important media classes for all four combinations of large and small maximum subscription levels and short and long durations. For example, small subscription levels and long durations are seen in streaming radio, small subscription levels and short durations are seen in streaming individual songs. Similarly, large subscription levels and short durations are seen in music videos, while large subscription levels and long durations are seen in feature length movies.

2.2.3 Adaptivity

Stream adaptivities are defined as the ratio between the minimum subscription level and the maximum subscription level. The minimum subscription level is the time–average instantaneous rate of the stream when encoded at the minimum resolution deemed useful by the provider, i.e., an encoding such that a lower resolution yields a stream with unacceptable perceived quality. Thus a stream with minimum subscription level, say S_{min}, and a maximum subscription level $S_{max} = S$, will have a stream adaptivity of $A = \frac{S_{min}}{S_{max}}$.

Adaptivities are random variables, denoted by A, with a common distribution F_A, and mean $\mathbb{E}[A] = \alpha$. The support of A is necessarily within $(0, 1]$. A known adaptivity is denoted by a. Adaptivities are assumed to be independent of both durations and maximum subscription levels. Note that a stream with maximum subscription level S and adaptivity A has a minimum subscription level of AS. In contrast with S and D, F_A may be discrete.

2.2.4 Offered subscription levels

A stream is offered at a set of *discrete* subscription levels, denoted by $\mathcal{S} = (S_k \mid AS = S_1 < \ldots < S_K = S)$, where $K \in \mathbb{Z}$ denotes the number of subscription levels available to clients, and S_k is the time–average rate corresponding to subscription/encoding k. That is, each encoding k has an instantaneous transmission rate $(B_k(t), 0 \leq t \leq D)$, and $S_k = \frac{1}{D}\int_0^D B_k(t)dt$. This abstraction is independent of the type of encoding used to create the subscription levels, e.g., hierarchical or simultaneous encoding. In hierarchical encoding, subscription level S_k corresponds to the sum of the first k layers, while for simultaneous encoding, S_k corresponds to the kth smallest encoding. We will focus in the sequel on the two subscription levels AS and S, the minimum and maximum subscription levels.

2.3 Network model

We let $\mathcal{L}$ denote the set of links, and the vector $\mathbf{c} = (c_l, l \in \mathcal{L})$ denote the capacities of those links. We assume this capacity is shared by rate adaptive streams. Let $\mathcal{R}$ denote the set of routes, where a route r is comprised of a set of links $\{l \in r\} = \{l \mid l \in r\}$. The vector $\boldsymbol{\lambda} = (\lambda_r, r \in \mathcal{R})$ denotes the arrival rate of new stream requests on each route. We assume all arrival processes are Poisson. The notation $\{r \ni l\} = \{r \mid l \in r\}$ denotes the set of routes incident on link l. We write $\nu_l = \sum_{r \ni l} \lambda_r$ for the arrival rate on link l, and $\boldsymbol{\nu} = (\nu_l, l \in \mathcal{L})$ for the corresponding vector of link arrival rates.

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

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

2.4 Quality of service

Modeling quality of service for multimedia streams is a difficult, and largely unsolved, problem. The Video Quality Experts Group recently performed a statistical analysis of nine proposed objective measures of video quality [24]. They found that none of the proposed models functioned adequately to replace subjective testing. In addition, the performance of the objective models were found to be statistically indistinguishable from one another.

Modeling quality of service for rate adaptive streams promises to be an even harder problem due to the dynamic changes in instantaneous rate. Because of this, any performance model necessitates making some possibly contentious assumptions. Below, we define three aspects of QoS which we feel are especially important: the expected normalized time-average subscription level, the rate of adaptation, and the blocking probability. We will introduce these three metrics in this section, then offer some discussion regarding other possible metric choices in Section 2.5.

2.4.1 Expected Normalized Time-Average Subscription Level

We define $(S(t), 0 \leq t \leq D)$ as the *subscription schedule* of a client, i.e., the instantaneous subscription level of the client at each time t that the client is active. The client is guaranteed to receive its minimum subscription level, and may in fact dynamically subscribe and unsubscribe various subscription levels throughout its tenure in the system. To model this we say $S(t) \in \mathcal{S}$ for each t. The actual values of $S(t)$ will depend on the *adaptation policy* used, which will be discussed in great detail in Chapter 3.

We define the time-average normalized subscription level of the client, denoted Q, as

$$Q = \frac{1}{D} \int_0^D \frac{S(t)}{S} dt, \tag{2.1}$$

Note that $Q \in [AS, 1]$, where $Q = AS$ corresponds to a client receiving its minimum subscription level throughout its duration, i.e., $S(t) = AS, 0 \leq t \leq D$, and $Q = 1$ corresponds to a client receiving its maximum subscription level throughout its duration, i.e., $S(t) = S, 0 \leq t \leq D$. We define the maximum *volume* of a stream as the number of bits comprising the maximum subscription level encoding, i.e., $V = SD$. We can interpret the time-average normalized subscription level as the fraction of the maximum stream volume which is received by the client.

We can assign each client a value for Q based on its stream parameters (S, D) and its subscription schedule $(S(t), 0 \leq t \leq D)$. Let $Q_{i,r}$ denote the assigned value for client i on route r. The customer average normalized time-average subscription level on route r, $\mathbb{E}^0[Q_r]$, is defined as

$$\mathbb{E}^0[Q_r] = \lim_{n_r \to \infty} \frac{1}{n_r} \sum_{i=1}^{n_r} Q_{i,r}, \tag{2.2}$$

where we use the notation $\mathbb{E}^0[\cdot]$ to denote that the expectation is taken as a client average.

We define the *overall* client average normalized time-average subscription level as

$$\mathbb{E}^0[Q] = \sum_{r \in \mathcal{R}} \frac{\lambda_r^a}{\lambda^a} \mathbb{E}^0[Q_r], \tag{2.3}$$

where λ_r^a is the mean rate at which new clients are *admitted* onto route r, and $\lambda^a = \sum_{r \in \mathcal{R}} \lambda_r^a$ is the mean rate at which new clients are admitted onto the network. Thus, the overall client average is simply the weighted sum of the client averages on each route, where the weights reflect the relative intensity of admissions on the route.

Our first, and primary, QoS metric is therefore the expected normalized time-average subscription level, i.e.,

$$\mathbb{E}^0[Q] = \mathbb{E}^0[\frac{1}{D}\int_0^D \frac{S(t)}{S}dt]. \tag{2.4}$$

This may be interpreted as the average fraction of the maximum stream volume received by a typical client.

2.4.2 Expected Rate of Adaptation

Rate of adaptation is defined as the time-average magnitude of the changes in subscription level experienced by a client over the course of its duration. It has been shown that client QoS is adversely affected by these changes [8]. Intuitively, a client with a high rate of adaptation undergoes frequent changes in subscription level, which results in undesirable “flickering” of the stream resolution.

We define the set of subscription level change times for a given client as $\mathcal{C} = \{t \in (0, D) \mid S(t^-) \neq S(t^+)\}$. We then define the rate of adaptation $R = \frac{1}{D}\sum_{t\in\mathcal{C}} |S(t^-) - S(t^+)|$. Our second system level QoS parameter is $\mathbb{E}^0[R]$, the time-average rate of adaptation seen by a typical client:

$$\mathbb{E}^0[R] = \mathbb{E}^0\Big[\frac{1}{D}\sum_{t\in\mathcal{C}} |S(t^-) - S(t^+)|\Big]. \tag{2.5}$$

As stated above, our primary metric of QoS will be the expected normalized time-average subscription level, and it is this quantity which we will seek to maximize when identifying the optimal adaptation policy. We will study the rate of adaptation of the optimal adaptation policy, however, and will show that the rate of adaptation will only be a problem for a relatively small class of streams under the optimal adaptation policy. Note that the cause of changes in the subscription level, which yields the rate of adaptation, is a consequence of dynamic adaptation. In the sequel we will study static admission control policies, wherein admitted streams are assigned a fixed subscription level which they maintain throughout their duration, i.e., the rate of adaptation is zero. We will show that static admission control policies achieve the same asymptotic normalized time-average subscription level as the optimal adaptation policy. Thus we can design the system to maximize our primary QoS metric, but without incurring any performance cost from dynamic adaptation.

2.4.3 Blocking Probability

The previous two QoS metrics are defined for all admitted streams, but, due to the assumption that the network will guarantee each stream sufficient bandwidth for transmission of its minimum subscription level, we will occasionally need to block arriving streams from the network during periods of congestion. To minimize the blocking probability we restrict ourselves to policies which admit as many streams as possible while respecting the minimum rates required by already admitted streams. Thus, we allow admissions even if that admission requires one or more admitted streams to reduce their subscription levels in order to accommodate the new stream. Under such an admission policy, using rate adaptive streams will always result in a lower blocking probability than will using non-adaptive streams. We can think of rate adaptive streams as trading off possible degradation in stream quality (i.e., through dynamic adaptation) for a lower system blocking probability.

In particular, a stream with parameters (d, s, a) will be admitted on route r at time t provided

$$\sum_{r' \ni l} \sum_{i=1}^{n_{r'}(t)} a_{i,r'} s_{i,r'} + as \leq c_l, \ \forall l \in r. \tag{2.6}$$

We define the stationary blocking probability for such a stream by

$$B(r, as) = 1 - \mathbb{P}\Big(\sum_{r' \ni l} \sum_{i=1}^{N_{r'}(t)} A_{i,r'} S_{i,r'} + as \leq c_l, \ \forall l \in r \Big), \tag{2.7}$$

where the probability is taken with respect to the *stationary* distribution of the network.

Note that the admission policy is independent of the adaptation policy. That is, the admission decision depends only on the minimum subscription levels of the admitted streams, while the decision of which streams to adapt depends on the offered subscription levels for each stream. An admission decision may necessitate adaptation, but doesn't specify which streams are to be adapted.

In the sequel we will primarily be concerned with networks operating in a low-blocking regime. Note that, for rate adaptive streams, assuming a low blocking regime does not imply the absence of network congestion. A low blocking regime simply means the average aggregate *minimum* load will not reach capacity. A regime with low network congestion means the average aggregate *maximum* load will not reach capacity.

2.5 Discussion of the QoS metrics

The above three QoS measures attempt to capture the most relevant contributing factors to a client's overall perception of stream quality, but they are by no means the only factors. In this section we address some possible criticisms of these metrics.

2.5.1 Normalizations and time-averages

We have defined our primary QoS metric to be the expected *normalized time-average* subscription level, which may be interpreted as the fraction of the maximum stream volume received by a client. The maximum stream volume, however, may vary by several orders of magnitude across streams. An audio stream of a three minute song may be on the order of several hundred kilobits while a video stream of a two hour movie may be on the order of tens of gigabits. Normalizing the number of bits received by a client by the maximum volume of the corresponding stream has the effect of equivocating the relative value of these two streams provided they receive the same fractional volume.

A valid criticism of this approach is that QoS may be better correlated with the volume of bits received rather than with the fractional volume, e.g., the video client may be more satisfied than the audio client even if the video client receives the minimum subscription level and the audio client receives the maximum subscription level due to the fact that the video client still received more information than the audio client. The corresponding metric might then be $Q' = \int_0^D S(t)dt$, i.e., the total volume of the received stream.

But equivocating stream volume with QoS has its own problems. Consider two video streams, one of an action sequence and one of a news clip, and assume the two streams are of the same duration. Encoding the action sequence will likely require a much higher bit rate than will encoding the news clip, yet it may well be the case that client satisfaction is the same for both. That is, it may be erroneous to assume the action sequence is of higher value to the client than the new clip solely on the basis of the fact that the action sequence requires more bits to encode. The fractional volume approach we have adopted avoids this complication by assuming all media information is of equal value to the client.

An alternative approach might be to consider a combined measure consisting of some weighting of the relative and absolute stream volumes, but identifying the appropriate weights would likely prove difficult. Instead, we address this issue in Chapter 6 where we discuss multiple service class networks for rate adaptive media. Here, each client is permitted to choose a service class where each class provides a different guarantee on the fractional volume transmitted for all streams in the class. As we will show in Chapter 3, optimizing for the average fractional volume results in an adaptation policy which discriminates against large volume streams, i.e., smaller volume streams are more likely to receive a higher fractional volume than larger volume streams. The multi-class generalization permits a large volume stream to ensure against such discrimination by choosing a service class with a strict

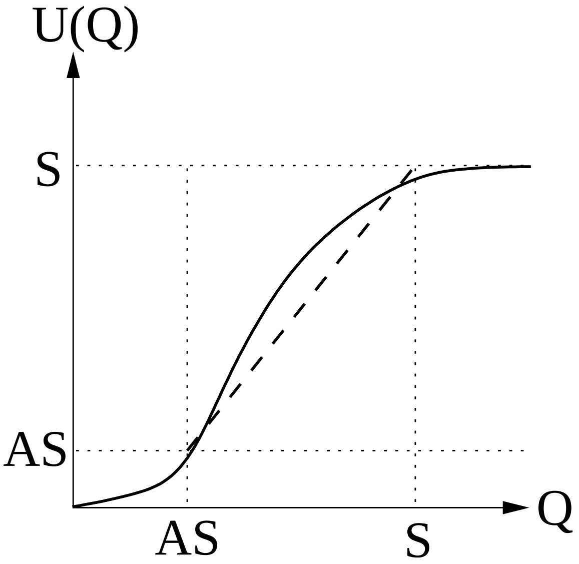

Figure 2.1: Illustration of a utility function for a rate adaptive stream

guarantee on the fractional volume. These classes would of course need to be priced so as to provide proper incentives for clients to choose a service class reflective of their actual QoS requirements.

2.5.2 Linear measures

Another approach to QoS is through the use of utility functions. One might propose, for example, a QoS metric $Q'' = u(\int_0^D S(t)dt)$, for some concave increasing function u with a convex neighborhood around zero. This approach is taken in [30]. The justification is that rate adaptive media has an intrinsic minimum and an intrinsic maximum, yet the adaptivity permits a graceful degradation in quality as a function of reduced resolution. Such a utility function is illustrated in Figure 2.1.

The problem with using such a utility function is that it would need to be parameterized to permit analysis, and no obvious choice of parameterization is available. Recent definitions of utility function classes have appeared in the literature for best-effort networks [20, 33] which maintain tractability without sacrificing generality, yet there are no known corresponding definitions appropriate for streaming media.

Recalling the recent conclusions of the Video Quality Experts Group [24], an accurate objective model of video quality has yet to be found, and so there is no compelling reason to adopt a nonlinear measure over our chosen linear one other than increased generality. Note that the linear measure is defined over the interval $[AS, S]$, i.e., over the range between the minimum and maximum subscription levels. Recall these were defined as the minimum resolution deemed useful the maximum resolution deemed necessary by the provider. Thus the linear measure captures the intrinsic minimum and maximum of u, as illustrated by the straight dashed line in Figure 2.1. The characteristic not captured is the concavity of u within between AS and S. If this concavity is not pronounced, the linear measure may well be a reasonable approximation.

2.5.3 Subscription levels and instantaneous encoding rates

We point out that the expected time-average normalized subscription level and expected rate of adaptation are defined in terms of the instantaneous subscription level $S(t)$. Recall that a subscription level S_k is the time average of the corresponding encoding, i.e., $S_k = \frac{1}{D}\int_0^D B_k dt$, where $(B_k(t), 0 \leq t \leq D)$ is the instantaneous rate of the encoding. One could propose that these metrics should instead be defined in terms of the instantaneous rate to account for the fact that media encodings are bursty.

Consider the illustration in Figure 2.2. The top panel shows the instantaneous encoding rates versus time for two different encodings, $B_1(t)$ and $B_2(t)$. The associated subscription levels S_1 and S_2 are also shown. The media is such that the instantaneous encoding rate is low for the first half of the media clip and is high for the second half. This might occur in video at a change from a pastoral to an action scene. The bottom panel shows a possible adaptation schedule for a client receiving the stream. For the first half of the clip the client receives the low subscription level, say because of network congestion, while for the second half of the clip the client receives the high subscription level. The instantaneous rate received by the client is labeled $B(t)$ and the dynamic subscription level is labeled $S(t)$. The client will be labeled as receiving a relatively low rate of adaptation because the change in the average subscription level is rather small, even though change in the instantaneous encoding rate is significant. Moreover, the example illustrates that provisioning capacity to provide adequate service for media streams can not be done on the basis of the mean rate alone. In particular, provisioning capacity for this stream based on the mean rate would incur heavy packet loss during the action scene.

The fact that media encodings exhibit large degrees of burstiness across multiple time scales has been extensively documented, e.g., [7]. The notion of effective bandwidths [16] has been proposed as a means of determining allocation strategies for bursty streams. The effective bandwidth of a stream is a function of the statistical characterization of the stream, the desired packet loss probability, the link capacity, and the number of streams sharing the link. The effective bandwidth lies somewhere between the mean and peak rate. Roughly speaking, the effective bandwidth will be near the peak rate when the link is multiplexing a small number of streams and the streams require a very loss packet loss probability, and the effective bandwidth approaches the mean as the number of multiplexed streams increases and the packet loss probability constraint is relaxed. The term statistical multiplexing gain refers to the reduction in the effective bandwidth as the number of multiplexed streams increases. Intuitively, this is just the Law of Large Numbers asserting that the dispersion of the aggregate load decreases as the average load is increased, i.e., it becomes increasingly unlikely that multiple streams will simultaneously burst and cause packet loss.

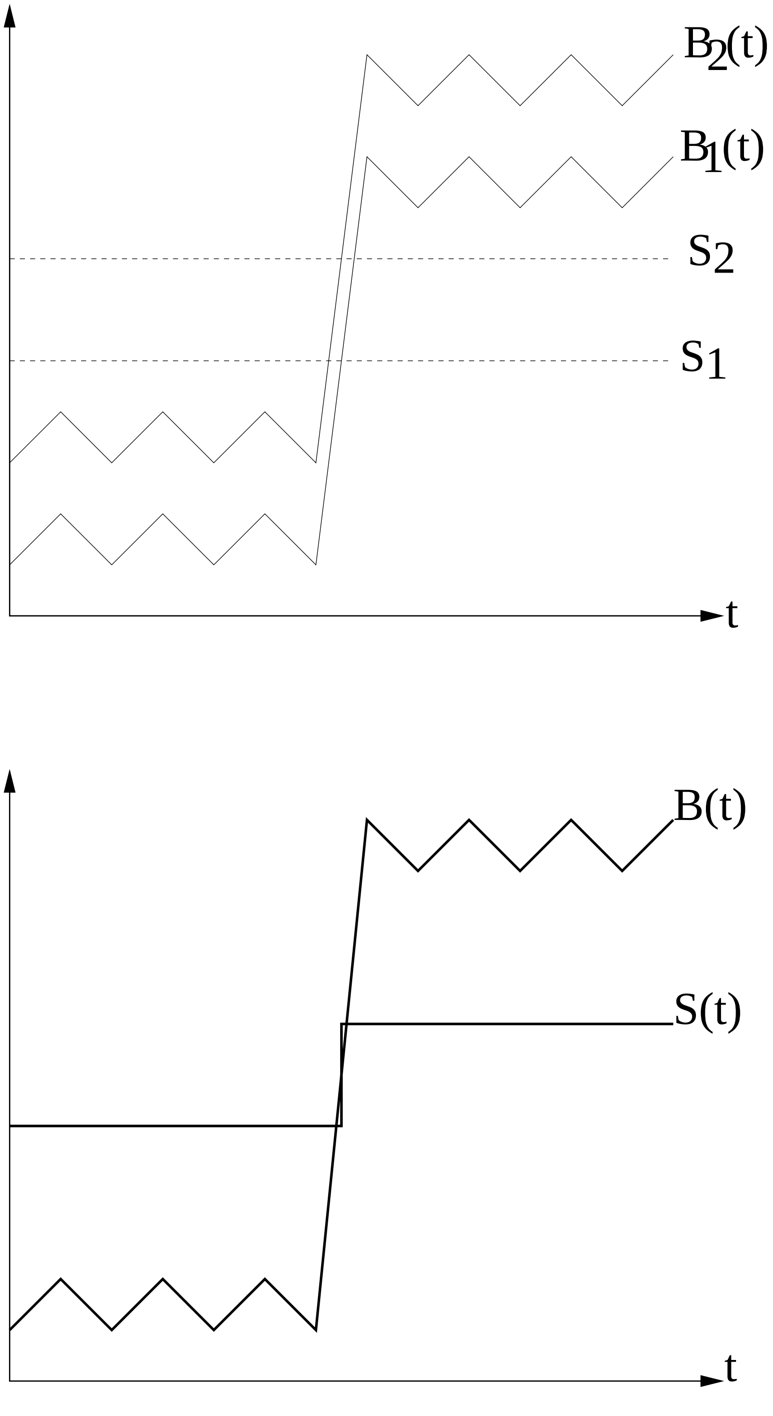

Figure 2.2: Illustration of subscription levels versus instantaneous encoding rates

In this work we are primarily interested in asymptotic expressions for QoS, i.e., how rate adaptive streams behave on large capacity links servicing large numbers of streams. The previous discussion suggests that the burstiness of media encodings, while significant for small capacity systems, may be of less importance on large capacity systems due to the statistical multiplexing gain. Thus, for large capacity systems, we *can* provision capacity for streams based on the mean rate, at least to a first-order approximation.

Chapter 3

Optimal Adaptation Policies

In this chapter we identify the adaptation policy which maximizes our primary Quality of Service metric, the client average normalized time-average subscription level, $\mathbb{E}^0[Q]$. We start out by defining the parameters used for simulation in Section 3.1. Section 3.2 will define the notion of a feasible adaptation policy. We introduce a network scaling in Section 3.3 appropriate for studying large capacity networks servicing large numbers of streams. In Section 3.4 we introduce a sub-optimal adaptation policy, which we call the fair share adaptation policy. We identify the optimal policy in Section 3.5 for two cases: *i*) stored media, where stream durations are typically known at the time of stream admission, and *ii*) live media, where stream durations are unknown. We apply the network scaling to the optimal adaptation policy in Section 3.6 and obtain closed form expressions for asymptotic optimal QoS.

3.1 Simulation parameters

As described in Section 2.2, our model of rate adaptive streams involves four parameters: *i*) a stream duration distribution, F_D, *ii*) a distribution on the maximum subscription level, F_S, *iii*) a distribution on stream adaptivity, F_A, and *iv*) a set of offered subscription levels $\mathcal{S}$. For purposes of optimization (this chapter) and admission control (Chapter 4), we will only be concerned with the first three of these parameters since optimal adaptation and optimal admission control depends on the set of offered subscription levels only through the minimum and maximum subscription levels. That is, intermediate subscription levels are irrelevant in that, under the optimal adaptation policy, clients will not make use of them. These intermediate subscription levels will be important in our discussion of distributed algorithms in Chapter 5.

Thus, for purposes of simulation in this chapter and the next, we need only specify the three distributions F_D, F_S, and F_A. We will see that the asymptotic Quality of Service under optimal adaptation and admission control does depend on the distributions of stream duration and maximum subscription level, but not on the distribution of stream adaptivity. That is, asymptotic QoS depends on the distribution F_A only through the mean $\mathbb{E}[A] = \alpha$, but no such insensitivity result holds for F_D and F_S.

3.1.1 Bounded exponential distribution

Bounded exponential distributions were chosen for both stream duration and maximum subscription level. A random variable X has a bounded exponential distribution with parameters $\mu \in \mathbb{R}^+$ and $M \in \mathbb{R}^+$ if its CDF is given by

$$F_X(x) = \begin{cases} \frac{1-e^{-\mu x}}{1-e^{-\mu M}}, & 0 \leq x \leq M \\ 1, & x > M \end{cases} . \tag{3.1}$$

The corresponding density of a bounded exponential distribution is

$$f_X(x) = \begin{cases} \frac{\mu e^{-\mu x}}{1-e^{-\mu M}}, & 0 \leq x \leq M \\ 0, & x > M \end{cases} \tag{3.2}$$

Note that $\lim_{M\to\infty} F_X$ yields an (unbounded) exponential distribution with parameter μ. The decision to use a bounded distribution was made solely because it greatly simplifies the numerical integration code required to compute the asymptotic QoS. For all intents and purposes the chosen bounds M will be large enough relative to the rate parameter μ so that the distribution is effectively unbounded. The first and second moments of the bounded exponential distribution are given by

$$\begin{aligned} \mathbb{E}[X] &= \frac{1}{\mu}\Big(\frac{1-(1+\mu M)e^{-\mu M}}{1-e^{-\mu M}}\Big) & (3.3) \\ \mathbb{E}[X^2] &= \frac{1}{\mu^2}\Big(\frac{1-(\frac{1}{2}\mu^2M^2+\mu M+1)e^{-\mu M}}{1-e^{-\mu M}}\Big). & (3.4) \end{aligned}$$

These equations can be used to calculate the variance of a bounded exponential random variable.

There is no compelling reason to use this choice of distribution over any other, and we do not claim the actual distributions of stream duration or maximum subscription level actually obey an exponential distribution. We will present summary statistics of asymptotic QoS for other distributions, namely uniform and bounded Pareto, to illustrate how the numerical results depend on the choice of distribution.

3.1.2 Duration and maximum subscription level distributions

The parameters for the stream duration distribution are $\mu_d = \frac{1}{180}$ and $M_d = 6000$ seconds. The values of μ_d and M_d are such that the resulting mean and variance are approximately that of an unbounded exponential random variable with the same rate parameter. Thus the mean stream duration is $\mathbb{E}[D] = \delta \approx 180$ seconds, and the variance is $Var(D) \approx (180)^2 = 32,400$.

The parameters for the maximum subscription level distribution are $\mu_s = \frac{10}{3}$ and $M_s = 10$ Mb/s. Again, the values of μ_s and M_s are such that the resulting mean and variance are approximately that of an unbounded exponential random variable with the same rate parameter. Thus the mean maximum subscription level is $\mathbb{E}[S] = \sigma \approx 0.3$ Mb/s, and the variance is $Var(S) \approx (0.3)^2 = 0.09$.

The CDF and PDF of these two distributions are plotted in Figures 3.1 and 3.2. Note that all axes are logarithmic.

3.1.3 Volume distributions

The maximum volume of a stream is the number of bits associated with the maximum subscription level, i.e, $V = SD$. Our expressions for asymptotic QoS will depend on the maximum volume distribution, $F_V(v) = \mathbb{P}(SD \leq v)$. Note that we have assumed S and D to be independent, so the CDF and PDF may be written

$$F_V(v) = \int_0^\infty F_S(\frac{v}{d})dF_D(d) = \int_0^\infty F_D(\frac{v}{s})dF_S(s) \tag{3.5}$$

$$f_V(v) = \int_0^\infty \frac{1}{d}f_S(\frac{v}{d})dF_D(d) = \int_0^\infty \frac{1}{s}f_D(\frac{v}{s})dF_S(s). \tag{3.6}$$

Given our choice of parameters on F_S and F_D it follows that the absolute maximum stream volume is $M_d M_s = 6000 \times 10 = 60,000$ Mb, and the mean stream volume is $\mathbb{E}[V] = \mathbb{E}[S]\mathbb{E}[D] = \sigma\delta = 0.3 \times 180 = 54$ Mb.

The expressions for asymptotic QoS will also depend on a related distribution, which we term the *typical time distribution.* For an arbitrary random variable X with CDF F_X and PDF f_X, we define the random variable $\hat{X}$ as having a CDF and PDF given by

$$F_{\hat{X}}(x) = \frac{1}{\mathbb{E}[X]}\int_0^x w dF_X(w) \tag{3.7}$$

$$f_{\hat{X}}(x) = \frac{1}{\mathbb{E}[X]}x f_X(x). \tag{3.8}$$

For the case of stream durations and stream volumes, the above CDF and PDF correspond to the distribution of a stream when the system is viewed at a typical time. The intuition is that you are more likely to see a larger volume stream than a smaller volume stream because larger volume streams will likely have longer durations than smaller volume streams. We will write $\hat{D}$ and $\hat{V}$ for stream durations and

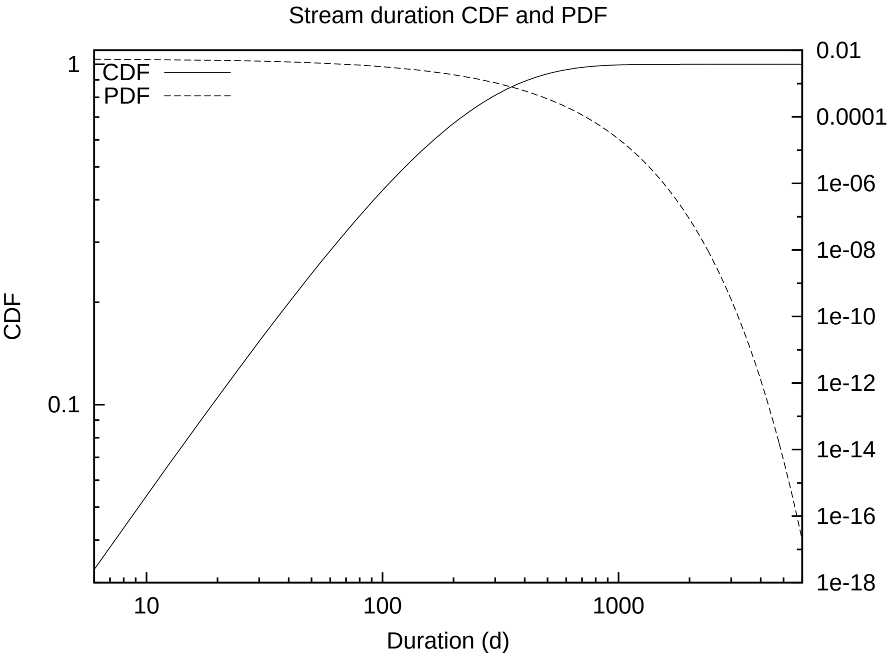

Figure 3.1: Stream duration CDF and PDF: $F_D(d)$ and $f_D(d)$.

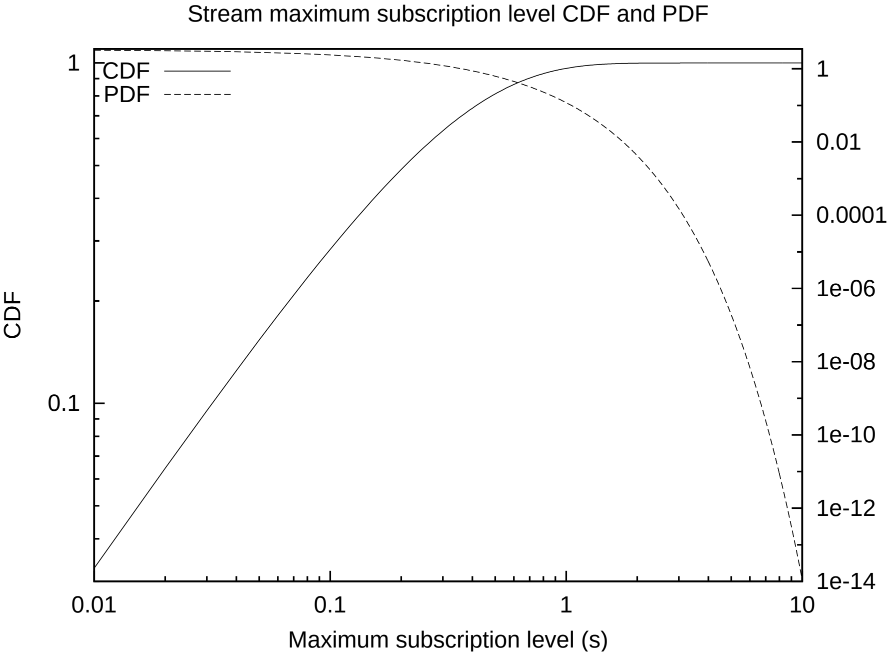

Figure 3.2: Stream maximum subscription level CDF and PDF: $F_S(s)$ and $f_S(s)$.

volumes when doing a typical time analysis, with the understanding that $\hat{D} \sim F_{\hat{D}}$ and $\hat{V} \sim F_{\hat{V}}$.

The CDF and PDF of the distributions for V and $\hat{V}$ are plotted in Figures 3.3 and 3.4. Note that all axes are logarithmic. The figures illustrates that larger volume streams are more prevalent when the system is viewed at a typical time compared with the actual stream volume distribution.

3.1.4 Adaptivity distribution

We have chosen the adaptivity distribution F_A to be uniform on $(\frac{1}{4}, \frac{3}{4})$ with mean $\mathbb{E}[A] = \alpha = \frac{1}{2}$. As stated earlier, the asymptotic QoS under the optimal adaptation and admission policies is *insensitive* to the distribution and depends only on the mean. The mean adaptivity has the interpretation that, on average, encoding media information at its highest useful resolution requires twice the bandwidth as encoding the media information at its lowest useful resolution.

