Channel Sharing by Rate-Adaptive Streaming Applications ⋆

Nikos Argiriou, Leonidas Georgiadis

Electrical and Computer Engineering Dept., Aristotle University of Thessaloniki, Thessaloniki, Greece.

Abstract

There are various techniques for adapting the transmission rate of an application while maintaining the perceived quality at the receiver at acceptable levels. Shared channel systems can use this rate adaptation capability to increase the number of concurrent applications in the system. This can be achieved by appropriately modifying the rate of the already running applications when a new connection arrives in the system. In this paper we present an analytical model for a class of algorithms for channel sharing by rate adaptive applications. We provide means for calculating performance measures related to the quality of reception of an application. We also present the design of algorithms that ensure fair channel sharing while keeping the application performance within acceptable levels.

Key words: Rate Adaptation, Streaming Applications, Perfromance Analysis, Internet QoS, Channel Sharing, HFC Networks, Cellular Networks.

1 Introduction

As network capacities increase, user demand for supporting multimedia applications increases as well. Multimedia applications have stringent Quality of Service (QoS) requirements (guaranteed bandwidth, delay, jitter, loss). To support these requirements, new standards have been and are being developed by international standards organizations (ATM, ITU-T, IETF etc.).

⋆ The results of this paper were presented at INFOCOM 2002, New York, June 23-27.

Email addresses: narg@egnatia.ee.auth.gr (Nikos Argiriou), leonid@eng.auth.gr (Leonidas Georgiadis).

The efficient transport of multimedia applications requires new network capabilities, such as appropriate link scheduling mechanisms, traffic shaping, new routing protocols etc. Besides, the capability of applications to adapt to changing network conditions provides a promising means for using network resources efficiently while providing the required application QoS.

Multimedia application adaptation can be done at several layers of the network protocol stack [16]. In this work we concentrate on adaptation at the application layer. Adaptation at this layer consist of the capability of the application to adjust its bandwidth (rate) requirements. This can be achieved by various coding techniques such as layered coding [4], [13] and adaptation of compression parameters [3], [2], [15], as well as bandwidth smoothing [3], 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. In particular, wavelet coding [15] is particularly well suited for continuous rate adaptation. Rate adaptation implies some variability in the perceived quality of the application. There is a relatively large class of applications that can tolerate this variability, such as, video teleconferencing, interactive training, low-cost information distribution such as news, and even some entertainment video.

Using rate adaptation applications can adapt their transmission rate to changing network conditions in order avoid congestion [8], [11], [12]. On the other hand, there are proposals where Variable Bit Rate (VBR) connections request Constant Bit Rate (CBR) service from the network depending on their current needs [5].

The ability of applications to adapt their transmission rate is particularly useful in shared-channel environments such as Hybrid Fiber Coax (HFC) networks and broadband wireless cellular networks. These channels are shared by a number of users. The advantage of channel sharing is that when the number of active users is small they can share all the available bandwidth and hence receive very good QoS. The disadvantage is that as the number of users that share the same bandwidth increases, if the system is left uncontrolled, the perceived quality of multimedia applications reduces significantly. However proper admission control, combined with application rate adaptation has the potential of guaranteeing acceptable quality of reception while achieving large system utilization. With this approach, 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 (see Figure 1), in order to reduce their rate accordingly.

