A Framework for Providing User Level Quality of Service Guarantees in Multi-Class Rate Adaptive Systems

Nikos Argiriou[1]: narg@egnatia.ee.auth.gr
Leonidas Georgiadis[2]: leonid@eng.auth.gr

Tel: *(+30) 2310996385*
Fax: *(+30) 2310 996312*

[1] *(corresponding author) PhD student, Electrical and Computer Engineering Dept., Aristotle University of Thessaloniki, Thessaloniki, Greece Electrical and Computer Engineering Dept., Aristotle University of Thessaloniki, Thessaloniki, Greece*

[2] *Professor, Electrical and Computer Engineering Dept., Aristotle University of Thessaloniki, Thessaloniki, Greece Electrical and Computer Engineering Dept., Aristotle University of Thessaloniki, Thessaloniki, Greece*

Contents

Manuscript

A Framework for Providing User Level Quality of Service Guarantees in Multi-Class Rate Adaptive Systems

Abstract: The problem of channel sharing by rate adaptive streams belonging to various classes is considered. Rate adaptation provides the opportunity for accepting more connections by adapting the bandwidth of connections that are already in the system. However, bandwidth adaptation must be employed in a careful manner in order to ensure that a) bandwidth is allocated to various classes in a fair manner (system perspective) and b) bandwidth adaptation does not affect adversely the perceived user quality of the connection (user quality). The system perspective aspect has been studied earlier. This paper focuses on the equally important user perspective. It is proposed to quantify user Quality of Service through measures capturing short and long term bandwidth fluctuations that can be implemented with the mechanisms of traffic regulators, widely used in networking for the purpose of controlling the traffic entering or exiting a network node. Furthermore, it is indicated how to integrate the user perspective metrics with the optimal algorithms for system performance metrics developed earlier by the authors. Simulation results illustrate the effectiveness of the proposed framework.

Keywords: Bandwidth Sharing, Rate-adaptive Streams, User Quality of Service, Traffic Regulator, Leaky bucket, Video performance metrics.

1. Introduction

As technology evolves, multimedia applications are getting more and more sophisticated. Despite their stringent Quality of Service (QoS) requirements (packet delay, delay jitter and throughput), their capability to adapt to changing network conditions, in combination with a proper admission control scheme, provides a promising means for using network bandwidth efficiently, thereby guaranteeing acceptable user QoS and achieving large system utilization.

In this paper we consider a communication channel, wired or wireless, whose bandwidth is shared by randomly arriving connections belonging to a number of classes. Connection bandwidth may be adapted within a given range,

hence giving the opportunity for bandwidth management in order to improve connection admission rate in overload situations. More specifically, when the number of active users is small, applications are admitted by the system with their maximum requested rate, while as the system load increases the application transmission rate is reduced, while still remaining within acceptable levels, so that more connections can be admitted. This process is facilitated by the existence of controllers (headend in HFC networks [1] and base stations in wireless cellular networks) that can convey feedback to the already running applications through the downstream channel, in order to reduce their rate accordingly. The adaptation can be achieved by various coding techniques such as layered coding [2], [3] and adaptation of compression parameters [4], [5], [6] as well as bandwidth smoothing [4], etc. Depending on the technique, rate adaptation can take one of a number of discrete values, or it can take any value within a specific range [7], [8]. Wavelet coding [5] is particularly well suited for continuous rate adaptation.

Several works addressed the problem of bandwidth adaptation management under various assumptions on channel characteristics and the bandwidth adaptation policies. In general, the design of bandwidth adaptation policies must take into account both system and user QoS requirements. As [9] explains in detail, by specifying an adaptation policy for all active streams the user perspective seeks to maximize each user's individual Quality of Service (QoS), while system perspective seeks to maximize a performance metric that is based an average QoS. While the system average QoS is very important, user QoS must also be taken into account, otherwise the system performance optimization leads to reduction of the user QoS for many streams in the system.

Most of the work up to now has focused on the design of policies based mainly on system QoS perspective [10], [11], [12], [13], [14], [15], [16], [17], [18], [19]. The introduction of time-average user metrics in many of them cannot provide a complete characterization of user performance, and only helps to study the effect on user perception of various policies oriented towards optimizing system-wide objectives. A work where individual user QoS is taken into account is [20], where a simple user QoS criterion is discussed referring to the minimum time-interval between two successive bandwidth changes.

User quality is an active research topic in multimedia evolution, mainly approached by two popular methods: subjective and objective quality assessment. While the subjective analysis [21] involves playing a sample audiovisual clip to a number of participants and taking their individual perception into account, the objective analysis [22], [23], [24], on the other side attempts to provide automated procedures without relying to human judgment. The problem with subjective quality assessment techniques is that they are based on individual perception which can vary significantly between a given set of individuals, while objective quality assessment techniques may not necessarily reflect the actual end-user

experience. There have been studies [25] that show that when objective and subjective quality assessment are performed simultaneously, the results are comparable. In the last two decades, a great deal of effort has been made to develop objective image and video quality assessment methods, which incorporate perceptual quality measures by considering Human Visual System (HVS) characteristics [26], [27], [28]. Advances in vision research have provided crucial information on the structure and the working mechanism of the human vision system, which have been adopted to design quality metrics [29], [30], [31]. Most current psychovisual quality metrics share the commonality of being based on multichannel vision models [32]. The current VQEG activities represent international standardization efforts towards an objective video quality metric, with delegations from ITU-T Study Groups 9 and 12 and ITU-R Study Group 11 [33]. However, there are still many issues to resolve and image and video quality assessment is an active research area.

In this paper we propose a framework for channel bandwidth management where both user QoS and system performance metrics are systematically taken into account. Specifically, it is proposed to quantify user QoS performance through metrics capturing short and long term individual user bandwidth fluctuations that can be implemented with the mechanisms of traffic regulators used for congestion control in networks. The proposed quantification is general and flexible enough to incorporate various user perception requirements that may arise from future research. Furthermore, it is indicated how to integrate the user perspective metrics with optimal algorithms for system performance metrics developed in [19] for multi-class systems.

The rest of the paper is organized as follows. In Section 2, the multi-class system model is presented. In Section 3, the performance metrics under consideration are described. Specifically, general user QoS metrics are introduced in Section 3.1 and the manner by which these metrics can be monitored is presented in Section 3.1.2; system performance metrics are introduced in Section 3.2. Section 4 presents a bandwidth adaptation policy that extends the policy proposed in [19] so that both user and system performance metrics are taken into consideration. Simulation results indicating the effectiveness of the approach are presented in Section 5. The main conclusions of the work are summarized in Section 6.

2. System model and Notation

Consider a communication channel of bandwidth B bps. Connection i arrives for transmission over the channel at time a_i and (provided that it is

accepted by the system) departs at time $d_i > a_i$. The connection holding time is defined as $h_i = d_i - a_i > 0$.

Connections belong to one of the classes in a set $\mathcal{C}$. Denote by c_i the class to which connection i belongs. Through appropriate compression techniques, transmission of connection i belonging to class c may take place at rates belonging to a set $\mathcal{B}_c \subseteq \left[\underline{B}_c, \overline{B}_c \right]$ where $\underline{B}_c$ and $\overline{B}_c$ are respectively the minimum and maximum bandwidth levels that connections in class c may take. Transmission rates of connection i may be adapted over time, but must belong to $\mathcal{B}_c$ for acceptable reception quality. Note that $\mathcal{B}_c$ may be a strict subset of $\left[\underline{B}_c, \overline{B}_c \right]$, e.g., may consist of only a discrete number of acceptable bandwidth levels.

Let $\mathcal{A}_c(t)$ be the set of class c connections that arrived and have been accepted by the system up to time t. Let also $\mathcal{N}_c(t)$ be the set of class c connections that are present in the system at time t. Define also, $\mathcal{A}(t) = \cup_{c \in \mathcal{C}} \mathcal{A}_c(t)$ and $\mathcal{N}(t) = \cup_{c \in \mathcal{C}} \mathcal{N}_c(t)$, that is, $\mathcal{A}(t)$ is the total number of connections accepted by the system and $\mathcal{N}(t)$ is the total number of ongoing connections.

Let c_i be the class to which connection i belongs. According to the previous definitions, if $b_i(t)$ is the bandwidth allocated to connection i at time t, $a_i \le t < d_i$, it must hold for any time t,

$$\sum_{i \in \mathcal{N}(t)} b_i(t) \le B, \tag{2.1}$$

$$b_i(t) \in \mathcal{B}_{c_i}, \text{ for all } i \in \mathcal{N}(t). \tag{2.2}$$

Inequality (2.1) expresses the fact that the total bandwidth allocated to the connections at any time t cannot exceed the bandwidth of the channel, while conditions (2.2) express the fact that each connection may adjust its rates within the allowable bounds of the class to which the connection belongs.

In order to operate the system, two policies must be defined: the Connection Admission Policy and the Bandwidth Adaptation Policy. The Connection Admission Policy decides whether to accept or reject a newly arriving connection, while the Bandwidth Adaptation Policy adjusts at any time t the bandwidth of the connections that are currently in the system. Hence, the Connection Admission Policy affects mainly the connection blocking probability, while the Bandwidth Adaptation Policy is responsible for allocating the channel bandwidth so that fairness is maintained among classes, while user Quality of Service is maintained.