3.1.5 Simulator implementation

We implemented a discrete event simulator in Perl to perform the simulations. All simulations are of a single communications link with a fixed capacity and no cross traffic. All streams are constant bit-rate streams, i.e., we have not modeled the variability in the instantaneous transmission rate inherent in many types of media encoding. Streams are characterized by their maximum subscription level, $S \sim F_S$, their duration, $D \sim F_D$, their adaptivity $A \sim F_A$, and their arrival time. A stream with minimum subscription level as is admitted if the sum of the minimum subscription levels of all active streams is less than $c - as$, i.e., $\sum_{i=1}^{N(t)} a_i s_i + as \leq c$. The system is static between arrival and departure events, hence arrivals and departures are the only events held in the event queue. Active clients are assigned subscription levels for the next inter-event time immediately following an arrival

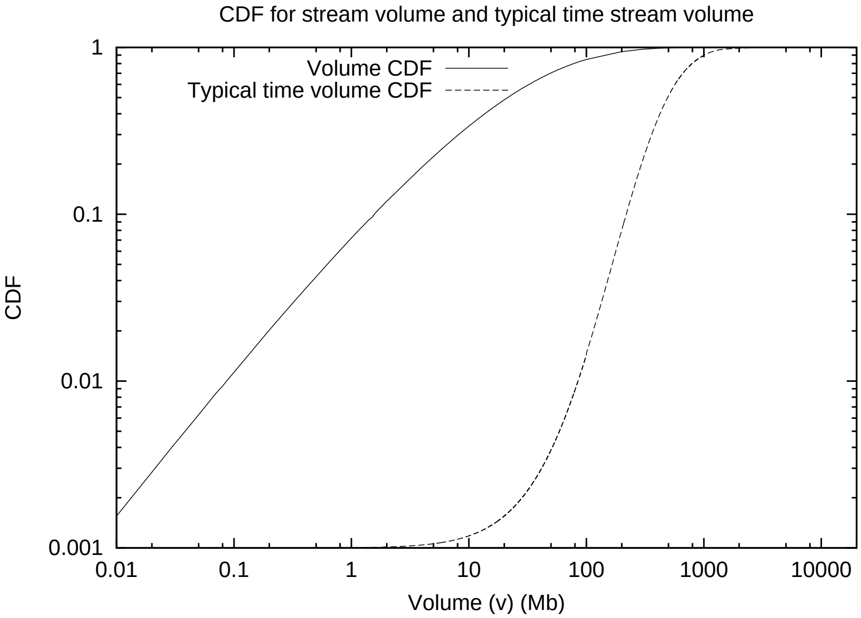

Figure 3.3: CDF for stream volume and typical time stream volume: $F_V(v)$ and $F_{\hat{V}}(v)$.

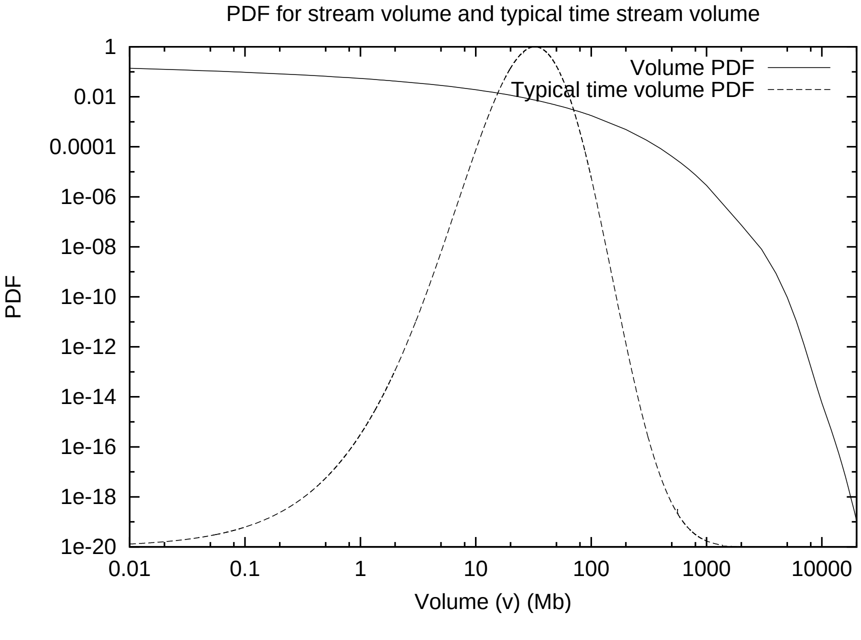

Figure 3.4: PDF for stream volume and typical time stream volume: $f_V(v)$ and $f_{\hat{V}}(v)$.

or departure event. These subscription levels are assigned according to the policy governing the particular simulation. At least 15,000 clients were simulated for each run.

The simulator is written in object oriented Perl with inheritance. Each class file corresponds to a distinct Perl module (pm) file. The parent module `Sim.pm` contains the primary simulation code. It's primary method is `run` which runs the simulation. The `run` method first creates arrays of random variables to create the clients, namely, inter-arrival times, stream durations, maximum subscription levels, and adaptivities. A first client is created, it's arrival time is inserted into the event queue, and an inter-arrival time is selected for the arrival of the next client. Then, while the number of clients that have arrived is less than the total number of clients to be simulated, the simulator continually selects the next event from the event queue. If the event is an arrival then a `arrival` method is called which checks to see if the client may be admitted, and generates the inter-arrival time until the next client arrival. If the event is a departure, then a `departure` method is called which computes the values of Q and R for the client and updates the other relevant state variables. The `adapt` method is called immediately after client admissions and client departures. A parent module `Client.pm` keeps track of the state information for each client including the subscription level which the `adapt` method assigned the client throughout the stream duration. The event is then deleted from the event queue and the next event is handled.

Child classes of `Sim.pm` and `Client.pm` were used to implement the various adaptation policies, admission control policies, and distributed algorithms described in this thesis. These include the following.

- `FairShareSim.pm`, `FairShareClient.pm`: the fair share adaptation policy, described in Subsection 3.4.1;
- `OptAdaptKnownSim.pm`, `OptAdaptKnownClient.pm`: the optimal adaptation

policy for known stream durations, described in Subsection 3.5.2;

- `OptAdaptUnknownSim.pm`, `OptAdaptUnknownClient.pm`: the optimal adaptation policy for unknown stream durations, described in Subsection 3.5.4;
- `OptAdmitSim.pm`, `OptAdmitClient.pm`: the optimal admission control policy, described in Subsection 4.2.2;
- `VolIndAlgSim.pm`, `VolIndAlgClient.pm`: the volume independent algorithm, described in Section 5.1;
- `VolDepAlgSim.pm`, `VolDepAlgClient.pm`: the volume dependent algorithm, described in Section 5.2;
- `MultiClassSim.pm`, `MultiClassClient.pm`: the multiple service class policy, described in Section 6.2.

Driver files were written to run the simulations required to obtain each of the figures; these are just standard Perl files. Several utility files were also written to handle data processing, such as `histogram.pl` which sorts the array of q and r data for each client by their volume into appropriately sized bins. The `gnuplot` program was used to create all of the plots. Confidence intervals are not provided on simulation results, but simulation results are intended only to provide qualitative insight into the policies and algorithms being discussed.

3.2 Feasible adaptation policies

An *adaptation policy* assigns a subscription level to each active stream on the network. We let π denote an adaptation policy, and write $(S^{\pi}(t), 0 \leq t \leq D)$ to denote the adaptation schedule of a client under the policy π. A *feasible* adaptation policy satisfies a stream constraint and a network constraint. In addition, we restrict the set of feasible adaptation policies to non-anticipatory policies, i.e., those policies which determine an allocation at each time based on information available at that time. Subsection 3.2.1 discusses the stream constraint and Subsection 3.2.2 discusses the network constraint.

3.2.1 Stream constraint

A feasible adaptation policy must assign each stream a subscription level that is offered by the content provider. Consider some arbitrary time t. Suppose there are $\mathbf{n}(t) = (n_r(t), r \in \mathcal{R})$ streams on the network, and the content provider of stream (i, r) has offered subscription levels $\mathcal{S}_{i,r}$. An adaptation policy satisfies the stream constraint if the subscription level assignment at time t, $\mathbf{s}^\pi(t) = (s^\pi_{i,r}(t), i = 1, \ldots, n_r(t), r \in \mathcal{R})$, obeys $s^\pi_{i,r}(t) \in \mathcal{S}_{i,r}$ for each stream (i, r).

3.2.2 Network constraint

A feasible adaptation policy must assign subscription levels such that the aggregate load on each link does not exceed the link capacity. We can write the network constraint as

$$\sum_{r \ni l} \sum_{i=1}^{n_r(t)} s^\pi_{i,r}(t) \leq c_l, \forall l \in \mathcal{L}. \tag{3.9}$$

Thus, for any time t, if the number of active streams on each route is $\mathbf{n}(t) = (n_r(t), r \in \mathcal{R})$ and the link capacities are $\mathbf{c} = (c_l, l \in \mathcal{L})$, then the network constraint requires the subscription level assigned to each stream at time t, $\mathbf{s}^\pi(t) = (s^\pi_{i,r}(t), i = 1, \ldots, n_r(t), r \in \mathcal{R})$ be such that the aggregate link loads not exceed the link capacities.

A possible critique of this constraint is that it does not guarantee the actual instantaneous load will not exceed network capacity. Recall the discussion in Subsection 2.5.3 where we pointed out the fact that subscription levels are *time-averages* of the associated encoding, and the actual instantaneous link load $\sum_{r \ni l} \sum_{i=1}^{n_r(t)} b_{i,r}(t)$

may exceed the aggregate subscription link load $\sum_{r \ni l} \sum_{i=1}^{n_r(t)} s_{i,r}(t)$. Effective bandwidths may be used to represent subscription levels, instead of time-averages, but recall that as the capacity is increased, the effective bandwidth approaches the time-average. Given that our focus is on large capacity networks, we will not pursue this point further, and will stick with the network constraint presented above.

3.3 Network scaling for rate adaptive streams

We introduce a network scaling to model the case where large numbers of streams share large capacity links. Specifically, we will investigate a sequence of networks, indexed by $m \in \mathbb{Z}^+$, where the m^{th} network has a vector of route arrival rates $\boldsymbol{\lambda}(m) = (\lambda_r(m), r \in \mathcal{R})$, and a vector of link capacities $\mathbf{c}(m) = (c_l(m), l \in \mathcal{L})$. The arrival rates will take the form $\lambda_r(m) = m\lambda_r$ for some scalar $\lambda_r \in \mathbb{R}^+$ and the link capacities will take the form $c_l(m) = \nu_l(m)\delta\gamma_l\sigma$ for some scalar $\gamma_l \in \mathbb{R}^+$.

We will interpret γ_l as the scaling parameter for link l and demonstrate the existence of three distinct scaling regimes parameterized by γ_l.

To motivate this interpretation we will first study the case of a single stream being transmitted over a single bottleneck link in Subsection 3.5.1, then consider the case of multiple streams on a single bottleneck link in Subsection 3.5.2, and finally introduce the full network scaling in Subsection 3.5.3.

3.3.1 Single stream, single link

Consider the trivial case of a single stream being served on a single link. Let the stream be typical, i.e., $S = \sigma$, $D = \delta$, $A = \alpha$. Thus the stream has a minimum subscription level of $\alpha\sigma$ and a maximum subscription level of σ. We parameterize the capacity of the link as $c = \gamma\sigma$.

Imagine we provision capacity on the link so that the capacity is insufficient to serve the stream at even its minimum subscription level. This implies $\gamma < \alpha$

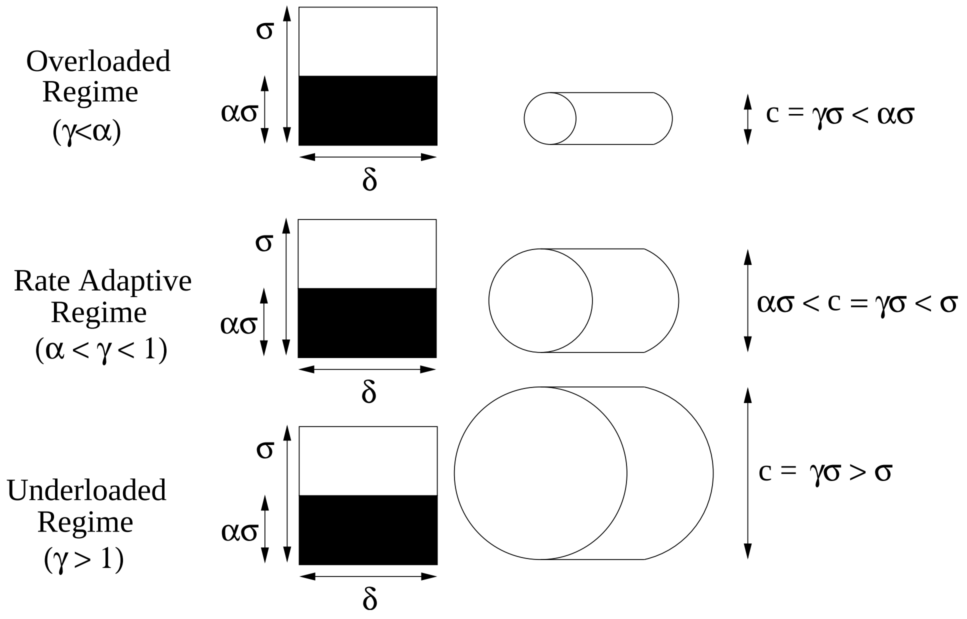

Figure 3.5: Illustration of three network scaling regimes parameterized by γ.

and we term this the overloaded regime. Imagine instead we provision capacity so that the capacity exceeds that required to serve the stream at its maximum subscription level. This means $\gamma > 1$ and we term this the underloaded regime. Finally, consider the case where the capacity exceeds the minimum subscription level but is inadequate to serve the stream at its maximum subscription level. This means $\alpha \leq \gamma \leq 1$ and we term this the rate adaptive regime. Figure 3.5 illustrates these three regimes.

3.3.2 Multiple streams, single link

The previous section illustrated how the capacity scaling parameter γ parameterizes three distinct capacity scaling regimes for the case of a single stream on a single link. Our network scaling sets the link capacity as a function of the expected number of streams traversing the link.

For this section we can drop the link and route subscripts so that the arrival rate is λ and the link capacity is c. Let $N(t)$ denote the number of active streams traversing the bottleneck link at a stationary time t, i.e., a time such that $N(t)$ is distributed according to its stationary distribution. Little's law states the average number of streams on the network, $\mathbb{E}[N(t)]$, is the product of the average stream arrival rate times the average stream duration, i.e., $\mathbb{E}[N(t)] = \lambda\delta$, assuming the blocking probability is low. We can write $N(m,t)$ for the number of streams at time t on the m^{th} link, i.e., when the arrival rate is $\lambda(m)$ and the link capacity is $c(m)$. As before, we have $\mathbb{E}[N(m,t)] = \lambda(m)\delta$.

We denote the average minimum link load, $\underline{\rho}(m)$, as the product of the average number of streams, $\lambda(m)\delta$, and the average minimum subscription level, $\alpha\sigma$, i.e., $\underline{\rho}(m) = \lambda(m)\delta\alpha\sigma$. Similarly, we denote the average maximum link load, $\bar{\rho}(m)$, as the product of the average number of streams and the average maximum subscription level σ, i.e., $\bar{\rho}(m) = \lambda(m)\delta\sigma$.

For the case when all streams are assumed to traverse at most one bottleneck link, the proposed scaling states that the arrival rate on the m^{th} link is $\lambda(m) = m\lambda$, and that the link capacity is $c(m) = (\lambda(m)\delta)(\gamma\sigma)$. Thus, the capacity is simply the product of the average number of streams on the link, $\lambda(m)\delta$, times some portion of the average maximum subscription level, $\gamma\sigma$.

The same three regimes parameterized in the previous section apply to this case as well. The overloaded regime is parameterized by $\gamma < \alpha$, which implies $c(m) < \underline{\rho}(m)$, i.e., the overloaded regime corresponds to the case where the provisioned capacity is insufficient to handle the average minimum load. The underloaded regime is parameterized by $\gamma > 1$, which implies $c(m) > \bar{\rho}(m)$, i.e., the underloaded regime corresponds to the case where the provisioned capacity is more than sufficient to handle the average maximum load. Finally, the rate adaptive regime, $\alpha \leq \gamma \leq 1$ corresponds to the case where $\underline{\rho}(m) \leq c(m) \leq \bar{\rho}(m)$, i.e., the provisioned capacity lies between the average minimum load and the average maximum load. As we will show in the sequel, it is the rate adaptive regime that is of primary interest.

3.3.3 General network

The general network scaling consists of defining scaling parameters on each link, i.e., $\boldsymbol{\gamma}(m) = (\gamma_l(m), l \in \mathcal{L})$. We then scale the arrival rate on each route according to $\lambda_r(m) = m\lambda_r$ and the capacity on each link according to $c_l(m) = \nu_l(m)\delta\gamma_l\sigma$, where $\nu_l(m) = \sum_{r \ni l} \lambda_r(m)$ is the aggregate mean arrival rate on link l in network m.

We can define a vector of average minimum loads, $\underline{\boldsymbol{\rho}}(m) = (\underline{\rho}_l(m), l \in \mathcal{L})$, where $\underline{\rho}_l(m) = \nu_l(m)\delta\alpha\sigma$, and a vector of average maximum loads, $\bar{\boldsymbol{\rho}}(m) = (\bar{\rho}_l(m), l \in \mathcal{L})$, where $\bar{\rho}_l(m) = \nu_l(m)\delta\sigma$. Just as in the previous section, we can say link l is in the overloaded regime if $\gamma_l < \alpha$, in the underloaded regime if $\gamma_l > 1$, and in the rate adaptive regime if $\alpha \leq \gamma_l \leq 1$. Again, these three regimes correspond to $c_l(m) < \underline{\rho}_l(m)$, $c_l(m) > \bar{\rho}_l(m)$, and $\underline{\rho}_l(m) \leq c_l(m) \leq \bar{\rho}_l(m)$ respectively.

3.4 Sub-optimal adaptation policies

In this section we introduce a sub-optimal feasible adaptation policy, which we term the fair-share adaptation policy, denoted π_f. Our purpose in introducing sub-optimal policies is to provide a baseline with which to compare the asymptotic QoS under the optimal adaptation policy. We will define the fair-share policy for a single link. This is sufficient for our purposes because we will only be able to obtain closed form expressions for asymptotic optimal QoS for a single link. Because the policy is only defined on a single link, we drop link and route subscripts for this section. Thus, the number of active streams is denoted $N(t)$ and the link capacity is denoted c.

3.4.1 Fair-share adaptation policy

The fair share adaptation policy for a single link uses the following subscription level assignment rule:

$$S_i^{\pi_f}(t) = \begin{cases} A_i S_i, & N(t) > \frac{c}{\alpha\sigma} \\ A_i S_i (1 - X(t)) + S_i X(t), & \frac{c}{\sigma} \leq N(t) \leq \frac{c}{\alpha\sigma} \\ S_i, & N(t) < \frac{c}{\sigma} \end{cases} \tag{3.10}$$

where

$$X(t) = \frac{\frac{c}{N(t)} - \alpha\sigma}{\sigma - \alpha\sigma}. \tag{3.11}$$

Note that $\frac{c}{\sigma} \leq N(t) \leq \frac{c}{\alpha\sigma}$ implies $X(t) \in [0, 1]$.

We term this policy a fair-share policy because all streams receive a subscription level that yields a constant proportional improvement above their minimum subscription level. That is, $\frac{S_i^{\pi_f}(t) - A_i S_i}{S_i - A_i S_i}$ is a constant. To see this, consider each of the three cases above. When $N(t) > \frac{c}{\alpha\sigma}$ we have a proportional improvement of zero since all streams receive their minimum subscription level. When $N(t) < \frac{c}{\sigma}$ we have a proportional improvement of one since all streams receive their maximum subscription level. When $\frac{c}{\sigma} \leq N(t) \leq \frac{c}{\alpha\sigma}$ we see the proportional improvement is

$$\frac{S_i^{\pi_f}(t) - A_i S_i}{S_i - A_i S_i} = \frac{A_i S_i (1 - X(t)) + S_i X(t) - A_i S_i}{S_i - A_i S_i} = X(t), \tag{3.12}$$

which is constant for a given $N(t)$.

We can interpret this policy as follows. Streams are assigned their minimum subscription level, $A_i S_i$, when the bandwidth per stream falls below the average minimum subscription level, i.e., when $\frac{c}{N(t)} < \alpha\sigma$. Streams are assigned their maximum subscription level, S_i, when the bandwidth per stream exceeds the average maximum subscription level, i.e., when $\frac{c}{N(t)} > \sigma$. When the bandwidth per stream lies between the average minimum subscription level and the average maximum subscription level, i.e., when $\alpha\sigma \leq \frac{c}{N(t)} \leq \sigma$, we assign streams a subscription level proportional to $X(t)$. Note that $X(t)$ is the relative difference between the actual fair share bandwidth, $\frac{c}{N(t)}$, and the average minimum subscription level $\alpha\sigma$. We say this difference is relative because it is normalized by the maximum difference, i.e., the difference between the average maximum subscription level, σ, and the average minimum subscription level, $\alpha\sigma$.

Finally, note that for the special case when all streams have the same minimum and maximum subscription levels, i.e., $A_i = \alpha, a.s.$ and $S_i = \sigma, a.s.$, the fair share policy becomes

$$S_i^{\pi_f}(t) = \begin{cases} \alpha\sigma, & \frac{c}{N(t)} < \alpha\sigma \\ \frac{c}{N(t)}, & \alpha\sigma \le \frac{c}{N(t)} \le \sigma \\ \sigma, & \frac{c}{N(t)} > \sigma \end{cases} . \qquad (3.13)$$

The above equation is clearly a fair share policy, thus we can interpret (3.10) as the appropriate generalization of (3.13) for streams with heterogeneous minimum and maximum subscription levels.

3.4.2 Feasibility of the fair-share adaptation policy

It bears comment that this policy is not necessarily feasible. Specifically, the policy as stated might feasibly violate the link capacity constraint or the stream constraint. Our purpose in studying this policy is to demonstrate the relative improvement in asymptotic QoS obtained under the optimal adaptation policy. Thus the QoS bounds obtained under this policy are optimistic in that enforcing the feasibility constraints can only reduce the asymptotic QoS. Thus our relative improvement in asymptotic QoS obtained under the optimal adaptation policy will be conservative.

The stream constraint requires that subscription level assignments correspond to encodings offered by the content provider, i.e., $S(t) \in \mathcal{S}$. We assume for purposes of this section that the content provider will dynamically adjust the encoding of the stream to whatever rate is required under the fair share policy. That is, we relax the assumption that $\mathcal{S}$ is a fixed discrete set, and assume a continuum of encodings are available, i.e., $\mathcal{S} = [AS, S]$.

The link constraint requires that the aggregate subscription assignments not exceed the link capacity, i.e., $\sum_{i=1}^{N(t)} S_i^{\pi_f}(t) \leq c$. Although the fair share policy does not ensure this constraint is satisfied, we can show that, under the link scaling, the asymptotic link utilization under the fair share policy is one, unless the system is over-provisioned. This is shown in the following lemma.

Lemma 1 *Under the link scaling, the asymptotic utilization for the fair share policy is*

$$\lim_{m\to\infty} \mathbb{E}\Big[\frac{1}{c(m)} \sum_{i=1}^{N(m,t)} S_i^{m,\pi_f}(t)\Big] = \begin{cases} 1, & \gamma \leq 1 \\ \frac{1}{\gamma}, & \gamma > 1 \end{cases} \tag{3.14}$$

where $c(m) = (\lambda(m)\delta)(\gamma\sigma)$ and $\lambda(m) = m\lambda$.

Proof of Lemma 1 We denote the aggregate load in the m^{th} link at time t under the fair share policy by $Y(m,t) = \sum_{i=1}^{N(m,t)} S_i^{m,\pi_f}(t)$. We can condition on $N(m,t)$ to obtain $\lim_{m\to\infty} \mathbb{E}[\frac{Y(m,t)}{c(m)}]$

$$\begin{aligned} &= \lim_{m\to\infty} \Big[\mathbb{E}[\frac{Y(m,t)}{c(m)} \mid N(m,t) > \frac{c(m)}{\alpha\sigma}]\mathbb{P}(N(m,t) > \frac{c(m)}{\alpha\sigma}) \qquad (3.15) \\ &+ \mathbb{E}[\frac{Y(m,t)}{c(m)} \mid \frac{c(m)}{\sigma} \leq N(m,t) \leq \frac{c(m)}{\alpha\sigma}]\mathbb{P}(\frac{c(m)}{\sigma} \leq N(m,t) \leq \frac{c(m)}{\alpha\sigma}) \\ &+ \mathbb{E}[\frac{Y(m,t)}{c(m)} \mid N(m,t) < \frac{c(m)}{\sigma}]\mathbb{P}(N(m,t) < \frac{c(m)}{\sigma})\Big]. \end{aligned}$$

We will refer to the three terms in the sum in the above equation as the overloaded regime term, the rate adaptive regime term, and the underloaded regime term. We consider three separate cases: $i)\gamma < \alpha$, $ii)\alpha \leq \gamma \leq 1$, $iii)\gamma > 1$.

Consider the first case, $\gamma < \alpha$, which corresponds to scaling the link in the overloaded regime. We will show that, for arbitrarily small $h > 0$,

$$\lim_{m\to\infty} \mathbb{P}(N(m,t) > \frac{c(m)}{\alpha\sigma}(1-h)) = 1, \ \gamma < \alpha,$$

i.e., when the link is provisioned into the overloaded regime, the asymptotic allocation under the fair share adaptation policy gives all streams their minimum subscription level. A little thought shows that, for $\epsilon(m) = \mathbb{E}[N(m,t)] - \frac{c(m)}{\alpha\sigma}(1-h)$,

$$\begin{aligned}
\lim_{m\to\infty} \mathbb{P}(N(m,t) > \frac{c(m)}{\alpha\sigma}(1-h)) &= 1 - \lim_{m\to\infty} \mathbb{P}(N(m,t) \leq \frac{c(m)}{\alpha\sigma}(1-h)) \\
&\geq 1 - \lim_{m\to\infty} \mathbb{P}(|N(m,t) - \mathbb{E}[N(m,t)]| \geq \epsilon(m)) \\
&\geq 1 - \lim_{m\to\infty} \frac{Var(N(m,t))}{\epsilon(m)^2},
\end{aligned}$$

where we have used Chebychev's inequality in the last step. Now consider $\epsilon(m)$. By Little's Law, the mean number of streams in m^{th} link is $\mathbb{E}[N(m,t)] = \lambda(m)(1-B(m))\delta$, where $B(m)$ denotes the blocking probability on link m. Using the link scaling, we have $\frac{c(m)}{\alpha\sigma}(1-h) = \frac{\lambda(m)\delta\gamma\sigma}{\alpha\sigma}(1-h)$. We can therefore write

$$\begin{aligned}
\epsilon(m) &= \mathbb{E}[N(m,t)] - \frac{c(m)}{\alpha\sigma}(1-h) \\
&= \lambda(m)(1-B(m))\delta - \frac{\lambda(m)\delta\gamma\sigma}{\alpha\sigma}(1-h) \\
&= \lambda(m)\delta\Big((1-B(m)) - \frac{\gamma}{\alpha}(1-h)\Big).
\end{aligned}$$

The variance $Var(N(m,t))$ can be bounded above by $\lambda(m)\delta$ as follows. If there were no capacity constraint then $N(m,t)$ would be a Poisson random variable with mean and variance equal to $\lambda(m)\delta$. The capacity constraint reduces the mean to $\lambda(m)(1-B(m))$ and reduces the variance as well. Thus, $\lambda(m)\delta$ is a valid upper bound. Finally, we claim the asymptotic blocking probability in the overloaded regime is $\lim_{m\to\infty} B(m) = 1 - \frac{\gamma}{\alpha}$. We argue loosely as follows. The minimum load arriving to the link is $\underline{\rho}(m) = \lambda(m)\delta\alpha\sigma$ and the link capacity is $c(m) = \lambda(m)\delta\gamma\sigma$. For $\gamma < \alpha$ we have $c(m) < \underline{\rho}(m)$ and hence the average fraction of load blocked by the link is $\frac{\underline{\rho}(m)-c(m)}{\underline{\rho}(m)} = 1-\frac{\gamma}{\alpha}$. Applying these observations allows

$$\begin{aligned}
\lim_{m\to\infty} \mathbb{P}(N(m,t) > \frac{c(m)}{\alpha\sigma}(1-h)) &\geq 1 - \lim_{m\to\infty} \frac{Var(N(m,t)}{\epsilon(m)^2} \\
&\geq 1 - \lim_{m\to\infty} \frac{\lambda(m)\delta}{(\lambda(m)\delta)^2((1-B(m)) - \frac{\gamma}{\alpha}(1-h))} \\
&= 1 - \lim_{m\to\infty} \frac{1}{(\lambda(m)\delta)((1-B(m)) - \frac{\gamma}{\alpha}(1-h))} \\
&= 1 - \lim_{m\to\infty} \frac{1}{\lambda(m)\delta(\frac{\gamma}{\alpha} - \frac{\gamma}{\alpha}(1-h))} = 1.
\end{aligned}$$

Thus, only the overloaded term in 3.15 is significant when $\gamma < \alpha$, and we can write

$$\begin{aligned}
\lim_{m\to\infty} \mathbb{E}[\frac{Y(m,t)}{c(m)}] &= \lim_{m\to\infty} \mathbb{E}[\frac{Y(m,t)}{c(m)} \mid N(m,t) > \frac{c(m)}{\alpha\sigma})\mathbb{P}(N(m,t) > \frac{c(m)}{\alpha\sigma}) \\
&= \lim_{m\to\infty} \mathbb{E}[\frac{Y(m,t)}{c(m)} \mid N(m,t) > \frac{c(m)}{\alpha\sigma}] \\
&= \lim_{m\to\infty} \mathbb{E}[\frac{1}{c(m)} \sum_{i=1}^{N(m,t)} S_i^{m,\pi_f}(t) \mid N(m,t) > \frac{c(m)}{\alpha\sigma}] \\
&= \lim_{m\to\infty} \mathbb{E}[\frac{1}{c(m)} \sum_{i=1}^{N(m,t)} A_i S_i].
\end{aligned}$$

Applying Wald's identity,

$$\begin{aligned}
\lim_{m\to\infty} \mathbb{E}[\frac{Y(m,t)}{c(m)}] &= \lim_{m\to\infty} \frac{1}{c(m)} \mathbb{E}[N(m,t)]\mathbb{E}[A_i S_i] \\
&= \lim_{m\to\infty} \frac{1}{c(m)} \lambda(m)(1-B(m))\delta\alpha\sigma \\
&= \lim_{m\to\infty} \frac{\lambda(m)(1-B(m))\alpha\sigma}{\lambda(m)\delta\gamma\sigma} \\
&= \lim_{m\to\infty} \frac{(1-B(m))\alpha}{\gamma} \\
&= \frac{(1-(1-\frac{\gamma}{\alpha}))\alpha}{\gamma} = 1.
\end{aligned}$$

Thus, the asymptotic utilization is one when $\gamma < \alpha$.