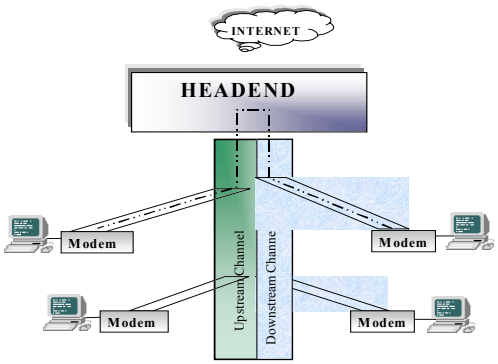

Fig. 1. CATV System

In this paper we address the issue of modeling and analysis of mechanisms for sharing a common channel by applications that can adapt their transmission rate within a certain range. Hence, the application rate is generally dependent on the system load. We concentrate on evaluating by analytical means metrics that affect the quality of reception of multimedia traffic. Several metrics related to the perceived quality of reception have been proposed and their development is still a subject of research [6], [7], [17], [14]. Experiments in [9] have shown that Perceptual Distortion Metrics (PDM) depend on the encoding method and increase as the average allocated rate to the connection increases, albeit nonlinearly. The study of particular PDMs and their relation to encoding methods is beyond the scope of this paper. Instead we concentrate on evaluating metrics related to allocated rate. These then can be combined with the particular encoding method to obtain the desired PDM. In addition to evaluating the average allocated rate, we develop methods for evaluating the frequency of rate adaptation that occur during the lifetime of a connection. While these are important metrics, there appears to be little literature on the subject. In [3] algorithms were designed to take into account quality of video reception under a different framework than ours, and their performance was studied by simulations and experiments. Our contribution consists in providing analytical means for evaluating the performance of a shared channel in the presence of rate adaptive applications.

The rest of the paper is organized as follows. In Section 2 we provide a model for channel sharing under a class of sharing algorithms and we analyze the system behavior under various performance measures. In Section 3 we examine specific channel sharing algorithms that fall into the class studied in Section 2 and provide numerical results. We conclude with Section 4 where we summarize our results and discuss directions for further research.

2 System Model and Analysis

We consider a communication link of capacity C. Connections arrive for transmission over the link as a Poisson process with rate λ. The connection holding times are exponential random variables independent of each other and independent of the arrival process. The average holding time of a connection is $1/\mu$. We define the system utilization $\rho = \lambda/\mu$. The assumption of exponential holding times, while not always realistic, allows for the detailed analysis of various performance metrics and provides insight to system operation. Moreover, we note that one of the main results of the paper, i.e., Theorem 1, holds for general holding time distributions.

The class of rate allocation algorithms that we consider is the following. If there are k connections in the system at time t, then the rate allocated to each connection is b_k. This assumes that each connection knows b_k. For HFC or wireless cellular systems this can be achieved by broadcasting b_k to the downstream channel. Alternatively, the sequence $\{b_k\}$ may be known to all connections a priori, in which case k needs to be broadcasted. Algorithms in this class provide the same rate to each connection in the system and are simple to implement. Moreover, as will be seen in Section 3, b_k can be chosen so that desirable performance metrics are obtained. More sophisticated algorithms where bandwidth is not necessarily allocated equitably among connections is a subject of further research.

We assume that b_k is decreasing with k. This assumption is not essential to the subsequent development and is made in order to simplify the notation. In any case, this is a reasonable assumption which is likely to hold in practical systems. Let $K_{\max}$ be the maximum number of connections that can be simultaneously active in the system. Clearly, the following inequality must hold

$$K_{\max} \leq \left\lfloor \frac{C}{b_{K_{\max}}} \right\rfloor .$$

If an arriving connection finds $K_{\max}$ connections in the system it is rejected. Otherwise, it is accepted. Hence, as far as connection admission is concerned we have an $M/M/K_{\max}$ loss system [18].

We are interested in deriving performance measures related to the manner in which rate is allocated to a connection throughout its holding time. One such measure is the mean rate allocated to a connection. One wishes to keep the mean allocated rate as high as possible. Another measure of interest is the frequency by which rate adaptation is taking place [3]. In Section 2.1 we examine in detail a method of deriving various metrics related to the mean allocated rate. In Section 2.2 corresponding measures are derived for the frequency of

rate adaptation.

2.1 Mean Allocated Rate

Assume that connection request c arrives at time $t = 0$ and is accepted (hence it finds $K \leq K_{\max} - 1$ other connections in the system). The connection holding time is H. Let $B_K(t)$ be the rate allocated to the connection at time t. The mean rate allocated to the connection while in the system is

$$\widehat{B}_K(H) = \frac{\int_0^H B_K(t)dt}{H}. \tag{1}$$

We are interested in evaluating

$$\overline{B} = E\left\{\widehat{B}_K(H)\right\},$$

as a measure of the overall performance of the connection as far as allocated rate is concerned. Another measure of interest is

$$\overline{B}_k = E\left\{\widehat{B}_K(H)\middle|\, K = k\right\},$$

which is useful when it is desirable to know the expected performance of the connection when it finds the system at a given state. Clearly, these two measures are related as follows.