In order to describe the proposed Connection Admission and Bandwidth

Adaptation policies, we need to discuss first the relevant metrics that capture system performance and user Quality of Service satisfaction.

3. Performance metrics

This work concentrates on metrics that are related to the way the system allocates bandwidth to various connections. We can distinguish the performance metrics in two broad classes:

1) Metrics that express *in detail* the manner in which bandwidth is allocated to each connection. These metrics are directly related to user perception as regards the manner in which bandwidth is allocated to the particular connection. These metrics are referred to as User QoS metrics.
2) Metrics that describe the *average* manner in which bandwidth is allocated to the connections belonging to a given class. These metrics are referred to as System Performance metrics. Based on the System Performance metrics, optimization problems related to fair allocation of bandwidth among classes can be defined.

3.1. User QoS Metrics

The main quantity of interest for our purposes is the manner by which the bandwidth allocated to connection i, $b_i(t)$, varies over time within its allowable boundaries. An appropriate metric that gives an overall - but coarse - indication of this variation is the *Scaled Mean Connection Bandwidth*, [9], [16], [17]. The scaled mean bandwidth allocated to connection i when a given Bandwidth Adaptation policy is employed, is defined as

$$\hat{b}_i = \frac{\int_{a_i}^{d_i} b_i(s)ds}{V_i}, \tag{3.1}$$

where $V_i = h_i \overline{B}_{c_i}$ and is called the application's volume. This measure is the average bandwidth allocated to the connection throughout its holding time, scaled by the best possible bandwidth the connection could achieve. This particular scaling is not crucial for the development that follows. Other scaling may be adopted, including no scaling at all, i.e., $\overline{B}_{c_i}$ may be set to 1 (one) for all connections.

The scaled mean bandwidth by itself does not give information on the small-scale variations of the bandwidth allocated to a connection, which might be important for user perception of quality of received information. In this respect, it

is of interest to define measures that indicate how these variations take place. In general, $b_i(t)$ is a step function, see Figure 1. Define for $a_i < t < d_i$, the jumps of the connection bandwidth when bandwidth adaptation occurs,

$$\Delta_i(t) = b_i(t) - b_i(t^-),$$

where $b_i(t^-) \equiv \lim_{s \to t} b_i(s)$. Then a general metric of the bandwidth variation that occurs at time t is of the form

$$g(\Delta_i(t)),$$

where $g(x)$ is a general cost function with $g(x) > 0$. Denote also as $\overline{g}(\Delta_i(t))$, its average value over the connection's holding time h_i,

$$\overline{g}(\Delta_i(t)) = \frac{\sum_{k=1}^{K_i} g(\Delta_i(t_k))}{h_i}, \tag{3.2}$$

where t_k, $a_i < t_k < t_{k+1} < d_i$, $k = 1,...,K_i$ are the times where the jumps of $b_i(t)$ occur.

Appropriate choices of $g(x)$ give various meaningful measures. Some of them are listed below:

- If

$$g(x) = \begin{cases} 1 & x \neq 0 \\ 0 & x = 0 \end{cases}, \tag{3.3}$$

then $g(\Delta_i(t))$ indicates that a unit cost incurs during any bandwidth change. In this case the quantity $\overline{g}(\Delta_i(t))$ expresses the *average number of bandwidth changes* during the connection's holding time. In case

$$g(x) = \begin{cases} 1 & x < 0 \\ 0 & x \geq 0 \end{cases},$$

then $g(\Delta_i(t))$ indicates that a unit cost incurs *only* when a bandwidth reduction occurs. As results in [34] demonstrate, frequent changes in connection bandwidth may have a deleterious effect in the user perception of quality of received information. This is also supported by results in by [35], [36], where for layered encoded video it is shown that the quality is influenced by the frequency of layer variations.

- When $g(x) = |x|$, then $g(\Delta_i(t))$ indicates that a cost equal to the magnitude of bandwidth adaptation incurs. Hence the quantity $\overline{g}(\Delta_i(t))$ expresses the *average size of bandwidth adaptations* during a connection's holding time. In the special case when

$$g(x) = \begin{cases} |x| & x < 0 \\ 0 & x \geq 0 \end{cases} \tag{3.4}$$

then $g(\Delta_i(t))$ indicates that a cost incurs *only when a bandwidth reduction occurs*, and this cost is equal to the magnitude of bandwidth adaptation.

- The choice $g(x) = |x|^2$ can be used if one needs to express the fact that large fluctuations are more detrimental than small ones.

Other reasonable forms of the function $g(x)$ can be given. It is expected that the form of the function is correlated with the end-user perception of reception quality. Also, it is quite possible that a combination of more than one of these functions might be appropriate. Since as was stated in the introduction the question of how user perception is related to various objective measures is still a largely unresolved issue, the specification is left in this generality. The framework to be proposed in the following sections can be applied with any choice of function, or even a combination of a number of these.

We note that in the recent work [16], [18], Weber and de Veciana studied the manner in which averages of some of these metrics in the form of (3.2) behave, under policies that operate based on optimizing system performance metrics and for single class systems**.** However, as far as end-user perception is concerned, short term variations in addition to averages should also be of importance. This issue is addressed in Sections 3.1.1 and 3.1.2.

3.1.1. Characterization of Bandwidth Fluctuation

As discussed in Section 3.1, from the point of view of this work the bandwidth variation of a connection at a given time t is characterized by $g(\Delta_i(t))$. The function $g(x)$ was left unspecified so that the framework can be general enough to incorporate specific metrics that may arise from studies related to user perception quality of received information. While for the moment no specific measure has been adopted in the literature, it is reasonable to expect that fluctuation of $g(\Delta_i(t))$ over time affects user perception. The simplest such measure is the average fluctuation described by the average in (3.2). However, this measure is a gross fluctuation indicator. A more detailed description that incorporates both short and long term fluctuations and is flexible enough to incorporate several details regarding the manner in which fluctuations occur si described below.

The main idea is to employ the concept of traffic regulator that has been used successfully in network design to shape the traffic entering and exiting from a network node [37], [38], [39]**.** From the point of view of current investigation, a traffic regulator can be thought of as a device that accepts as input a quantity $I(t)$, which may be arbitrary, and provides an output $G(t)$ whose fluctuations

are guaranteed to be constrained in a specific manner. As applied to networks, $I(t)$ represents information flow (in bits or packets) that arrives to the regulator up to time t, and $G(t)$ represents the information flow that is allowed to exit from the regulator. However, this does not need to be the case, and it will not be in the framework to which the concept is applied in the current work.

To be more specific, let a function $f(\tau) \geq 0, \quad \tau \geq 0$ be given. For technical reasons $f(\tau)$ can be assumed to be subadditive (i.e., $f(s) + f(t) \geq f(s+t)$ for all $s, \ t \geq 0$, see [38]). This function represents the desirable upper bound on the fluctuations of the quantity $G(t)$ exiting the regulator in the following sense.

A function $G(t)$ is called $f-$ constrained if for all $t \geq 0$,

$$G(t) - G(s) \leq f(t-s), 0 \leq s \leq t \tag{3.5}$$

A regulator whose output $G(t)$ is f-constrained is called f-constraining regulator.

According to this definition, a regulator that is $f-$ constraining ensures that at any time interval $[s,t]$ the fluctuation of the quantity $G(t)$ exiting from the regulator does not exceed $f(t-s)$. The best known such regulator is the one called Leaky Bucket, corresponding to the case where

$$f(\tau) = \sigma + \rho\tau, \ \tau \geq 0.$$

The Leaky Bucket is characterized by the parameters (σ, ρ) where $\sigma \geq 0$ represents a bound on the short-term fluctuations and is called burstiness, while $\rho > 0$ represents an upper bound on the long term average of the quantity $G(t)$. There are well known methods to implement leaky buckets [40], [41], [42]**.** Moreover, it is well known that by combining leaky buckets in series or in parallel, more general regulators can be obtained with piecewise-linear constraining functions of the form

$$f(\tau) = \min\{\sigma_1 + \rho_1\tau, \ \sigma_2 + \rho_2\tau, \ ...., \ \sigma_L + \rho_L\tau\}.$$

Thus in general, leaky buckets can be used to enforce piecewise linear, concave constraining functions as Figure 2 shows.

In the framework under consideration, for a given connection i, $G_i(t)$ represents a metric of the total bandwidth fluctuation of the connection up to time t and is meaningful only during the connection's holding time, i.e., in the interval $[a_i, d_i]$. More specifically, recall that $t_k, \ a_i < t_k < t_{k+1} < d_i, \ k = 1,...,K_i$ are the times where the jumps of connection i bandwidth occur. Then define for $a_i \leq t \leq d_i, \ G_i(a_i) = 0,$

$$G_i(t) = \sum_{t_k \le t} g(\Delta_i(t_k)).$$

The constraining function $f_i(\tau)$ for the connection is prespecified and it is required that the system operate so that $G_i(t)$ satisfies (3.5).