Consider the case when the link is provisioned in the rate adaptive regime, i.e., when $\alpha \leq \gamma \leq 1$. We will show that

$$\lim_{m\to\infty} \mathbb{P}(\frac{c(m)}{\sigma} \leq N(m,t) \leq \frac{c(m)}{\alpha\sigma}) = 1, \ \alpha \leq \gamma \leq 1,$$

i.e., when the link is provisioned in the rate-adaptive regime, the asymptotic fair share allocation always gives a proportional improvement of $\lim_{m\to\infty} X(m,t)$. We define $\epsilon(m) = \max\{\mathbb{E}[N(m,t)]-\frac{c(m)}{\sigma}, \frac{c(m)}{\alpha\sigma}-\mathbb{E}[N(m,t)]\}$. Recalling that the blocking probability in the rate adaptive regime goes to zero exponentially fast, we have $\mathbb{E}[N(m,t)] = \lambda(m)\delta$. Applying the rate adaptive scaling to $\epsilon(m)$ gives $\epsilon(m) = \lambda(m)\delta \max\{1-\gamma, \frac{\gamma}{\alpha}-1\}$. A little thought shows

$$\begin{aligned}\lim_{m\to\infty} \mathbb{P}(\frac{c(m)}{\sigma} \leq N(m,t) \leq \frac{c(m)}{\alpha\sigma}) &\geq 1 - \lim_{m\to\infty} \mathbb{P}(|N(m,t)-\mathbb{E}[N(m,t)]| > \epsilon(m)) \\ &\geq 1 - \lim_{m\to\infty} \frac{Var(N(m,t))}{\epsilon(m)^2},\end{aligned}$$

where we have again applied Chebychev's inequality in the last step. Recalling that, under the assumed low blocking regime, the variance of $N(m,t)$ is $\lambda(m)\delta$, we see that

$$\begin{aligned}\lim_{m\to\infty} \mathbb{P}(\frac{c(m)}{\sigma} \leq N(m,t) \leq \frac{c(m)}{\alpha\sigma}) &\geq 1 - \lim_{m\to\infty} \frac{\lambda(m)\delta}{(\lambda(m)\delta)^2(\max\{1-\gamma,\frac{\gamma}{\alpha}-1\})^2} \\ &\geq 1 - \lim_{m\to\infty} \frac{1}{(\lambda(m)\delta)(\max\{1-\gamma,\frac{\gamma}{\alpha}-1\})^2} = 1,\end{aligned}$$

Thus, only the rate adaptive term in 3.15 is significant when $\alpha \leq \gamma \leq 1$, and we can write

$$
\begin{aligned}
& \lim_{m\to\infty} \mathbb{E}[\frac{Y(m,t)}{c(m)}] \\
= & \lim_{m\to\infty} \mathbb{E}[\frac{Y(m,t)}{c(m)} \mid \frac{c(m)}{\sigma} \leq N(m,t) \leq \frac{c(m)}{\alpha\sigma}]\mathbb{P}(\frac{c(m)}{\sigma} \leq N(m,t) \leq \frac{c(m)}{\alpha\sigma}) \\
= & \lim_{m\to\infty} \mathbb{E}[\frac{Y(m,t)}{c(m)} \mid \frac{c(m)}{\sigma} \leq N(m,t) \leq \frac{c(m)}{\alpha\sigma}] \\
= & \lim_{m\to\infty} \mathbb{E}[\frac{1}{c(m)} \sum_{i=1}^{N(m,t)} S_i^{m,\pi_f}(t) \mid \frac{c(m)}{\sigma} \leq N(m,t) \leq \frac{c(m)}{\alpha\sigma}] \\
= & \lim_{m\to\infty} \mathbb{E}[\frac{1}{c(m)} \sum_{i=1}^{N(m,t)} A_i S_i(1 - X(m,t)) + S_i X(m,t)] \\
= & \lim_{m\to\infty} \frac{1}{c(m)} \mathbb{E}[\sum_{i=1}^{N(m,t)} A_i S_i + X(m,t) \sum_{i=1}^{N(m,t)} S_i(1-A_i)] \\
= & \Big[\lim_{m\to\infty} \frac{1}{c(m)} \mathbb{E}[\sum_{i=1}^{N(m,t)} A_i S_i]\Big] + \Big[\lim_{m\to\infty} \frac{1}{c(m)} \mathbb{E}[X(m,t) \sum_{i=1}^{N(m,t)} S_i(1-A_i)]\Big]
\end{aligned}
$$

Applying the definition of $X(m,t)$ gives

$$
\begin{aligned}
= & \Big[\lim_{m\to\infty} \frac{1}{c(m)} \mathbb{E}[\sum_{i=1}^{N(m,t)} A_i S_i]\Big] + \Big[\lim_{m\to\infty} \frac{1}{c(m)} \mathbb{E}[\frac{\frac{c(m)}{N(m,t)} - \alpha\sigma}{\sigma - \alpha\sigma} \sum_{i=1}^{N(m,t)} S_i(1-A_i)]\Big] \\
= & \Big[\lim_{m\to\infty} \frac{1}{c(m)} \mathbb{E}[\sum_{i=1}^{N(m,t)} A_i S_i]\Big] + \Big[\lim_{m\to\infty} \frac{1}{\sigma(1-\alpha)} \mathbb{E}[\frac{1}{N(m,t)} \sum_{i=1}^{N(m,t)} S_i(1-A_i)]\Big] \\
- & \Big[\lim_{m\to\infty} \frac{1}{c(m)} \frac{\alpha}{1-\alpha} \mathbb{E}[\sum_{i=1}^{N(m,t)} S_i(1-A_i)]\Big].
\end{aligned}
$$

Applying Wald's identity to the first and third terms, and the Law of Large Numbers to the second term yields:

$$\begin{aligned}
&= \Big[\lim_{m\to\infty} \frac{1}{c(m)}\mathbb{E}[N(m,t)]\mathbb{E}[AS]\Big] + \Big[\frac{1}{\sigma(1-\alpha)}\mathbb{E}[S(1-A)]\Big] \\
&- \Big[\lim_{m\to\infty} \frac{1}{c(m)}\frac{\alpha}{1-\alpha}\mathbb{E}[N(m,t)]\mathbb{E}[S(1-A)]\Big] \\
&= \Big[\lim_{m\to\infty} \frac{\lambda(m)\delta\alpha\sigma}{\lambda(m)\delta\gamma\sigma}\Big] + \Big[1\Big] - \Big[\lim_{m\to\infty} \frac{\lambda(m)\delta\sigma(1-\alpha)\alpha}{\lambda(m)\delta\gamma\sigma(1-\alpha)}\Big] \\
&= \Big[\frac{\alpha}{\gamma}\Big] + \Big[1\Big] - \Big[\frac{\alpha}{\gamma}\Big] = 1.
\end{aligned}$$

Thus, the asymptotic utilization is one when $\alpha \le \gamma \le 1$.

Consider the third case, $\gamma > 1$, which corresponds to scaling the link in the underloaded regime. We will show that

$$\lim_{m\to\infty} \mathbb{P}(N(m,t) < \frac{c(m)}{\sigma}) = 1,\ \gamma > 1,$$

i.e., when the link is provisioned in the underloaded regime, the asymptotic allocation under the fair share adaptation policy gives all streams their maximum subscription level. Define $\epsilon(m) = \frac{c(m)}{\sigma} - \mathbb{E}[N(m,t)]$. Noting that $\mathbb{E}[N(m,t)] = \lambda(m)\delta$ and $\frac{c(m)}{\sigma} = \lambda(m)\delta\gamma$ yields $\epsilon(m) = \lambda(m)\delta(\gamma - 1)$. A little thought shows

$$\begin{aligned}
\lim_{m\to\infty} \mathbb{P}(N(m,t) < \frac{c(m)}{\sigma}) &= 1 - \lim_{m\to\infty} \mathbb{P}(N(m,t) \ge \frac{c(m)}{\sigma}) \\
&\ge 1 - \lim_{m\to\infty} \mathbb{P}(|N(m,t) - \mathbb{E}[N(m,t)]| \ge \epsilon(m)) \\
&\ge 1 - \lim_{m\to\infty} \frac{Var(N(m,t))}{\epsilon(m)^2},
\end{aligned}$$

where we have used Chebychev's inequality in the last step. Noting that $Var(N(m,t)) = \lambda(m)\delta$ because of the low blocking probability for $\gamma > 1$, we see that

$$\begin{aligned}
\lim_{m\to\infty} \mathbb{P}(N(m,t) < \frac{c(m)}{\sigma}) &\ge 1 - \lim_{m\to\infty} \frac{\lambda(m)\delta}{(\lambda(m)\delta)^2(\gamma-1)^2} \\
&\ge 1 - \lim_{m\to\infty} \frac{1}{(\lambda(m)\delta)(\gamma-1)^2} = 1.
\end{aligned}$$

Thus, only the underloaded term in 3.15 is significant when $\gamma > 1$, and we can write

$$\begin{aligned}
\lim_{m\to\infty} \mathbb{E}[\frac{Y(m,t)}{c(m)}] &= \lim_{m\to\infty} \mathbb{E}[\frac{Y(m,t)}{c(m)} \mid N(m,t) < \frac{c(m)}{\sigma})\mathbb{P}(N(m,t) < \frac{c(m)}{\sigma}) \\
&= \lim_{m\to\infty} \mathbb{E}[\frac{Y(m,t)}{c(m)} \mid N(m,t) < \frac{c(m)}{\sigma}] \\
&= \lim_{m\to\infty} \mathbb{E}[\frac{1}{c(m)} \sum_{i=1}^{N(m,t)} S_i^{m,\pi_f}(t) \mid N(m,t) < \frac{c(m)}{\sigma}] \\
&= \lim_{m\to\infty} \mathbb{E}[\frac{1}{c(m)} \sum_{i=1}^{N(m,t)} S_i].
\end{aligned}$$

Applying Wald's identity,

$$\begin{aligned}
\lim_{m\to\infty} \mathbb{E}[\frac{Y(m,t)}{c(m)}] &= \lim_{m\to\infty} \frac{1}{c(m)} \mathbb{E}[N(m,t)]\mathbb{E}[S_i] \\
&= \lim_{m\to\infty} \frac{1}{c(m)} \lambda(m)\delta\sigma \\
&= \lim_{m\to\infty} \frac{\lambda(m)\delta\sigma}{\lambda(m)\delta\gamma\sigma} = \frac{1}{\gamma} < 1.
\end{aligned}$$

Thus, the asymptotic utilization is $\frac{1}{\gamma}$ when $\gamma > 1$. This is the best we can hope for since the system is over-provisioned. ■

3.4.3 Asymptotic QoS of the fair-share adaptation policy

We next investigate the asymptotic expected time-average normalized subscription level under the fair share adaptation policy. The following lemma shows that this quantity is linear in the scaling parameter when the link is provisioned in the rate adaptive regime. We write $\mathbb{E}^0[Q^{m,\pi_f}]$ to denote the time-average normalized subscription level in the m^{th} link, i.e., when the link arrival rate is $\lambda(m)$ and the link capacity is $c(m)$, under the fair-share adaptation policy for a typical client. The asymptotic quantity, denoted q^{γ,π_f}, is then given by $q^{\gamma,\pi_f} = \lim_{m\to\infty} \mathbb{E}^0[Q^{m,\pi_f}]$.

Lemma 2 *The asymptotic expected time-average normalized subscription level under the fair share adaptation policy is*

$$q^{\gamma,\pi_f} = \begin{cases} \alpha, & \gamma < \alpha \\ \gamma, & \alpha \le \gamma \le 1 \\ 1, & \gamma > 1 \end{cases} . \tag{3.16}$$

Proof of Lemma 2 Recall the definition of the expected time-average normalized subscription level:

$$\mathbb{E}^0[Q^{m,\pi_f}] = \mathbb{E}^0[\frac{1}{D}\int_0^D \frac{S^{m,\pi_f}(t)}{S}dt].$$

We condition on D and obtain

$$\begin{aligned} \mathbb{E}^0[Q^{m,\pi_f}] &= \mathbb{E}^0[\frac{1}{D}\int_0^D \frac{S^{m,\pi_f}(t)}{S}dt] \\ &= \int_0^\infty \mathbb{E}^0[\frac{1}{D}\int_0^D \frac{S^{m,\pi_f}(t)}{S}dt \mid D = d]dF_D(d) \\ &= \int_0^\infty \mathbb{E}^0[\frac{1}{d}\int_0^d \frac{S^{m,\pi_f}(t)}{S}dt \mid D = d]dF_D(d) \\ &= \int_0^\infty \frac{1}{d}\int_0^d \mathbb{E}^0[\frac{S^{m,\pi_f}(t)}{S} \mid D = d]dtdF_D(d) \\ &= \int_0^\infty \frac{1}{d}\int_0^d \mathbb{E}^0[\frac{S^{m,\pi_f}(t)}{S}]dtdF_D(d) \\ &= \mathbb{E}^0[\frac{S^{m,\pi_f}(t)}{S}] \end{aligned}$$

where the t in the last equation is to be understood as a typical time. The above development is simply a proof of the intuitive idea that the expected time-average normalized subscription level under the fair-share adaptation policy is simply the expected normalized subscription level at a typical time. This follows because the subscription assignment under the fair share policy is independent of the stream duration.

Analogously to (3.15), we condition on $N(m,t)$ to obtain $\lim_{m\to\infty}\mathbb{E}[\frac{S^{m,\pi_f}(t)}{S}]$

$$\begin{aligned} &= \lim_{m\to\infty}\Big[\mathbb{E}[\frac{S^{m,\pi_f}(t)}{S} \mid N(m,t) > \frac{c(m)}{\alpha\sigma})\mathbb{P}(N(m,t) > \frac{c(m)}{\alpha\sigma}) \qquad (3.17)\\ &+ \mathbb{E}[\frac{S^{m,\pi_f}(t)}{S} \mid \frac{c(m)}{\sigma} \le N(m,t) \le \frac{c(m)}{\alpha\sigma}]\mathbb{P}(\frac{c(m)}{\sigma} \le N(m,t) \le \frac{c(m)}{\alpha\sigma})\\ &+ \mathbb{E}[\frac{S^{m,\pi_f}(t)}{S} \mid N(m,t) < \frac{c(m)}{\sigma}]\mathbb{P}(N(m,t) < \frac{c(m)}{\sigma})\Big]. \end{aligned}$$

As in Lemma 1, we will refer to the three terms in the sum in the above equation as the overloaded regime term, the rate adaptive regime term, and the underloaded regime term. We consider three separate cases: $i)\gamma < \alpha$, $ii)\alpha \le \gamma \le 1$, $iii)\gamma > 1$.

Consider the first case, $\gamma < \alpha$, which corresponds to scaling the link in the overloaded regime. We proved in Lemma 1 that, for arbitrarily small $h > 0$,

$$\lim_{m\to\infty}\mathbb{P}(N(m,t) > \frac{c(m)}{\alpha\sigma}(1-h)) = 1,\ \gamma < \alpha,$$

i.e., when the link is provisioned in the overloaded regime, the asymptotic allocation under the fair share adaptation policy gives each stream its minimum subscription level. Thus only the overloaded regime term in (3.17) is significant and we can write

$$\begin{aligned} q^{\gamma,\pi_f} &= \lim_{m\to\infty}\mathbb{E}^0[Q^{m,\pi_f}]\\ &= \lim_{m\to\infty}\mathbb{E}^0[\frac{S^{m,\pi_f}(t)}{S} \mid N(m,t) > \frac{c(m)}{\alpha\sigma}]\\ &= \lim_{m\to\infty}\mathbb{E}[\frac{AS}{S}] = \alpha. \end{aligned}$$

Thus, when the link is provisioned in the overloaded regime, the client average asymptotic normalized subscription level is α, the average normalized minimum subscription level. This confirms the intuition that a feasible allocation policy will typically grant streams their minimum subscription level when the system is overloaded.

Consider the second case, $\alpha \leq \gamma \leq 1$, which corresponds to scaling the link in the rate adaptive regime. We proved in Lemma 1 that

$$\lim_{m\to\infty} \mathbb{P}(\frac{c(m)}{\sigma} \leq N(m,t) \leq \frac{c(m)}{\alpha\sigma}) = 1,\ \alpha \leq \gamma \leq 1,$$

i.e., when the link is provisioned in the rate adaptive regime, the asymptotic fair share allocation always gives a proportional improvement of $\lim_{m\to\infty} X(m,t)$. Thus only the rate adaptive term in (3.17) is significant and we can write

$$\begin{aligned}
q^{\gamma,\pi_f} &= \lim_{m\to\infty} \mathbb{E}^0[Q^{m,\pi_f}] \\
&= \lim_{m\to\infty} \mathbb{E}^0[\frac{S^{m,\pi_f}(t)}{S} \mid \frac{c(m)}{\sigma} \leq N(m,t) \leq \frac{c(m)}{\alpha\sigma}] \\
&= \lim_{m\to\infty} \mathbb{E}[A(1-X(m,t)) + X(m,t)] \\
&= \lim_{m\to\infty} \Big[\mathbb{E}[A] + \mathbb{E}[(1-A)X(m,t)]\Big] \\
&= \alpha + (1-\alpha) \lim_{m\to\infty} \mathbb{E}[X(m,t)] \\
&= \alpha + (1-\alpha) \lim_{m\to\infty} \mathbb{E}[\frac{\frac{c(m)}{N(m,t)} - \alpha\sigma}{\sigma - \alpha\sigma}] \\
&= \alpha + (1-\alpha) \lim_{m\to\infty} \frac{c(m)}{\sigma(1-\alpha)} \mathbb{E}[\frac{1}{N(m,t)}] - \alpha \\
&= \lim_{m\to\infty} \frac{c(m)}{\sigma} \mathbb{E}[\frac{1}{N(m,t)}].
\end{aligned}$$

Recall that $N(m,t)$ is Poisson. One can show that

$$\lim_{m\to\infty} \mathbb{E}[\frac{1}{N(m,t)}] - \frac{1}{\mathbb{E}[N(m,t)]} = 0.$$

Applying these results yields

$$\begin{aligned}
q^{\gamma,\pi_f} &= \lim_{m\to\infty} \frac{c(m)}{\sigma} \frac{1}{\mathbb{E}[N(m,t)]} \\
&= \lim_{m\to\infty} \frac{\lambda(m)\delta\gamma\sigma}{\sigma\lambda(m)\delta} = \gamma.
\end{aligned}$$

Thus, when the link is provisioned in the rate adaptive regime, the asymptotic fair share allocation yields a client average asymptotic normalized subscription level equal to the capacity scaling parameter.

Consider the third case, $\gamma > 1$, which corresponds to scaling the link in the underloaded regime. We proved in Lemma 1 that

$$\lim_{m\to\infty} \mathbb{P}(N(m,t) < \frac{c(m)}{\sigma}) = 1,\ \gamma > 1,$$

i.e., when the link is provisioned in the underloaded regime, the asymptotic allocation under the fair share adaptation policy gives each stream its maximum subscription level. Thus only the underloaded regime term in (3.17) is significant and we can write

$$\begin{aligned} q^{\gamma,\pi_f} &= \lim_{m\to\infty} \mathbb{E}^0[Q^{m,\pi_f}] \\ &= \lim_{m\to\infty} \mathbb{E}^0[\frac{S^{m,\pi_f}(t)}{S} \mid N(m,t) < \frac{c(m)}{\sigma}] \\ &= \lim_{m\to\infty} \mathbb{E}[\frac{S}{S}] = 1. \end{aligned}$$

Thus, when the link is provisioned in the underloaded regime, the client average asymptotic normalized subscription level is 1, the average normalized maximum subscription level. This confirms the intuition that a feasible allocation policy will typically grant streams their maximum subscription level when the system is underloaded.

■

The important fact about the fair-share adaptation policy is that the asymptotic expected normalized subscription level equals the scaling parameter when the link capacity is scaled in the rate adaptive regime. We will use this result to evaluate the improvement in asymptotic expected normalized subscription level under the optimal adaptation policy.

3.5 Optimal adaptation policy

Having introduced a sub-optimal adaptation policy, we now turn our attention to identifying the optimal adaptation policy. The optimal adaptation policy is defined as the adaptation policy which maximizes the expected time-average normalized subscription level over all feasible adaptation policies. Throughout this chapter we will frequently use the term quality of service to refer to the expected time average normalized subscription level, but, as described in Chapter 2, our definition of quality of service actually also includes the rate of adaptation and the blocking probability. This terminology is adapted solely for the sake of brevity, but it does emphasize the fact that the normalized subscription level is our *primary* metric. In the sequel we will demonstrate that, for large capacity networks, the rate of adaptation under the optimal policy is zero, and, when the network is not provisioned in the overloaded regime, so too is the blocking probability. Thus the normalized subscription level is the only asymptotically non-trivial metric.

We will investigate two different models: a model for stored media and a model for live media. The stream duration for stored media is typically known a priori, and hence its stream volume is known. For live media, however, the stream duration may well be unknown, and hence the stream volume is unknown. We will identify the optimal adaptation policies for each of these models. The optimal adaptation policy for stored media will be denoted π_k, to reflect the fact that the stream duration is *known*, and the optimal adaptation policy for live media will be denoted π_u, to reflect the fact that the stream duration is *unknown*.

We study the optimal adaptation policy for stored media on a general network in Subsection 3.5.1 and specialize the result to a single link in Subsection 3.5.2. We study the optimal adaptation policy for live media on a general network in Subsection 3.5.3 and specialize the result to a single link in Subsection 3.5.4. Subsection 3.5.5 will present some simulation results comparing the Quality of Service under the two policies.

3.5.1 Optimal adaptation policy for stored media, general network

We define the *instantaneous QoS of a stream*, $Q_{i,r}(t)$, as

$$Q_{i,r}(t) = \begin{cases} \frac{S_{i,r}(t)}{S_{i,r}D_{i,r}}, & B_{i,r} \le t \le B_{i,r} + D_{i,r} \\ 0, & \text{otherwise} \end{cases}, \tag{3.18}$$

for $B_{i,r}$ the arrival time of stream (i, r). We define the *instantaneous aggregate QoS*, $Q_{agg}(t)$, as the sum of the instantaneous QoS for all active streams, i.e.,

$$Q_{agg}(t) = \sum_{r \in \mathcal{R}} \sum_{i=1}^{N_r(t)} Q_{i,r}(t). \tag{3.19}$$

The following lemma demonstrates that maximizing the expected time average normalized subscription level, $\mathbb{E}^0[Q]$, is equivalent to maximizing the expected instantaneous aggregate QoS at a typical time, $\mathbb{E}[Q_{agg}(t)]$.

Lemma 3 *The expected time-average normalized subscription level, $\mathbb{E}^0[Q]$, is proportional to the expected instantaneous aggregate QoS at a typical time, $\mathbb{E}[Q_{agg}(t)]$. In particular,*

$$\mathbb{E}^0[Q] = \frac{1}{\lambda^a}\mathbb{E}[Q_{agg}(t)], \tag{3.20}$$

where $\lambda^a = \sum_{r \in \mathcal{R}} \lambda_r^a$ is the aggregate rate at which streams are admitted *into the network, and λ_r^a is the rate of admissions on route r.*

Proof of Lemma 3.

We define the *time-average instantaneous aggregate QoS* as

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

where the second equality follows by ergodicity, and $\mathbb{E}[\cdot]$ denotes expectation with respect to the stationary distribution. Similarly, we can define the *client average QoS on route* r as

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

where we again have the second equality by ergodicity, and $\mathbb{E}^0[\cdot]$ denotes expectation with respect to a client average. Note that *i*) the admission policy is independent of the adaptation policy, and *ii*) in terms of admission, the system is a stochastic knapsack with continuous sizes [27]. Thus there is a blocking probability on each route, B_r, and the rate at which streams are admitted into the system on route r is $\lambda_r^a = \lambda_r(1 - B_r)$. Let the expected QoS of a typical stream be

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

where $\lambda^a = \sum_{r \in \mathcal{R}} \lambda_r^a$. This can be thought of as choosing a random client by conditioning on the probability of choosing a client from a given route.

Brumelle's Theorem [32] relates the average cost per client, say $\mathbb{E}^0[X]$, to the average cost received per unit of time by the system, say $\mathbb{E}[Y(t)]$, by $\mathbb{E}[Y(t)] = \lambda \mathbb{E}^0[X]$, where λ is the rate at which new clients are admitted into the system. Thus Brumelle's theorem relates a client average quantity, $\mathbb{E}^0[X]$ to an instantaneous system quantity, $\mathbb{E}[Y(t)]$. Brumelle's Theorem is a generalization of Little's Law which relates a client average quantity, namely the average client delay, say $\mathbb{E}^0[D]$, to an instantaneous system quantity, namely, the average number of clients in the system, $\mathbb{E}[N(t)]$, by $\mathbb{E}[N(t)] = \lambda \mathbb{E}^0[D]$.

Applying Brumelle's Theorem to the client average QoS and the expected instantaneous aggregate QoS yields

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

■

The previous lemma proves that maximizing the client average QoS is equivalent to maximizing the expected instantaneous aggregate QoS at a stationary time t. The following theorem states the optimal adaptation policy for stored media on a general network.

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

$$\begin{aligned} \max_{\mathbf{s}(t)} \quad & q_{agg}(t) = \sum_{r\in\mathcal{R}} \sum_{i=1}^{n_r(t)} \frac{s_{i,r}(t)}{s_{i,r} d_{i,r}} \\ s.t. \quad & \sum_{r\ni l} \sum_{i=1}^{n_r(t)} s_{i,r}(t) \le c_l \; \forall l \in \mathcal{L}, \\ & s_{i,r}(t) \in \mathcal{S}_{i,r}, i = 1, ..., n_r(t), \forall r \in \mathcal{R}. \end{aligned} \tag{3.21}$$

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

$$\frac{q_{agg}^{\pi_k}(t) - q_{agg}^{\tilde{\pi}_k}(t)}{q_{agg}^{\pi_k}(t)} \le \frac{\kappa_k}{n(t)}, \tag{3.22}$$

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

Proof of Theorem 1.

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

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

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

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

over all feasible $\mathbf{s}(t) = (s_{i,r}(t), i = 1, ..., n_r(t), r \in \mathcal{R})$, where we can assume the random variables $N_r(t)$ and $D_{i,r}$ are known because they are in $\sigma(t)$. Feasible $\mathbf{s}(t)$

requires $s_{i,r}(t) \in \mathcal{S}_{i,r}$ and that the link capacity constraints be obeyed. This yields (3.21).

We next prove the existence of an allocation $\mathbf{s}^{\tilde{\pi}_k}(t)$ with $s_{i,r}^{\tilde{\pi}_k}(t) \in \{a_{i,r}s_{i,r},\ s_{i,r}\}$ that is nearly optimal, with a bound given by (3.22). We denote the value of the objective under an allocation $\mathbf{s}(t)$ as

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

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

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

The capacity constraints will be written

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

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

$$\begin{aligned} \mathbf{s}(t) \in \{\mathbf{as}, \mathbf{s}\} \quad &\Rightarrow \quad s_{i,r}(t) \in \{a_{i,r}s_{i,r}, s_{i,r}\}, \\ \mathbf{s}(t) \in \mathcal{S} \quad &\Rightarrow \quad s_{i,r}(t) \in \mathcal{S}_{i,r}, \\ \mathbf{s}(t) \in [\mathbf{as}, \mathbf{s}] \quad &\Rightarrow \quad a_{i,r}s_{i,r} \leq s_{i,r}(t) \leq s_{i,r}. \end{aligned}$$

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

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

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

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

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

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

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

Combining these observations yields

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

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

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

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

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

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

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

$\blacksquare$

The first part of the theorem demonstrates that each stream is weighted inversely by its volume $v_{i,r} = s_{i,r}d_{i,r}$, i.e., the product of its maximum subscription level and its duration. The intuition is that the system is able to maximize the client average QoS by granting higher QoS to clients consuming fewer network resources. The second part of the theorem illustrates the existence of a near-optimal allocation such that all streams use either their minimum or maximum subscription level. Thus, for networks supporting large numbers of streams we may achieve a close to optimal solution by only using the minimum and maximum subscription levels.

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

3.5.2 Optimal adaptation policy for stored media, single link

To gain some further intuition for the optimal adaptation policy, we will consider the special case of a single link. As before, we drop the link and route subscripts and let $N(t)$ denote the number of active streams traversing the link at time t, and c denote the link capacity. For the single link case we are able to write down the allocation $\mathbf{s}^{\tilde{\pi}_k}(t)$ from Theorem 1 in closed form.

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

$$s_i^{\tilde{\pi}_k}(t) = \begin{cases} s_i, i = 1, \ldots, \bar{n}-1 \\ a_i s_i, i = \bar{n}, \ldots, n(t) \end{cases} \tag{3.23}$$

where

$$\bar{n} = \max\Big\{ m \mid \sum_{i=1}^{m-1} s_i + \sum_{i=m}^{n(t)} a_i s_i \le c \Big\}. \tag{3.24}$$

Proof of Corollary 1. The integer program (3.21) for the case of a single bottleneck is

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

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

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

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

■

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

3.5.3 Optimal adaptation policy for live media, general network

In this section we assume stream durations are unknown at the time of stream admission. We denote the optimal adaptation policy under this assumption by π_u, and the approximate optimal adaptation policy by $\tilde{\pi}_u$. Although stream durations are unknown, we can use the current age of a stream at some time t to infer the stream volume. In particular, let $l_{i,r}(t) = t - b_{i,r}$ denote the age of stream (i, r) at time t, where $b_{i,r}$ is the arrival time of the stream. It turns out that the quantity we need to estimate for the optimal adaptation policy for unknown stream durations is $\mathbb{E}[\frac{1}{D} \mid D \geq l(t)]$. We define the *expected volume of a stream at time t, $v(t)$*, as $v(t)^{-1} = \frac{1}{s}\mathbb{E}[\frac{1}{D} \mid D > l(t)]$.

Theorem 2 *The adaptation policy π_u that maximizes $\mathbb{E}^0[Q]$ when stream durations are unknown is the instantaneous bandwidth allocation $\mathbf{s}^{\pi_u}(t)$ at each time t resulting from the solution of (3.21) with the quantity $\frac{1}{d_{i,r}}$ replaced with $\mathbb{E}[\frac{1}{D} \mid D > l_{i,r}(t)]$, where $l_{i,r}(t)$ is the current age of stream (i, r) at time t. There exists a feasible allocation $\mathbf{s}^{\tilde{\pi}_u}(t)$ with $s^{\tilde{\pi}_u}_{i,r}(t) \in \{a_{i,r}s_{i,r},\ s_{i,r}\}$ for all $i = 1, \ldots, n_r(t)$ and all $r \in \mathcal{R}$ such that the value of the objective under $\mathbf{s}^{\tilde{\pi}_u}(t)$ is nearly optimal. In particular,*

$$\frac{q^{\pi_u}_{agg}(t) - q^{\tilde{\pi}_u}_{agg}(t)}{q^{\pi_u}_{agg}(t)} \leq \frac{\kappa_u}{n(t)}, \tag{3.25}$$

for $\kappa_u < \infty$.

Proof of Theorem 2. The approach used to prove Theorem 1 applies here as well. The difference is that the filtration $\sigma(t)$ doesn't include the durations of the active streams. We can recover the current ages $l_{i,r}$ of the active streams from the arrival times $b_{i,r}$ as $\{l_{i,r} = t - b_{i,r} \mid b_{i,r} \leq t\}$. This yields

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

The same considerations on feasible $\mathbf{s}(t)$ apply here yielding the same equation as (3.21), with $\frac{1}{d_{i,r}}$ replaced by $\mathbb{E}[\frac{1}{D} \mid D > l_{i,r}]$. Obtaining the bound on (3.25) is similar to the proof for the bound on (3.22), yielding

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

■

For the case of unknown stream durations we see that streams are weighted according to their expected inverse volume at time t, i.e., $\frac{1}{s_{i,r}}\mathbb{E}[\frac{1}{D} \mid D > l_{i,r}(t)]$, as opposed to being weighted according to their inverse volume $\frac{1}{s_{i,r}d_{i,r}}$ as in the case for known stream durations.

3.5.4 Optimal adaptation policy for live media, single link

The solution to (3.21) when there is at most one bottleneck link per route and stream durations are unknown is to sort streams traversing a given bottleneck by *"expected"* volume.

Corollary 2 *Consider a bottleneck link traversed by $n(t)$ active streams, labeled in order of increasing expected volume $v_1(t)^{-1} > ... > v_n(t)^{-1}$. The allocation $\mathbf{s}^{\tilde{\pi}_u}(t)$ for the case of at most one bottleneck link per route and unknown stream durations is given by (3.23) and (3.24) with $\mathbf{s}^{\tilde{\pi}_u}(t))$ replacing $\mathbf{s}^{\tilde{\pi}_k}(t))$.*

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

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

3.5.5 Simulation results for adaptation policies

In this subsection we present some simulation results illustrating the quality of service under the fair share adaptation policy, the optimal adaptation policy for known stream durations, and the optimal adaptation policy for unknown stream durations.

Figure 3.6 is a histogram of the time-average normalized subscription level versus stream volume, and Figure 3.7 is a histogram of the rate of adaptation versus stream volume. Note that the x axis is on a logarithmic scale. The stream volume PDF, $f_V(v)$ is also plotted. Recall the mean volume is $\mathbb{E}[V] = 54$ Mb. Also recall the average adaptivity is $\mathbb{E}[A] = \alpha = \frac{1}{2}$, so the range of feasible expected time-average normalized subscription levels is $\mathbb{E}^0[Q] \in [\alpha, 1] = [\frac{1}{2}, 1]$. The figures are simulations of a single link with arrival rate $\lambda = 5.0$ and capacity $c = 202.5$ Mb/s. Simulations were run for about 15,000 clients. The average number of active streams is $\mathbb{E}[N(t)] = \lambda\delta = 5.0 \times 180 = 900$. The capacity corresponds to scaling the network with a capacity scaling parameter $\gamma = \frac{c}{\lambda\delta\sigma} = \frac{202.5}{5.0\times180\times0.3} = \frac{3}{4}$. The histograms use 100 bins, where each bin has a width corresponding to one percent of the volume distribution. That is, if bin j covers the interval $[v_j^{min}, v_j^{max}]$, then $\mathbb{P}(v_j^{min} \leq V \leq v_j^{max}) = 0.01$ for $j = 1, \ldots 100$. The histograms are therefore plots of $\mathbb{E}^0[Q \mid V = v]$ and $\mathbb{E}^0[R \mid V = v]$ versus v, respectively.

Several points are noteworthy. Consider first the fair share adaptation policy. As proved in Lemma 2, the asymptotic expected time-average normalized subscription level under the fair share policy equals the scaling parameter when the network is provisioned in the rate adaptive regime. The top plot illustrates that the normal-

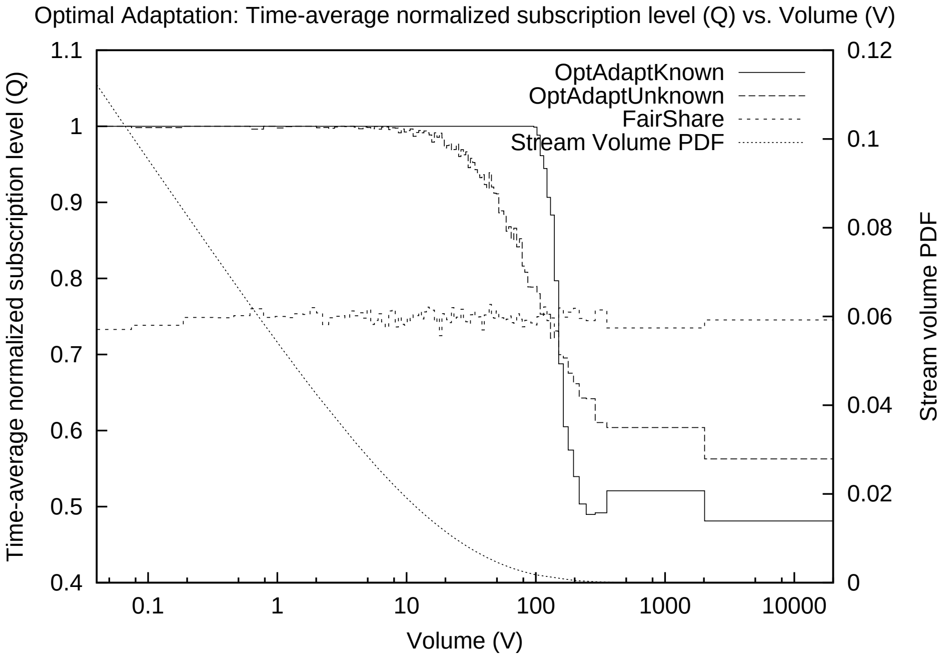

Figure 3.6: Histogram of $\mathbb{E}^0[Q \mid V = v]$ versus v for three adaptation policies.