$$\begin{aligned}\overline{B} &= E\left\{\widehat{B}_K(H)\right\} \\ &= E\left\{E\left\{\widehat{B}_K(H)\middle|\, K\right\}\right\} \\ &= \sum_{k=0}^{K_{\max}-1} \overline{B}_k q_k,\end{aligned} \tag{2}$$

where $q_k,\ 0 \leq k \leq K_{\max} - 1$, is the probability that the number of connections in the system found by connection c upon arrival is k. In steady state, it is known from the Poisson Arrivals See Time Averages (PASTA) property, [18], that the distribution of the number of connections found in the system by an arriving connection is equal to the time average distribution, π_k, of the number of connections in the system. Since the system under consideration behaves as a loss system as far as connection admission is concerned, we have that

$$\pi_k = \frac{\rho^k/k!}{\sum_{l=0}^{K_{\max}} \rho^l/l!},\ 0 \leq k \leq K_{\max}. \tag{3}$$

Since connection c is also accepted by the system, the number of connections

that it finds in the system must be at most $K_{\max} - 1$. Hence, we have

$$\begin{aligned} q_k &= \frac{\pi_k}{\sum_{l=0}^{K_{\max}-1} \pi_l} \\ &= \frac{\rho^k/k!}{\sum_{l=0}^{K_{\max}-1} \rho^l/l!}, \ 0 \leq k \leq K_{\max} - 1. \end{aligned} \tag{4}$$

Observe that q_k is the distribution of connections in a loss system that accepts up to $K_{\max} - 1$ connections.

When the holding time of a connection is known a priori, the following measures may also be of interest

$$\begin{aligned} \overline{B}(h) &= E\left\{ \widehat{B}_K(H) \middle| H = h \right\} \\ \overline{B}_k(h) &= E\left\{ \widehat{B}_K(H) \middle| K = k, \ H = h \right\}, \end{aligned}$$

which are related as follows

$$\overline{B}(h) = \sum_{k=0}^{K_{\max}-1} \overline{B}_k(h) q_k. \tag{5}$$

We start by providing means for evaluating $\overline{B}_k(h)$ and $\overline{B}_k$. Let us define

$$I_K(H) = \int_0^H B_K(t) dt.$$

Hence, $\widehat{B}_K(H) = I_K(H)/H$. Denote $\overline{I}_k(h) \triangleq E\left\{ I_K(H) | K = k, \ H = h \right\}$.

The analysis proceeds by setting up the basic recurrences of system evolution and then using Laplace transform approach to extract average values of basic performance metrics. More specifically, the analysis results in a system of equations, with the unknowns representing Laplace transforms, whose dimensionality is the maximum number of connections that can be admitted to the channel. By transform inversion, one then obtains the numerical quantities needed to evaluate performance.

Consider the first time, X_K, (after time $t = 0$) at which a new connection arrives or one of the K connections found by c upon arrival departs. It is well known that given K, X_K is an exponential random variable with rate $\lambda_{X_K} = \lambda + K\mu$. The form of $I_K(H)$ depends on whether X_K occurs before or after connection c leaves the system. Indeed,

- If $X_K \geq H$ (see Figure 2, (a)), then no rate adaptation occurs during the connection's holding time and hence $I_K(H) = b_{K+1}H$ (note that since there are in total $K + 1$ connections in the system, the connection is allocated rate b_{K+1}).

- If $X_K < H$ then up to time X_K the connection is allocated rate b_{K+1} and hence we can write

$$I_K(H) = b_{K+1}X_k + G_K(H),$$

where $G_K(H) = \int_{X_K}^{H} B_K(t)dt$. The statistics of $G_K(H)$ depend on whether a departure or an arrival occurs at time X_K.

If a departure occurs at X_K (hence $K \geq 1$, see Figure 2 (b)), then the number of connections in the system, other than connection c, becomes $K-1$. In this case, due to the statistical assumptions about connection arrival and holding times, $G_K(H)$ has the same distribution as $I_{K-1}(H - X_K)$.

If an arrival occurs at X_K and $K \leq K_{\max} - 2$ (see Figure 2 (c)), then the new connection is accepted by the system. By a similar reasoning we conclude that $G_K(H)$ has the same distribution as $I_{K+1}(H - X_K)$.

If an arrival occurs at X_K and $K = K_{\max} - 1$ (see Figure 2 (d)) then the new connection is rejected and $G_K(H)$ has the same distribution as $I_K(H - X_K)$.