3.1.2. Satisfaction of User Quality of Service

In order to guarantee the user Quality of Service satisfaction, for each connection i, the following actions are taken.

- One or more metrics $g_i^{(l)}(x),\ 1 \le l \le L_i,$ are specified for the connection bandwidth fluctuation $x = \Delta_i(t)$.
- Constraining functions $f_i^{(l)}(\tau),\ 1 \le l \le L_i,$ for each of the metrics are specified. These functions may be specific to the class to which the connection belongs.
- Every time t_n when the possibility for bandwidth adaptation for connection i exists (to be determined by the adaptive algorithm described in Section 4.2), the system must ensure that all $G_i^{(l)}(t),\ 1 \le L_i$ satisfy (3.5). This can be easily implemented by employing $f_i^{(l)}$ – constraining regulators with the following provisions. a) The input to the regulator is assumed infinite (i.e., there is always the possibility for bandwidth adaptation), b) the regulator outputs a bandwidth adaptation only at the times t_n and c) the system checks that the value of the new bandwidth is within the allowable rates $\mathcal{B}_{c_i}$ and if necessary accepts a modified value for the new connection bandwidth. This last part will be discussed in more detail in Section 4.2. The following example clarifies the manner in which the user QoS is guaranteed.

Example. For simplicity in the notation the connection index i is neglected in this example. Let $g(x)$ be the metric used and let an (σ, ρ) Leaky Bucket regulator be used as a constraining function for the bandwidth fluctuations of a connection. This can be viewed as token bucket of size $\sigma,$ where tokens are accumulated at rate ρ, as long as the amount of tokens is smaller than σ (tokens generated when the bucket is full are discarded). If at time t_n the possibility for bandwidth adaptation exists, the content of the bucket is observed. The amount of the tokens at t_n determines the largest amount of adaptation $g(\Delta(t_n))$ that may occur. The system may pick any amount smaller than or equal to the amount specified by the tokens in the bucket (in order to

ensure that the resulting rates are in $\mathcal{B}_{c_i}$ and possibly satisfy other optimization objectives - see Section 4.2**).** Next, the system updates appropriately the amount of tokens in the bucket in order to ensure that only as many as necessary are taken.

To clarify the previous discussion, consider two options for the regulations. In the first option, a single Leaky Bucket controls the frequency of rate adaptation, while in the second two Leaky Buckets are employed, one controlling the rate of adaptation and the other the bandwidth reduction rate.

- *A single Leaky Bucket controlling the frequency of rate adaptation*

Assume that $L = 1,\ g^{(1)}(x) = 1$ if $x \neq 1$ and $g^{(1)}(0) = 0$. If the content of the bucket is smaller than 1 at time t_n, then no bandwidth adaptation for the connection will be allowed. If the content of the bucket is larger that 1, then it is allowable to change the bandwidth of the connection. However, it is up to the Bandwidth Adaptation Policy to decide (based on system-wide optimization) whether this adaptation will actually occur. If the adaptation does occur, then the content of the bucket is reduced by 1.

- *Two Leaky Buckets, one controlling the frequency of rate adaptation and the other the bandwidth variation rate.*

Assume now that $L = 2,\ g^{(1)}(x)$ is as defined before, and $g^{(2)}(x) = |x|$ if $x < 0$ and $g^{(2)}(x) = 0$ otherwise. There are now two Leaky Buckets (σ_1, ρ_1) and (σ_2, ρ_2). At time t_n the system operates as follows, see Figure 3.

1. If the content of the (σ_1, ρ_1) bucket is smaller than one, no bandwidth adaptation for the connection is allowed.
2. If the content of the (σ_1, ρ_1) bucket is larger than one, then a bandwidth adaptation may be allowed. The content of the (σ_2, ρ_2) bucket indicates the magnitude of this adaptation. Let the content of the (σ_2, ρ_2) bucket be α. Then because of the form of $g^{(2)}(x)$, the bandwidth of the connection cannot become smaller than $b(t_n) - \alpha$, but can be increased by as much as possible.
3. To decide the actual bandwidth allocated to the connections, the Bandwidth Adaptation policy will take into account the minimum possible bandwidth adaptation $b(t_n) - \alpha$ for the connection. However, the Bandwidth Adaptation Policy may decide to decrease $b(t_n)$ by only β, where $0 < \beta < \alpha$. In this

case, the content of the (σ_1, ρ_1) bucket will be reduced by 1, while the content of the (σ_2, ρ_2) bucket will be reduced by β. Note that if instead the bandwidth of the connection was *increased* by an amount $\gamma > 0$, then the content of the (σ_2, ρ_2) bucket will not be reduced; this is so, since in this case $\Delta(t_n) > 0$ and hence $g^{(2)}(\Delta(t_n)) = 0$.

3.2. System Performance Metrics

The overall performance of a class should express the average bandwidth that has been allocated to connections of this class. Hence, an appropriate measure to consider is the Class Average Scaled Bandwidth $\widehat{B}_c$, defined as

$$\widehat{B}_c = \lim_{t\to\infty} \frac{\sum_{i\in\mathcal{A}_c(t)} \hat{b}_i}{A_c(t)},$$

where $A_c(t) \triangleq |\mathcal{A}_c(t)|$ is the number of class c connections that have been accepted by the system by time t. Of course, one would like to have $\widehat{B}_c$ as high as possible. However, in the model under consideration there are several classes in the system and making $\widehat{B}_c$ simultaneously high for all classes may not be possible. Hence the issue of fair bandwidth allocation among classes arises. A good overview of system design based on optimization problems related to fairness is given in [43]. In general, in this case one attempts to maximize a reward function of the form

$$\widehat{B} = \sum_{i\in\mathcal{C}} \phi_c\left(\widehat{B}_c\right), \tag{3.6}$$

where $\phi_c(x)$ are concave functions expressing the satisfaction, reward or utility for obtaining average scaled bandwidth x. Various choices of $\phi_c(x)$ provide various fairness criteria. Some of the most common ones are described below.

- *Linear utilities.* In this case, $\phi_c(x) = r_c x$.
- *Proportional Fairness.* In this case, $\phi_c(x) = \log(x)$. This allocation has several important properties discussed in [43]. Intuitively, since the $\log(x)$ function increases very slowly for large x, this type of utilities express the fact that the satisfaction received by a given increase in bandwidth is higher if the already allocated bandwidth is smaller.
- *Max-min Fairness.* Intuitively here one attempts to maximize the bandwidth of the classes with the minimal allocated bandwidth, while splitting evenly whatever bandwidth remains to the rest of the classes. While this problem cannot be directly translated in the form (3.6) it can be

well approximated by choosing

$$\phi_c(x) = c - g(x)^m,$$

where c and m are constants and $g(x)$ is a differentiable, decreasing, convex and positive function.

Another important performance metric is the *System Blocking Probability*, that is, the percentage of arriving connections that cannot be accepted by the system. One would like to keep the system blocking probability as low as possible. The Connection Admission policy attempts to keep blocking probability within acceptable levels.

In [17] a general algorithm for optimizing performance metrics of the form (3.6) is presented, under a given Connection Admission policy and without taking into account user QoS metrics. As will be described in Section 4.1, in this case the System Blocking Probability can be made independent of the Bandwidth Adaptation policy. However, when user QoS metrics are also taken into account, the blocking probability will also depend on the Bandwidth Adaptation policy. Hence in the current study we will examine the dependence of blocking probability on user QoS metrics.

4. Bandwidth Adaptation Policy

In this section a Bandwidth Adaptation Policy designed to incorporate both system performance objectives and user specific QoS requirements is proposed. The Bandwidth Adaptation Policy proposed in [19] does not account for user-specific QoS requirements. As will be seen however, this policy can be appropriately modified in order to incorporate such requirements. Section 4.1 outlines for ease of reference the main steps of the policy proposed in [19], while Section 4.2 indicates how to incorporate user-specific QoS requirements in the policy.

4.1. A Policy for System Performance Optimization

In this section the policy developed in [19] is outlined. This policy will be extended in Section 4.2 to incorporate both user and system performance metrics.

There is an extensive literature on the design of Connection Admission policies when the connection bandwidth requirements are fixed. A good reference is [44], where a nice collection of several Connection Admission Policies and their analysis can be found, along with a large number of other references. In the current investigation connection bandwidths can be adapted. The class of

Connection Admission policies adopted in [19] operate based on the minimum acceptable bandwidth levels ($\underline{B}_i$). More specifically, the following general Connection Admission Policy is adopted.

Acceptable Connection Admission Policy. Any policy π designed for connections with fixed bandwidth requirements may be employed. Whenever a new connection arrives to the system, the policy π admits or rejects the connection using as connection bandwidths the minimum acceptable connection bandwidth levels $\underline{B}_i$.

Hence, under an acceptable Connection Admission policy the decision to accept or reject a new connection depends only on the *minimal* bandwidth that can be allocated to all the connections (including the newly arrived one). Note that this does not mean that all the connections will necessarily receive the minimal bandwidth. It is the task of the Bandwidth Allocation Policy to distribute appropriately the available channel bandwidth to the connections. However, since acceptance of rejection of a newly arriving connection depends only on the minimal bandwidths that can be allocated, it can be seen that the number of connections admitted in the system (and hence the blocking probability) is independent of the Bandwidth Adaptation Policy.