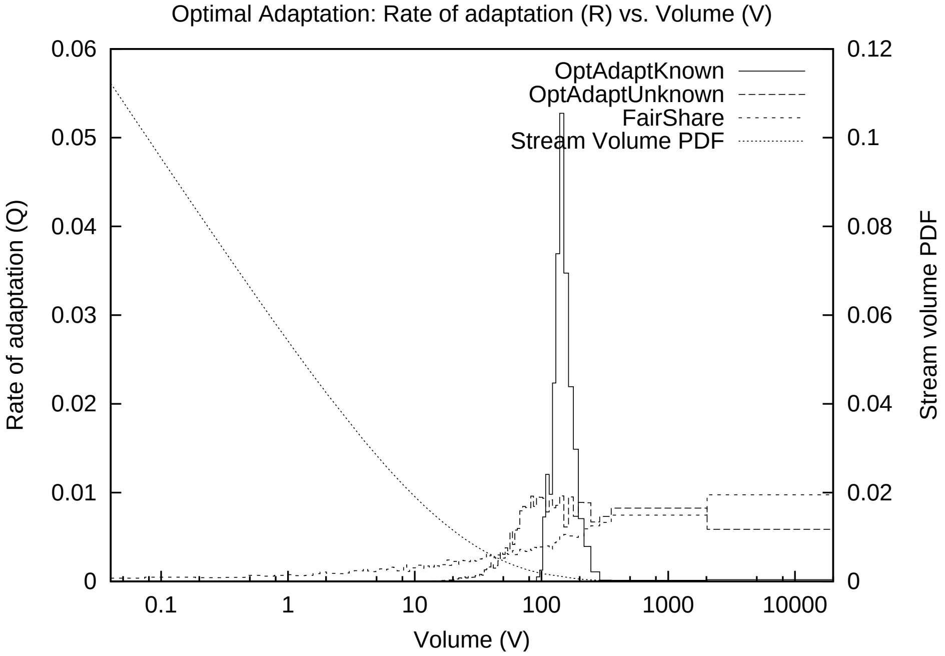

Figure 3.7: Histogram of $\mathbb{E}^0[R \mid V = v]$ versus v for three adaptation policies.

ized subscription level indeed equals the scaling parameter $\gamma = 0.75$. It also confirms that *all* clients, regardless of their stream volume, receive the same time-average normalized subscription level under the fair share policy. The rate of adaptation under the fair share policy is fairly low, but seems to be monotonically increasing in the stream volume.

Now consider the optimal adaptation policies. The top figure illustrates a fairly dramatic improvement in the time-average normalized subscription level under the optimal adaptation policy for almost all stream volumes. There is a sharp drop off around $v = 200$ Mb such that streams with volumes larger than this end up worse off than under the fair share policy. The volume CDF, however, shows that $P(V > 200) \approx 0.06$, thus it is only the largest 6% of the streams which are so penalized. So we see that roughly 94% of the clients receive an increase in QoS from 0.75 to 1 while 6% of the clients receive a decrease in QoS from 0.75 to 0.5. The top figure also illustrates that the optimal adaptation policies discriminate among streams according to volume, granting precedence to smaller volume streams over larger volume streams. The optimal adaptation policy for known stream durations achieves a very sharp volume discrimination between large and volume streams, while the optimal adaptation policy for unknown stream durations achieves a more gradual discrimination due to the fact that less information is available about the stream volumes.

Consider the rate of adaptation under the optimal adaptation policy for known stream durations. The plot illustrates that small and large volume streams achieve a near-zero rate of adaptation, meaning these streams receive a constant subscription level throughout their duration. The plot also illustrates that streams with volume near $v = 200$ experience a much higher rate of adaptation. Intuitively, if congestion arises, streams with these intermediate volumes are the first to be adapted to their minimum subscription level. Similarly, when capacity is available, the optimal adaptation policy chooses these streams with intermediate volumes as the first to be adapted to their maximum subscription level.

3.6 Asymptotic optimal QoS

In this section we apply the network scaling to the optimal adaptation policy. Subsection 3.6.1 applies the network scaling to the optimal adaptation policy on a general network, and Subsection 3.6.2 applies the network scaling to the optimal adaptation policy on a single link. Finally, Subsection 3.6.2 presents some computational results of the asymptotic expected time-average normalized subscription level under the optimal adaptation policy, for the case of known stream durations on a single link.

3.6.1 Asymptotic optimal QoS, general network

We apply the network scaling to the optimal adaptation policy for known stream durations on a general network, as presented in Subsection 3.5.1. The following theorem states that the asymptotic optimal adaptation policy is the solution, at each time t, of a nonlinear optimization problem over a vector of instantaneous volume thresholds $\mathbf{v}^{\gamma,\pi_k}(t) = (v_r^{\gamma,\pi_k}(t), r \in \mathcal{R})$. We write $\mathbf{s}^{\gamma,\pi_k}(t)$ for the asymptotic optimal instantaneous subscription level assignment at time t when the network is provisioned according to a network scaling vector $\gamma = (\gamma_l, l \in \mathcal{L})$. The instantaneous volume thresholds on each route function such that, at each time t, stream (i,r) with volume $v_{i,r} = s_{i,r} d_{i,r}$ receives a subscription level

$$s_{i,r}^{\gamma,\pi_k}(t) = \begin{cases} s_{i,r}, & v_{i,r} < v_r^{\gamma,\pi_k}(t) \\ a_{i,r} s_{i,r}, & \text{else} \end{cases} . \tag{3.26}$$

If a stream's volume is less than the instantaneous volume threshold for its route, then its subscription level assignment at time t is its maximum subscription level.

Similarly, if its volume exceeds the instantaneous volume threshold for its route, then its subscription level assignment at time t is its minimum subscription level.

Theorem 3 *Under the network scaling, the asymptotic optimal adaptation policy for known stream durations on a general network is such that the asymptotic optimal instantaneous subscription level assignment is of the form given in (3.26). At a typical time, t, the expected optimal instantaneous volume threshold vector, $\mathbf{v}^{\gamma,\pi_k}(t)$, is the solution of the following nonlinear program:*

$$\begin{aligned} \max_{\mathbf{v}(t)} \quad & \sum_{r\in\mathcal{R}} \lambda_r F_V(v_r(t)) + \alpha\lambda_r \bar{F}_V(v_r(t)) \\ s.t. \quad & \sum_{r\ni l} \lambda_r F_{\hat{V}}(v_r(t)) + \alpha\lambda_r \bar{F}_{\hat{V}}(v_r(t)) \leq \gamma_l \nu_l, \forall l \in \mathcal{L}. \end{aligned} \tag{3.27}$$

Moreover, the above constraints comprise a convex set and the objective is a concave function, ensuring the existence of a unique optimum.

Proof of Theorem 3.

Under the network scaling, the optimization problem (3.21) for the m^{th} network is

$$\begin{aligned} \max_{\mathbf{s}(m,t)} \quad & G^m(\mathbf{s}(m,t)) = \sum_{r\in\mathcal{R}} \sum_{i=1}^{n_r(m,t)} \frac{s_{i,r}(m,t)}{s_{i,r}d_{i,r}} \\ s.t. \quad & \sum_{r\ni l} \sum_{i=1}^{n_r(m,t)} s_{i,r}(m,t) \leq c_l(m) \; \forall l \in \mathcal{L}, \\ & s_{i,r}(m,t) \in \mathcal{S}_{i,r}, i = 1, ..., n_r(m,t), \forall r \in \mathcal{R}. \end{aligned}$$

Let $\mathbf{s}^{\pi_k}(m,t)$ and $\mathbf{s}^{\tilde{\pi}_k}(m,t)$ denote the allocations $\mathbf{s}^{\pi_k}(t)$ and $\mathbf{s}^{\tilde{\pi}_k}(t)$, defined in Theorem 1, on the m^{th} network. As shown in Theorem 1,

$$\lim_{m\to\infty} \frac{G^m(\mathbf{s}^{\pi_k}(m,t)) - G^m(\mathbf{s}^{\tilde{\pi}_k}(m,t))}{G^m(\mathbf{s}^{\tilde{\pi}_k}(m,t))} = 0.$$

Thus for arbitrary $\epsilon > 0$ there exists an $m(\epsilon) \in \mathbb{Z}$ such that the proportional difference in the objective under allocations $\mathbf{s}^{\pi_k}(m,t)$ and $\mathbf{s}^{\tilde{\pi}_k}(m,t)$ is less than ϵ for all $m > m(\epsilon)$.

We next point out that, for two streams (i,r) and (j,r) on route r, if $s_{i,r}d_{i,r} < s_{j,r}d_{j,r}$ then necessarily $s^{\pi_k}_{i,r}(m,t) \geq s^{\pi_k}_{j,r}(m,t)$, since otherwise we could increase the value of the objective by swapping bandwidth from j to i. Thus, only two rates are necessary in the limit, and the allocation given to streams on a given route is non-increasing in stream volume. This implies all streams with volume less than some (route dependent) threshold are given their maximum subscription level and all streams with volume above that threshold are given their minimum subscription level. We can therefore transform the problem of finding the optimal subscription level assignment to one of finding the optimal volume thresholds for each route:

$$\begin{aligned}
\max_{\mathbf{v}(m,t)} \quad & \sum_{r\in\mathcal{R}} \Big(\sum_{i=1}^{n_r(m,t)} \frac{\mathbb{I}(v_{i,r} \leq v_r(m,t))}{d_{i,r}} + \sum_{i=1}^{n_r(m,t)} \frac{a_{i,r}\mathbb{I}(v_{i,r} > v_r(m,t))}{d_{i,r}} \Big) \\
s.t. \quad & \sum_{r\ni l} \Big(\sum_{i=1}^{n_r(m,t)} s_{i,r}\mathbb{I}(v_{i,r} \leq v_r(m,t)) + \sum_{i=1}^{n_r(m,t)} a_{i,r}s_{i,r}\mathbb{I}(v_{i,r} > v_r(m,t)) \Big) \leq c_l(m),
\end{aligned}$$

$\forall l \in \mathcal{L}$. We can normalize the objective by dividing by $\lambda(m) = \sum_{r\in\mathcal{R}} \lambda_r(m)$ without affecting the optimal instantaneous volume threshold vector. This yields

$$\begin{aligned}
\max_{\mathbf{v}(m,t)} \quad & \frac{1}{\lambda(m)} \sum_{r\in\mathcal{R}} \Big(\sum_{i=1}^{n_r(m,t)} \frac{\mathbb{I}(v_{i,r} \leq v_r(m,t))}{d_{i,r}} + \sum_{i=1}^{n_r(m,t)} \frac{a_{i,r}\mathbb{I}(v_{i,r} > v_r(m,t))}{d_{i,r}} \Big) \qquad (3.28) \\
s.t. \quad & \sum_{r\ni l} \Big(\sum_{i=1}^{n_r(m,t)} s_{i,r}\mathbb{I}(v_{i,r} \leq v_r(m,t)) + \sum_{i=1}^{n_r(m,t)} a_{i,r}s_{i,r}\mathbb{I}(v_{i,r} > v_r(m,t)) \Big) \leq c_l(m),
\end{aligned}$$

$\forall l \in \mathcal{L}$. Consider the above optimization problem for a typical time t. We are interested in finding the expected value of the objective and the left hand side of the constraint. At a typical time the stream durations are stretched, i.e., $D \sim F_{\hat{D}}$. This just means we are more likely to see longer duration streams than shorter duration streams at a typical time. We therefore replace $D_{i,r}$ with $\hat{D}_{i,r}$. The expected value of the first term in the objective is given by

$$\begin{aligned}
& \frac{1}{\lambda(m)}\mathbb{E}\Big[\sum_{i=1}^{N_r(m,t)} \frac{\mathbb{I}(S_{i,r}\hat{D}_{i,r} \le v_r(m,t))}{\hat{D}_{i,r}}\Big] \\
= \quad & \frac{1}{\lambda(m)}\mathbb{E}[N_r(m,t)]\mathbb{E}[\frac{1}{\hat{D}}\mathbb{I}(S\hat{D} \le v_r(m,t))] \\
= \quad & \frac{\lambda_r(m)}{\lambda(m)}\delta\int_0^\infty\int_0^\infty \frac{1}{x}\mathbb{I}(xy \le v_r(m,t))f_{\hat{D}}(x)dxf_S(y)dy \\
= \quad & \frac{\lambda_r(m)}{\lambda(m)}\delta\int_0^\infty\int_0^\infty \frac{1}{x}\mathbb{I}(xy \le v_r(m,t))x\frac{1}{\delta}f_D(x)dxf_S(y)dy \\
= \quad & \frac{\lambda_r(m)}{\lambda(m)}\int_0^\infty\int_0^\infty \mathbb{I}(xy \le v_r(m,t))f_D(x)dxf_S(y)dy \\
= \quad & \frac{\lambda_r(m)}{\lambda(m)}\int_0^\infty\Big[\int_0^{\frac{v_r(m,t)}{y}} f_D(x)dx\Big]f_S(y)dy \\
= \quad & \frac{\lambda_r(m)}{\lambda(m)}\int_0^\infty F_D(\frac{v_r(m,t)}{y})f_S(y)dy \\
= \quad & \frac{\lambda_r(m)}{\lambda(m)}F_V(v_r(m,t)) \\
= \quad & \frac{\lambda_r}{\lambda}F_V(v_r(m,t)).
\end{aligned}$$

A similar analysis of the second term in the objective yields

$$\frac{1}{\lambda(m)}\mathbb{E}\Big[\sum_{i=1}^{N_r^m(t)} \frac{A_{i,r}\mathbb{I}(S_{i,r}D_{i,r} > v_r)}{D_{i,r}}\Big] = \alpha\frac{\lambda_r}{\lambda}\bar{F}_V(v_r).$$

Next consider the first sum in the left hand side of the constraint. The expected value is given by

$$
\begin{aligned}
& \mathbb{E}\Big[\sum_{r\ni l}\sum_{i=1}^{N_r(m,t)} S_{i,r}\mathbb{I}(S_{i,r}\hat{D}_{i,r} \leq v_r(m,t))\Big] \\
= & \sum_{r\ni l}\mathbb{E}[N_r(m,t)]\mathbb{E}[S\mathbb{I}(S\hat{D} \leq v_r(m,t))] \\
= & \sum_{r\ni l}\lambda_r(m)\delta\int_0^\infty\int_0^\infty y\mathbb{I}(xy \leq v_r(m,t))f_S(y)dyf_{\hat{D}}(x)dx \\
= & \sum_{r\ni l}\lambda_r(m)\delta\int_0^\infty\int_0^\infty y\mathbb{I}(xy \leq v_r(m,t))f_S(y)dyx\frac{1}{\delta}f_D(x)dx \\
= & \sum_{r\ni l}\lambda_r(m)\int_0^\infty\int_0^\infty xy\mathbb{I}(xy \leq v_r(m,t))f_S(y)dyf_D(x)dx \\
= & \sum_{r\ni l}\lambda_r(m)\int_0^\infty\int_0^\infty w\mathbb{I}(w \leq v_r(m,t))f_S(\frac{w}{x})\frac{1}{x}dwf_D(x)dx \\
= & \sum_{r\ni l}\lambda_r(m)\int_0^\infty w\mathbb{I}(w \leq v_r(m,t))\Big[\int_0^\infty \frac{1}{x}f_S(\frac{w}{x})f_D(x)dx\Big]dw \\
= & \sum_{r\ni l}\lambda_r(m)\int_0^\infty w\mathbb{I}(w \leq v_r(m,t))f_{SD}(w)dw \\
= & \sum_{r\ni l}\lambda_r(m)\int_0^{v_r(m,t)} wf_{SD}(w)dw \\
= & \sum_{r\ni l}\lambda_r(m)\mathbb{E}[V]F_{\hat{V}}(v_r(m,t)) \\
= & \sum_{r\ni l}\lambda_r(m)F_V(v_r(m,t))\mathbb{E}[V \mid V \leq v_r(m,t)]
\end{aligned}
$$

A similar argument shows the expected value of the second term in the constraint equals

$$
\mathbb{E}\Big[\sum_{r\ni l}\sum_{i=1}^{N_r(m,t)} A_{i,r}S_{i,r}\mathbb{I}(S_{i,r}\hat{D}_{i,r} \geq v_r(m,t))\Big] = \sum_{r\ni l}\lambda_r(m)\alpha\bar{F}_V(v_r(m,t))\mathbb{E}[V \mid V > v_r(m,t)].
$$

It is easily seen from the definition of $F_{\hat{V}}$ that

$$\begin{aligned} F_V(v)\mathbb{E}[V \mid V \leq v] &= \mathbb{E}[V]F_{\hat{V}}(v) \\ \bar{F}_V(v)\mathbb{E}[V \mid V > v] &= \mathbb{E}[V]\bar{F}_{\hat{V}}(v). \end{aligned}$$

Thus we can write the constraint as

$$\mathbb{E}[V]\sum_{r \ni l} \lambda_r(m)(F_{\hat{V}}(v_r(m,t)) + \alpha\bar{F}_{\hat{V}}(v_r(m,t))) \leq c_l(m),\ \forall l \in \mathcal{L}.$$

Dividing both sides by $c_l(m)$ and applying the network scaling yields

$$\frac{\mathbb{E}[V]}{\nu_l(m)\delta\gamma_l\sigma}\sum_{r \ni l} \lambda_r(m)(F_{\hat{V}}(v_r(m,t)) + \alpha\bar{F}_{\hat{V}}(v_r(m,t))) \leq 1,\ \forall l \in \mathcal{L}$$

Noting that $\nu_l(m)$ and $\lambda_r(m)$ are both linear in m allows

$$\sum_{r \ni l} \lambda_r(F_{\hat{V}}(v_r(m,t)) + \alpha\bar{F}_{\hat{V}}(v_r(m,t))) \leq \nu_l\gamma_l,\ \forall l \in \mathcal{L}$$

Combining the expressions for the expected value of the objective and the expected value of the constraint yields that the volume threshold at a typical time t on the m^{th} network solves a nonlinear optimization problem with an expected objective and constraint given by

$$\begin{aligned} \max_{\mathbf{v}(m,t)} \quad & \sum_{r \in \mathcal{R}} \lambda_r(F_V(v_r(m,t)) + \alpha\bar{F}_V(v_r(m,t))) \\ s.t. \quad & \sum_{r \ni l} \lambda_r(F_{\hat{V}}(v_r(m,t)) + \alpha\bar{F}_{\hat{V}}(v_r(m,t))) \leq \nu_l\gamma_l,\ \forall l \in \mathcal{L}. \end{aligned}$$

The important point is that the expected value of is independent of m, which means under the network scaling the asymptotic expected instantaneous volume threshold vector is independent of m. Thus, $v(t)$ is the solution to

$$\begin{aligned} \max_{\mathbf{v}(t)} \quad & \frac{1}{\lambda}\sum_{r \in \mathcal{R}} \lambda_r(F_V(v_r(t)) + \alpha\bar{F}_V(v_r(t))) \\ s.t. \quad & \sum_{r \ni l} \lambda_r(F_{\hat{V}}(v_r(t)) + \alpha\bar{F}_{\hat{V}}(v_r(t))) \leq \nu_l\gamma_l,\ \forall l \in \mathcal{L}. \end{aligned}$$

Finally, we prove that the above nonlinear optimization problem consists of maximizing a concave objective over a convex set, which guarantees the existence of a unique optimum. We first simplify the above problem by substituting $\bar{F}_V = 1 - F_V$ and $\bar{F}_{\hat{V}} = 1 - F_{\hat{V}}$. Collecting terms and keeping only the terms in the objective containing functions of v yields

$$\begin{array}{cl} \max\limits_{\mathbf{v}(t)} & \sum\limits_{r \in \mathcal{R}} \lambda_r F_V(v_r(t)) \\ s.t. & \sum\limits_{r \ni l} \lambda_r F_{\hat{V}}(v_r(t)) \leq \dfrac{\nu_l \gamma_l - \nu_l \alpha}{1 - \alpha}, \forall l \in \mathcal{L}. \end{array}$$

We divide both sides of the constraint by ν_l and define $\lambda_{rl} = \frac{\lambda_r}{\nu_l}$, the normalized mean arrival rate of route r on link l. We also define $\xi_l = \frac{\gamma_l - \alpha}{1 - \alpha}$. This allows us to write the optimization problem as

$$\begin{array}{cl} \max\limits_{\mathbf{v}(t)} & \sum\limits_{r \in \mathcal{R}} \lambda_r F_V(v_r(t)) \\ s.t. & \sum\limits_{r \ni l} \lambda_{rl} F_{\hat{V}}(v_r(t)) \leq \xi_l, \forall l \in \mathcal{L}. \end{array}$$

We use the change of variables $y_r(t) = F_{\hat{V}}(v_r(t)) \in [0, 1]$ so that the problem becomes

$$\begin{array}{cl} \max\limits_{\mathbf{y}(t)} & \sum\limits_{r \in \mathcal{R}} \lambda_r F_V(F_{\hat{V}}^{-1}(y_r(t))) \\ s.t. & \sum\limits_{r \ni l} \lambda_{rl} y_r(t) \leq \xi_l, \forall l \in \mathcal{L}. \end{array}$$

Note that the constraints are linear so that the feasible region is a convex set. We next show the objective is a concave increasing function. For brevity we write y_r for $y_r(t)$. Consider

$$\begin{aligned} \frac{\partial}{\partial y_r} \sum_{r' \in \mathcal{R}} \lambda_{r'} F_V(F_{\hat{V}}^{-1}(y_{r'})) &= \lambda_r \frac{\partial}{\partial y_r} F_V(F_{\hat{V}}^{-1}(y_r)) \\ &= \lambda_r \frac{\partial F_V(F_{\hat{V}}^{-1}(y_r))}{\partial F_{\hat{V}}^{-1}(y_r)} \frac{\partial F_{\hat{V}}^{-1}(y_r)}{\partial y_r} \\ &= \lambda_r f_V(F_{\hat{V}}^{-1}(y_r)) \frac{\partial F_{\hat{V}}^{-1}(y_r)}{\partial y_r}. \end{aligned}$$

We can get an expression for $\frac{\partial F_{\hat{V}}^{-1}(y_r)}{\partial y_r}$ by differentiating the identity $F_{\hat{V}}(F_{\hat{V}}^{-1}(y)) = y$:

$$\begin{aligned}
\frac{d}{dy} F_{\hat{V}}(F_{\hat{V}}^{-1}(y)) &= \frac{d}{dy} y \\
\frac{dF_{\hat{V}}(F_{\hat{V}}^{-1}(y))}{dF_{\hat{V}}^{-1}(y)} \frac{dF_{\hat{V}}^{-1}(y)}{dy} &= 1 \\
f_{\hat{V}}(F_{\hat{V}}^{-1}(y)) \frac{dF_{\hat{V}}^{-1}(y)}{dy} &= 1 \\
\frac{dF_{\hat{V}}^{-1}(y)}{dy} &= \frac{1}{f_{\hat{V}}(F_{\hat{V}}^{-1}(y))} \\
\frac{dF_{\hat{V}}^{-1}(y)}{dy} &= \frac{\mathbb{E}[V]}{f_V(F_{\hat{V}}^{-1}(y)) F_{\hat{V}}^{-1}(y))}.
\end{aligned}$$

Substituting into the prior expression yields

$$\frac{\partial}{\partial y_r} \sum_{r' \in \mathcal{R}} \lambda_{r'} F_V(F_{\hat{V}}^{-1}(y_{r'})) = \frac{\lambda_r \mathbb{E}[V]}{F_{\hat{V}}^{-1}(y_r))} > 0.$$

Thus, the first derivative of the objective is positive. Taking the second derivative of the objective yields

$$\begin{aligned}
\frac{\partial}{\partial y_r^2} \sum_{r' \in \mathcal{R}} \lambda_{r'} F_V(F_{\hat{V}}^{-1}(y_{r'})) &= \lambda_r \mathbb{E}[V] \frac{\partial}{\partial y_r} \frac{1}{F_{\hat{V}}^{-1}(y_r))} \\
&= -\frac{\lambda_r \mathbb{E}[V]}{F_{\hat{V}}^{-1}(y_r))^2} \frac{\partial}{\partial y_r} F_{\hat{V}}^{-1}(y_r) \\
&= -\frac{\lambda_r \mathbb{E}[V]}{F_{\hat{V}}^{-1}(y_r))^2 f_{\hat{V}}(F_{\hat{V}}^{-1}(y))} < 0.
\end{aligned}$$

Thus the second derivative of the objective is negative. This means the objective is a concave increasing function in each y_r.

■

Thus, for large capacity networks, provisioning the network according to the rate adaptive scaling implies the optimal adaptation policy for known stream durations assigns an instantaneous volume threshold on each route. Streams with volumes less than the instantaneous volume threshold for their route receive their minimum subscription level and streams with volumes larger than the instantaneous volume threshold for their route receive their maximum subscription level.

3.6.2 Asymptotic optimal QoS, single link

In this section we consider the asymptotic expected time-average normalized subscription level under the optimal adaptation policy for known stream durations on a single link, denoted $q^{\gamma,\pi_k} = \lim_{m\to\infty} \mathbb{E}^0[Q^{m,\pi_k}]$. We obtain a closed form expression which we interpret as the rate-adaptive asymptotic analogue of Erlang's blocking probability equation for loss networks. This expression can be used to explicitly compute how the asymptotic expected time-average normalized subscription level under the optimal adaptation policy varies as a function of the system parameters. The expression is given in the following theorem.

Theorem 4 *The asymptotic expected time-average normalized subscription level under the optimal adaptation policy for known stream durations on a single link is*

$$q^{\gamma,\pi_k} = \begin{cases} \alpha, & \gamma < \alpha \\ 1-(1-\alpha)\bar{F}_V(F_{\hat{V}}^{-1}(\frac{\gamma-\alpha}{1-\alpha})), & \alpha \le \gamma \le 1 \\ 1, & \gamma > 1 \end{cases} . \tag{3.29}$$

Proof of Theorem 4. Let Q^{m,π_k} denote the QoS of a typical stream in the m^{th} scaling of the link capacity under the optimal adaptation policy π_k. Similarly, let $S^{m,\pi_k}(t)$ denote the instantaneous allocation to a typical stream at some time t after that stream's admission:

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

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

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

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

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

for the t in the RHS understood to be a typical time. This allows

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

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

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

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

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

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

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

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

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

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

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

so that

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

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

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

By Wald's identity,

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Note that $\mathbb{I}(\mathbb{E}[Z(m,t,s,d)] > \gamma)$ is equivalent to $\mathbb{I}(sd > F_{\hat{V}}^{-1}(\xi))$ for $\xi = \frac{\gamma-\alpha}{1-\alpha}$.

Notice that for $\gamma < \alpha$, i.e., when the link is provisioned in the overloaded regime, the asymptotic probability the stream is adapted is one. To see this, note that $\gamma < \alpha$ implies $\xi < 0$, and hence

$$\lim_{m\to\infty} p(m,t,s,d) = \mathbb{I}(F_{\hat{V}}(sd) > \xi) = 1.$$

Similarly, for $\gamma > 1$, i.e., when the link is provisioned in the underloaded regime, the asymptotic probability the stream is adapted is zero. To see this, note that $\gamma > 1$ implies $\xi > 1$, and hence

$$\lim_{m\to\infty} p(m,t,s,d) = \mathbb{I}(F_{\hat{V}}(sd) > \xi) = 0.$$

Substituting these values into the integral yields

$$\begin{aligned} q^{\gamma,\pi_k} &= 1 - (1-\alpha)\int_0^\infty \int_0^\infty (1) dF_D(d) dF_S(s) = \alpha, \; \gamma < \alpha \\ q^{\gamma,\pi_k} &= 1 - (1-\alpha)\int_0^\infty \int_0^\infty (0) dF_D(d) dF_S(s) = 1, \; \gamma > 1. \end{aligned}$$

Finally, consider the case when the link is provisioned in the rate adaptive regime, i.e., when $\alpha \leq \gamma \leq 1$. Substituting this into the integral yields

$$q^{\gamma,\pi_k} = 1 - (1-\alpha)\int_0^\infty \int_0^\infty \mathbb{I}(sd > F_{\hat{V}}^{-1}(\xi))dF_D(d)dF_S(s),\ \alpha \leq \gamma \leq 1.$$

This can be simplified as follows

$$\begin{aligned}
q^{\gamma,\pi_k} &= 1 - (1-\alpha)\int_0^\infty \int_0^\infty \mathbb{I}(d > \frac{F_{\hat{V}}^{-1}(\xi)}{s})dF_D(d)dF_S(s),\ \alpha \leq \gamma \leq 1 \\
&= 1 - (1-\alpha)\int_0^\infty \int_{\frac{F_{\hat{V}}^{-1}(\xi)}{s}}^\infty dF_D(d)dF_S(s),\ \alpha \leq \gamma \leq 1 \\
&= 1 - (1-\alpha)\int_0^\infty \bar{F}_D(\frac{F_{\hat{V}}^{-1}(\xi)}{s})dF_S(s),\ \alpha \leq \gamma \leq 1 \\
&= 1 - (1-\alpha)\bar{F}_V(F_{\hat{V}}^{-1}(\xi)),\ \alpha \leq \gamma \leq 1.
\end{aligned}$$

■

The above expression for the asymptotic QoS depends on three quantities: the scaling parameter, γ, the average adaptivity, α, and the distribution of stream volume, F_V. Note that asymptotic QoS depends on the distribution F_A only through its mean $\mathbb{E}[A] = \alpha$, i.e., asymptotic QoS is insensitive to the adaptivity distribution under the optimal adaptation policy. This equation can be thought of as an asymptotic analogue to Erlang's blocking probability equations for loss networks. For networks supporting multimedia streams, a loss network is simply a network where streams are admitted at a fixed subscription level and are guaranteed a transmission rate sufficient to successfully transmit that subscription level over the network. Note that, because there is no dynamic adaptation, the relevant QoS parameter is just the blocking probability. For the simple case of a single link where all streams have a fixed subscription level, say σ, Erlang's blocking probability equation is just

$$E(\rho, C) = \frac{\frac{\rho^C}{C!}}{\sum_{n=0}^{C} \frac{\rho^n}{n!}}, \tag{3.30}$$

where $\rho = \lambda\delta$ and $C = \lfloor \frac{c}{\sigma} \rfloor \in \mathbb{Z}^+$ is the maximum number of streams which can be admitted into the system. Thus, $E(\rho, C) = \mathbb{P}(N(t) = C)$, i.e., the probability an arriving stream arrives to find all available circuits are filled, and is therefore blocked. This equation allows network designers to provision links so as to satisfy some desired blocking probability. For rate adaptive streams the blocking probability is negligible (provided the link is not provisioned in the overloaded regime) because active streams can adapt their subscription levels to free up capacity for newly arriving streams. Instead, the relevant QoS metric is expected time average normalized subscription level, and the expression in Theorem 4 can be used by network designers to provision links so as to satisfy some desired constraint on this quantity.

Note that the asymptotic QoS under the optimal adaptation policy equals the asymptotic QoS under the fair share policy when the link is provisioned in either the overloaded or underloaded regime. That is, when the link is overloaded, i.e., $\gamma < \alpha$, the asymptotic QoS under both policies is $q^\gamma = \alpha$, the average minimum normalized subscription level. Similarly, when the link is underloaded, i.e., $\gamma > 1$, the asymptotic QoS under both policies is $q^\gamma = 1$, the average maximum normalized subscription level. More importantly, the QoS under these regimes is independent of the scaling parameter, γ. Thus the asymptotic marginal increase in QoS obtained by a marginal increase in capacity is zero, i.e., $\frac{\partial q^\gamma}{\partial \gamma} = 0$ for $\gamma < \alpha$ and $\gamma > 1$. This suggests that the rate adaptive regime, i.e., $\alpha \leq \gamma \leq 1$, is the appropriate target design regime. The asymptotic marginal increase in QoS obtained by a marginal increase in capacity is strictly positive, and the optimal adaptation policy achieves a strictly positive improvement in asymptotic QoS over the fair share policy. This improvement is studied in the next subsection.

3.6.3 Computational results for asymptotic optimal QoS

In this subsection we plot the expression for the asymptotic expected time-average normalized subscription level under the optimal adaptation policy, given in Theorem

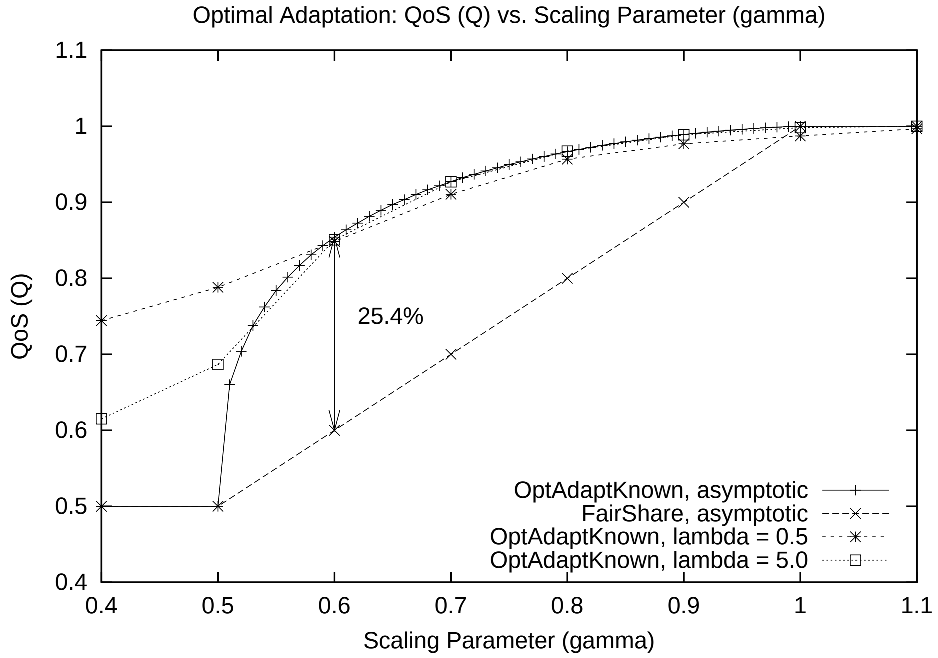

Figure 3.8: Plot of the asymptotic QoS q^{π_k} versus the capacity scaling parameter γ.