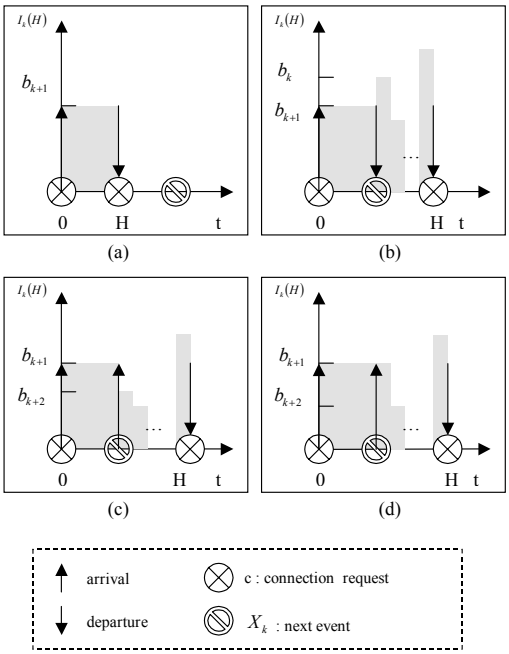

Fig. 2. Evolution of $I_K(H)$

Let us define

$$\overline{I}_k(h, x_k) = E\{\overline{I}_K(H)|K = k, H = h, X_k = x_k\}.$$

From the previous discussion we conclude the following:

- For $1 \leq k \leq K_{\max} - 2$,

$$\overline{I}_k(h, x_k) = \begin{cases} b_{k+1}h & x_k \geq h \\ b_{k+1}x_k + \overline{I}_{k+1}(h - x_k) & \text{arrival at } x_k < h \\ b_{k+1}x_k + \overline{I}_{k-1}(h - x_k) & \text{depart. at } x_k < h \end{cases} \tag{6}$$

- For $k = 0$ (no departure can occur at this state),

$$\overline{I}_k(h, x_k) = \begin{cases} b_{k+1}h & x_k \geq h \\ b_{k+1}x_k + \overline{I}_{k+1}(h - x_k) & \text{arrival at } x_k < h \end{cases} \tag{7}$$

- For $k = K_{\max} - 1$ (if an arrival occurs at this state and $X_K \leq h$, it is rejected since there are $K_{\max}$ active connections in the system),

$$\overline{I}_k(h, x_k) = \begin{cases} b_{k+1}h & x_k \geq h \\ b_{k+1}x_k + \overline{I}_k(h - x_k) & \text{arrival at } x_k < h \\ b_{k+1}x_k + \overline{I}_{k-1}(h - x_k) & \text{depart. at } x_k < h \end{cases} \tag{8}$$

Note that the events "arrival occurs at X_k" and "departure occurs at X_k" have respective probabilities

$$p_a = \frac{\lambda}{\lambda + k\mu}, \;\; p_d = \frac{k\mu}{\lambda + k\mu}.$$

Moreover, due to the statistical assumptions about the arrival and departure processes, these events are independent of the event $\{X_k \leq h\}$.

Let us define

$$\begin{aligned} f_{X_k}(x) &= (\lambda + k\mu)\, e^{-(\lambda + k\mu)x}, \\ F_{X_K}(x) &= 1 - e^{-(\lambda + k\mu)x}, \; F^c_{X_k} = e^{-(\lambda + k\mu)x}. \end{aligned}$$

Also, denote by $\otimes$ the convolution between two functions $g_1(h)$, $g_2(h)$, $h \geq 0$, i.e.,

$$(g_1 \otimes g_2)(x) = \int_0^h g_1(h - x) g_2(x) dx.$$

Averaging the recursive formulas (6), (7), (8), with respect to X_k, and setting for convenience $\overline{I}_{-1}(h) \equiv 0$, we have for $0 \leq k \leq K_{\max} - 2$,

$$\begin{aligned}\overline{I}_k(h) &= \int_0^\infty \overline{I}_k(h,x) f_{X_k}(x)dx \\ &= b_{k+1} h F^c_{X_k}(h) + b_{k+1}\int_0^h x f_{X_k}(x)dx \\ &+ \frac{\lambda}{\lambda+k\mu}\left(\overline{I}_{k+1}\otimes f_{X_k}\right)(h) \\ &+ \frac{k\mu}{\lambda+k\mu}\left(\overline{I}_{k-1}\otimes f_{X_k}\right)(h), \end{aligned} \tag{9}$$

and for $k = K_{\max} - 1$

$$\begin{aligned}\overline{I}_k(h) &= b_{k+1} h F^c_{X_k}(h) + b_{k+1}\int_0^h x f_{X_k}(x)dx \\ &+ \frac{\lambda}{\lambda+k\mu}\left(\overline{I}_{k}\otimes f_{X_k}\right)(h) \\ &+ \frac{k\mu}{\lambda+k\mu}\left(\overline{I}_{k-1}\otimes f_{X_k}\right)(h). \end{aligned} \tag{10}$$