Given the Connection Admission Policy described above, the Bandwidth Adaptation Policy is designed to optimize system performance objectives of the type (3.6). For the description below, $\phi_c(x)$, $c \in \mathcal{C}$ are given reward functions. In order to describe the policy it is useful to extend somewhat the notion of scaled mean connection bandwidth defined in Section 2. Specifically, this definition is extended for all times $t \geq 0$, as follows.

$$\hat{b}_i(t) = \begin{cases} 0 & t < a_i \\ \frac{1}{h_i \overline{B}_i} \int_{a_i}^{t} b_i(s)ds & a_i \leq t < d_i \\ \hat{b}_i & t \geq d_i \end{cases}. \tag{4.1}$$

Next, the performance of class c at time t is defined as the average of the performance measures of all connections that have been admitted by the system up to time t, that is,

$$\widehat{B}_c(t) = \frac{\sum_{i \in \mathcal{A}_c(t)} \hat{b}_i(t)}{A_c(t)}. \tag{4.2}$$

For simplicity in the description of the Bandwidth Adaptation policy it is assumed that the Connection Admission policy is Complete Sharing, that is, the policy accepts a new connection if and only if the sum of minimal acceptable bandwidth levels of all connections in the system, if the new connection is accepted, does not exceed the channel bandwidth. Any other acceptable Connection Admission policy may be employed under the provision that additional constraints are included in the optimization problem (4.4) below - see [19].

Each time t_n a new connection arrives or a connection already in the system departs, perform the following.

- Compute the derivatives

$$r_c(t_n) = \left.\frac{d\phi_c(b)}{db}\right|_{b=\widehat{B}_c(t_n))} . \tag{4.3}$$

- Allocate to connection $i \in \mathcal{N}(t_n)$ bandwidth $b_i(t_n) = b_i^*$, where b_i^* is the solution to the following optimization problem.

$$\begin{aligned} \max & \left\{ \sum_{c\in\mathcal{C}(t_n)} \sum_{i\in\mathcal{N}_c(t_n)} r_c(t_n) \frac{b_i}{V_i} \frac{t_n}{A_c(t_n)} \right\} \\ & \sum_{c\in\mathcal{C}(t_n)} \sum_{i\in\mathcal{N}_c(t_n)} b_i \le B \\ & b_i \in \mathcal{B}_{c_i}, \ \ i \in \mathcal{N}(t_n). \end{aligned} \tag{4.4}$$

Note that the policy does not rely on system statistics and that (4.4) represents a Linear Programming (LP) optimization problem which can be solved very efficiently if $\mathcal{B}_c$ is the whole interval $\left[\underline{B}_c, \overline{B}_c\right]$ [16], [19].

For easy reference, the combination of the two policies described above (Connection Admission and Bandwidth Adaptation) is referred to as the System Performance Oriented (SPO) policy.

Notes:

1. In practice, in order for the system to adapt easier to statistical parameter changes, it will be appropriate to replace the average in (4.2) with a weighted average, or an average over a finite window. The same holds for the quantities $t_n / A_c(t_n)$, which are in effect time averages representing $1/\lambda_c$, the inverse of the class arrival rates.
2. In case the connection holding times h_i are random variables with mean $H_i = E[h_i]$ and their exact values are unknown to the system, an operational policy is obtained by setting $V_i = H_i \underline{B}_i$ in the optimization problem (4.4). Also, the quantities $\widehat{B}_c(t)$ can be appropriately updated by taking into account the fact that the connection holding times are known upon connection departure. These modifications result in policies that perform fairly well with respect to the optimal when the connection holding times are known [19].

4.2. A Policy Satisfying both User and System Performance Objectives

As was discussed in Section 3.1.2 in order to satisfy the user QoS objectives, each connection is associated with (at least one) metric $g_i(x)$ and constraining function $f_i(x)$. Combined, these functions determine at each time the amount by which the bandwidth of a connection may change relative to the value it currently has. Hence the bandwidth that connection i may take at decision instant t_n does not necessarily belong to the whole set $\mathcal{B}_{c_i}$ any more, but to a subset $\mathcal{B}_i(t_n) \subseteq \mathcal{B}_{c_i}$; in particular, if $\mathcal{B}_i(t_n)$ contains only one element, then no bandwidth adaptation of connection i is allowed at time t_n. Let $\underline{B}_i(t_n)$ be the minimal of the bandwidths in the set $\mathcal{B}_i(t_n)$.

The rule to incorporate in order to take into account both user and system performance objectives in the design is simple: the Connection Admission and Bandwidth Adaptation policies operate as in Section 4.1 *after replacing* $\mathcal{B}_{c_i}$ and $\underline{B}_i$ with $\mathcal{B}_i(t_n)$, $\underline{B}_i(t_n)$ respectively. Specifically, the Connection Admission policy bases its decisions on the minimal connection bandwidths that can be allocated at time t_n, $\underline{B}_i(t_n)$, while the Bandwidth Adaptation Policy determines the bandwidths that are allocated to the connections as solution to the following optimization problem.

$$\begin{aligned} &\max\left\{\sum_{c\in\mathcal{C}(t_n)}\sum_{i\in\mathcal{N}_c(t_n)} r_c(t_n)\frac{b_i}{V_i}\frac{t_n}{A_c(t_n)}\right\} \\ &\sum_{c\in\mathcal{C}(t_n)}\sum_{i\in\mathcal{N}_c(t_n)} b_i \le B, \\ &b_i \in \mathcal{B}_i(t_n),\ i\in\mathcal{N}(t_n), \end{aligned} \tag{4.5}$$

where $r_c(t_n)$ is given by (4.3). Note that after the solution $\{b_i^*\}_{i\in\mathcal{N}(t_n)}$ of the optimization problem (4.5) is determined, the actual bandwidth adaptation for each connection can be computed. As discussed in Section 3.1.2 (see the example in that section) this in turn determines the actual amount of regulator output needed for the update and hence the exact amount of tokens is taken out of the bucket.

The steps taken at decision instant t_n are described by the following Algorithm. The steps are also described pictorially in Figure 4.

1) If a time t_n a new connection j arrives, set $\mathcal{B}_j(t_n) = \mathcal{B}_{c_j}$, $\underline{B}_j = \underline{B}_j(t_n)$. For a connection i that is already in the system determine based on the connection regulator(s) the set $\mathcal{B}_i(t_n)$ and $\underline{B}_i(t_n)$.

2) (*Connection Admission Policy*) If at t_n a new connection arrives, decide to accept the connection based on the acceptable policy π that operates with

connection bandwidths $\underline{B}_i(t_n)$.

3) (*Bandwidth Adaptation Policy*) Determine the bandwidth $\{b_i^*\}_{i\in\mathcal{N}(t)}$ allocated to the connections as solutions to Linear Program (4.5).

4) Update the connection regulators based on the actual bandwidth adaptation that takes place for each connection.

Note that since $\underline{B}_i(t_n)$ depends now on the manner in which bandwidth adaptation has taken place earlier, the number of connections admitted by the system through (4.5) is dependent on these bandwidth adaptations. This is in contrast to the case where the Bandwidth Adaptation Policy was designed to optimize only system performance objectives.

5. Simulation Results

Simulations to investigate the performance of the proposed policy as well as the influence of the regulator constraints on system performance were performed. The simulation setup is as follows. It is assumed that applications belong to four classes, $|\mathcal{C}| = 4$. Class c is characterized by four parameters $\{[\underline{B}_c, \overline{B}_c],\ f_c,\ g_c(x),\ (\sigma_c, r_c)\}$ where:

- $[\underline{B}_c, \overline{B}_c]$ denotes the bandwidth requirements of class c connections, in Mbps. A connection in class c may be transmitted at any rate in the interval $[\underline{B}_c, \overline{B}_c]$. Define the bandwidth range as $BR_c = \overline{B}_c - \underline{B}_c$. A connection may take any value within this range.
- f_c is the probability distribution of class c connection holding times whose average value is H_c. For the simulations we chose the triangular distribution:

$$f_c^t(h_c) = \begin{cases} \frac{h_c - H_c + aH_c}{(aH_c)^2} & H_c - aH_c \le h_c < H_c \\ \frac{H_c + aH_c - h_c}{(aH_c)^2} & H_c \le h_c \le H_c + aH_c \\ 0 & \text{otherwise} \end{cases}$$

where $a = 0.3$. For these distributions the connection holding times are concentrated around their mean. This is a reasonable assumption since the definition of a class implies often that connections in the class have similar characteristics. The connection holding times can be known (streaming stored multimedia content), or unknown random variables where the

system knows only their mean value (teleconferencing, video telephony etc.). In the following tests it is assumed that the holding times are unknown (see note 2 at the end of Section 4.1).

- $g_c(x)$ is the adaptation function chosen in order to control a specific user performance metric. Experiments with two values for the bandwidth adaptation function were performed. In the first set of experiments, the function has the form of (3.4) that is, the amount of bandwidth decrease is controlled. In the second set of experiments, the function has the form of (3.3) that is, the frequency of bandwidth adaptation is controlled.
- (σ_c, ρ_c) is the leaky bucket adopted as constraining function for $g_c(x)$.