4, for the special case of a single link servicing streams with known stream durations.

Figure 3.8 exhibits q^{γ,π_k} versus γ. Also shown are two plots of simulation results illustrating the convergence to the asymptotic QoS. Recall the scaling regimes have transitions at $\gamma = \alpha = 0.5$ and $\gamma = 1$. The two simulation results use $\lambda = 0.5$ and $\lambda = 5.0$ respectively, and a link capacity $c = \gamma\lambda\sigma\delta$. A plot of q^{γ,π_f}, the asymptotic QoS under the fair share adaptation policy, is also given. The label indicates that at $\gamma = 0.6$ the optimal policy achieves an asymptotic QoS 25.4% greater than under fair share, i.e., a 42% increase in performance.

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

Figure 3.8 exhibits q^{γ,π_k} when F_S and F_D are the bounded exponential distributions described in Section 3.1. We investigated several other probability distributions for F_S and F_D, several of which are plotted in Figure 3.9. In addition to the bounded exponential distribution, we also studied the uniform distribution and the bounded Pareto distribution. To facilitate comparison we kept the means of all the distributions the same, namely, $\mathbb{E}[S] = 0.3$ (Mb/s) and $\mathbb{E}[D] = 180$ (seconds). The uniform distributions we used are $S \sim Uni(\frac{1}{10}, \frac{5}{10})$ and $D \sim Uni(60, 300)$. The bounded Pareto distributions we used are $S \sim Par(1.382, \frac{1}{10}, 10)$ and $D \sim Par(1.382, 60, 6000)$, where $Par(\alpha_p, m, M)$ means a Pareto distribution over (m, M) with an exponent of α_p. Table 3.6.3 shows the distributions, the variance of the corresponding volume $V = SD$, and five summary statistics over the rate adaptive regime ($\alpha = \frac{1}{2} \leq \gamma \leq 1$). The statistics we considered are

$$\begin{aligned}
z_a &= \frac{1}{1-\alpha}\int_{\alpha}^{1} q^{\gamma,\pi_k} d\gamma, \\
z_b &= \frac{1}{1-\alpha}\int_{\alpha}^{1} (q^{\gamma,\pi_k} - q^{\gamma,\pi_f}) d\gamma, \\
z_c &= \frac{1}{1-\alpha}\int_{\alpha}^{1} \frac{q^{\gamma,\pi_k} - q^{\gamma,\pi_f}}{\gamma} d\gamma, \\
z_d &= \max_{\alpha\leq\gamma\leq 1} \{q^{\gamma,\pi_k} - q^{\gamma,\pi_f}\}, \\
z_e &= \max_{\alpha\leq\gamma\leq 1} \{\frac{q^{\gamma,\pi_k} - q^{\gamma,\pi_f}}{\gamma}\}.
\end{aligned}$$

These correspond to the average QoS, the average increase in QoS above fair share, the average improvement in QoS above fair share, the maximum increase in QoS above fair share, and the maximum improvement in QoS above fair share. The distribution pairs in the table not plotted in Figure 3.9 are bounded above and

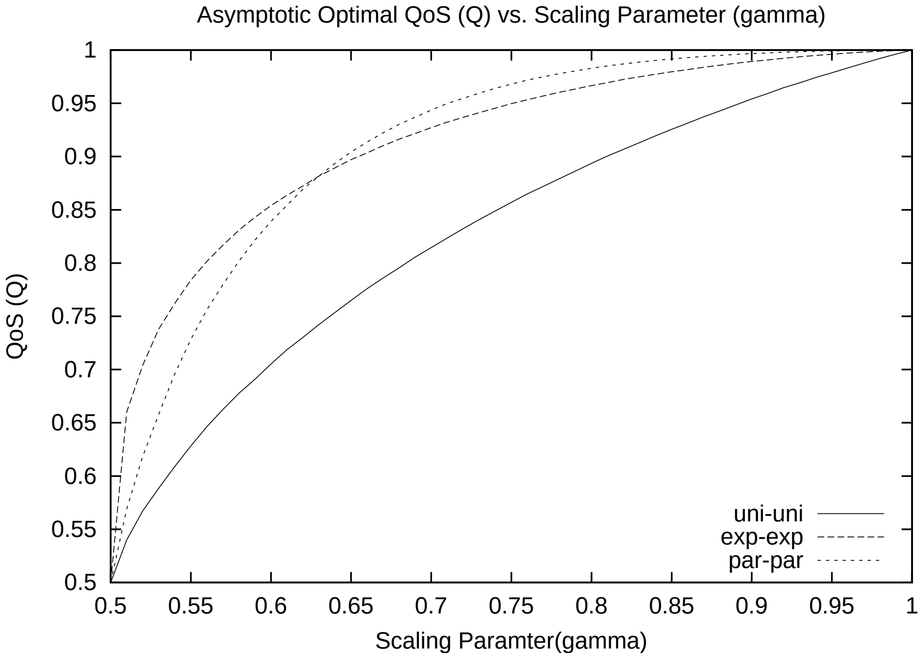

Figure 3.9: Plot of the asymptotic QoS q^{γ,π_k} versus the capacity scaling parameter γ for various distributions F_S and F_D.

below by one of the plots shown.

Surprisingly, there is not a direct correlation between the variance $Var(V)$ and any of the five summary statistics considered above. The optimal adaptation policy achieves a higher QoS than the fair share adaptation policy by exploiting the variation in stream volume of the active streams. Thus one might conjecture that stream volume distributions with high variance would outperform distributions with lower variance. The table illustrates that no such trend is apparent.

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

Table 3.1: Performance metrics for the optimal adaptation policy for several choices of simulation distributions.

Chapter 4

Optimal Admission Control Policies

The preceding chapter on optimal adaptation policies demonstrated that maximizing the QoS requires

- possibly adjusting each active stream's subscription level at each time t that a new stream is admitted or an active stream departs, and
- knowledge of all system parameters, i.e., knowledge of each stream's volume (or expected volume for unknown stream durations), and all link capacities.

These requirements are quite restrictive because they require centralized control, i.e., some central server would need to collect stream information each time the network state changed (arrivals or departures), and disseminate a new subscription level assignment to each active stream. On anything other than trivially small networks such centralized control is infeasible. This chapter addresses the first of these concerns by proposing an admission control policy which obtains the same asymptotic QoS as optimal adaptation, but without the need for dynamically adjusting subscription levels. Chapter 5 addresses the second concern by proposing a class of distributed algorithms which attempt to regain much of the increase in QoS obtained under optimal adaptation and admission, but without knowledge of system parameters.

In this chapter we identify an optimal admission control policy wherein streams are admitted at a subscription level which they maintain throughout their duration, i.e., no dynamic adaptation. Intuitively, large volume streams are admitted at their minimum subscription level and small volume streams are admitted at their maximum subscription level. We will prove the existence of an optimal volume threshold on each route which is used to differentiate between small and large volume streams on that route. Intuitively, these volume thresholds are set based on the expected congestion level of the route, i.e., under-provisioned links have a low volume threshold, meaning all but the smallest volume streams are admitted at their minimum subscription level, while over-provisioned links have a high volume threshold, meaning all but the largest volume streams are admitted at their maximum subscription level. Figure 4 illustrates the idea. A large volume stream with volume V exceeding the volume threshold v^* is admitted at its minimum subscription level which it then maintains throughout its duration, while a small volume stream with volume less than the threshold is admitted at its maximum subscription level, which it maintains throughout its duration.

Section 4.1 provides some motivation as to why admission control serves as an adequate replacement for dynamic adaptation on large capacity links. Section 4.2 identifies the asymptotic optimal admission control policy for a general network and for a single link. Section 4.3 presents some simulation results illustrating the QoS under the asymptotic optimal admission control policy. Finally, Section 4.4 presents some issues surrounding implementing optimal admission control for rate adaptive streams.

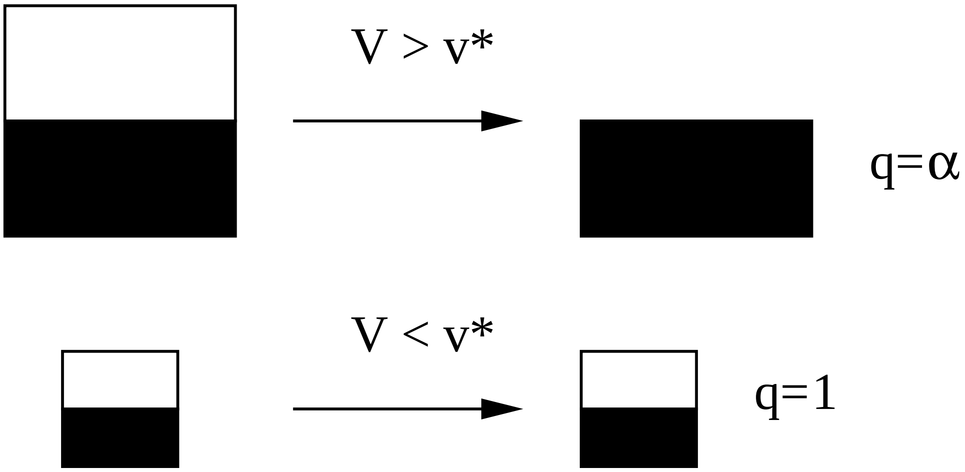

Figure 4.1: Illustration of the volume dependent admission control policy.

4.1 Motivation behind the admission control policies

Theorem 3 proved that the asymptotic optimal dynamic adaptation policy for stored media on a general network is a threshold policy. In particular, at each time t, there is an instantaneous optimal volume threshold vector, $\mathbf{v}^{\pi_k}(m,t)$, such that streams with volumes less than the instantaneous volume threshold receive their maximum subscription level and streams with volumes greater than the instantaneous volume threshold receive their minimum subscription level. The volume thresholds are *instantaneous* in that they depend on the instantaneous number of streams, $\mathbf{n}(t)$, and the characteristics of each active stream. Thus, a stream's instantaneous subscription level under the optimal adaptation policy may change if the instantaneous volume threshold changes. That is, if $v_r^{\pi_k}(m,t_1), v_r^{\pi_k}(m,t_2)$ are the optimal instantaneous volume thresholds on route r at times $t_1 < t_2$, and stream (i,r) has a volume $v_{i,r}$ such that $v_r^{\pi_k}(m,t_1) < v_{i,r} < v_r^{\pi_k}(m,t_2)$, then the stream would receive its minimum subscription level at time t_1 and its maximum subscription level at time t_2.

Consider the simulation results plotted in Figures 4.2 and 4.3, showing histograms for $\mathbb{E}^0[Q^{\pi_k} \mid V = v]$ and $\mathbb{E}^0[R^{\pi_k} \mid V = v]$ versus stream volume for three different sets of arrival rates and link capacities. The three sets correspond to $(\lambda, c) = (0.05, 2.025)$, $(0.5, 20.25)$, and $(5.0, 202.5)$, where c is in Mb/s. All three sets correspond to a network scaling with scaling parameter $\gamma = \frac{3}{4}$. The expected number of active streams for the three simulations are $\mathbb{E}[N(t)] = \lambda\delta = 9$, 90, 900, respectively. We refer to these as the small, medium, and large simulations respectively. As we increase the arrival rate and link capacity from small to large, we see in Figure 4.2 that there is an increasingly pronounced volume threshold distinguishing the time-average normalized subscription level between small and large volume streams. Similarly, we seen in Figure 4.3 that increasing the arrival rate and link capacity from small to large has the effect of increasing the rate of adaptation for streams with volumes in an increasingly narrow interval. This leads to the conclusion that, under the network scaling, the asymptotic instantaneous volume threshold is a constant. That is, $\lim_{m\to\infty} \mathbf{v}^{\pi_k}(m, t) = \mathbf{v}^{\pi_k}$, which is independent of t.

This observation motivates our discussion of admission control policies as a replacement for dynamic adaptation. Since the asymptotic optimal dynamic adaptation policy grants a constant subscription level to all streams, depending on whether their stream volume is above or below a fixed volume threshold, it follows that we can assign a fixed subscription level using the same threshold at the time of admission, without compromising the asymptotic optimal QoS.

4.2 Asymptotic optimal admission control policies

We parameterize a set of two-class admission policies as follows. Upon admission each stream is assigned to an adaptation class based on its volume. Streams admitted to class 1 receive their full subscription level and streams admitted to class 2 receive their minimum subscription level. We will show that two classes suffice to obtain an asymptotic QoS equal to that obtained under the dynamic adaptation

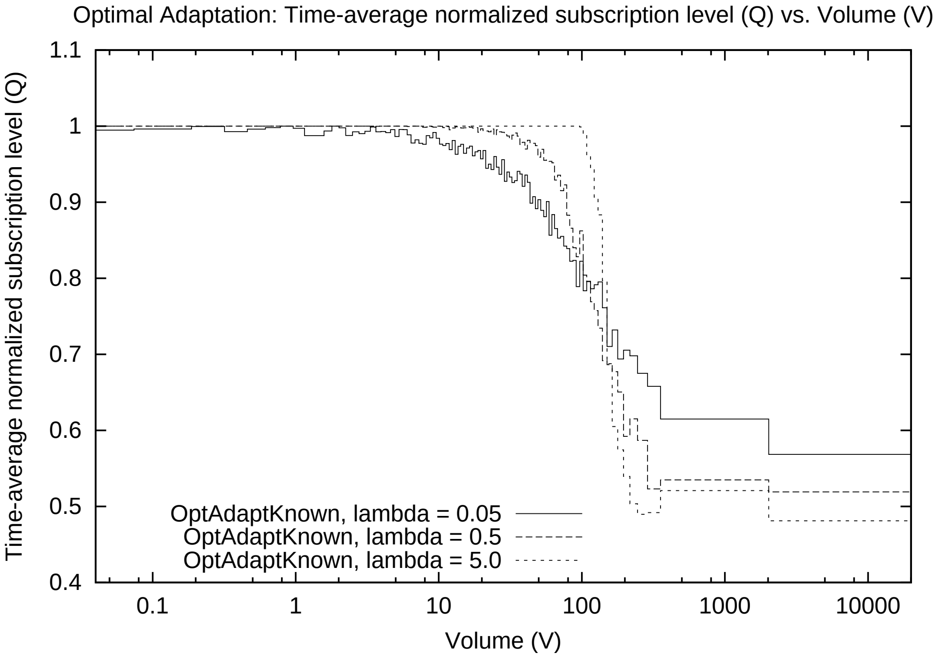

Figure 4.2: Histogram of $\mathbb{E}^0[Q^{\pi_k} \mid V = v]$ versus v for three different sets of arrival rate and link capacity.

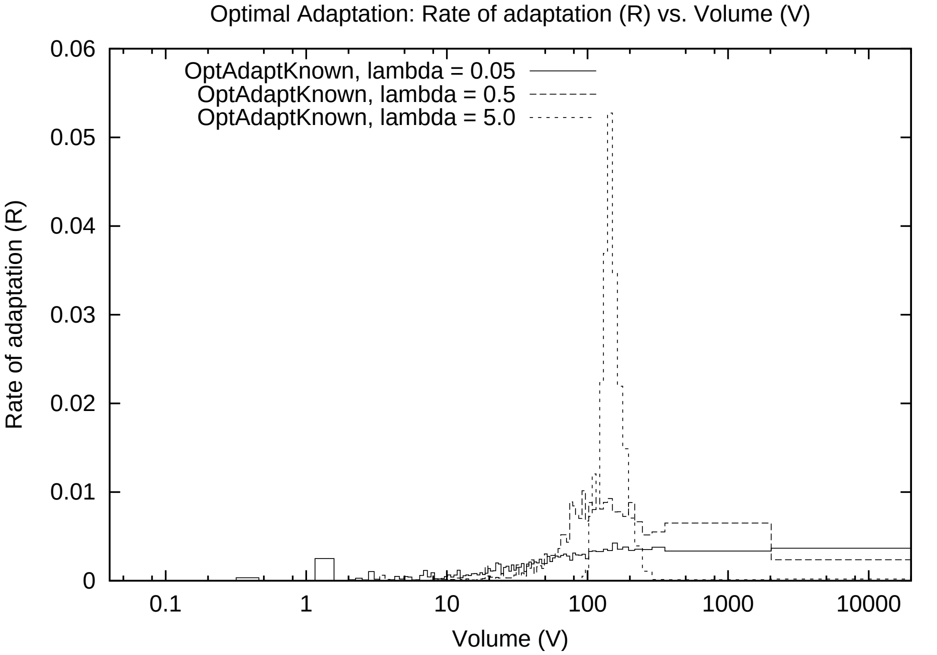

Figure 4.3: Histogram of $\mathbb{E}^0[R^{\pi_k} \mid V = v]$ versus v for three different sets of arrival rate and link capacity.

policy. Note that the class of admission policies can be thought of as a subset of the class of adaptation policies, where adaptation decisions are only made at the time of a stream's admission into the network. Thus, showing that an admission policy achieves an asymptotic QoS equal to that under the optimal adaptation policy implies that the admission policy is optimal.

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

$$\sum_{r \ni l} \sum_{k=1}^{2} \sum_{i=1}^{n_{k,r}(t)} \beta_k s_{i,k,r} + (1-\beta_k)a_{i,k,r}s_{i,k,r} + \beta_k s + (1-\beta_k)as \leq c_l, \; \forall l \in r', \quad (4.1)$$

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

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

$$\begin{aligned} \lambda_{1,r}(m) &= \lambda_r(m)\mathbb{P}(V \leq v_r), \quad (4.2) \\ \lambda_{2,r}(m) &= \lambda_r(m)\mathbb{P}(V > v_r). \end{aligned}$$

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

$$\lim_{m\to\infty} \frac{1}{c_l(m)} \sum_{r \ni l} \Big(\lambda_{1,r}(m)\mathbb{E}[V \mid V \leq v_r] + \alpha\lambda_{2,r}(m)\mathbb{E}[V \mid V > v_r]\Big) \leq 1, \forall l \in \mathcal{L}. \quad (4.3)$$

It is shown in [27] that blocking is asymptotically zero for this case, although convergence is $O(\frac{1}{\sqrt{c}})$. Our objective is to maximize the asymptotic client average normalized subscription level which, under the assumed asymptotic zero blocking regime, is given by

$$\lim_{m\to\infty} \frac{\sum_{r\in\mathcal{R}} \lambda_{1,r}(m) + \alpha\lambda_{2,r}(m)}{\sum_{r\in\mathcal{R}} \lambda_r(m)}. \tag{4.4}$$

Thus we need to identify the optimal set of volume thresholds $\mathbf{v}$ that maximizes the asymptotic normalized subscription level (4.4) subject to the asymptotic utilization being bounded by 1 on each link (4.3).

The following two subsections identify the asymptotic optimal admission control policy for a general network and a single link. In each case we show that the asymptotic QoS under the optimal admission control policy equals that obtained under the optimal dynamic adaptation policy.

4.2.1 Asymptotic optimal admission control, general network

The following theorem proves that the asymptotic expected normalized subscription level under the asymptotic optimal admission policy equals that under the asymptotic optimal dynamic adaptation policy on a general network. It also proves that the optimal thresholds are decreasing in the asymptotic congestion on the route.

Theorem 5 *The asymptotic optimal static volume threshold vector, which maximizes (4.4) subject to the constraints (4.3), is the same as the asymptotic optimal instantaneous volume threshold vector, which solves (3.27).*

Moreover, the optimal volume thresholds have the form

$$v_r^{\gamma,\pi_a} = \frac{\mathbb{E}[V]}{\sum_{l\in r} z_l^{\gamma,\pi_a}} \tag{4.5}$$

where $\mathbf{z}^{\gamma,\pi_a} = (z_l^{\gamma,\pi_a}, l \in \mathcal{L})$ *is a vector of optimal Lagrange multipliers for the constraints given in (3.27).*

Proof of Theorem 5.

The optimal static volume threshold vector is the solution to the following optimization problem:

$$\begin{aligned}
\max_{\mathbf{v}} \quad & \lim_{m\to\infty} \frac{\sum_{r\in\mathcal{R}} \lambda_{1,r}(m) + \alpha\lambda_{2,r}(m)}{\sum_{r\in\mathcal{R}} \lambda_r(m)} \\
s.t. \quad & \lim_{m\to\infty} \frac{1}{c_l(m)} \sum_{r\ni l} \Big(\lambda_{1,r}(m)\mathbb{E}[V \mid V \le v_r] + \alpha\lambda_{2,r}(m)\mathbb{E}[V \mid V > v_r] \Big) \le 1, \forall l \in \mathcal{L}.
\end{aligned}$$

Note that $\lambda_{1,r}(m) = \lambda_r(m)F_V(v_r)$ and $\lambda_{2,r}(m) = \lambda_r(m)\bar{F}_V(v_r)$. Also note that $F_V(v_r)\mathbb{E}[V \mid V \le v_r] = \mathbb{E}[V]F_{\hat{V}}(v_r)$ and $\bar{F}_V(v_r)\mathbb{E}[V \mid V > v_r] = i\mathbb{E}[V]\bar{F}_{\hat{V}}(v_r)$. This allows us to rewrite the optimization problem as:

$$\begin{aligned}
\max_{\mathbf{v}} \quad & \lim_{m\to\infty} \frac{\sum_{r\in\mathcal{R}} \lambda_r(m)(F_V(v_r) + \alpha\bar{F}_V(v_r))}{\sum_{r\in\mathcal{R}} \lambda_r(m)} \\
s.t. \quad & \lim_{m\to\infty} \frac{\mathbb{E}[V]}{c_l(m)} \sum_{r\ni l} \lambda_r(m)(F_{\hat{V}}(v_r) + \alpha\bar{F}_{\hat{V}}(v_r)) \le 1, \forall l \in \mathcal{L}.
\end{aligned}$$

Applying the network scaling and multiplying the objective by $\lambda = \sum_{r\in\mathcal{R}} \lambda_r$ yields

$$\begin{aligned}
\max_{\mathbf{v}} \quad & \sum_{r\in\mathcal{R}} \lambda_r(F_V(v_r) + \alpha\bar{F}_V(v_r)) \\
s.t. \quad & \sum_{r\ni l} (F_{\hat{V}}(v_r) + \alpha\bar{F}_{\hat{V}}(v_r)) \le \gamma_l, \forall l \in \mathcal{L}.
\end{aligned}$$

This is seen to be the same optimization problem as (3.27).

Thus, the asymptotic optimal admission control policy sets the volume threshold vector equal to the *expected* instantaneous volume threshold vector at a typical time under the optimal dynamic adaptation policy. We require, however, a stronger result to prove the asymptotic QoS under optimal admission control equals that under optimal dynamic adaptation. Intuitively, setting the admission control volume threshold vector at the *expected* instantaneous volume threshold vector does not rule out the possibility that dynamic adaptation obtains a higher QoS than admission control by virtue of the fact that the instantaneous adaptation vector is *dynamic*.

We next prove that the *asymptotic* instantaneous volume threshold vector under optimal dynamic adaptation is a constant (equaling its mean).

In the proof of Theorem 3 we argued that, for large enough m, the optimization problem (3.21) is equivalent to the optimization problem (3.28). We will show that, as we let $m \to \infty$, the objective and the constraint become constants, independent of t. The proof is just a straightforward application of the law of large numbers. Consider the limiting value of the objective, given by

$$\lim_{m\to\infty} \frac{1}{\lambda(m)} \sum_{r\in\mathcal{R}} \Big(\sum_{i=1}^{N_r(m,t)} \frac{\mathbb{I}(S_{i,r}\hat{D}_{i,r} \le v_r(m,t)) + A_{i,r}\mathbb{I}(S_{i,r}\hat{D}_{i,r} > v_r(m,t))}{\hat{D}_{i,r}} \Big).$$

We write this as a product of limits:

$$\begin{aligned} = \quad & \sum_{r\in\mathcal{R}} \Big(\lim_{m\to\infty} \frac{N_r(m,t)}{\lambda(m)} \Big) \\ \times \quad & \Big(\lim_{m\to\infty} \frac{1}{N_r(m,t)} \sum_{i=1}^{N_r(m,t)} \frac{\mathbb{I}(S_{i,r}\hat{D}_{i,r} \le v_r(m,t)) + A_{i,r}\mathbb{I}(S_{i,r}\hat{D}_{i,r} > v_r(m,t))}{\hat{D}_{i,r}} \Big) \end{aligned}$$

The first term in the product is easily seen to equal $\frac{\lambda_r \delta}{\lambda}$. By the law of large numbers, the second term equals

$$\mathbb{E}\Big[\frac{\mathbb{I}(S\hat{D} \le v_r(t)) + A\mathbb{I}(S\hat{D} > v_r(t))}{\hat{D}} \Big].$$

In the proof of Theorem 3 we computed this expectation to be $\frac{1}{\delta}(F_V(v_r(t)) + \alpha\bar{F}_V(v_r(t)))$. Combining this result yields the asymptotic value of the objective to be

$$\frac{1}{\lambda} \sum_{r\in\mathcal{R}} \lambda_r (F_V(v_r(t)) + \alpha\bar{F}_V(v_r(t))).$$

Now consider the limiting value of the constraint, given by

$$\lim_{m\to\infty} \frac{1}{c_l(m)} \sum_{r\ni l} \Big(\sum_{i=1}^{N_r(m,t)} S_{i,r}\mathbb{I}(S_{i,r}\hat{D}_{i,r} \le v_r(m,t)) + A_{i,r}S_{i,r}\mathbb{I}(S_{i,r}\hat{D}_{i,r} > v_r(m,t)) \Big)$$

for each $l \in \mathcal{L}$. We again write this as a product of limits:

$$\begin{aligned} &= \sum_{r \ni l} \Big(\lim_{m \to \infty} \frac{N_r(m,t)}{c_l(m)} \Big) \\ &\times \Big(\lim_{m \to \infty} \frac{\sum_{i=1}^{N_r(m,t)} S_{i,r} \mathbb{I}(S_{i,r} \hat{D}_{i,r} \le v_r(m,t)) + A_{i,r} S_{i,r} \mathbb{I}(S_{i,r} \hat{D}_{i,r} > v_r(m,t))}{N_r(m,t)} \Big) \end{aligned}$$

The first term in the product is easily seen to equal $\frac{\lambda_r}{\nu_l \gamma_l \sigma}$ by applying the definition of the network scaling and noting that $\mathbb{E}[N(m,t)] = \lambda_r(m)\delta$. By the law of large numbers, the second term equals

$$\mathbb{E}\Big[S\mathbb{I}(S\hat{D} \le v_r(t)) + AS\mathbb{I}(S\hat{D} > v_r(t)) \Big].$$

In the proof of Theorem 3 we computed this expectation to be $\sigma F_{\hat{V}}(v_r(t)) + \alpha\sigma \bar{F}_{\hat{V}}(v_r(t))$. Combining this result yields the asymptotic value of the constraint to be

$$\sum_{r \ni l} \lambda_r (F_{\hat{V}}(v_r(t)) + \alpha \bar{F}_{\hat{V}}(v_r(t)) \le \nu_l \gamma_l, \ \forall l \in \mathcal{L}.$$

Combining the asymptotic value of the objective with the asymptotic value of the constraint, we see the asymptotic optimal instantaneous volume threshold is given by the solution to

$$\begin{aligned} \max_{\mathbf{v}(t)} \quad & \frac{1}{\lambda} \sum_{r \in \mathcal{R}} \lambda_r (F_V(v_r(t)) + \alpha \bar{F}_V(v_r(t))) \\ s.t. \quad & \sum_{r \ni l} \lambda_r (F_{\hat{V}}(v_r(t)) + \alpha \bar{F}_{\hat{V}}(v_r(t)) \le \nu_l \gamma_l, \ \forall l \in \mathcal{L}. \end{aligned}$$

But the solution to this problem is independent of t, hence the asymptotic optimal instantaneous volume threshold is a constant. Moreover, the optimization problem is seen to be the same as maximizing (4.4) subject to (4.3). Hence, the asymptotic QoS under the optimal admission control policy equals the asymptotic QoS under the optimal dynamic adaptation policy.

We next prove that the optimal volume thresholds have a form $v_r^{\gamma,\pi_a} = \frac{\mathbb{E}[V]}{\sum_{l \in r} z_l^{\gamma,\pi_a}}$ where $\mathbf{z}^{\gamma,\pi_a} = (z_l^{\gamma,\pi_a}, l \in \mathcal{L})$ is a vector of optimal Lagrange multipliers on the constraints. Recall from the proof of Theorem 3 that we can transform the optimization problem to

$$\begin{aligned} \max_{\mathbf{y}} \quad & \sum_{r\in\mathcal{R}} \lambda_r F_V(F_{\hat{V}}^{-1}(y_r)) \\ s.t. \quad & \sum_{r\ni l} \lambda_{rl} y_r \leq \xi_l, \forall l \in \mathcal{L}, \end{aligned}$$

using the change of variable $y_r = F_{\hat{V}}(v_r)$. We can write the Lagrangian, $\mathcal{L}(\mathbf{y}, \mathbf{z})$, as

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

Taking derivatives with respect to v_r and simplifying yields

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

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

■

We interpret $\sum_{l\in r} z_l^{\gamma,\pi_a}$ as the route cost, using the standard interpretation of Lagrange multipliers as quantifying the cost of the associated constraint. Thus, the result states that congested routes have smaller volume thresholds, i.e., on congested routes we can afford to admit a smaller fraction of streams at their maximum subscription level. Similarly, uncongested routes have larger volume thresholds, i.e., on uncongested routes we can afford to admit a larger fraction of streams at their maximum subscription level.

4.2.2 Asymptotic optimal admission control, single link

We specialize the results of the previous section to the case of a single link. This allows us to obtain a closed form expression for the optimal volume threshold. The following theorem identifies the optimal volume threshold on a single link and shows that the resulting asymptotic QoS using this threshold equals the asymptotic QoS under the optimal dynamic adaptation policy. We denote the optimal admission policy as π_a. We write v^{γ,π_a} for the optimal volume threshold and denote the asymptotic expected normalized subscription level under this policy as q^{γ,π_a}.

Theorem 6 *The asymptotic optimal admission policy, π_a, that achieves asymptotic zero blocking for the special case of single bottleneck links has a volume threshold*

$$v^{\gamma,\pi_a} = \begin{cases} 0, & \gamma \leq \alpha \\ F_{\hat{V}}^{-1}(\frac{\gamma-\alpha}{1-\alpha}), & \alpha < \gamma < 1 \\ \infty, & \gamma \geq 1 \end{cases} . \tag{4.6}$$

The asymptotic normalized subscription level under this policy is

$$q^{\gamma,\pi_a} = \begin{cases} \alpha, & \gamma \leq \alpha \\ 1-(1-\alpha)\bar{F}_V(v^{\gamma,\pi_a}), & \alpha < \gamma \leq 1 \\ 1, & \gamma \geq 1 \end{cases} . \tag{4.7}$$

Proof of Theorem 6. For the case of a single link the asymptotic constraint from Theorem 5 becomes

$$\lambda(F_{\hat{V}}(v) + \alpha\bar{F}_{\hat{V}}(v) \leq \lambda\gamma.$$

This is easily simplified to yield

$$F_{\hat{V}}(v) \leq \frac{\gamma-\alpha}{1-\alpha}.$$

. For $\gamma < \alpha$, the RHS is negative, and hence the constraint requires $v^{\gamma,\pi_a} = 0$. The interpretation is that when we provision the link in the overloaded regime, we must admit all streams at their minimum subscription level. Similarly, for $\gamma > 1$, the RHS is greater than one, and hence the constraint requires $v^{\gamma,\pi_a} = \infty$. The interpretation is that when we provision the link in the underloaded regime, we can admit all streams at their maximum subscription level. Next consider when $\alpha \leq \gamma \leq 1$. The asymptotic objective for a single link is to maximize $F_V(v) + \alpha\bar{F}_V(v)$ which is equivalent to maximizing $F_V(v)$, which is by definition increasing in v. Hence the optimal v occurs when the constraint is binding, i.e., $v^{\gamma,\pi_a} = F_{\hat{V}}^{-1}(\frac{\gamma-\alpha}{1-\alpha})$.

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

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

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

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

This allows:

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

When $\gamma < \alpha$ we have $\bar{F}_V(v^{\gamma,\pi_a}) = \bar{F}_V(0) = 1$, and hence $q^{\gamma,\pi_a} = \alpha$. When $\gamma > 1$ we have $\bar{F}_V(v^{\gamma,\pi_a}) = \bar{F}_V(\infty) = 0$, and hence $q^{\gamma,\pi_a} = 1$.

■

Substituting the optimal volume threshold v^{γ,π_a} into the expression for asymptotic QoS under the optimal admission control policy, q^{γ,π_a} yields

$$q^{\gamma,\pi_a} = \begin{cases} \alpha, & \gamma \leq \alpha \\ 1-(1-\alpha)\bar{F}_V(F_{\hat{V}}^{-1}(\frac{\gamma-\alpha}{1-\alpha})), & \alpha < \gamma \leq 1 \\ 1, & \gamma \geq 1 \end{cases}, \tag{4.8}$$

which is the same expression as given for q^{γ,π_k}. Note v^{γ,π_a} is increasing in γ, i.e., as we increase the link capacity relative to the load we can admit a higher fraction of the streams at their maximum subscription level.