Taking the Laplace transform of (9) and (10) we have respectively,

$$\begin{aligned}\overline{\mathbf{I}}_k(s) &= b_{k+1}\mathbf{F}^d_k(s) + \frac{\lambda}{\lambda+k\mu}\overline{\mathbf{I}}_{k+1}(s)\mathbf{F}_k(s) \\ &+\frac{k\mu}{\lambda+k\mu}\overline{\mathbf{I}}_{k-1}(s)\mathbf{F}_k(s),\ 0\le k\le K_{\max}-2, \end{aligned} \tag{11}$$

$$\begin{aligned}\overline{\mathbf{I}}_k(s) &= b_{k+1}\mathbf{F}^d_k(s) + \frac{\lambda}{\lambda+k\mu}\overline{\mathbf{I}}_{k}(s)\mathbf{F}_k(s) \\ &+\frac{k\mu}{\lambda+k\mu}\overline{\mathbf{I}}_{k-1}(s)\mathbf{F}_k(s),\ k = K_{\max}-1, \end{aligned} \tag{12}$$

where $\overline{\mathbf{I}}_k(s)$ is the Laplace transform of $\overline{I}_k(h)$ and

$$\mathbf{F}_k(s) = \frac{\lambda+k\mu}{s+\lambda+k\mu}, \tag{13a}$$

$$\mathbf{F}^d_k(s) = \frac{1}{s\left(s+\lambda+k\mu\right)}. \tag{13b}$$

For each s, (11) and (12) constitute a system of $K_{\max}$ equations in $K_{\max}$ unknowns, $\overline{\mathbf{I}}_k(s)$, $0\le k\le K_{\max}-1$. Hence $\overline{\mathbf{I}}_k(s)$ can be calculated. By numerically inverting the Laplace transforms, one can determine $\overline{I}_k(h)$, $0\le k\le$

$K_{\max} - 1$. It is then straightforward to compute

$$\overline{B}_k(h) = E\left\{ \widehat{B}_K(H) \middle| K = k,\ H = h \right\} = \frac{\overline{I}_k(h)}{h}.$$

Consider now $\overline{B}_k = E\left\{ \widehat{B}_K(H) \middle| K = k \right\}$. We have

$$\begin{aligned}
\overline{B}_k &= E\left\{ \frac{I_K(H)}{H} \middle| K = k \right\} = \\
&= E\left\{ E\left\{ \frac{I_K(H)}{H} \middle| H = h, K = k \right\} \middle| K = k \right\} \\
&= \int_0^\infty \frac{\overline{I}_k(h)}{h} \mu e^{-\mu h} dh, \qquad (14)
\end{aligned}$$

where we used the fact that K (the number of connections in the system when connection c arrives) is independent of the holding time of connection c. Hence, the computation can be performed once $\overline{I}_k(h)$ has been determined. However, this process is tedious since it requires that we first invert the Laplace transform of $\overline{I}_k(h)$ and next compute the integral. An alternative more efficient process can be obtained, which requires the integration of the Laplace transform directly, thus avoiding the inverse Laplace transform. This can be done as follows.