The channel capacity is $L = 500$ Mb and the system runs for $N = 10000$ connection requests. The total connection arrival rate λ is varying between $(\lambda_{\min}, \lambda_{\max}) = (0.198, 0.271)$ arrivals/sec. This set of rates is chosen so that the system operates with acceptable blocking probability (less that 0.01) when no user QoS is required (and hence the SPO policy is applied). An arriving connection has equal probability of belonging to any of the four classes, i.e., the connection arrival rate for class c is $\lambda_c = \lambda / 4$.

The values chosen for the experiments are shown in the following table 1. We also experimented with two types of system performance functions (3.6), linear and a concave (proportionally fair) utility function:

$$\varphi\left(\widehat{B}_c^{\pi}(t)\right) = \begin{cases} r_c \widehat{B}_c^{\pi}(t) & \text{linear with } r_c = 1 \\ \log\left(\widehat{B}_c^{\pi}(t)\right) & \text{concave} \end{cases}.$$

All simulations are implemented using the OMNET++ discrete event simulation system [45].

5.1. Control of Bandwidth Decrease

In this section it is assumed that it is desirable to control the connection bandwidth decrease. In order to implement this requirement, experiments were performed with the following values for the leaky buckets.

$$\sigma_c(k) \in \{\sigma_c(1), \sigma_c(2), \sigma_c(3)\} =$$

$$\{1.1 * BR_c, 1.3 * BR_c, 2.35 * BR_c\},\ \text{Mb}$$

$$\rho_c(k) = \rho = 2\ \text{Mb/sec},$$

where k is a scaling parameter used to denote the strictness of the user constraints in a decreasing order. For two leaky buckets with the same bound ρ on the long-term average rate, the one with the larger value of $\sigma_c(k)$ allows larger short-term

fluctuations of the connection bandwidth function and is therefore less strict. This should have implications on system performance; the simulations intend to quantify these implications.

The charts show curves for different values of the scaling parameter k (leaky bucket controls only the bandwidth decrease with burstiness $\sigma_c(k)$) in comparison to the SPO policy.

System performance is expressed by the system reward function and the blocking probability. Figure 5(a) and Figure 5(b) show the results for the linear case. From Figure 5(a), it is seen that the leaky bucket constraints may have a significant effect on the blocking probability. When the constraint is strictest ($k = 1$), then the degradation in blocking probability is large for large values of λ. However, as the constraints become looser, the blocking probability quickly comes close to the one obtained by the SPO policy. This shows the importance of introducing the burstiness parameter σ_c : this parameter should be set to the maximum possible value allowed by user-perception considerations. On the other hand, from Figure 5(b), it is seen that the use of leaky buckets introduces a small increase, larger for $k = 1$, on the system reward function relative to the SPO policy. Since the SPO policy optimizes the system performance metric when no QoS constraints are imposed, this may seem unexpected at first. However, it should be taken into account that the SPO policy is optimal among *all* Bandwidth Adaptation policies under a *given* Connection Admission policy that operates based on the minimal allocatable connection bandwidths $\underline{B}_i$. Introducing user QoS constraints enforces the Connection Admission policy to operate based on the time-varying bandwidths $\underline{B}_i(t_n)$ and hence more connections may be rejected. As a result the connections accepted by the system have more bandwidth available to share and hence the system performance metric increases. Similar conclusions are drawn form Figure 5(c) and Figure 5(d) for the case of concave system reward function.

The charts in Figure 5(e) and Figure 5(f) show the maximum connection bandwidth increase and decrease, when the system operates under the SPO policy (Figure 5(e)), and under the policy that operates with leaky bucket parameter corresponding to $k = 1$ (Figure 5(f)). The system performance in these figures is based on the linear rewards function with $r_c = 1$ and the total arrival rate is $\lambda_{\max}$. In these charts, the x axis represents the connection order number, after the connections are sorted in increasing order, according to their average volume, $H_c\overline{B}_c$. Hence, all class 1 connections have lower index than class 2 connections, all connections from class 2 have lower index than class 3 connections etc. Connections that don't incur any adaptations during their duration in the system have a zero footprint at the x axis. In Figure 5(e) where the SPO policy is used and hence no user QoS constraints are taken into account, we see that the

connections can decrease their bandwidth to any value within their range. On the other hand, in Figure 5(f) we see that the bandwidth decrease is effectively controlled due to the use of the leaky buckets. Note that bandwidth increases in the two figures are similar. As explained in Section 3.1if it is desirable to also control bandwidth increases, this can be done by the use of appropriate functions $g_c(x)$. Another interesting observation from these figures is that applications with smaller volume incur more adaptations. This is due to the choice of system performance metric. Indeed, it can be seen from the solution to (4.5) that under the chosen metric, applications with smaller volume are the first candidates for bandwidth adaptation provided that the arrival rates to all classes are the same and $r_c(t_n) = 1$. Of course, this behavior may be altered if so desired, by changing the scaling factors in V_i. Similar conclusions are drawn form Figure 5(g), Figure 5(h) for the case of the convex policy.

5.2. Control of Frequency of Rate Adaptation

In this section it is assumed that it is desirable to control the frequency of rate adaptation of a connection. In order to implement this requirement, experiments were performed with the following values for the leaky buckets.

$$\sigma_c(k) \in \{\sigma_c(1), \sigma_c(2), \sigma_c(3)\} =$$

$$\{12.5, 15, 18.8\},\ \text{adaptations}$$

$$\rho = 0.075\ \text{adapt/sec}$$

where k is the scaling parameter. Now the leaky bucket size $\sigma(k)$ shows the maximum number of adaptations in a short interval, while the leaky bucket rate ρ shows the maximum on the average rate of adaptations. The conclusions are quite similar to the case that the adaptation function $g_c(x)$ represents the bandwidth decrease.

System performance is similarly expressed by the system reward function and the blocking probability. Figure 6(a) and Figure 6(b) show the results for the linear rewards, and Figure 6(c) and Figure 6(d) for the convex rewards. Similar conclusions as in Section 5.1 hold.

Figure 6(e) and Figure 6(f) show the average frequency of rate adaptation per connection, when the system operates under the SPO policy (Figure 6(e)), and under the policy that operates with leaky bucket parameter corresponding to $k = 1$ (Figure 6(f)). The system performance in these figures is based on the linear rewards function and the total arrival rate is $\lambda_{\max}$. As in Section 5.1 the x axis represents the connection order number, after the connections are sorted in

increasing order, according to their average volume. We see again that the user QoS constraints are effectively controlled by the use of leaky buckets.

Similar conclusions are drawn form Figure 6(g) and Figure 6(h) for the case of the convex policy.

6. Conclusions

The problem of channel sharing by rate-adaptive multi-class streams with user specific Quality of Service constraints was considered. In order to control the quality of end user perception of received information, the concept of constraining appropriately chosen bandwidth adaptation functions was considered. It was showed how this concept can be combined with earlier work in [19] to provide a policy that is oriented towards optimizing system performance, without violating user QoS requirements. Simulation results showed the usefulness and efficiency of the proposed framework.

The bandwidth adaptation regulators required by the framework are very general. Moreover, more than one of them can be used to control the quality of each connection. Regarding the computational cost of implementing the proposed framework, it consists of two parts: a) the cost of implementing the regulators per connection and b) the cost of solving the Linear Programming optimization problem (4.5). In the context of network traffic control, efficient methods for implementing regulators of the Leaky Bucket type have been proposed and can also be used in the current framework. The Linear Programming optimization can be solved very efficiently if the connection bandwidths can be varied in a continuous manner. The complexity increases if the connection bandwidths may take discrete values, in which case reliance to approximate optimization algorithms may be appropriate. Whether the implementation complexity is acceptable or not depends on the anticipated rate of connection arrivals and the computational power of the system. The exploration of algorithms that achieve good tradeoffs between performance and complexity may be worthwhile.

The issue of which combination of regulators is the appropriate one for effecting acceptable user perception is an open problem. The proper answer to this question requires extensive experimentation regarding the relation of user perception to bandwidth adaptation regulators. However, it seems reasonable that user perception is affected by both the frequency of rate of bandwidth adaptation and the amount of bandwidth change that occurs during an adaptation. The proposed framework is general enough to control both of these parameters.

7. References

1. C. Adjih, N. Argiriou, M. Chaudier, E. Deberdt, F. Dumontet, L. Georgiadis, P. Jacquet, An architecture for IP quality of service provisioning over CATV networks, in: EMMSEC, Stockholm, Sweden, 1999.

2. A. Eleftheriadis, D. Anastasiou, Optimal data partitioning of MPEG-2 coded video, in: First IEEE International Conference on Image Processing, Austin, Texas, pp. 273-277, 1994.

3. P. Pancha, M. Zarki, Prioritized transmission of variable bit rate MPEG video, in: IEEE GLOBECOM, Orlando, FL, USA, pp. 1135-1139, 1992.