4.3 Simulation results for asymptotic optimal admission control policies

In this section we present some simulation results showing the quality of service on links with finite capacity and arrival rate under the asymptotic optimal admission control policy.

Figure 4.4 contains a plot of the asymptotic QoS, $q^{\pi_a} = q^{\pi_k}$ versus the scaling parameter γ. Also shown are two simulation plots under policy π_a for $\lambda = 0.5$ and $\lambda = 5.0$. The plots illustrate the convergence in the expected normalized subscription level under the network scaling. Recall our simulations use $\alpha = 0.5$.

Figure 4.5 contains simulation results of the blocking probability under the optimal admission policy and the optimal adaptation policy for $\lambda = 0.5$ and $\lambda = 5.0$. Recall that the asymptotic blocking probability is zero for $\gamma \geq \alpha$, and is $1 - \frac{\gamma}{\alpha}$ for $\gamma < \alpha$. Both policies illustrate convergence to these values. The plots also illustrate a critical difference between optimal adaptation and optimal admission control, namely, the convergence rate of the blocking probability. Under dynamic adaptation the blocking probability converges to zero exponentially fast for $\gamma \geq \alpha$. This is because dynamic adaptation permits active streams to reduce their subscription levels to free up capacity for newly arriving streams. Under the admission control policy, however, the blocking probability converges to zero as $O(\frac{1}{\sqrt{c}})$ for $\gamma \geq \alpha$. The slower convergence rate exists because, under the admission control policy, active streams are *not* permitted to reduce their subscription level to free up capacity for newly arriving streams. The system functions as a loss network which is critically loaded for $\alpha \leq \gamma \leq 1$ [27].

The asymptotic optimal admission control policy therefore achieves the same asymptotic expected normalized subscription level as that obtained under optimal dynamic adaptation, i.e., $q^{\pi_a} = q^{\pi_k}$. Moreover, the rate of adaptation under admission control is by definition zero. The penalty of using admission control over dynamic adaptation is that, although both achieve asymptotic zero blocking, the rate at which the blocking probability converges to zero is much slower for optimal admission control than for optimal dynamic adaptation.

4.4 Implementation issues for asymptotic optimal admission control policies

At the beginning of this chapter we described two basic problems with implementing the optimal adaptation policies of Chapter 3: *i*) the need for centralized control to disseminate optimal subscription level assignments each time the network state changed, and *ii*) knowledge of all system parameters.

The optimal admission control policies have their own implementation difficulties, but they are not as insurmountable as those for dynamic adaptation. Optimal admission control requires each link maintain its instantaneous link load, and that each route maintain its optimal volume threshold. The instantaneous link load is defined as $\rho_l(t) = \sum_{i=1}^{n_l(t)} s_{i,l}^{\gamma,\pi_a}$, where $n_l(t)$ is the number of streams traversing link l at time t, and $\mathbf{s}_l^{\gamma,\pi_a} = (s_{i,l}^{\gamma,\pi_a}, i = 1, \ldots, n_l(t))$ is the subscription level of each active

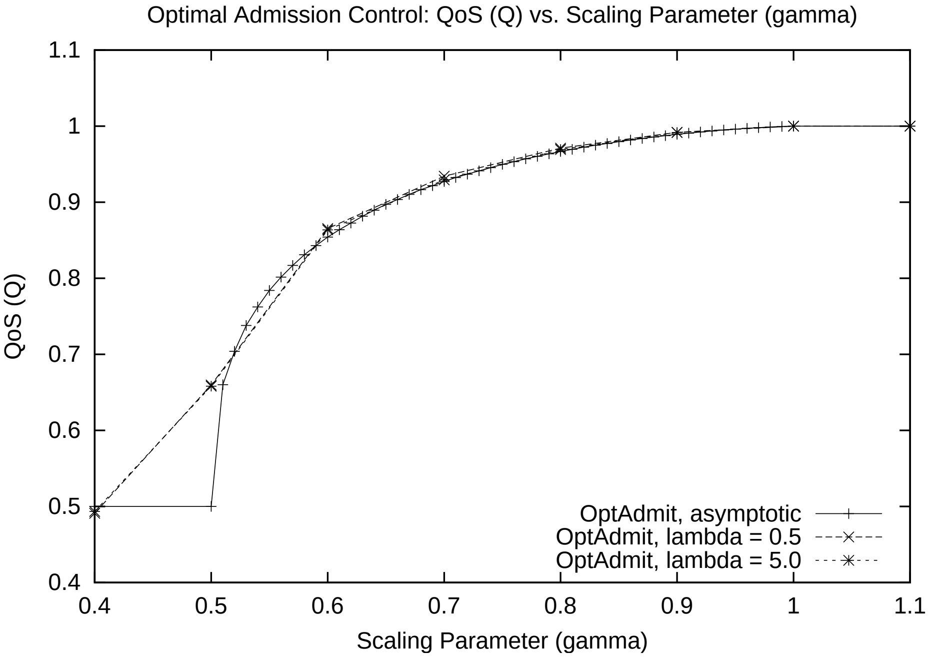

Figure 4.4: Plot of the asymptotic QoS, q^{π_a}, versus the scaling parameter γ.

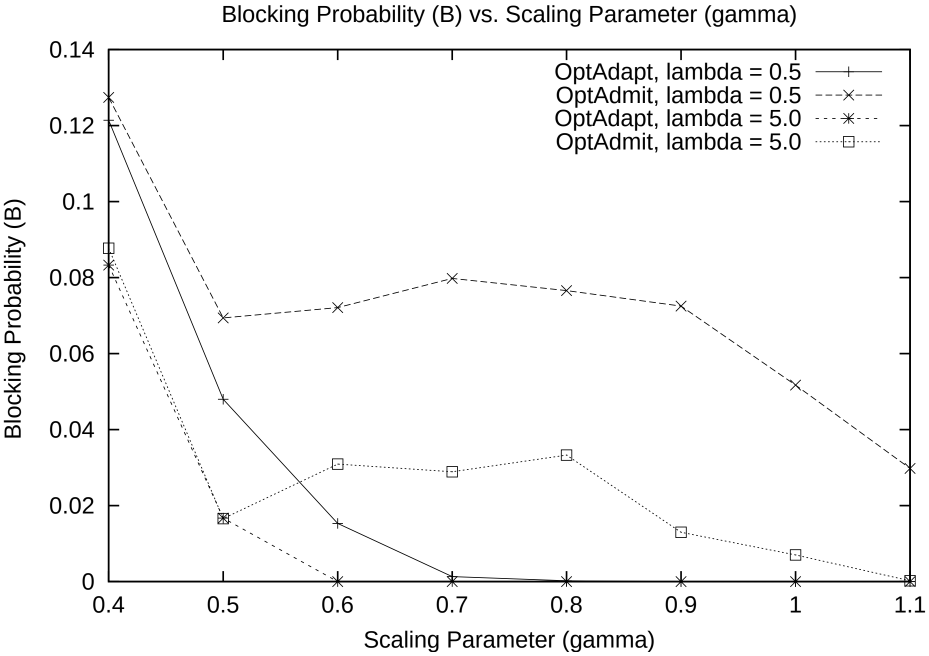

Figure 4.5: Plot of the blocking probabilities under policies π_k and π_a versus the scaling parameter γ.

stream under the optimal admission control. The optimal route volume threshold, denoted v_r^{γ,π_k}, must be identified for each route using the knowledge of the stream volume distribution, F_V, the effective network scaling vector $\boldsymbol{\gamma} = (\gamma_l, l \in \mathcal{L})$, and the average route arrival rates $\boldsymbol{\lambda} = (\lambda_r, r \in \mathcal{R})$. A stream (i,r) requesting service on route r with minimum and maximum subscription levels, $a_{i,r}s_{i,r}$ and $s_{i,r}$, and volume $v_{i,r}$ would first identify its appropriate subscription level as

$$s_{i,r}^{\gamma,\pi_a} = \begin{cases} s_{i,r}, & v_{i,r} < v_r^{\gamma,\pi_a} \\ a_{i,r}s_{i,r}, & \text{else} \end{cases} . \tag{4.9}$$

It would then query each link $l \in r$ and check to see that there is sufficient capacity available at each link comprising the route, i.e.,

$$\rho_l(t) + s_{i,r}^{\gamma,\pi_a} \leq c_l, \forall l \in r. \tag{4.10}$$

If sufficient capacity is available along each link comprising the route then the stream is admitted, else the stream is blocked.

The implementation difficulties with the above approach are as follows.

- Link state needs to be maintained. Note this is much less restrictive than the requirement that network wide state needs to be maintained at a centralized server, but it still poses a significant challenge.

- The optimal admission control policy achieves its high quality of service by assuming stream distributional properties as well as route arrival rates and link capacities are available. These are used to compute the optimal route volume thresholds.

- The thresholds are optimal only when the network traffic is stationary. This assumption is especially troublesome seeing as network traffic is known to exhibit non-stationarities on multiple time scales.

Chapter 5

Distributed Algorithms

Optimal adaptation is not practical because it requires a centralized server to collect network wide state information and disseminate subscription level assignments each time the network state changes. Optimal admission control may not be practical because it requires that *i*) link state information be maintained, *ii*) stream and network distributional properties are known, and *iii*) network traffic is stationary.

To overcome this drawback we propose a class of distributed algorithms which attempt to attain the QoS obtained under optimal adaptation and admission control but without making any assumptions regarding network state, knowledge of distributional properties, or traffic stationarity. These algorithms are similar in spirit to the receiver driven layered multicast (RLM) algorithms proposed in [17], where streams react to congestion signals from the network by reducing their subscription level, and periodically probe the network to determine if sufficient capacity is available to increase their subscription level. Our algorithms differs from RLM in that the size of the subscription level changes and the time between probes depends on stream volume. In particular, small volume streams are more aggressive in pursuing available bandwidth than are large volume streams.

We define a class of volume independent distributed algorithms in Section

5.1, and a class of volume dependent distributed algorithms in Section 5.2. Finally, we present some simulation results illustrating the quality of service under both classes of algorithms in Section 5.3.

5.1 Volume independent distributed algorithms

The volume independent algorithm is presented as Algorithm 1. A client sets its initial subscription level to one of the intermediate subscription levels offered by the content provider, i.e., the subscription level closest to $\frac{a_{i,r}s_{i,r}+s_{i,r}}{2}$. The client then starts a timer τ. If the client receives no congestion signal from the network for τ seconds then the client increases the subscription level to the next largest one available, and resets the timer. If at any time the client does receive a congestion signal from the network it immediately decreases its subscription level to the next smallest one available, and resets the timer. Thus, the timer corresponds to the time that must pass since the last subscription level *change* before the client may subscribe to a higher subscription level.

Our definition of congestion signal is loose. Congestion signals could be taken as packet loss or receiving packets with the explicit congestion notification bit (ECNB) set in the TCP header. The actual implementation of congestion signaling is not important for our purposes.

Algorithm 1 Pseudo code for volume independent adaptation algorithm.

```
set initial subscription level to intermediate level
while active do
  if $\tau$ seconds passed since last subscription change then
    increase subscription by one level
  end if
  if receive congestion notification from network then
    decrease subscription by one level
  end if
end while
```

5.2 Volume dependent distributed algorithms

The volume dependent algorithm is presented as Algorithm 2. We presume each stream has knowledge of its own volume, v, as well as knowledge of the mean stream volume $\mathbb{E}[V]$. For this section we define *small* streams as those streams having volumes smaller than the mean, i.e., $v < \mathbb{E}[V]$, and *large* streams as those streams having volumes larger than the mean, i.e., $v > \mathbb{E}[V]$.

The volume dependent algorithm differs from the volume independent algorithm in that *i*) small and large streams have different initial subscription levels, *ii*) the timer $\tau(s,d)$ depends on the stream volume, and *iii*) small and large streams increase and decrease their subscription levels differently. The idea is for small streams to be more aggressive than large streams in pursuing available capacity, and for small streams to be more reluctant than large streams in reducing their subscription level during congestion. Small streams start out by setting their initial subscription level to the maximum subscription level offered by the provider, while large streams set their initial subscription level to the corresponding minimum subscription level. Each stream maintains a timer $\tau(s,d)$ which functions identically to the timer τ described in the volume independent algorithm. The difference is that the value of the timer depends on the stream volume. Small streams use a short timer while large streams use a long timer. This permits small streams to attempt subscription level increases more frequently than large streams. When a small stream receives congestion notification from the network it decreases its current subscription level to the next lower subscription level, while a large stream decreases its subscription level to the lowest offered subscription level. Thus small streams are reluctant to give up their subscription levels while large streams do so readily. Finally, when the timer indicates a stream may increase its subscription level, small streams increase their current subscription levels to the maximum subscription level, while large streams increase their current subscription levels to the next larger one offered. Thus by combining differentiated initial subscription levels, timers, congestion response, and subscription increase behavior, we are able to achieve the volume discrimination seen under the optimal adaptation and admission policies.

Algorithm 2 Pseudo code for volume dependent adaptation algorithm.

```
if $v < \mathbb{E}[V]$ then
    set initial subscription level as high as possible
else
    set initial subscription level as low as possible
end if
while active do
    if $\tau(s, d)$ seconds passed since last subscription change then
        if $v < \mathbb{E}[V]$ then
            increase subscription level as much as possible
        else
            increase subscription level as little as possible
        end if
    end if
    if receive congestion notification from network then
        if $v < \mathbb{E}[V]$ then
            decrease subscription level as little as possible
        else
            decrease subscription level as much as possible
        end if
    end if
end while
```

Figures 5.2 and 5.1 illustrate the behavior of the volume independent and the volume dependent distributed adaptation algorithms. Figure 5.2 shows a stream with three subscription levels available. The client starts out at its intermediate subscription level. Some τ seconds pass without receiving congestion notification from the network so the stream increases its subscription level to the next highest level. Then a sequence of two congestion signals force the client to reduce its subscription level to the intermediate, and then to the lowest subscription level. Two intervals of length τ pass after each of which the client increases its subscription level to the intermediate, and then to the maximum subscription level.

Figure 5.1 shows a large client and a small client, each of which has three subscription levels available. The large client sets the initial subscription level to the lowest subscription level, waits some long period of time $\tau(v)$ before increasing to the intermediate subscription level, and then another long period of time $\tau(v)$ before increasing to the maximum subscription level. The large client then receives a congestion signal at which time it reduces its subscription level back to the minimum one offered.

The bottom figure in Figure 5.1 shows that a small client sets its initial subscription level to the maximum subscription level. The client receives two congestion signals, after each of which it reduces its subscription level by one. The client then waits some short period of time $\tau(v)$ before increasing its subscription level back to the maximum.

5.3 Simulation results for distributed algorithms

The simulation results in this section use the same distributions for stream duration, maximum subscription level, and adaptivity are as before. In addition, we assume content providers make five encodings of each stream available, where the intermediate three encodings are equally spaced between the minimum and maximum subscription level. In particular, the set of offered subscription levels is assumed to take the form $\mathcal{S} = \{AS, \frac{S}{4}(3A+1), \frac{S}{2}(A+1), \frac{S}{4}(A+3), S\}$.

Implementing the algorithms also requires setting the parameter τ for the volume independent algorithm and the function $\tau(s,d)$ for the volume dependent algorithm. The timer for the volume dependent algorithm is set using the equation $\tau(s,d) = \frac{\kappa v d}{\mathbb{E}[V]}$, with κ a constant. This formula implies the number of subscription increase attempts (disregarding congestion resetting the timers) is inversely proportional to the stream volume, i.e., $\frac{d}{\tau(s,d)} = \frac{\mathbb{E}[V]}{\kappa v}$. We chose $\kappa = 0.0277$,

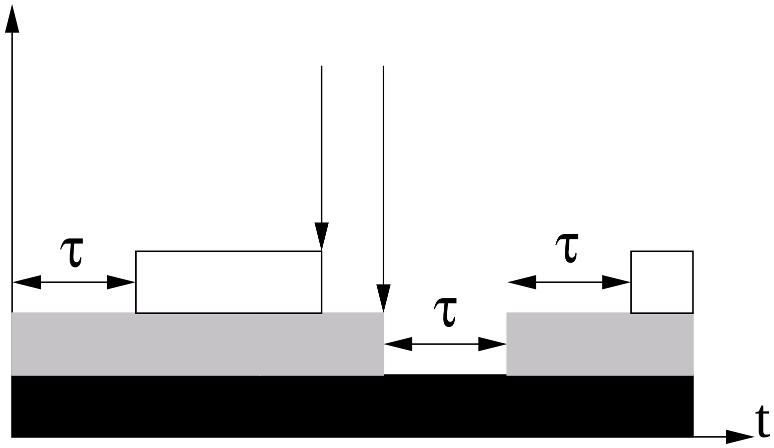

Figure 5.1: Illustration of volume independent adaptation algorithm.

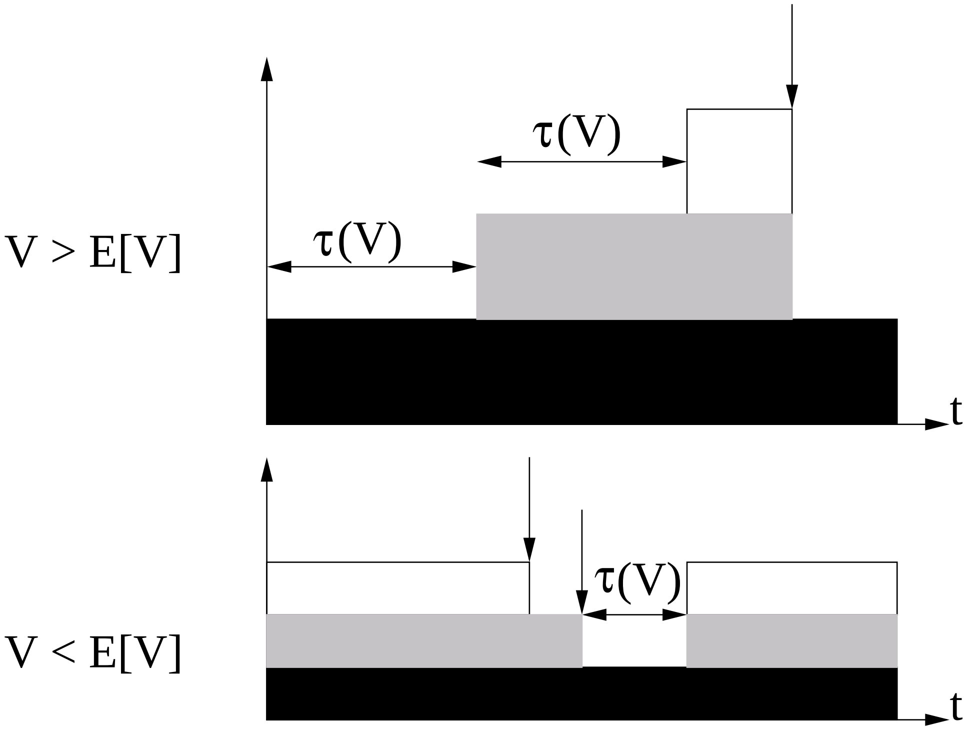

Figure 5.2: Illustration of volume dependent adaptation algorithm.

which, for the bounded exponential distributions for D used for simulation yields $\mathbb{E}[\tau(S, D)] = \kappa \frac{\mathbb{E}[D^2]}{\mathbb{E}[D]} = 10$ seconds. To keep the comparison fair, we chose $\tau = 10$ seconds for the volume independent algorithm.

Figure 5.3 and 5.3 show histograms of the time-average normalized subscription level and the rate of adaptation versus stream volume for the volume independent algorithm and the volume independent algorithm. Note that the x axis is on a logarithmic scale. The stream volume PDF, $f_V(v)$ is also plotted. Recall the mean volume is $\mathbb{E}[V] = 54$ Mb. Also recall the average adaptivity is $\mathbb{E}[A] = \alpha = \frac{1}{2}$, so the range of feasible expected time-average normalized subscription levels is $\mathbb{E}^0[Q] \in [\alpha, 1] = [\frac{1}{2}, 1]$. The plots are simulations of a single link with arrival rate $\lambda = 5.0$ and capacity $c = 202.5$ Mb/s. Simulations were run for about 15,000 clients. The average number of active streams is $\mathbb{E}[N(t)] = \lambda\delta = 5.0 \times 180 = 900$. The capacity corresponds to scaling the network with a capacity scaling parameter $\gamma = \frac{c}{\lambda\delta\sigma} = \frac{202.5}{5.0 \times 180 \times 0.3} = \frac{3}{4}$. The histogram bins were constructed as described for Figures 3.6 and 3.7.

Figure 5.3 shows the volume dependent algorithm achieves the goal of granting higher subscription levels to smaller volume streams and lower subscription levels to larger volume streams. The volume independent algorithm is seen to yield an expected time-average normalized subscription level that is indeed independent of the stream volume. Comparing this plot with Figure 3.6, which uses the same simulation parameters to plot the expected time-average normalized subscription level under the optimal and fair share adaptation policies, shows that the volume dependent algorithm functions similarly to the optimal adaptation policy, while the volume independent algorithm functions similarly to the fair share adaptation policy.

Figure 5.3 shows that the rate of adaptation is acceptably small for almost all of the streams. Large volume streams, which are correlated with long duration streams, have a higher rate of adaptation under both algorithms because they are afforded more opportunities to increase or decrease their subscription levels.

We next present some simulation results illustrating the expected time average normalized subscription level as a function of the link capacity scaling parameter γ for the two algorithms. Figure 5.3 shows this quantity for the volume dependent and the volume independent algorithms. The asymptotic optimal QoS, $q^{\gamma,\pi_k} = q^{\gamma,\pi_a}$ is also shown for purposes of comparison. The figure illustrates the volume dependent algorithm gets fairly close to the optimal QoS, while the volume independent algorithm lags behind. The figure illustrates that the volume independent algorithm is roughly linear in the scaling parameter for $\alpha \leq \gamma \leq 1$, while the volume dependent algorithm exhibits a concave increase in QoS in this regime. The label indicates that for $\gamma = 0.6$ the volume dependent algorithm achieved a QoS 16.8% greater than that achieved by the volume independent algorithm, which corresponds to a performance improvement of 28%.

5.4 Implementation issues for distributed volume dependent adaptation algorithms

We present these algorithms as mere proofs of concept, and without any claims of sound engineering design with respect to parameter choice. We feel our experiments lend credence to the point, however, that making subscription adaptation algorithms volume dependent can achieve a moderate increase in QoS over volume independent algorithms. In principle, implementation is fairly straightforward. Some estimate of the expected stream volume, $\mathbb{E}[V]$ would need to be included in client-side code in order to allow client software to determine if the current stream is small or large. In addition, the timer functions $\tau(s, d)$ should be set to yield an acceptable rate of subscription level increase attempts. Finally, a viable congestion notification mechanism would need to be implemented.

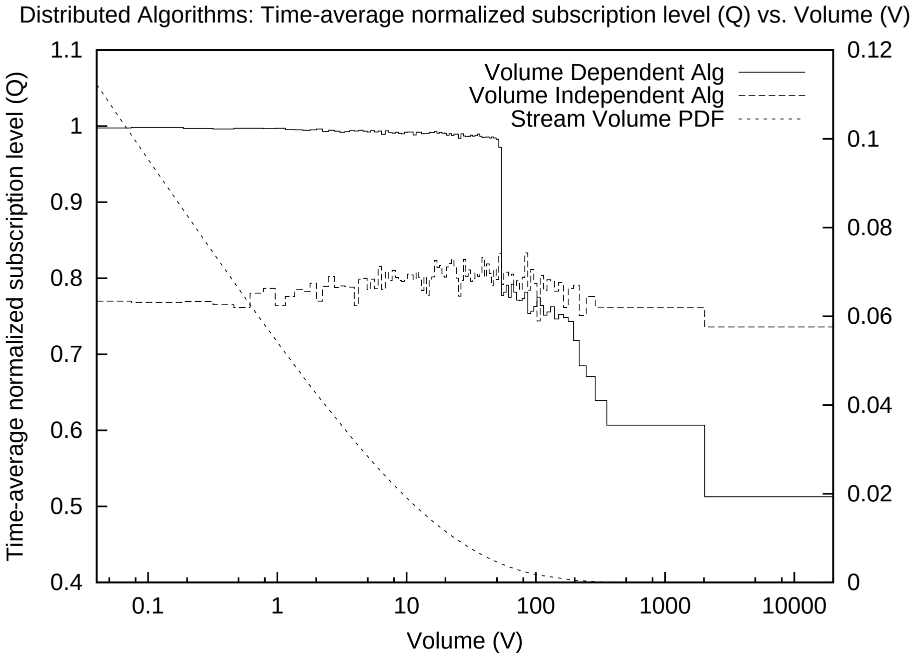

Figure 5.3: Histogram of $\mathbb{E}^0[Q \mid V = v]$ versus v for the volume independent and volume dependent distributed adaptation algorithms.

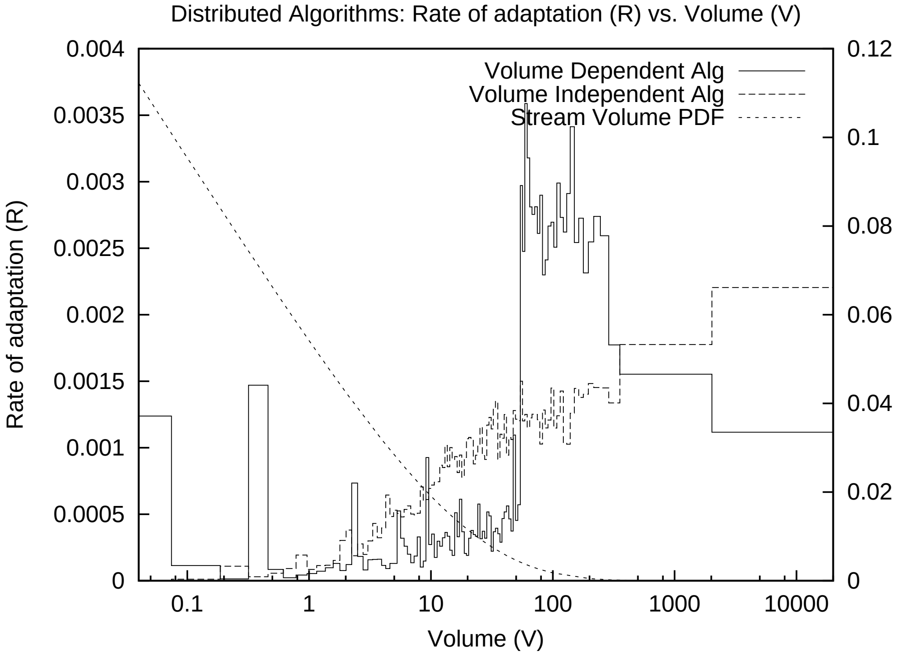

Figure 5.4: Histogram of $\mathbb{E}^0[R \mid V = v]$ versus v for the volume independent and volume dependent distributed adaptation algorithms.

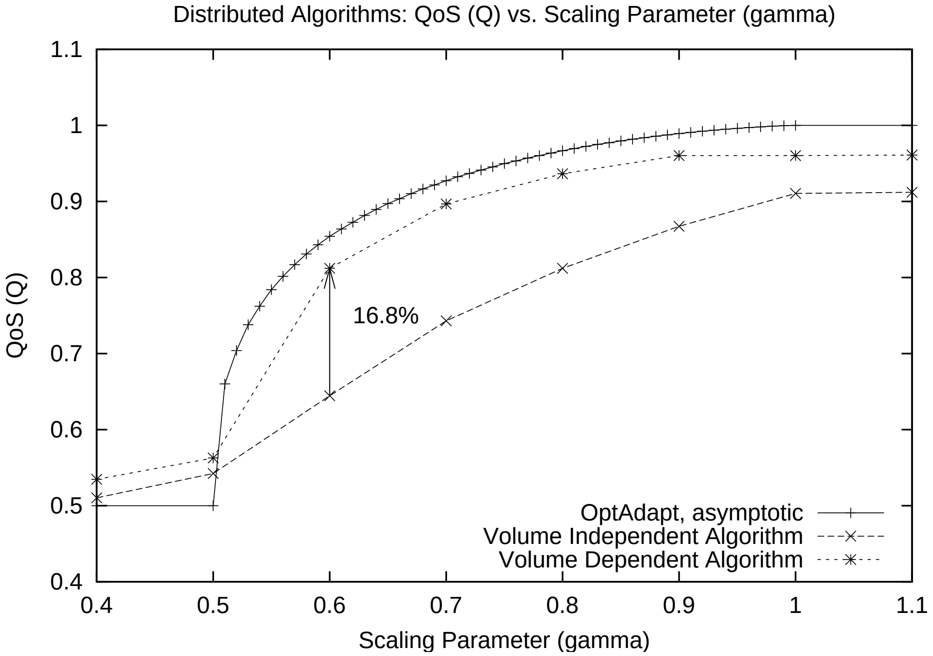

Figure 5.5: Simulation results of $\mathbb{E}^0[Q]$ under the two distributed adaptation algorithms versus the capacity scaling parameter γ.

Chapter 6

Multiple Service Classes

The previous three chapters on optimal adaptation, optimal admission control, and volume dependent distributed algorithms share the characteristic that the client average normalized subscription level is maximized by granting higher subscription levels to small volume streams and lower subscription levels to large volume streams. The intuition is clear: we can maximize a client average by granting higher service to clients requesting fewer network resources and lower service to clients requesting larger network resources.

We emphasize the fact that *all* clients receive satisfactory QoS under volume discrimination. That is, even though large volume streams receive a lower normalized subscription level than small volume streams, they still are guaranteed an *acceptable* stream resolution. We have defined the minimally acceptable stream resolution in terms of the *content provider's* estimation of minimally acceptable stream quality. In practice different clients may have different subjective definitions of what constitutes minimally acceptable quality. That is, instead of defining the minimally acceptable stream resolution in terms of the maximum compression of media information using a given encoding algorithm, it might actually be more realistic to define the minimally acceptable stream resolution on a client by client basis. We might abstract this notion by saying client (i, r) has a minimally acceptable adaptivity parameter $a_{i,r}$, which can be interpreted as a client's individual *QoS requirement*.

To emphasize this distinction between provider and client definitions of minimally acceptable stream resolutions, consider the example of a content provider offering a certain media stream with adaptivity a and maximum subscription level s. The adaptivity a is such that the provider's assessment of the minimally acceptable stream encoding has an average instantaneous rate of as. Now consider two clients, denoted i and j, with individual adaptivity parameters a_i and a_j, such that $a < a_i < a_j < 1$. Both clients have a subjective minimum adaptivity exceeding that decided upon by the content provider; client i requires a minimum subscription level of $a_i s$ and client j requires a minimum subscription level of $a_j s$. The point is that optimal adaptation, which gives large volume streams the minimally acceptable stream resolution, as determined by the content provider, may prove unsatisfactory to clients with more stringent QoS requirements.

In this chapter we propose a network architecture comprising K service classes, where each service class k carries a distinct normalized subscription level α_k. Streams selecting service class k are guaranteed a minimum normalized subscription level of α_k, but may receive higher subscription levels if sufficient capacity is available. For simplicity we will restrict our attention in this chapter to the single link case, but the development may be generalized to an arbitrary network. We also assume the link admits as many clients as possible while still respecting all of the QoS guarantees of the already admitted clients. Note that we assume clients are allowed to *select* their own service class, which means we will need to price the service classes in order to ensure client selection corresponds to actual client QoS requirements.

Formally, we assume a price vector $\mathbf{u} = (u_k, k = 1, \ldots, K)$, and a set of demand functions $\boldsymbol{\lambda}(\mathbf{u}, \boldsymbol{\alpha}) = (\lambda_k(\mathbf{u}, \boldsymbol{\alpha}), k = 1, \ldots, K)$, where $\lambda_k(\mathbf{u}, \boldsymbol{\alpha})$ is the mean arrival rate of streams requesting service in class k under a price vector $\mathbf{u}$ and service class QoS guarantees $\boldsymbol{\alpha}$. We take the price vector and service class QoS guarantee vector to be exogenous parameters, hence the arrival rates for each class are fixed. We will therefore write the per class arrival rates as $\boldsymbol{\lambda} = (\lambda_k, k = 1, \ldots, K)$.

We are interested in identifying the optimal adaptation policy for a multiple class network architecture, where our criterion for optimality is still to maximize the *overall* client average normalized subscription level. Thus it will still be optimal to discriminate among streams based on volume. Note that, from a client perspective, a small volume client will be more likely to choose a service class with a less stringent QoS guarantee, while a large volume client will be more likely to choose a service class with a more stringent QoS guarantee. That is, small volume clients, who stand to benefit from volume discrimination, do not require the additional QoS protection afforded by multiple service classes, while large volume clients, who stand to suffer under volume discrimination, *will* require the additional QoS protection.

To model this fact we assume that the stream volume distribution is class-specific. Formally, we model a service class k by specifying a distribution F_{D_k} and F_{S_k} on the duration and maximum subscription level for class k streams. This yields a class-specific volume distribution

$$F_{V_k}(v) = \int_0^\infty F_{S_k}(\frac{v}{d}) dF_{D_k}(d) = \int_0^\infty F_{D_k}(\frac{v}{s}) dF_{S_k}(s). \tag{6.1}$$

We label the class specific average durations and maximum subscription levels as $\mathbb{E}[D_k] = \delta_k$ and $\mathbb{E}[S_k] = \sigma_k$. In summary, a class k consists of:

- a guarantee on the normalized subscription level, α_k,
- an arrival rate, λ_k,
- a duration distribution F_{D_k} with mean $\mathbb{E}[D_k] = \delta_k$, and
- a maximum subscription level distribution F_{S_k} with mean $\mathbb{E}[S_k] = \sigma_k$.