Define

$$\widetilde{I}_k(h) = \overline{I}_k(h) e^{-\mu h}.$$

From (9) it follows that for $0 \le k \le K_{\max} - 2$,

$$\begin{aligned}
\widetilde{I}_k(h) &= b_{k+1} h F_k^c(h) e^{-\mu h} + b_{k+1} e^{-\mu h} \int_0^h x f_X(x) dx \\
&+ \frac{\lambda}{\lambda + k\mu} \left(\widetilde{I}_{k+1} \otimes \left(f_{X_k} e^{-\mu x} \right) \right)(h) \qquad (15) \\
&+ \frac{k\mu}{\lambda + k\mu} \left(\widetilde{I}_{k-1} \otimes \left(f_{X_k} e^{-\mu x} \right) \right)(h),
\end{aligned}$$

while for $k = K_{\max} - 1$, (10) becomes

$$\widetilde{I}_k(h) = b_{k+1} h F_k^c(h) e^{-\mu h} + b_{k+1} e^{-\mu h} \int_0^h x f_X(x) dx + \frac{\lambda}{\lambda + k\mu} \left(\widetilde{I}_k \otimes \left(f_{X_k} e^{-\mu x} \right) \right)(h) + \frac{k\mu}{\lambda + k\mu} \left(\widetilde{I}_{k-1} \otimes \left(f_{X_k} e^{-\mu x} \right) \right)(h). \tag{16}$$

Taking the Laplace Transform of (15) and (16) we get

$$\widetilde{\mathbf{I}}_k(s) = b_{k+1} \widetilde{\mathbf{F}}_k^d(s) + \frac{\lambda}{\lambda + k\mu} \widetilde{\mathbf{I}}_{k+1}(s) \widetilde{\mathbf{F}}_k(s) + \frac{k\mu}{\lambda + k\mu} \widetilde{\mathbf{I}}_{k-1}(s) \widetilde{\mathbf{F}}_k(s), \ 0 \leq k \leq K_{\max} - 2, \tag{17}$$

and

$$\widetilde{\mathbf{I}}_k(s) = b_{k+1} \widetilde{\mathbf{F}}_k^d(s) + \frac{\lambda}{\lambda + k\mu} \widetilde{\mathbf{I}}_k(s) \widetilde{\mathbf{F}}_k(s) + \frac{k\mu}{\lambda + k\mu} \widetilde{\mathbf{I}}_{k-1}(s) \widetilde{\mathbf{F}}_k(s), \ k = K_{\max} - 1, \tag{18}$$

where

$$\widetilde{\mathbf{F}}_k(s) = \frac{\lambda + k\mu}{s + \lambda + (k+1)\mu}, \tag{19a}$$

$$\widetilde{\mathbf{F}}_k^d(s) = \frac{1}{(s+\mu)\left(s + \lambda + (k+1)\mu\right)}. \tag{19b}$$

Now, if the Laplace transform of $f(h)$ is $\mathbf{F}(s)$, then it holds [10],

$$\int_0^\infty \mathbf{F}(s) ds = \int_0^\infty \frac{f(h)}{h} dh.$$

Therefore, from (14) we conclude that

$$\overline{B}_k = \mu \int_0^\infty \widetilde{\mathbf{I}}_k(s) ds. \tag{20}$$

Hence, the algorithm for computing $\overline{B}_k$ is the following.

(1) For a given s, $\widetilde{\mathbf{I}}_k(s)$ are calculated by the solution of the linear system of equations (17) and (18).
(2) Since we can compute $\widetilde{\mathbf{I}}_k(s)$ for any s from step 1, we can use a numerical method of integration to compute $\overline{B}_k$ based on (20).

The following example illustrates the process developed above.

Example 1.

Let us consider the case $K_{\max} = 2$. Then, we have from (11) and (12).

$$\begin{aligned}\begin{bmatrix}\overline{\mathbf{I}}_0(s)\\ \overline{\mathbf{I}}_1(s)\end{bmatrix} &= \begin{bmatrix}1 & -\frac{\lambda}{s+\lambda}\\ -\frac{\mu}{s+\lambda+\mu} & \frac{s+\mu}{s+\lambda+\mu}\end{bmatrix}^{-1}\begin{bmatrix}\frac{b_1}{s(s+\lambda)}\\ \frac{b_2}{s(s+\lambda+\mu)}\end{bmatrix}\\ &= \begin{bmatrix}\frac{s+\mu}{s^2(s+\lambda+\mu)}b_1 + \frac{\lambda}{s^2}\frac{b_2}{s+\lambda+\mu}\\ \frac{\mu}{(s+\lambda+\mu)s^2}b_1 + \frac{1}{s^2}(s+\lambda)\frac{b_2}{s+\lambda+\mu}\end{bmatrix}.\end{aligned}$$