4. N. G. Duffield, K. K. Ramakrishnan, A. R. Reibman, SAVE: an algorithm for smoothed adaptive video over explicit rate networks, IEEE/ACM Transactions on Networking Vol. 6, No. 6, pp. 717-728, 1998.

5. D. Taubman, A. Zakhor, A common framework for rate and distortion based scaling of highly scalable compressed video, IEEE Transactions on Circuits and Systems for Video Technology Vol. 6, No. 4, pp. 329-354, 1996.

6. ANSI t1.801.03-1996: Digital transport of one-way Video Telephony signals - parameters for objective performance assessment, 1996.

7. L. Delgrossi, C. Halstrinck, D. B. Henhmann, R. G. Herrtwich, J. Krone, C. Sandvoss, C. Vogt, Media scaling for audiovisual communication with the Heidelberg transport system, in: Proceedings ACM Multimedia '93, Anaheim, USA, pp. 99-104, 1993.

8. A. Eleftheriadis, D. Anastasiou, Meeting arbitrary QoS constraints using dynamic rate shaping of code digital video, in: Fifth International Workshop on Network and Operating System Support for Digital Audio and Video, Durham, New Hampshire, USA, pp. 89-100, 1995.

9. S. Weber, G. de Veciana, Telecommunications Network Design, Kluwer Academic Publisher, Ch. Asymptotic Analysis of Rate Adaptive Multimedia Streams, pp. 167-192, 2002.

10. P. P. Demestichas, V. P. Demesticha, Y. I. Manolessos, G. D. Stamoulis and M. E. Theologou, QoS Management by Means of Application

Control, Journal of Network and Systems Management Vol. 7, No. 2, pp. 177-197, 1999.

11. M. Mahajan, M. Parashar, Managing QoS for Multimedia Applications in the Differentiated Services Environment, Journal of Network and Systems Management Vol. 11, No. 4, pp. 469-498, 2003.

12. A. K. Talukdar, B. R. Badrinath, A. Acharya, Rate adaptation schemes in networks with mobile hosts, in: ACM/IEEE MOBICOM, Dallas, Texas, pp. 169-180, 1998.

13. T. Kwon, S. Kim, Y. Choi, N. Naghshineh, Threshold-type call admission control in Wireless/Mobile multimedia networks using prioritized adaptive framework, IEEE Electronics Letters Vol. 36, No. 9, pp. 852-854, 2000.

14. T. Kwon, Y. Choi, S. K. Das, Bandwidth adaptation algorithms for adaptive multimedia services in mobile cellular networks, KLUWER Wireless Personal Communications Vol. 22, No. 3, pp. 337-357, 2002.

15. T. Kwon, Y. Choi, C. Bisdikian, M. Naghshineh, QoS provisioning in Wireless/Mobile multimedia networks using an adaptive framework, ACMWireless Networks Vol. 9, No. 1, pp. 51-59, 2003.

16. S. Weber, G. Veciana, Rate adaptive multimedia streams: Optimization, admission control, and distributed algorithms, IEEE/ACMTransactions on Networking Vol. 13, No. 6, pp. 1275-1288, 2005.

17. N. Argiriou, L. Georgiadis, Channel sharing by rate-adaptive streaming applications, Performance Evaluation Vol. 55, No. 3-4, pp. 211-229, 2004.

18. S. Weber, G. Veciana, Flow-level QoS for a dynamic load of rate adaptive sessions sharing a bottleneck link, Comput. Networks Vol. 51, No. 8, pp. 1981-1997, 2007.

19. N. Argiriou, L. Georgiadis, Channel sharing by multi-class rate adaptive streams: Performance region and optimization, Computer Networks Vol. 51, No. 6, pp. 1616-1629, 2007.

20. V. Bharghavan, K. Lee, S. Lu, S. Ha, D. Dwyer, The TIMELY adaptive resource management architecture, IEEE Personal Communications Magazine Vol. 5, No. 4, pp. 20-31, 1998.

21. ITU-500-R recommendation BT.500-8: Methodology for the subjective assessment of the quality of television pictures, 1998.

22. ITU-T recommendation P.862: Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs, 2001.

23. ITU-T recommendation P.861: Objective quality measurements of telephone band and (300-3400 hz) speech codecs, 1998.

24. ITU-T recommendation J.144: Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference, 2004.

25. A. Clark, Modeling the effects of burst packet loss and recency on subjective voice quality, in: Proceedings of IP Telephony Workshop, pp. 123-127, 2001.

26. J. Lubin, The Use of Psychophysical Data and Models in the Analysis of Display System Performance, A.B.Watson, Cambridge, MA:MIT Press, pp. 163-178, 1993.

27. S. Daly, The Visible Differences Predictor: An Algorithm for the Assessment of Image Fidelity, A. B. Watson, Cambridge, MA:MIT Press, pp. 179-206, 1993.

28. P. B. E. al, Quality meter and digital television applications, in: Proceedings SPIE Visual Communications and Image Processing, Vol. 4067, pp. 780-790, 2000.

29. C. V. D. B. Lambrecht, Perceptual models and architectures for video coding applications, Ph.D. thesis, Ecole Polytechnique Federale de Lausanne, Switzerland, 1996.

30. S. Winkler, A perceptual distortion metric for digital color video, in: Proceedings SPIE Human Vision and Electronic Imaging Conference, Vol. 3644, pp. 175-184, 1999.

31. A. B.Watson, Toward a perceptual video quality metric, in: Proceedings SPIE Human Vision and Electronic Imaging III, Vol. 3299, pp. 139-147, 1998.

32. Z. Yu, H. R. Wu, Human visual system based objective digital video quality metrics, in: International Conference on Signal Processing 2000 of 16th IFIP World Computer Congress, Vol. II, pp. 1088-1095, 2000.

33. Video Quality Experts Group, Final report from the video quality experts group on the validation of objective models of video quality assessment, March 2000.

34. B. Girod, Psychovisual aspects of image communications, Signal Processing Vol. 28, pp. 239-251, 1992.

35. S. Nelakuditi, R. Harinath, E. Kusmierek, Z. Zhang, Providing smoother quality layered video stream, in: Proceedings of the 10th International Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV), Chapel Hill, NC, 2000.

36. M. Zink, J. Schmitt, R. Steinmetz, Retransmission scheduling in layered video caches, in: Proceedings of IEEE International Conference on Communications 2002 (ICC 2002), IEEE, New York, USA, pp. 2474-2478, 2002.

37. R. L. Cruz, A calculus for network delay, part i: Network elements in isolation, IEEE Transactions on Information Theory Vol. 37, pp. 114-131, 1991.

38. C.-S. Chang, Performance Guarantees in Communication Networks, Springer Verlag, 2000.

39. J. Y. L. Boudec, P. Thiran, Network Calculus, A Theory of Deterministic. Queuing Systems for the Internet, Springer, 2001.

40. N. Li, S. C. Liew, Video compression with output traffic conforming to leaky bucket network access control, in: IEEE International Conference Of Image Processing, 1996.

41. O. Heckmann, F. Rohmer, J. Schmitt, The token bucket allocation and reallocation problems, Tech. Rep. TR-KOM-2001-12, Darmstadt University of Technology, December 2001.

42. C. Dovrolis, M. Vedam, P. Ramanathan, The selection of the token bucket parameters in the IETF guaranteed service class, Tech. rep., University of Wisconsin-Madison, Madison, WI 537061691, USA, 1998.

43. J.-Y. Boudec, Rate adaptation, congestion control and fairness: A tutorial, Tech. rep., Ecole Polytechnique Federale de Lausanne (EPFL), December 2000.

44. K.W. Ross, Multiservice Loss Models for Broadband Telecommunication Networks, Springer Verlag, 1995.

45. A. Varga, The OMNET++ discrete event simulation system, in: ESM'01, the 15th European Simulation Multiconference, SCS-European Publishing House, Prague, Czech Republic, pp. 319-324, 2001.