In Section 6.1 we generalize our link capacity scaling from Section 3.3 to handle multiple service classes. Section 6.2 will identify the optimal adaptation policy for multiple service classes and identify the asymptotic per-class client average normalized subscription level. In Section 6.3 we identify the asymptotically optimal admission control policy for multiple service classes. Finally, Section 6.4 will present a computation based case study illustrating how the QoS with a multiple service class network architecture differs from the QoS with a single service class architecture.

6.1 Capacity scaling for multiple service classes

Let $\mathbf{N}(t) = (N_k(t), k = 1, \ldots, K)$ denote the instantaneous number of streams active at time t in each service class. We denote the minimum and maximum offered loads for class k as $\underline{\rho}_k = (\lambda_k \delta_k)(\alpha_k \sigma_k)$ and $\bar{\rho}_k = (\lambda_k \delta_k)\sigma_k$ respectively. Thus, the minimum offered load for class k is the product of the average number of class k streams, i.e., $\mathbb{E}[N_k(t)] = \lambda_k \delta_k$ (in a low blocking regime), times the class k average minimum subscription level $\alpha_k \sigma_k$. Similarly, the maximum offered load is the product of the average number of class k streams times the class k average maximum subscription level. We define the overall minimum and maximum offered loads as $\underline{\rho} = \sum_{k=1}^{K} \underline{\rho}_k$ and $\bar{\rho} = \sum_{k=1}^{K} \bar{\rho}_k$ respectively. We define the overall adaptivity as the ratio of the overall minimum offered load over the overall maximum offered load, i.e., $\alpha = \underline{\rho}/\bar{\rho}$.

Consider a sequence of links, indexed by m, where we linearly scale the the arrival rate vector and the link capacity as follows. The arrival rate vector on the m^{th} link is $\boldsymbol{\lambda}(m) = (\lambda_k(m), k = 1, \ldots, K)$ and $\lambda_k(m) = m\lambda_k$. We denote the per class and overall minimum and maximum offered loads on the m^{th} link as $\underline{\rho}_k(m), \bar{\rho}_k(m), \underline{\rho}(m), \bar{\rho}(m)$ respectively. Note that the overall adaptivity α is independent of m under this scaling. The link capacity on the m^{th} link is $c(m) = \gamma \bar{\rho}(m)$ for some $\gamma > 0$.

We define three capacity scaling regimes, parameterized by γ, analogous to the three regimes defined for the single service class scaling. These regimes are, as before, i) the overloaded regime, parameterized by $\gamma < \alpha$, ii) the rate adaptive regime, parameterized by $\alpha \leq \gamma \leq 1$, and iii) the underloaded regime, parameterized by $\gamma > 1$. In the overloaded regime the provisioned capacity is inadequate to handle the overall minimum offered load, i.e., $c(m) = \gamma\bar{\rho}(m) < \alpha\bar{\rho}(m) = \underline{\rho}(m)$. In the underloaded regime the provisioned capacity exceeds the overall maximum offered load, i.e., $c(m) = \gamma\bar{\rho}(m) > \bar{\rho}(m)$. Finally, in the rate adaptive regime the provisioned capacity lies between the overall minimum and maximum offered loads. As before, the rate adaptive regime is the primary regime of interest.

6.2 Optimal adaptation policy for multiple service classes

In this section we identify the optimal adaptation policy for multiple service classes. Our objective, as before, is to maximize the *overall* client average time-average normalized subscription level, $\mathbb{E}^0[Q]$. The following theorem states the optimal adaptation policy is to sort all the active streams by stream volume and grant the maximum subscription level to as many small volume streams as possible while respecting the per-class minimum subscription level guarantees. We denote the optimal multi-class adaptation policy as π_m. We denote the i^{th} active stream of class k as (i, k). We write $\mathbf{n}(t) = (n_k(t), k = 1, \ldots, K)$ for the number of active streams in each class, and $n(t) = \sum_{k=1}^{K} n_k(t)$ for the total number of active streams on the link. We denote the instantaneous subscription level assignment under the optimal multi-class adaptation policy as $\mathbf{s}^{\pi_m}(t) = (s_{i,k}^{\pi_m}(t), i = 1, \ldots, n_k(t), k = 1, \ldots, K)$. We assume that content providers match their offered subscription levels to the set of service classes offered by the network, i.e., $\mathcal{S} = (\alpha_1 s, \ldots, \alpha_K s, s)$.

Theorem 7 *The optimal multi-class adaptation policy, π_m, which maximizes the overall client average time-average normalized subscription level, $\mathbb{E}^0[Q]$, is the instantaneous allocation, $\mathbf{s}^{\pi_m}(t) = (s_{i,k}^{\pi_m}(t), i = 1, \ldots, n_k(t), k = 1, \ldots, K)$, which* *solves the following integer programming problem:*

$$\begin{aligned} \max_{\mathbf{s}(t)} \quad & q_{agg}(t) = \sum_{k=1}^{K} \sum_{i=1}^{n_k(t)} \frac{s_{i,k}(t)}{s_{i,k} d_{i,k}} \qquad (6.2) \\ s.t. \quad & \sum_{k=1}^{K} \sum_{i=1}^{n_k(t)} s_{i,k}(t) \le c, \\ & s_{i,k}(t) \in \mathcal{S}_{i,k}, \forall i = 1, \ldots, n_k(t), k = 1, \ldots, K. \end{aligned}$$

There exists a near optimal multi-class adaptation policy, denoted $\tilde{\pi}_m$, with an instantaneous allocation $\mathbf{s}^{\tilde{\pi}_m}(t)$ with $s_{i,k}^{\tilde{\pi}_m}(t) \in \{\alpha_k s_{i,k}, s_{i,k}\}$ for each $i = 1, \ldots, n_k(t)$ and each $k = 1, \ldots, K$. In particular,

$$\frac{q^{\pi_m}(t) - q^{\tilde{\pi}_m}(t)}{q^{\pi_m}(t)} \le \frac{\kappa_m}{n(t)}, \qquad (6.3)$$

for $\kappa_m < \infty$.

Sort the active streams by volume, indexed by j, so that stream (i,k) is labeled j if stream (i,k) has the j^{th} smallest volume out of all the active streams. Thus, $s_1 d_1 < \ldots < s_{n(t)} d_{n(t)}$. The allocation under the near optimal multi-class adaptation policy is given by

$$s_j^{\tilde{\pi}_m}(t) = \begin{cases} s_j, & j = 1, \ldots, \bar{n} - 1 \\ a_j s_j, & j = \bar{n}, \ldots, n(t) \end{cases}, \qquad (6.4)$$

where $\bar{n}$ equals

$$\bar{n} = \max\{m \mid \sum_{j=1}^{m-1} s_j + \sum_{j=m}^{n(t)} a_j s_j \le c\}. \qquad (6.5)$$

Proof of Theorem 7. Let $B_{i,k}$ denote the random arrival time of stream i on class k. Define the *instantaneous QoS* of stream (i,k) as

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

Define the *class k aggregate instantaneous QoS* at time t as

$$Q_{agg,k}(t) = \sum_{k=1}^{N_k(t)} Q_{i,k}(t).$$

Define the *class k expected aggregate instantaneous QoS* as

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

where the RHS follows by ergodicity, and the t in the LHS is understood to be a typical time. Define the *class k client average QoS* as

$$\mathbb{E}^0[Q_k] = \lim_{n_k\to\infty} \frac{1}{n_k}\sum_{i=1}^{n_k}\int_{-\infty}^{\infty} Q_{i,k}(t)dt,$$

where the RHS again follows by ergodicity. Straightforward application of Brumelle's Theorem [32] shows that

$$\mathbb{E}[Q_{agg,k}(t)] = \lambda_k^a \mathbb{E}^0[Q_k],$$

where λ_k^a is the rate at which class k clients are admitted onto the link.

Define the *overall aggregate instantaneous QoS* at time t as

$$Q_{agg}(t) = \sum_{k=1}^{K} Q_{agg,k}(t).$$

The *expected aggregate instantaneous QoS* is defined as

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

where the RHS follows by ergodicity and t in the LHS is again understood to be a typical time. Define the *overall client average QoS* as

$$\mathbb{E}^0[Q] = \sum_{k=1}^{K} \frac{\lambda_k^a}{\lambda^a}\mathbb{E}^0[Q_k],$$

where $\lambda^a = \sum_{k=1}^{K} \lambda_k^a$ denotes the overall admission rate. This implies

$$\mathbb{E}[Q_{agg}(t)] = \sum_{k=1}^{K} \mathbb{E}[Q_{agg,k}(t)] = \sum_{k=1}^{K} \lambda_k^a \mathbb{E}^0[Q_k] = \lambda^a \mathbb{E}^0[Q].$$

The conclusion is that maximizing the overall client average QoS is equivalent to maximizing the overall aggregate instantaneous QoS.

We define the filtration $\sigma(t)$ representing the information available at time t, which includes stream arrival times, stream durations, stream maximum subscription levels, and stream adaptivities. Maximizing the overall aggregate instantaneous QoS at each time t means our objective can be written

$$q_{agg}(t) = \mathbb{E}[Q_{agg}(t) \mid \sigma(t)] = \sum_{k=1}^{K} \sum_{n=1}^{n_k(t)} \frac{s_{i,k}(t)}{s_{i,k} d_{i,k}}.$$

Applying the capacity constraints and the stream constraints means that the optimal adaptation policy for multiple service classes is the solution, at each time t, of the following integer programming problem:

$$\begin{array}{ll} \max\limits_{\mathbf{s}(t)} & q_{agg}(t) = \sum\limits_{k=1}^{K} \sum\limits_{i=1}^{n_k(t)} \dfrac{s_{i,k}(t)}{s_{i,k} d_{i,k}} \\ s.t. & \sum\limits_{k=1}^{K} \sum\limits_{i=1}^{n_k(t)} s_{i,k}(t) \leq c, \\ & s_{i,k}(t) \in \mathcal{S}_{i,k}, \forall i = 1, \ldots, n_k(t), k = 1, \ldots, K \end{array}$$

This proves the first part of the theorem.

Consider a relaxation of the above problem, where we relax the assumption that the set of offered subscription levels is a discrete set. In particular, we assume $\mathcal{S}_{i,k} = [\alpha_k s_{i,k}, s_{i,k}]$. We apply the change of variables $x_{i,k}(t) = \frac{s_{i,k}(t) - \alpha_k s_{i,k}}{1 - \alpha_k}$ and the constraint relaxation to obtain

$$\begin{array}{ll} \max\limits_{\mathbf{x}(t)} & \sum\limits_{k=1}^{K} \sum\limits_{i=1}^{n_k(t)} \dfrac{(1 - \alpha_k) x_{i,k}(t)}{s_{i,k} d_{i,k}} \\ s.t. & \sum\limits_{k=1}^{K} \sum\limits_{i=1}^{n_k(t)} (1 - \alpha_k) x_{i,k}(t) \leq c', \\ & 0 \leq x_{i,k}(t) \leq 1, \forall i = 1, \ldots, n_k(t), k = 1, \ldots, K \end{array}$$

where $c' = c - \sum_{k=1}^{K} \sum_{i=1}^{n_k(t)} \alpha_k s_{i,k}$. This is seen to be a relaxation of a knapsack problem, where the item values are $v_{i,k} = \frac{1 - \alpha_k}{s_{i,k} d_{i,k}}$ and the item weights are $w_{i,k} = 1 - \alpha_k$. The solution is to sort items by maximum value per unit weight, i.e.,

$\frac{v_{i,k}}{w_{i,k}} = \frac{1}{s_{i,k}d_{i,k}}$. Thus, the solution of the relaxation is to sort streams by their volume.

As was shown in the proof of Theorem 1, the solution to the relaxed problem has the characteristic that at most one stream receives a subscription level intermediary between its minimum and maximum subscription level. A little thought shows that $\bar{n}$ in (6.5) identifies that stream. Let $G(\mathbf{s}^{\tilde{\pi}_m}(t))$ denote the value of the objective under the allocation given in (6.4). Let $G(\mathbf{s}^{\pi_m}(t))$ denote the value of the objective under the allocation solving the (non-relaxed) integer programming problem stated in the theorem. Finally, let $G(\mathbf{s}(t))$ denote the value of the objective under the allocation solving the relaxed linear programming problem stated above. Note that $G(\mathbf{s}^{\tilde{\pi}_m}(t)) < G(\mathbf{s}^{\pi_m}(t)) < G(\mathbf{s}(t))$ since $\mathbf{s}^{\tilde{\pi}_m}(t)$ is in the feasible set of both the relaxed and non-relaxed problem, and since the relaxed problem necessarily has a higher objective value than the non-relaxed problem. We can therefore bound the difference in the objective under allocations $\tilde{\pi}_m$ and π_m as

$$G(\mathbf{s}^{\pi_m}(t)) - G(\mathbf{s}^{\tilde{\pi}_m}(t)) \leq G(\mathbf{s}(t)) - G(\mathbf{s}^{\tilde{\pi}_m}(t)) \leq \bar{s},$$

where $\bar{s}$ is the maximum possible subscription level. Thus the difference in the objective is a constant. The relative difference is easily shown to be bounded as

$$\frac{G(\mathbf{s}^{\pi_m}(t)) - G(\mathbf{s}^{\tilde{\pi}_m}(t))}{G(\mathbf{s}^{\pi_m}(t))} \leq \frac{\bar{d}\bar{s}}{\underline{\alpha}n(t)} = \frac{\kappa_m}{n(t)},$$

where $\bar{d}$ is the maximum possible stream duration and $\underline{\alpha} = \min_{1\leq k\leq K}\{\alpha_k\}$. Thus, the difference goes to zero for links servicing large numbers of streams.

■

The theorem just confirms the intuition that offering streams minimum quality of service guarantees doesn't change the form of the optimal adaptation policy. This is because the optimal adaptation policy weights streams by their maximum volume irrespective of their individual adaptivities.

The next theorem identifies the asymptotic quality of service under the optimal adaptation policy for each service class. We define the asymptotic expected time-average normalized subscription level for class k streams under the multi-class capacity scaling as $q_k^{\gamma,\pi_m} = \lim_{m\to\infty} \mathbb{E}^0[Q_k^{m,\pi_m}]$, where $\mathbb{E}^0[Q_k^{m,\pi_m}]$ is the expected value of Q on the m^{th} link under policy π_m for class k.

Theorem 8 *The asymptotic expected time-average normalized subscription level for class k streams under the optimal adaptation policy for multiple service classes is $q_k^{\gamma,\pi_m} = \alpha_k$ for $\gamma < \alpha$, $q_k^{\gamma,\pi_m} = 1$ for $\gamma > 1$, and*

$$q_k^{\gamma,\pi_m} = 1 - (1-\alpha_k)\int_0^\infty \int_0^\infty \mathbb{I}\Big(\sum_{k'=1}^{K}(\bar{\rho}_{k'} - \underline{\rho}_{k'})F_{\hat{V}_{k'}}(sd) > \gamma\bar{\rho} - \underline{\rho}\Big)dF_{D_k}(d)dF_{S_k}(s), \tag{6.6}$$

for $\alpha \le \gamma \le 1$.

Proof of Theorem 8. Let Q_k^{m,π_m} denote the QoS of a typical class k stream in the m^{th} scaling of the link capacity under the optimal multi-class adaptation policy π_m. Similarly, let $S_k^{m,\pi_m}(t)$ denote the instantaneous allocation to a typical class k stream at some time t after that stream's admission:

$$q_k^{\gamma,\pi_m} = \lim_{m\to\infty} \mathbb{E}^0[Q_k^{m,\pi_m}] = \lim_{m\to\infty} \mathbb{E}^0[\frac{1}{D}\int_0^D \frac{S_k^{m,\pi_m}(t)}{S}dt].$$

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

$$q_k^{\gamma,\pi_m} = \lim_{m\to\infty}\int_0^\infty \int_0^\infty \mathbb{E}^0[\frac{1}{D}\int_0^D \frac{S_k^{m,\pi_m}(t)}{S}dt \mid D = d, S = s]dF_{D_k}(d)dF_{S_k}(s).$$

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

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

for the t in the RHS understood to be a typical time. This allows

$$q_k^{\gamma,\pi_m} = \lim_{m\to\infty}\int_0^\infty \int_0^\infty \mathbb{E}^0[\frac{S_k^{m,\pi_m}(t)}{S} \mid D = d, S = s]dF_{D_k}(d)dF_{S_k}(s).$$

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

$$\mathbb{E}^0[\frac{S_k^{m,\pi_m}(t)}{S} \mid D = d, S = s] = 1 - (1-\alpha_k)p(m,t,s,d).$$

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

$$q_k^{\gamma,\pi_m} = 1 - (1-\alpha_k)\int_0^\infty \int_0^\infty \lim_{m\to\infty} p(m,t,s,d) dF_{D_k}(d) dF_{S_k}(s).$$

We focus now on $\lim_{m\to\infty} p(m,t,s,d)$. Let $\mathbf{N}(m,t) = (N_k(m,t), k = 1,\ldots,K)$ denote the number of active streams in each class on the m^{th} link at a typical time t. The event that a stream with volume sd is adapted at a typical time t is equivalent to the event

$$\sum_{k=1}^{K}\sum_{i=1}^{N_k(m,t)} S_{i,k}\mathbb{I}(S_{i,k}\hat{D}_{i,k} \le sd) + s + \sum_{k=1}^{K}\sum_{i=1}^{N_k(m,t)} \alpha_k S_{i,k}\mathbb{I}(S_{i,k}\hat{D}_{i,k} > sd) \ge c(m)$$

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

$$= \mathbb{P}(\sum_{k=1}^{K}\sum_{i=1}^{N_k(m,t)} S_{i,k}\mathbb{I}(S_{i,k}\hat{D}_{i,k} \le sd) + s + \sum_{k=1}^{K}\sum_{i=1}^{N_k(m,t)} \alpha_k S_{i,k}\mathbb{I}(S_{i,k}\hat{D}_{i,k} > sd) \ge c(m))$$

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

$$Z(m,t,s,d) = \frac{1}{\bar{\rho}(m)}\Big(\sum_{k=1}^{K}\sum_{i=1}^{N_k(m,t)} S_{i,k}\mathbb{I}(S_{i,k}\hat{D}_{i,k} \le sd) + \sum_{k=1}^{K}\sum_{i=1}^{N_k(m,t)} \alpha_k S_{i,k}\mathbb{I}(S_{i,k}\hat{D}_{i,k} > sd)\Big)$$

so that

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

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

$$\mathbb{E}[Z(m,t,s,d)] = \frac{1}{\bar{\rho}(m)}\mathbb{E}\Big[\sum_{k=1}^{K}\sum_{i=1}^{N_k(m,t)} S_{i,k}\big(\mathbb{I}(S_{i,k}\hat{D}_{i,k} \leq sd)] + \alpha_k \mathbb{I}(S_{i,k}\hat{D}_{i,k} > sd)\big)\Big].$$

By Wald's identity,

$$\mathbb{E}[\sum_{k=1}^{K}\sum_{i=1}^{N_k(m,t)} S_{i,k}\mathbb{I}(S_{i,k}\hat{D}_{i,k} \leq sd)] = \sum_{k=1}^{K}\mathbb{E}[N_k(m,t)]\mathbb{E}[S_k\mathbb{I}(S_k\hat{D}_k \leq sd)].$$

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

$$\begin{aligned}
\mathbb{E}[S_k\mathbb{I}(S_k\hat{D}_k \leq sd)] &= \int_0^\infty\int_0^\infty x\mathbb{I}(xy \leq sd)dF_{\hat{D}_k}(y)dF_{S_k}(x) \\
&= \int_0^\infty x\Big[\int_0^{\frac{sd}{x}} dF_{\hat{D}_k}(y)\Big]dF_{S_k}(x) \\
&= \int_0^\infty x\Big[\int_0^{\frac{sd}{x}} \frac{1}{\mathbb{E}[D_k]}ydF_{D_k}(y)\Big]dF_{S_k}(x).
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}[S_k\mathbb{I}(S_k\hat{D}_k \leq sd)] &= \frac{1}{\mathbb{E}[D_k]}\int_0^\infty\int_0^{sd} zdF_{D_k}(\frac{z}{x})\frac{1}{x}dF_{S_k}(x) \\
&= \frac{1}{\mathbb{E}[D_k]}\int_0^{sd}\Big[\int_0^\infty \frac{z}{x}f_{D_k}(\frac{z}{x})f_{S_k}(x)dx\Big]dz \\
&= \frac{1}{\mathbb{E}[D_k]}\int_0^{sd} z\Big[f_{V_k}(z)\Big]dz \\
&= \frac{\mathbb{E}[V_k]}{\mathbb{E}[D_k]}\int_0^{sd}\frac{z}{\mathbb{E}[V_k]}dF_{V_k}(z) \\
&= \sigma_k F_{\hat{V}_k}(sd).
\end{aligned}$$

A similar argument shows that $\mathbb{E}[\alpha_k S_k\mathbb{I}(S_k\hat{D}_k > sd)] = \alpha_k\sigma_k\bar{F}_{\hat{V}_k}(sd)$. We combine the above results and note that the $m's$ cancel to obtain

$$\mathbb{E}[Z(m,t,s,d)] = \frac{1}{\bar{\rho}}\sum_{k=1}^{K}\bar{\rho}_k(F_{\hat{V}_k}(sd) + \alpha_k\bar{F}_{\hat{V}_k}(sd)).$$

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

$$Z(m,t,s,d) = \frac{1}{\bar{\rho}(m)}\sum_{k=1}^{K}\sum_{i=1}^{N_k(m,t)} W_{i,k}$$

for $W_{i,k} = S_{i,k}(1 - (1-\alpha_k)\mathbb{I}(S_{i,k}\hat{D}_{i,k} \geq sd))$. and thereby obtain

$$\begin{aligned}
Var(Z(m,t,s,d)) &= \frac{1}{(\bar{\rho}(m))^2} Var(\sum_{k=1}^{K} \sum_{i=1}^{N_k(m,t)} W_{i,k}) \\
&= \frac{1}{(\bar{\rho}(m))^2} \sum_{k=1}^{K} Var(\sum_{i=1}^{N_k(m,t)} W_{i,k}) \\
&= \frac{1}{(\bar{\rho}(m))^2} \sum_{k=1}^{K} \Big[\mathbb{E}[N_k(m,t)]Var(W_k) + \mathbb{E}[W_k]^2 Var(N_k(m,t))\Big] \\
&= \frac{1}{(\bar{\rho}(m))^2} \sum_{k=1}^{K} \Big[\lambda_k(m)\delta_k Var(W_k) + \lambda_k(m)\delta_k \mathbb{E}[W_k]^2\Big] \\
&= \frac{1}{(\bar{\rho}(m))^2} \sum_{k=1}^{K} \lambda_k(m)\delta_k \mathbb{E}[W_k^2] \\
&\leq \frac{1}{(\bar{\rho}(m))^2} \sum_{k=1}^{K} \lambda_k(m)\delta_k \mathbb{E}[S_k^2] \\
&= \frac{1}{m(\bar{\rho})^2} \sum_{k=1}^{K} \lambda_k \delta_k \mathbb{E}[S_k^2].
\end{aligned}$$

The second equality follows because the random variables $N_k(m,t)$ and $N_{k'}(m,t)$ are independent (in a low blocking regime).

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

$$\mathbb{P}(Z(m,t,s,d) \geq \gamma - \frac{s}{\bar{\rho}(m)}) \leq \mathbb{P}(|Z(m,t,s,d) - \mathbb{E}[Z(m,t,s,d)]| > \epsilon(m))$$

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

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

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

$$\lim_{m\to\infty} \mathbb{P}(Z(m,t,s,d) \geq \gamma - \frac{s}{\bar{\rho}(m)}) = 0$$

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

$$\lim_{m\to\infty} \mathbb{P}(Z(m,t,s,d) \geq \gamma - \frac{s}{\bar{\rho}(m)}) = 1$$

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

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

Note that $\mathbb{I}(\mathbb{E}[Z(m,t,s,d)] > \gamma)$ is equivalent to

$$\mathbb{I}\Big(\sum_{k=1}^{K} \bar{\rho}_k (F_{\hat{V}_k}(sd) + \alpha_k \bar{F}_{\hat{V}_k}(sd)) > \gamma\bar{\rho}\Big)$$

This is easily simplified to

$$\mathbb{I}\Big(\sum_{k=1}^{K} (\bar{\rho}_k - \underline{\rho}_k) F_{\hat{V}_k}(sd) > \gamma\bar{\rho} - \underline{\rho}\Big).$$

Notice that for $\gamma < \alpha$, i.e., when the link is provisioned in the overloaded regime, the indicator function is satisfied for all values sd. To see this, note that $\gamma < \alpha$ implies $\gamma\bar{\rho} - \underline{\rho} < 0$, and so the indicator requires a sum of positive numbers exceed zero. Thus, $\lim_{m\to\infty} p(m,t,s,d) = 1$. Similarly, for $\gamma > 1$, i.e., when the link is provisioned in the underloaded regime, the asymptotic probability the stream is adapted is zero. To see this, note that

$$\sum_{k=1}^{K} (\bar{\rho}_k - \underline{\rho}_k) F_{\hat{V}_k}(sd) \leq \sum_{k=1}^{K} (\bar{\rho}_k - \underline{\rho}_k) = \bar{\rho} - \underline{\rho}.$$

Hence, the indicator function is never satisfied because it requires a sum of numbers bounded above by $\bar{\rho} - \underline{\rho}$ exceed a number $\gamma\bar{\rho} - \underline{\rho} > \bar{\rho} - \underline{\rho}$. Thus, $\lim_{m\to\infty} p(m,t,s,d) = 0$. Substituting these values into the integral yields

$$\begin{aligned} q_k^{\gamma,\pi_m} &= 1 - (1-\alpha_k) \int_0^\infty \int_0^\infty (1) dF_{D_k}(d) dF_{S_k}(s) = \alpha_k, \ \gamma < \alpha \\ q_k^{\gamma,\pi_m} &= 1 - (1-\alpha_k) \int_0^\infty \int_0^\infty (0) dF_{D_k}(d) dF_{S_k}(s) = 1, \ \gamma > 1. \end{aligned}$$

Finally, consider the case when the link is provisioned in the rate adaptive regime, i.e., when $\alpha \leq \gamma \leq 1$. Substituting this into the integral yields the equation given in the theorem.

■

6.3 Asymptotic optimal admission control for multiple service classes

In this section we identify the asymptotic optimal admission control policy for multiple service classes. This policy is the multi-class analogue of the optimal admission control policy developed in Chapter 4.

Theorem 9 *The asymptotic optimal admission control policy for multi-service networks, denoted, π_{am}, is a volume threshold policy with optimal volume threshold $v^{\gamma,\pi_{am}}$. The optimal threshold is zero when $\gamma < \alpha$ and is infinite when $\gamma > 1$. For $\alpha \leq \gamma \leq 1$ the optimal volume threshold solves*

$$\sum_{k=1}^{K} (\bar{\rho}_k - \underline{\rho}_k) F_{\hat{V}_k}(v^{\gamma,\pi_{am}}) = \gamma\bar{\rho} - \underline{\rho}. \tag{6.7}$$

The optimal admission control policy for client i on class k is to assign the client a subscription level

$$s_{i,k}^{\gamma,\pi_{am}} = \begin{cases} s_{i,k}, & v_{i,k} \leq v^{\gamma,\pi_{am}} \\ \alpha_k s_{i,k}, & else \end{cases}. \tag{6.8}$$

The asymptotic expected normalized subscription level for class k streams under policy π_{am} equals that obtained under the optimal multi-class adaptation policy π_m.

Proof of Theorem 9. Note that the set of feasible admission control policies is a subset of the set of feasible adaptation policies. Thus, if we identify an admission policy achieving the same asymptotic QoS as the optimal adaptation policy, then that policy is necessarily optimal. All we need show then is that the asymptotic QoS under the volume threshold policy given in the theorem yields an asymptotic QoS equal to that under the optimal multi-class adaptation policy.

When $\gamma < \alpha$ the optimal threshold is zero, hence the asymptotic QoS for class k clients is $q_k^{\gamma,\pi_{am}} = \alpha_k = q_k^{\gamma,\pi_m}$. Similarly, for $\gamma > 1$ the optimal threshold is infinite, hence the asymptotic QoS for class k clients is $q_k^{\gamma,\pi_{am}} = 1 = q_k^{\gamma,\pi_m}$. Consider the case for the rate adaptive regime, $\alpha \le \gamma \le 1$. The asymptotic QoS for class k clients is 1 for clients with volumes less than $v^{\gamma,\pi_{am}}$ and α_k for clients with volumes exceeding $v^{\gamma,\pi_{am}}$. We can therefore write the asymptotic QoS as

$$q^{\gamma,\pi_{am}} = F_{V_k}(v^{\gamma,\pi_{am}}) + \alpha_k \bar{F}_{V_k}(v^{\gamma,\pi_{am}}).$$

Simple rearranging yields

$$q^{\gamma,\pi_{am}} = 1 - (1-\alpha_k)\bar{F}_{V_k}(v^{\gamma,\pi_{am}}).$$

We can apply the definition of $F_V(v)$ to obtain

$$\begin{aligned} q^{\gamma,\pi_{am}} &= 1-(1-\alpha_k)\int_0^\infty \bar{F}_{D_k}(\frac{v^{\gamma,\pi_{am}}}{s})dF_{S_k}(s) \\ &= 1-(1-\alpha_k)\int_0^\infty \Big[\int_{\frac{v^{\gamma,\pi_{am}}}{s}}^\infty dF_{D_k}(d)\Big]dF_{S_k}(s) \\ &= 1-(1-\alpha_k)\int_0^\infty\int_0^\infty \mathbb{I}(d > \frac{v^{\gamma,\pi_{am}}}{s})dF_{D_k}(d)dF_{S_k}(s) \\ &= 1-(1-\alpha_k)\int_0^\infty\int_0^\infty \mathbb{I}(sd > v^{\gamma,\pi_{am}})dF_{D_k}(d)dF_{S_k}(s). \end{aligned}$$

Now note that, by the definition of $v^{\gamma,\pi_{am}}$, the condition in the indicator function may be written as

$$\mathbb{I}(sd > v^{\gamma,\pi_{am}}) \Leftrightarrow \mathbb{I}(\sum_{k=1}^K(\bar{\rho}_k - \underline{\rho}_k)F_{\hat{V}_k}(sd) > \gamma\bar{\rho} - \underline{\rho}).$$

We may therefore write the asymptotic QoS as

$$1-(1-\alpha_k)\int_0^\infty\int_0^\infty \mathbb{I}(\sum_{k=1}^K(\bar{\rho}_k - \underline{\rho}_k)F_{\hat{V}_k}(sd) > \gamma\bar{\rho} - \underline{\rho})dF_{D_k}(d)dF_{S_k}(s).$$

This is the same expression for the asymptotic QoS under the optimal multi-class adaptation policy, q^{γ,π_m}. This proves the theorem.

■

Note that the optimal volume threshold is independent of the service class k, i.e., the network discriminates among streams based on their stream volume, irrespective of their service class.

6.4 Case study: multiplexing audio and video streaming clients

In this section we apply the theorems on optimal multi-class adaptation policies to a case study which investigates how to optimally share link capacity among audio and video streaming clients. The case study will investigate three separate link architectures and compare the resulting QoS. The three link architectures are

- **Single service class sharing.** Audio and video streams share the link capacity and are jointly adapted according to the optimal single class adaptation policy of Chapter 3. All streams, audio and video, are assumed to have a fixed stream adaptivity of $\alpha_1 = \frac{1}{4}$.
- **Two service class sharing.** The network offers two service classes: a free service class with a guaranteed normalized subscription level of α_1 and a premium service class with a guaranteed normalized subscription level of $\alpha_2 = \frac{3}{4}$. Audio clients will generally subscribe to the free class, relying on volume discrimination to grant them acceptable QoS, while video clients will generally subscribe to the premium class, in order to provide them with QoS protection from volume discrimination.
- **Partitioning.** We partition the link capacity *evenly* into two separate channels, one for audio and one for video. All audio streams share the audio

channel and all video streams share the video channel. The streams on each channel are adapted according to the optimal single class adaptation policy of Chapter 3. As in the previous case, audio streams are guaranteed a normalized subscription level of α_1 and video streams are guaranteed a normalized subscription level of α_2.

The three link architectures are illustrated in Figure 6.1.

6.4.1 Simulation parameters

We let 1 denote the free service class for audio streams and 2 denote the premium service class for video streams.

The audio clients have an arrival rate $\lambda_1 = 1$, a typical stream duration of $\delta_1 = 180$ seconds (i.e., the typical length of a song), an average maximum subscription level of $\sigma_1 = 0.1$ Mb/s, and a service class guarantee of $\alpha_1 = \frac{1}{4}$. The distributions for audio client stream durations and maximum subscription levels are both (effectively unbounded) exponentials. Thus, $F_{D_1}(d) = 1 - \exp\left(-\frac{d}{\delta_1}\right)$ and $F_{S_1}(s) = 1 - \exp\left(-\frac{s}{\sigma_1}\right)$.