Inverting the Laplace transform we get,

$$\begin{aligned}\overline{I}_0(h) &= \frac{b_1}{\lambda+\mu}\left(1-e^{-h(\lambda+\mu)}\right)\\ &+\frac{\mu b_1}{(\lambda+\mu)^2}\left(-1+(\lambda+\mu)h+e^{-h(\lambda+\mu)}\right)\\ &+\frac{\lambda b_2}{(\lambda+\mu)^2}\left(-1+(\lambda+\mu)h+e^{-h(\lambda+\mu)}\right),\end{aligned} \tag{21}$$

and

$$\begin{aligned}\overline{I}_1(h) &= \frac{\mu}{(\lambda+\mu)^2}\left(-1+(\lambda+\mu)h+e^{-h(\lambda+\mu)}\right)b_1\\ &+\frac{b_2}{\lambda+\mu}\left(1-e^{-h(\lambda+\mu)}\right)\\ &+\frac{\lambda b_2}{(\lambda+\mu)^2}\left(-1+(\lambda+\mu)h+e^{-h(\lambda+\mu)}\right).\end{aligned} \tag{22}$$

Hence,

$$\overline{B}_0(h) = \frac{\overline{I}_0(h)}{h},\ \overline{B}_1(h) = \frac{\overline{I}_1(h)}{h},$$

and

$$\overline{B}(h) = \overline{B}_0(h)q_0 + \overline{B}_1(h)q_1 = b_1\frac{1}{(\rho+1)} + b_2\frac{\rho}{(1+\rho)}. \tag{23}$$

Similarly, from (15) and (16) we have

$$\begin{aligned}\begin{bmatrix}\widetilde{\mathbf{I}}_0(s)\\ \widetilde{\mathbf{I}}_1(s)\end{bmatrix} &= \begin{bmatrix}1 & -\frac{\lambda}{s+\lambda+\mu}\\ -\frac{\mu}{s+\lambda+2\mu} & \frac{s+2\mu}{s+\lambda+2\mu}\end{bmatrix}^{-1}\begin{bmatrix}\frac{b_1}{(s+\mu)(s+\lambda+\mu)}\\ \frac{b_2}{(s+\mu)(s+\lambda+2\mu)}\end{bmatrix}\\ &= \begin{bmatrix}\frac{s+2\mu}{(s+\mu)^2(s+\lambda+2\mu)}b_1 + \frac{\lambda}{(s+\mu)^2}\frac{b_2}{s+\lambda+2\mu}\\ \frac{\mu}{(s+\lambda+2\mu)(s+\mu)^2}b_1 + \frac{s+\lambda+\mu}{(s+\mu)^2}\frac{b_2}{s+\lambda+2\mu}\end{bmatrix}.\end{aligned}$$

Integrating the previous Laplace transforms we get from (20)

$$\begin{aligned}\overline{B}_0 &= \mu \int_0^\infty \widetilde{\mathcal{I}}_0(s)ds \\ &= b_1 \frac{1}{(\rho+1)} + b_2 \frac{\rho}{(1+\rho)} \\ &+ (b_1 - b_2)\frac{\rho \ln(\rho+2)}{\rho^2+2\rho+1},\end{aligned} \tag{24}$$

and

$$\begin{aligned}\overline{B}_1 &= \mu \int_0^\infty \widetilde{\mathcal{I}}_1(s)ds \\ &= b_1 \frac{1}{(\rho+1)} + b_2 \frac{\rho}{(1+\rho)} \\ &- (b_1 - b_2)\frac{\ln(\rho+2)}{\rho^2+2\rho+1}.\end{aligned} \tag{25}$$

Of course, we could obtain (24) and (25) from (14) but this would be more tedious. Finally, from (24), (25) and (2) we have

$$\overline{B} = b_1 \frac{1}{(\rho+1)} + b_2 \frac{\rho}{(1+\rho)}. \tag{26}$$

We observe from (21), (22), (23) and (26) that

$$\lim_{h\to\infty} \frac{\overline{I}_0(h)}{h} = \lim_{h\to\infty} \frac{\overline{I}_1(h)}{h} = \overline{B}(h) = \overline{B} = \frac{1}{1+\rho} b_1 + \frac{\rho}{1+\rho} b_2.$$

In fact, these formulas hold for general $K_{\max}$ and general holding time distributions. This is expressed in the following theorem.