Tables

table 1: Simulation Data

	class 1	class 2	class 3	class 4
Average holding time H_c (sec)	200	240	300	360
Bandwidth $\left(\underline{B}_c, \overline{B}_c\right)$ (Mb)	$(2, 20)$	$(4, 30)$	$(6, 40)$	$(8, 50)$
Bandwidth range BR_c	18	26	34	42
Arrival rate (conn/sec) λ_c	$0.25 * \lambda$	$0.25 * \lambda$	$0.25 * \lambda$	$0.25 * \lambda$
Application volume V_c	4000	7200	12000	18000

Figures

Table of Figures

Figure 1 A new representation for user adaptation

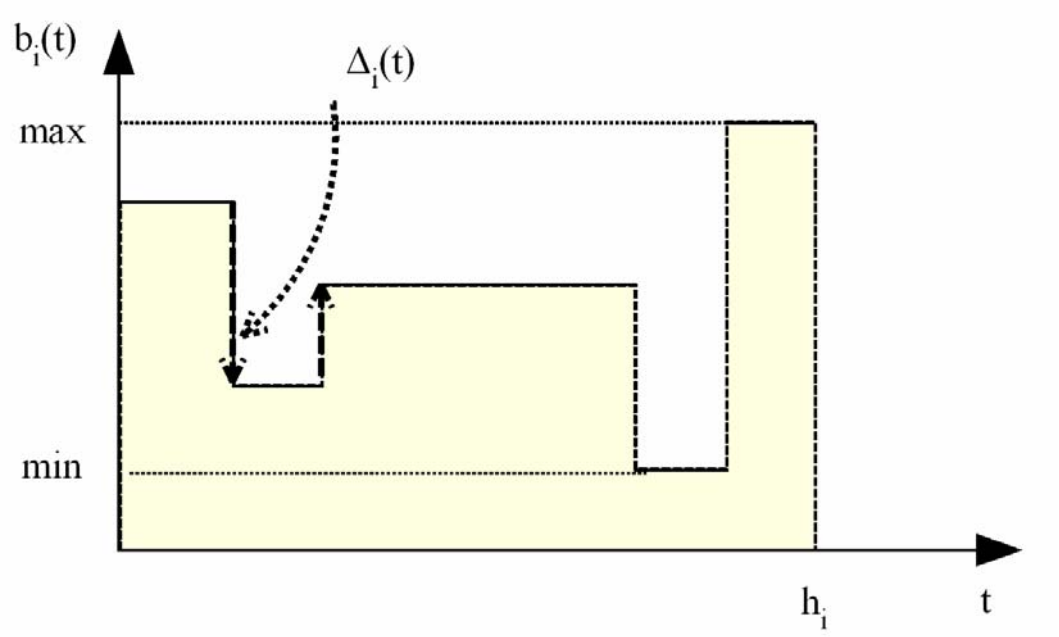


Figure 2 Multiple Leaky Bucket lead to linear or concave envelopes

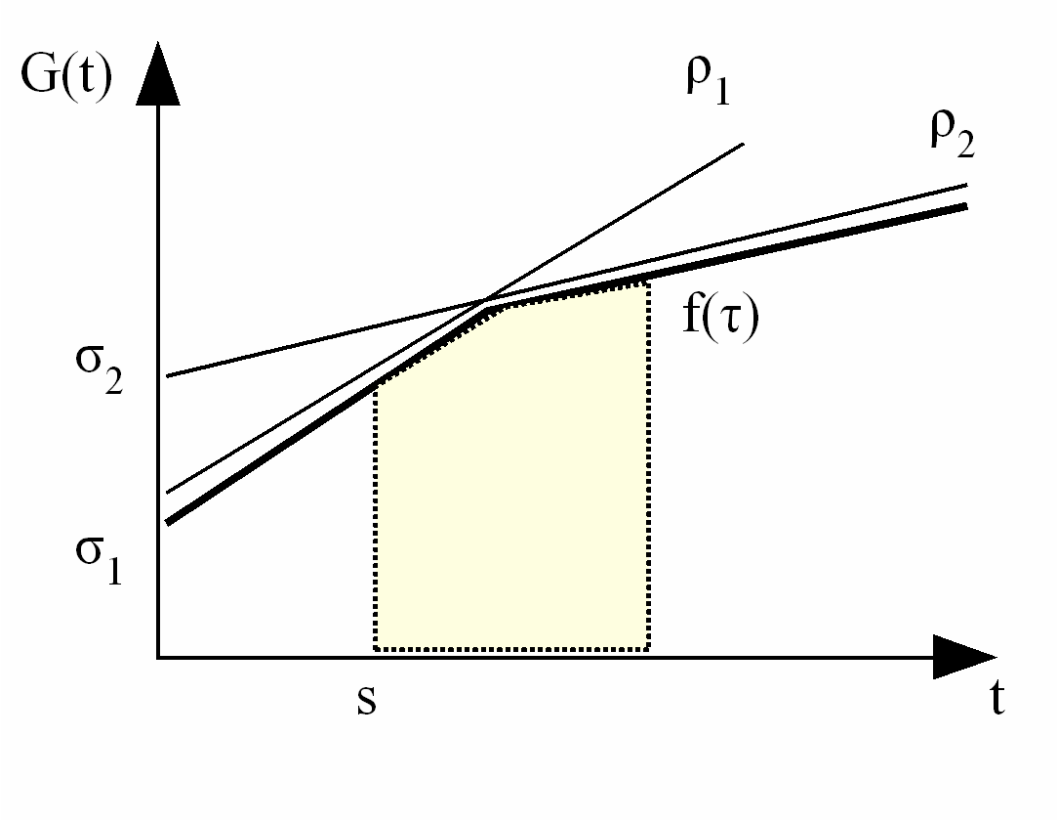


Figure 3 Simple example with two leaky buckets

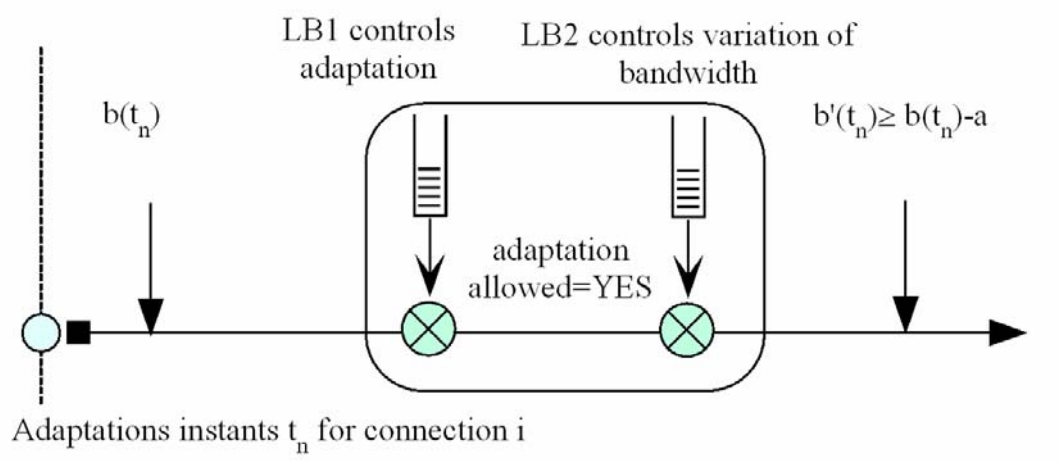


Figure 4 Combined Connection Admission and Bandwidth Adaptation Algorithm

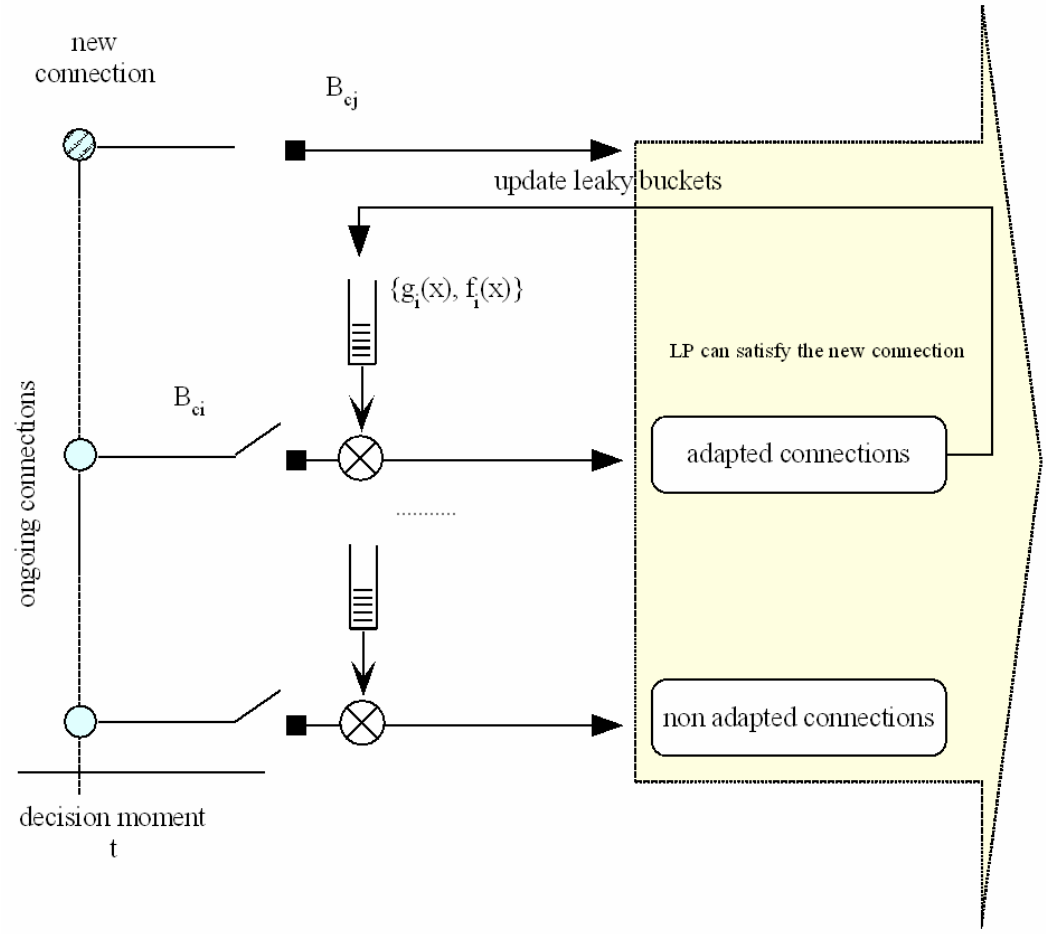


Figure 5 System performance (linear and concave reward functions) and user QoS when the regulators control bandwidth decrease.