The video clients have an arrival rate of $\lambda_2 = 0.01$, where the arrival rate is determined by the demand function evaluated at the price charged by the network to use the premium class. The typical video stream duration is $\delta_2 = 1800$ seconds (e.g., a half-hour television program), and the average maximum subscription level is $\sigma_2 = 1.0$ Mb/s. The video content provider offers a minimal stream encoding with an adaptivity of $\alpha_2 = \frac{1}{4}$ (just like the audio streams), but the premium service class provides a QoS guarantee of $\alpha_2 = \frac{3}{4}$. Thus, video clients with a low willingness to pay will choose class 1, and likely receive a low-quality video stream with adaptivity $\frac{1}{4}$, while video clients with a high willingness to pay will choose class 2, and likely receive a high-quality video stream with adaptivity $\frac{3}{4}$. The distributions for video client stream durations and maximum subscription levels are both (effectively unbounded)

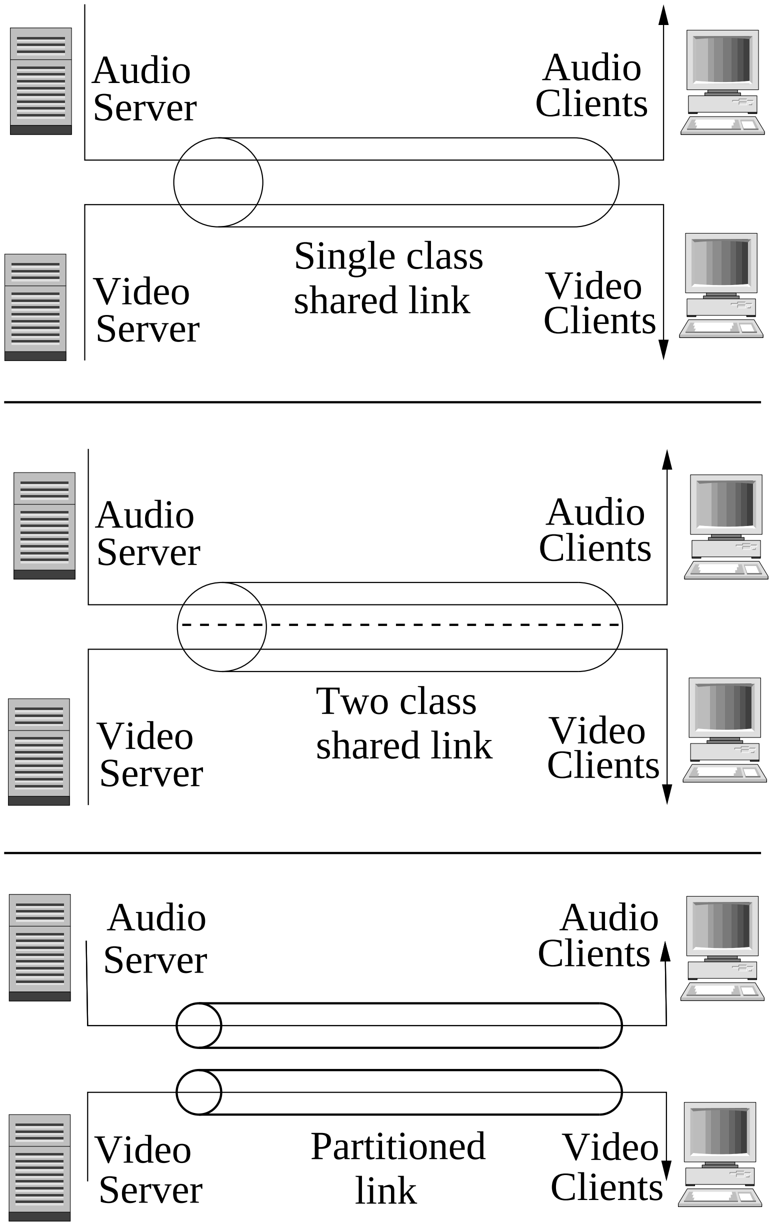

Figure 6.1: Illustration of three link architectures for the case study.

exponentials. Thus, $F_{D_2}(d) = 1 - \exp(-\frac{d}{\delta_2})$ and $F_{S_2}(s) = 1 - \exp(-\frac{s}{\sigma_2})$.

Note that the minimum offered load for the audio clients class is $\underline{\rho}_1 = (\lambda_1\delta_1)(\alpha_1\sigma_1) = (1 \times 180)(\frac{1}{4} \times \frac{1}{10}) = 4.5$ Mb/s, and the maximum offered load for the audio clients is $\bar{\rho}_1 = (\lambda_1\delta_1)(\sigma_1) = (1 \times 180)(\frac{1}{10}) = 18$ Mb/s. The minimum offered load for the video clients when no premium service class is offered (e.g., the first link architecture with a single service class) is $\underline{\rho}_2 = (\lambda_2\delta_2)(\alpha_2\sigma_2) = (\frac{1}{100} \times 1800)(\frac{1}{4} \times 1) = 4.5$ Mb/s. The minimum offered load for the video clients when a premium service class is offered (e.g., the second and third link architectures) is $\underline{\rho}_2 = (\lambda_2\delta_2)(\alpha_2\sigma_2) = (\frac{1}{100} \times 1800)(\frac{3}{4} \times 1) = 13.5$ Mb/s. The maximum offered load for the video clients (for all three scenarios) is $\bar{\rho}_2 = (\lambda_2\delta_2)(\sigma_2) = (\frac{1}{100} \times 1800)(1) = 18$ Mb/s. Thus the overall minimum offered load is 9 Mb/s when no premium service class is offered, and is 18 Mb/s when the premium service class is offered. The overall maximum offered load is 36 Mb/s. Note that the maximum offered load is 18 Mb/s for both the audio and video streams. This is the rational behind partitioning the bandwidth evenly into two channels for the third link architecture.

The different link architectures will have different scaling regimes because of the different service classes and adaptation policies. These regimes are illustrated in Figure 6.2. Consider the first link architecture where both audio and video are guaranteed a normalized subscription level of $\alpha_1 = \frac{1}{4}$, and the optimal single service class adaptation policy is used. The overall adaptivity for this link architecture is $\frac{\underline{\rho}}{\bar{\rho}} = \frac{9}{36} = \frac{1}{4}$, hence the rate adaptive regime for both audio and video streams corresponds to $\frac{1}{4} \leq \gamma \leq 1$. Consider the second link architecture where audio streams have a guaranteed normalized subscription level of $\alpha_1 = \frac{1}{4}$ and video streams have a guaranteed normalized subscription level of $\alpha_2 = \frac{3}{4}$. The overall adaptivity for this link architecture is $\frac{\underline{\rho}}{\bar{\rho}} = \frac{18}{36} = \frac{1}{2}$, hence the rate adaptive regime for both audio and video streams corresponds to $\frac{1}{2} \leq \gamma \leq 1$. The rate adaptive regime is smaller for this scenario than for the previous scenario because the video streams are now

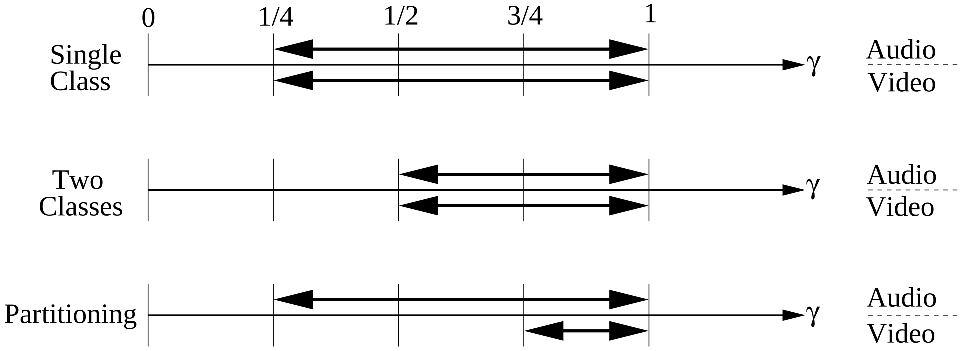

Figure 6.2: Illustration of the rate-adaptive regimes for audio and video streams under each of the three link architectures.

guaranteed a larger normalized subscription level. Consider finally the third link architecture, where audio and video streams use separate channels, audio streams are guaranteed a normalized subscription level of $\alpha_1 = \frac{1}{4}$ and video streams are guaranteed a normalized subscription level of $\alpha_2 = \frac{3}{4}$. Because they are on separate channels, the rate adaptive regime for the audio streams corresponds to $\frac{1}{4} \leq \gamma \leq 1$, while the rate adaptive regime for the video streams corresponds to $\frac{3}{4} \leq \gamma \leq 1$.

We can therefore think of the different link architectures as having different trade offs between blocking probabilities and QoS guarantees. The first link architecture, with a single service class and no premium QoS guarantee for video streams, has the smallest overloaded regime, $0 \leq \gamma < \frac{1}{4}$, which corresponds to the regime with non-zero asymptotic blocking. The second link architecture, with two service classes sharing the link capacity and a premium QoS guarantee for video streams, has a common overloaded regime of $0 \leq \gamma \leq \frac{1}{2}$. Roughly speaking, the increased width of the overloaded regime interval, i.e., $\frac{1}{4} \leq \gamma \leq \frac{1}{2}$ is the cost of increasing the QoS guarantee of video streams from $\frac{1}{4}$ to $\frac{3}{4}$. The third link architecture, with two separate channels for audio and video streams, and a premium QoS guarantee for video streams, has different scaling regimes for the two channels. The audio channel is only overloaded for $0 \leq \gamma \leq \frac{1}{4}$, but the video channel is overloaded for $0 \leq \gamma \leq \frac{3}{4}$.

6.4.2 Computation results

In this section we compute the asymptotic QoS under for audio and video streams under each of the three link architectures. Consider the first link architecture where audio and video streams share the same channel and share the same normalized subscription level guarantee of $\alpha_1 = \frac{1}{4}$. This is equivalent to a multiple service class model where all service classes have the same QoS guarantee. The asymptotic normalized subscription level for the audio (class 1) and video (class 2) streams is therefore given by

$$q_1^{\gamma,1} = 1 - (1-\alpha_1)\bar{F}_{V_1}(F_{\hat{V}_1}(v_1^*(\gamma)) \tag{6.9}$$

$$q_2^{\gamma,1} = 1 - (1-\alpha_1)\bar{F}_{V_2}(F_{\hat{V}_2}(v_1^*(\gamma)) \tag{6.10}$$

where $v_1^*(\gamma)$ is the v that solves

$$(\bar{\rho}_1 - \underline{\rho}_1)F_{\hat{V}_1}(v) + (\bar{\rho}_2 - \underline{\rho}_2)F_{\hat{V}_2}(v) = \gamma\bar{\rho} - \underline{\rho}. \tag{6.11}$$

For the first link architecture we have $\bar{\rho}_1 = 18$, $\underline{\rho}_1 = 4.5$, $\bar{\rho}_2 = 18$, $\underline{\rho}_2 = 4.5$, $\bar{\rho} = 36$ and $\underline{\rho} = 9$. Substituting these values and simplifying, we obtain that $v_1^*(\gamma)$ is the v that solves

$$F_{\hat{V}_1}(v) + F_{\hat{V}_2}(v) = \frac{8}{3}\gamma - \frac{2}{3}. \tag{6.12}$$

Consider next the second link architecture where audio and video streams share the same channel, but audio streams have a QoS guarantee of $\alpha_1 = \frac{1}{4}$ and video streams have a QoS guarantee of $\alpha_2 = \frac{3}{4}$. The asymptotic normalized subscription level for the audio and video streams is now

$$q_1^{\gamma,2} = 1 - (1-\alpha_1)\bar{F}_{V_1}(F_{\hat{V}_1}(v_2^*(\gamma)) \tag{6.13}$$

$$q_2^{\gamma,2} = 1 - (1-\alpha_2)\bar{F}_{V_2}(F_{\hat{V}_2}(v_2^*(\gamma)) \tag{6.14}$$

where $v_2^*(\gamma)$ solves (6.11) but with $\underline{\rho}_2 = 13.5$ and $\underline{\rho} = 18$. Substituting these values and simplifying, we obtain that $v_2^*(\gamma)$ is the v that solves

$$F_{\hat{V}_1}(v) + \frac{1}{3}F_{\hat{V}_2}(v) = \frac{8}{3}\gamma - \frac{4}{3}. \tag{6.15}$$

Finally, consider the third link architecture where audio and video streams share different channels, audio streams have a QoS guarantee of $\alpha_1 = \frac{1}{4}$ and video streams have a QoS guarantee of $\alpha_2 = \frac{3}{4}$. The asymptotic normalized subscription level for audio and video streams is then given by

$$q_1^{\gamma,3} = 1 - (1-\alpha_1)\bar{F}_{V_1}(F_{\hat{V}_1}(\frac{\gamma - \alpha_1}{1-\alpha_1}) \tag{6.16}$$

$$q_2^{\gamma,3} = 1 - (1-\alpha_2)\bar{F}_{V_2}(F_{\hat{V}_2}(\frac{\gamma - \alpha_2}{1-\alpha_2}). \tag{6.17}$$

Figures 6.3 and 6.4 plot the asymptotic normalized subscription levels for audio and video streams respectively under each of the three link architectures. That is, Figure 6.3 plots the asymptotic QoS for audio streams under a single class link architecture, $q_1^{\gamma,1}$, a two class link architecture, $q_1^{\gamma,2}$, and a partitioned architecture, $q_1^{\gamma,3}$, while Figure 6.4 plots the asymptotic QoS for video streams under a single class link architecture, $q_2^{\gamma,1}$, a two class link architecture, $q_2^{\gamma,2}$, and a partitioned architecture, $q_2^{\gamma,3}$.

Consider the first link architecture, i.e., the single service class. Both audio and video streams receive the minimum QoS guarantee, i.e., $\frac{1}{4}$, in the overloaded regime, i.e., $0 < \gamma \leq \frac{1}{4}$. As we increase γ in the rate adaptive regime, i.e., $\frac{1}{4} \leq \gamma \leq 1$, the audio streams show a rapid increase in QoS up to the maximum normalized subscription level of 1 at around $\gamma = 0.6$, while the increase for the video streams is rather minimal. This is because under the single service class with common QoS guarantees, the optimal adaptation policy grants preferential treatment to the smaller (audio) streams and discriminates against the larger (video) streams. As we increase γ above 0.6 we see the video streams showing a faster increase in QoS. This is because the audio streams are already receiving their maximum subscription level and so the full benefit of increasing the capacity goes to the video streams.

Next consider the second link architecture, i.e., the two service classes. Both audio and video streams receive their respective minimum QoS guarantees ($\frac{1}{4}$ and $\frac{3}{4}$ respectively) in the overloaded regime, i.e., $0 < \gamma \leq \frac{1}{2}$. As we increase γ in the

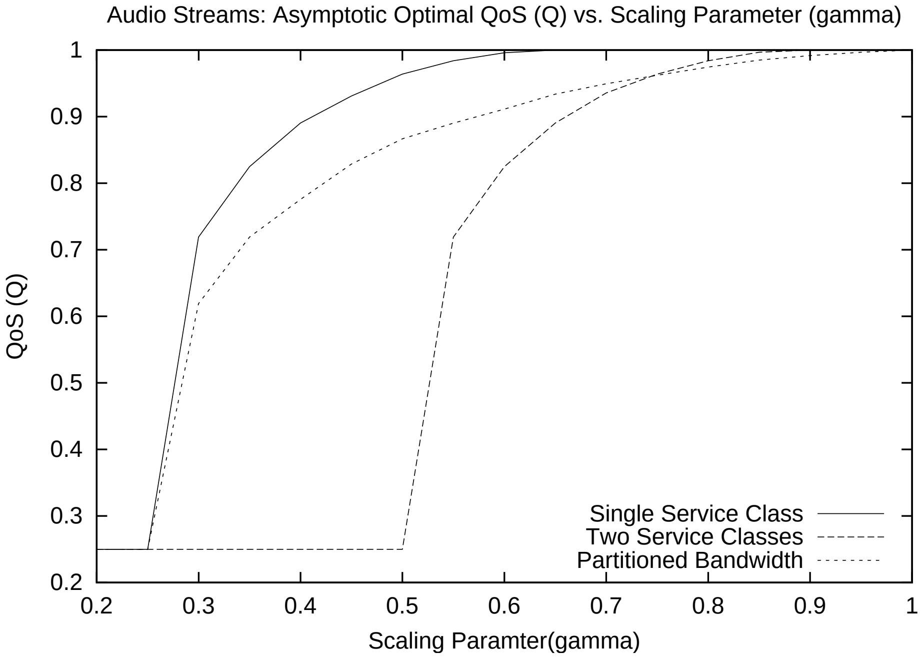

Figure 6.3: Plot of the asymptotic QoS for audio clients versus the scaling parameter γ under the three scenarios.

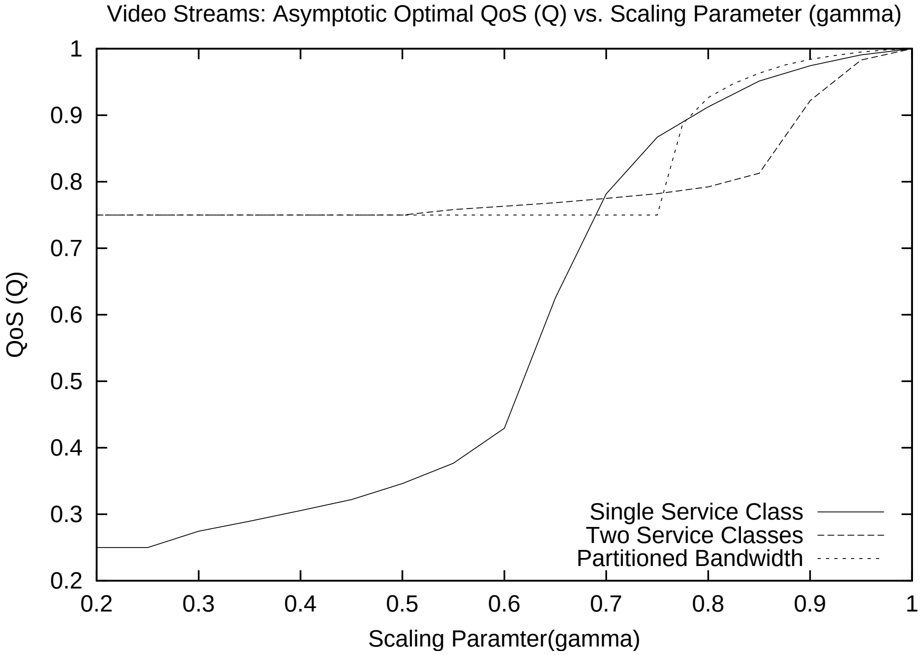

Figure 6.4: Plot of the asymptotic QoS for video clients versus the scaling parameter γ under the three scenarios.

rate adaptive regime, i.e., $\frac{1}{2} \leq \gamma \leq 1$, we see the audio streams again show a rapid increase in QoS up to the maximum normalized subscription level of 1 at around $\gamma = 0.9$, while the increase for the video streams is again rather minimal. Again, this is because the optimal adaptation policy discriminates against the video streams. As we increase γ above 0.9 we see the video streams showing a faster increase in QoS. Note that for $\gamma > 0.7$ the single service class architecture actually gives better average QoS to video streams than does the two class architecture. This may appear somewhat counter-intuitive considering the two class architecture grants a better QoS guarantee to video streams than does the single class architecture. The explanation is that the average video QoS is higher for the single class architecture than for the two class architecture above $\gamma > 0.7$ because the single class architecture can adapt the largest video streams down to a normalized subscription level of $\frac{1}{4}$ while the two class architecture can only adapt the largest video streams down to a normalized subscription level of $\frac{3}{4}$. The two class architecture therefore must adapt *more* video streams than the one class architecture so that the *overall* average QoS for video streams is lower.

Finally consider the third link architecture, i.e., partitioned channels. We see the audio streams receive their minimum normalized subscription level of $\frac{1}{4}$ in the overloaded regime, i.e., $0 < \gamma < \frac{1}{4}$. Similarly, video streams receive their minimum subscription level of $\frac{3}{4}$ in the overloaded regime, i.e., $0 < \gamma < \frac{3}{4}$. As we increase γ in the rate adaptive regime for the audio streams, i.e., $\frac{1}{4} \leq \gamma \leq 1$, the audio streams show a rapid increase in QoS, which, for most of the rate adaptive regime, lies between the audio QoS under the single and two class architectures. The audio QoS under partitioning is strictly below the audio QoS under a single service class because, under the single service class, there are always video streams to adapt before audio streams, while there are no video streams available to adapt under the partitioned architecture. Moreover, the audio QoS under partitioning is higher than under the two class architecture for $\frac{1}{4} \leq \gamma \leq \frac{3}{4}$ because, under the partitioning architecture, we are blocking video streams in this regime. Once we are not blocking video streams under the partitioning architecture, i.e., $\frac{3}{4} \leq \gamma \leq 1$, the audio QoS is equivalent under the partitioning and two-class architecture. Next consider the video QoS under the partitioning architecture. We see that the video QoS is better under a two class architecture than under partitioning for $\frac{1}{2} < \gamma \leq \frac{3}{4}$; this is because video streams are still being blocked under partitioning in this regime, and so all video streams receive their minimum subscription level. Once video streams are not being blocked under partitioning, i.e., $\frac{3}{4} \leq \gamma \leq 1$, the video QoS is higher under partitioning than under a two class architecture. This is because video QoS under the two class architecture is still being sacrificed for higher audio QoS in this regime due to the optimal volume discrimination policy.

In summary, the link designer may use Figures 6.3 and 6.4 to design how to implement an adaptation policy to meet a combined target of low blocking and high normalized subscription levels. If the link designer can afford to provision the link above $\gamma > \frac{3}{4}$ then a partitioning strategy obtains the maximum QoS for both audio and video classes and has zero asymptotic blocking as well. If the link designer can only afford to provision the link in the regime $\frac{1}{2} \leq \gamma \leq \frac{3}{4}$, then the link designer might choose a two class architecture because obtains asymptotic zero blocking for both audio and video and provides a strong QoS guarantee for video streams. If low blocking is more important than offering a QoS guarantee and the link designer can only afford to provision the link in the regime $\frac{1}{4} \leq \gamma \leq \frac{1}{2}$, then the link designer might choose a single class link architecture.

Chapter 7

Conclusion

This thesis presents a thorough theoretical analysis of system-level design issues for optimally supporting rate adaptive multimedia streams on communication networks. This chapter will summarize the important findings in each chapter and comment on possible extensions.

7.1 Optimal adaptation

Our primary contributions with respect to adaptation policies are as follows.

- *Network scaling.* We have identified the appropriate network scaling through which to study asymptotic network performance for rate adaptive streams. In particular, the three scaling regimes, parameterized by the capacity scaling parameter γ, suggests the rate adaptive regime, i.e., $\alpha \leq \gamma \leq 1$, is the target design regime. This is the target regime because the system experiences low blocking and yet a marginal increase in capacity results in a marginal increase in QoS.

- *Optimal adaptation policy.* We have identified the optimal adaptation policy on a general network for both stored media and live media, i.e., π_k and π_u.

We have shown that optimal adaptation requires the solution, at each time t, of an integer programming problem.

- *Two rate near-optimal adaptation policies.* We have proved the existence of near-optimal adaptation policies, i.e., $\tilde{\pi}_k$ and $\tilde{\pi}_u$, which make use of only the minimum and maximum subscription levels. This implies that, under optimal adaptation, there is little benefit for content providers to offer additional encodings at intermediate resolutions between the minimum and maximum.
- *Asymptotic optimal QoS.* We have shown that, for large capacity networks, the integer programming problem yielding the optimal subscription level allocation is equivalent to a nonlinear programming problem over a set of instantaneous volume thresholds. For the case of stored media on a single link we are able to identify closed form expressions for asymptotic QoS under optimal adaptation. These expressions are the asymptotic rate-adaptive analog of Erlang's blocking probability equation for loss networks.

Although we are able to identify the optimal adaptation policy for live media on a general network, we are unable to extend the analysis beyond this point. In particular, it appears difficult to demonstrate that the optimal adaptation policy for live media on large capacity networks is equivalent to the solution of a nonlinear programming problem over a class of volume thresholds. Similarly, we are unable to identify closed form expressions for asymptotic optimal QoS under the optimal adaptation policy for live media.

7.2 Optimal admission control

Regarding admission control, our primary contributions are as follows.

- *Optimal admission control policy.* The primary contribution is the identification of the optimal admission control policy as the solution to a nonlinear programming problem over a set of volume threshold vectors.

- *Equivalent asymptotic QoS.* The salient point about optimal admission control is that it achieves an asymptotic QoS equaling that obtained under optimal dynamic adaptation. In particular, the asymptotic expected normalized subscription levels are equivalent under optimal admission control and optimal dynamic adaptation. Optimal admission control guarantees a zero rate of adaptation, but at the expense of a blocking probability with slow convergence to zero.

We have discussed some of the implementation issues surrounding optimal admission control, i.e., link state, knowledge of system parameters and distributions, and stationary traffic. A possible extension of this work would be to develop adaptive admission control policies which are robust to errors in system parameter estimates and function under a non-stationary traffic workload. Intuitively, we should be able to identify online algorithms which monitor stream properties to obtain increasingly accurate system parameter estimates, and which monitor network load to react to traffic non-stationarities on multiple time scales.

7.3 Distributed algorithms

Our analysis of our proposed class of distributed volume dependent adaptation algorithms suggests the following findings.

- *Volume independent algorithms.* Volume independent algorithms are in a sense equivalent to the fair-share adaptation policy presented in Chapter 3 in that they achieve a similar QoS.

- *Volume dependent algorithms.* Our proposed volume dependent algorithms attempt to attain the increase in QoS seen under the optimal adaptation and admission policies by allowing small volume streams to more aggressively pursue network capacity than large volume streams. Our simulation results suggest that volume dependent algorithms may be a practical distributed solution for rate adaptive multimedia streams.

The obvious extension of this work would be to implement the algorithms for actual media streams on actual networks and test their performance. This will be the subject of future work.

7.4 Multiple service classes

The motivation to investigate multiple service classes was that each of the previous chapters resulted in policies which discriminated against large volume streams. Clients wishing to view such streams with a guaranteed minimum QoS of their choosing might benefit from such service classes. The network might indeed offer such services to recover costs, provided clients would be willing to pay for them. Our primary findings include the following.

- *Multi-class capacity scaling.* We proposed a generalization of the single-class capacity scaling of Chapter 3 appropriate for links offering multiple service classes. As before, the rate adaptive regime is the primary regime of interest.

- *Optimal multi-class adaptation policy.* We have identified the optimal multi-class adaptation policy on a single link as the solution to an integer programming problem. We have also shown the existence of a near-optimal solution which makes use of only two subscription levels. Finally, we developed a closed form expression for the asymptotic QoS under the optimal multi-class adaptation policy.

- *Optimal admission control.* We identified the optimal admission control policy for links offering multiple service classes as a volume threshold policy, similar to that offered in Chapter 4. We have shown the asymptotic QoS under the optimal multi-class admission policy equals that under the optimal multi-class adaptation policy.

One obvious generalization here is to extend the multi-class analysis to a general network and to develop the appropriate distributed algorithms. A second, perhaps more interesting extension, is to study pricing mechanisms for the various service classes to ensure clients properly self-select their appropriate service class. This topic falls under the rubric of incentive and congestion pricing for information services. It would be interesting to apply some of the results in this field to develop a pricing mechanism appropriate for rate adaptive streaming media service classes.

7.5 Other applications of rate adaptation

Stated generally, the principle of rate adaptation is that a QoS target may be obtained by selectively degrading the service quality of some clients in order to give superior service to other clients. If the QoS target is to maximize the overall client average QoS, then the principle of rate adaptation suggests selectively degrading the service quality of clients consuming more network resources in order to give superior service to clients consuming fewer network resources. One example of this is the optimal adaptation policy which discriminates against large volume clients and gives preference to small volume clients. A second example of this is the optimal admission policy which sets a volume threshold inversely proportional to the route cost, i.e., clients traversing multiple congested links are more likely to receive their minimum subscription level than clients traversing a small number of uncongested links. If the QoS target is to satisfy heterogeneous QoS requirements for different client classes, then the principle of rate adaptation suggests selectively degrading clients with strict QoS requirements by less than clients with less stringent QoS requirements.

This principle can be extended to other realistic communications scenarios. One example is wireless communications where clients adjust their transmission power in proportion to the distance to the next available base station or relay node. Clients transmitting at a high power level in order to reach their desired relay node might cause unacceptable interference levels for other clients located nearby. The principle of rate adaptation suggests giving priority to clients capable of transmitting at lower power levels, i.e., closer distances, in order to maximize the overall system capacity.

Another example might be an ISP offering heterogeneous service guarantees in the form of an inexpensive residential subscription level and a more expensive business subscription level. The ISP might promise business customers strict priority over residential customers in that business customer streams are adapted only after all residential customers have already been adapted. That is, the ISP selectively degrades low priority residential customer streams before degrading business customer streams.

Bibliography

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

[2] Alan Bain and Peter 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] Philip Chou, Alexander Mohr, Albert Wang, and Sanjeev Mehrotra. Error control for receiver-driven layered multicast of audio and video. *IEEE Transactions on Multimedia*, 3(1), March 2001.

[4] Lisette Cullinane and Jitendra Marwaha. Streaming media: when will it take off? *Telephony Online*, November 2001. `www.telephonyonline.com`.

[5] Jean-Claude Delcroix and Jonathan Green-Armytage. Planning for the bandwidth that applications will need. Technical report, Gartner Group, March 2002. `www.gartner.com`.

[6] Sally Floyd and Kevin Fall. Promoting the use of end-to-end congestion control in the Internet. *IEEE/ACM Transactions on Networking*, 7(4):458–472, 1999.

[7] Mark W. Garrett and Walter Willinger. Analysis, modeling and generation of self-similar VBR video traffic. In *SIGCOMM*, pages 269–280, 1994.

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

[9] 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.

[10] 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.

[11] DiffServ Working Group. Differentiated services, February 2003. `www.ietf.org/html.charters/diffserv-charter.html`.

[12] IntServ Working Group. Integrated services, September 2000. `www.ietf.org/html.charters/intserv-charter.html`.

[13] Motion Picture Experts Group. The mpeg home page, 2003. `mpeg.telecomitalialab.com/`.

[14] Koushik Kar, Saswati Sarkar, and Leandros Tassiulas. Optimization based rate control for multirate multicast sessions. In *IEEE Conference on Computer Communications (Infocom)*, pages 123–132, 2001.

[15] Gerry Kaufhold. Slow economy driving streaming onto corporate networks. Technical Report #MB0107SM, Cahner's In-Stat Group, August 2001. `www.instat.com`.

[16] Frank 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.

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

[18] Art Mena and John Heidemann. An empirical study of real audio traffic. In *IEEE Conference on Computer Communications (Infocom)*, pages 101–110, Tel-Aviv, Israel, March 2000. IEEE.

[19] Jupiter Media Metrix, January 2001. `www.jup.com`.

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

[21] Nielsen Netratings, October 2001. `www.nielsen-netratings.com`.

[22] Bryan Porter. Video streaming at work to be worth 140m by 2006. *New Media Zero*, January 2002. `www.newmediazero.com`.

[23] Reza Rejaie, Mark Handley, and Deborah Estrin. Quality adaptation for congestion controlled video playback over the internet. In *SIGCOMM*, pages 189–200, 1999.

[24] Ann Marie Rohaly, Philip Corriveau, and John Libert. Video quality experts group: Current results and future directions. In *Proc. International Conference on Visual Communications and Image Processing*, June 2000.

[25] Bill Rose and Warren Kurtzman. Broadband revolution 2: the media world of speedies. Technical report, Arbitron Inc, and Coleman Insights, June 2001. `www.arbitron.com`.

[26] Bill Rose and Joe Lenski. Internet and multimedia 10. Technical report, Arbitron/Edison Media Research, 2003. `www.arbitron.com`.

[27] Keith Ross. *Multiservice loss models for broadband telecommunications networks*. Springer-Verlag, 1995.

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

[29] Schulzrinne, Casner, Frederick, and Jacobson. *RTP: a transport protocol for real-time applications.* The Internet Engineering Taskforce, March 2003. `www.ietf.org/internet-drafts/draft-ietf-avt-rtp-new-12.txt`.

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

[31] Brett Vickers, Celio Albuquerque, and Tatsuya Suda. Source-adaptive multilayered multicast algorithms for real-time video distribution. Technical Report Technical Report ICS-TR 99-45, University of California, Irvine, June 1999.

[32] Jean Walrand. *An introduction to queueing networks.* Prentice Hall, 1988.

[33] Hengqing Ye. Stability of a bandwidth allocation model for tcp internet, 2002. Submitted.

Vita

Steven Patrick Weber, son of Thomas and Dianne Weber, was born and raised in Milwaukee, Wisconsin. He attended Marquette University, graduating with a bachelors degree in computer engineering in 1996. He completed his masters degree in the Department of Electrical and Computer Engineering at the University of Texas at Austin in 1999 under the supervision of Professor Gustavo de Veciana. Since 1999 he has been a Ph.D. candidate in the Department of Electrical and Computer Engineering at the University of Texas at Austin.

Permanent Address: 1500 Royal Crest Drive, #333
Austin TX, 78741

This dissertation was typeset with LaTeX 2$_\varepsilon$[1] by the author.

[1] LaTeX 2$_\varepsilon$ is an extension of LaTeX. LaTeX is a collection of macros for TeX. TeX is a trademark of the American Mathematical Society. The macros used in formatting this dissertation were written by Dinesh Das, Department of Computer Sciences, The University of Texas at Austin, and extended by Bert Kay and James A. Bednar.