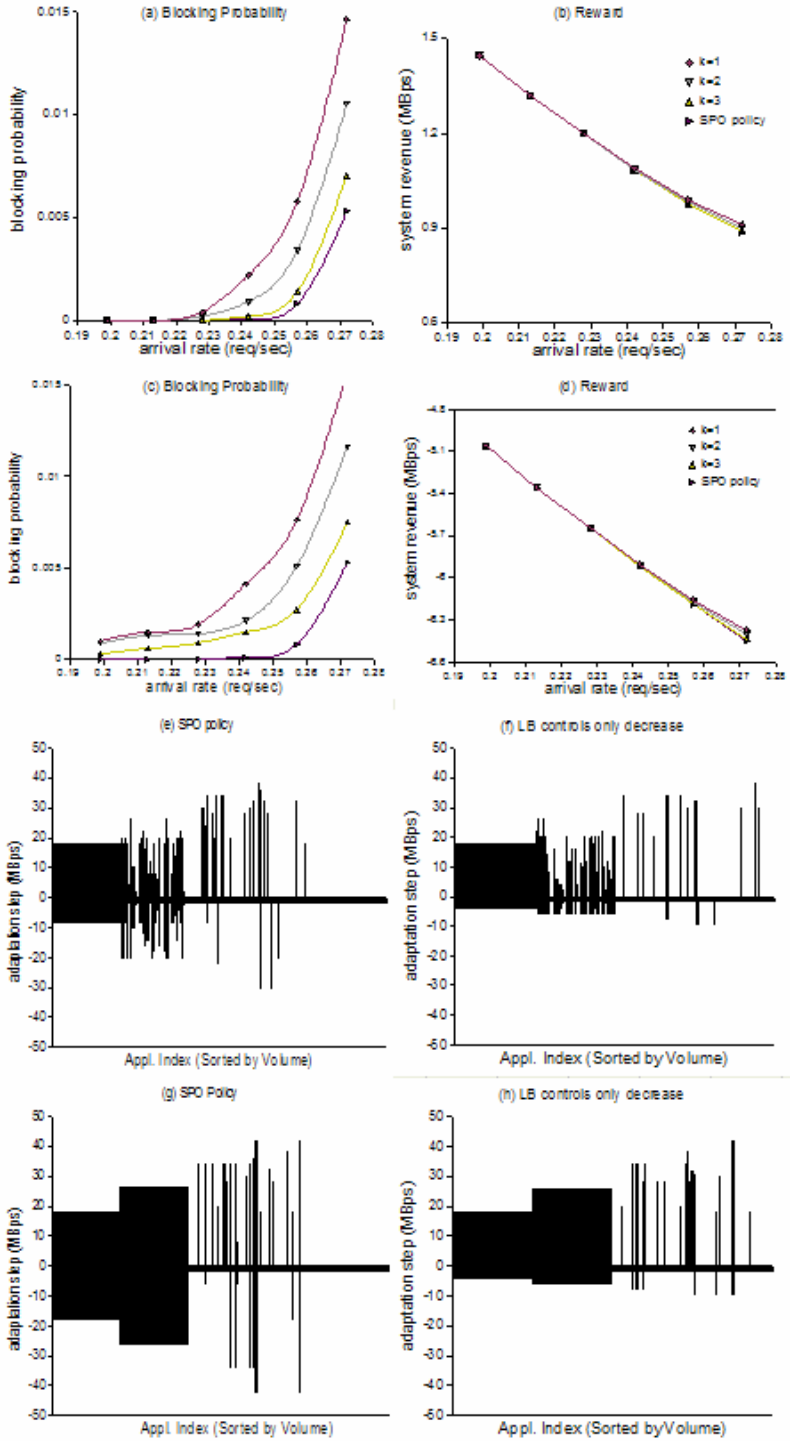
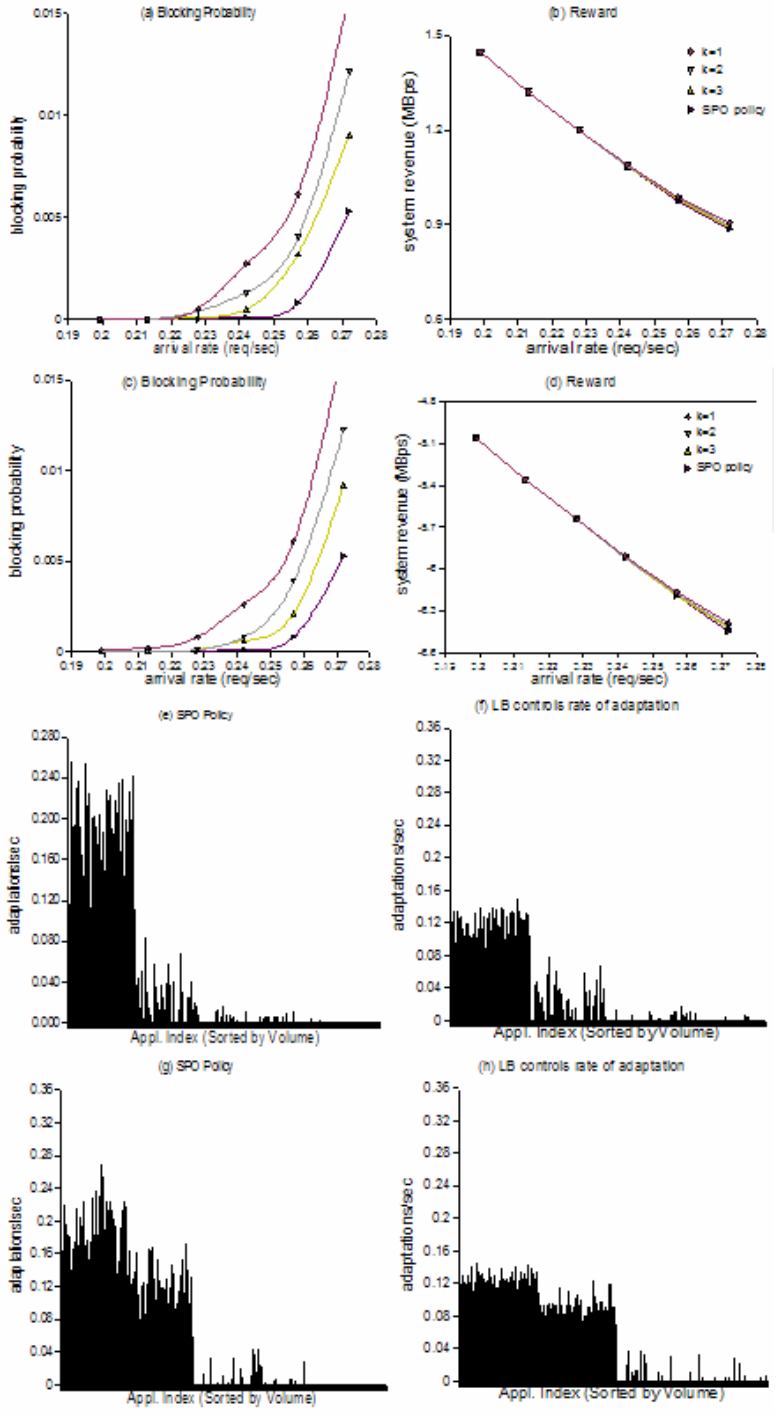


Figure 6 System performance (linear and concave reward functions) and user QoS when the regulators control frequency of rate adaptation.



Biography

Nikolaos G. Argiriou was born in Larissa, Greece on September 1, 1973. He received the Diploma degree in Electrical Engineering from the Dept. of Electrical Engineering, Telecommunication Division, Aristotle University of Thessaloniki, Greece, in 1996. He worked as a researcher, on secure medical image transmission over networks, at the Image Processing Lab at the same university during 1996-97. During 1998-2000 he was a researcher for the European Project Esprit Catserver concerning the use of advanced Quality of Service techniques in CATV networks. He is currently pursuing his Ph.D. degree at Aristotle University of Thessaloniki.
His current research interests are in the development and implementation of QoS techniques for wired and wireless networks.

Leonidas Georgiadis received the Diploma degree in electrical engineering from Aristotle University, Thessaloniki, Greece, in 1979, and his M.S. and Ph.D degrees both in electrical engineering from the University of Connecticut, in 1981 and 1986 respectively. From 1981 to 1983 he was with the Greek army.
From 1986 to 1987 he was Research Assistant Professor at the University of Virginia, Charlottesville. In 1987 he joined IBM T. J. Watson Research Center, Yorktown Heights as a Research Staff Member. Since October 1995, he has been with the Telecommunications Department of Aristotle University, Thessaloniki, Greece. His interests are in the area of wireless networks, high speed networks, routing, scheduling, congestion control, modeling and performance analysis.