Theorem 1 *For $0 \leq k \leq K_{\max} - 1$, any $h > 0$ and general holding time distributions the following relations hold.*

$$\lim_{h\to\infty} \overline{B}_k(h) = \overline{B} = \overline{B}(h) = \sum_{k=0}^{K_{\max}-1} b_{k+1} q_k.$$

The proof of this theorem is given in the Appendix. According to the theorem, the evaluation of $\overline{B}$ ($\overline{B}(h)$) is straightforward and does not need the knowledge of $\overline{B}_k$ ($\overline{B}_k(h)$) as is implied by (2) ((5)). Moreover, the theorem states that the average rate allocated to a connection is equal to the expected rate allocated to the connection at the moment it arrives to the system and independent of its holding time. The latter is an interesting and desirable property of the studied method of bandwidth allocation. Note that this property is not true if we condition on the number of connections found on arrival, i.e., for $\overline{B}_k(h)$. As the theorem states, the average rate allocated to an accepted connection that

finds k other connections in the system is only asymptotically (as $h \to \infty$) equal to the average rate allocated to a connection upon arrival, which is intuitively clear.

2.2 Frequency of Rate Adaptation

Assume again that connection c with holding time H finds K other connections in the system upon arrival. Let $S_K(H)$ be the number of times rate adaptation takes place during the lifetime of a connection. The frequency of rate adaptation is then

$$\widehat{R}_K(H) = \frac{S_K(H)}{H}.$$

As in Section 2.1 we develop formulas for the calculation of analogous measures related to average adaptation rates. These measures are

$$\begin{aligned}
\overline{R} &= E\left\{\widehat{R}_K(H)\right\}, \\
\overline{R}_k &= E\left\{\widehat{R}_K(H)\middle|\, K=k\right\}, \\
\overline{R}(h) &= E\left\{\widehat{R}_K(H)\middle|\, H=h\right\}, \\
\overline{R}_k(h) &= E\left\{\widehat{R}_K(H)\middle|\, K=k,\ H=h\right\}.
\end{aligned}$$

Let us define

$$r_k^+ = \begin{cases} 1 \text{ if } b_{k+1} \neq b_k \\ 0 \quad \text{otherwise} \end{cases}, \ 1 \leq k \leq K_{\max} - 1, \tag{27a}$$

$$r_k^- = \begin{cases} 1 \text{ if } b_{k-1} \neq b_k \\ 0 \quad \text{otherwise} \end{cases}, \ 2 \leq k \leq K_{\max}. \tag{27b}$$

It will be convenient to also define

$$r_{K_{\max}}^+ = 0, \ r_1^- = 0.$$

Observe that if connection c is in the system together with k other connections and there is a new arrival, then rate adaptation will take place if $b_{k+2} \neq b_{k+1}$ (recall that before the new arrival the rate allocated to connection c is b_{k+1}). Similarly, if one of the k connections leaves the system, rate adaptation will take place if $b_k \neq b_{k+1}$. Based on this observation and following the reasoning of section 2.1, we have for $\overline{S}_k(h) \triangleq E\left\{S_K(H)|K=k, H=h\right\}$ and $\overline{S}_k(h, x_k) = E\left\{\overline{S}_K(H)|K=k, H=h, X_k=x_k\right\}$:

- For $1 \leq k \leq K_{\max} - 2$,

$$\overline{S}_k(h, x_k) = \begin{cases} 0 & x_k > h \\ r^+_{k+1} + \overline{S}_{k+1}(h - x_k) & \text{arrival at } x_k \leq h \\ r^-_{k+1} + \overline{S}_{k-1}(h - x_k) & \text{depart. at } x_k \leq h \end{cases} \tag{28}$$

- For $k = 0$

$$\overline{S}_k(h, x_k) = \begin{cases} 0 & x_k > h \\ r^+_{k+1} + \overline{S}_{k+1}(h - x_k) & \text{arrival at } x_k \leq h \end{cases} \tag{29}$$

- For $k = K_{\max} - 1$,

$$\overline{S}_k(h, x_k) = \begin{cases} 0 & x_k > h \\ \overline{S}_k(h - x_k) & \text{arrival at } x_k \leq h \\ r^-_{k+1} + \overline{S}_{k-1}(h - x_k) & \text{depart. at } x_k \leq h \end{cases} \tag{30}$$

These recursions are similar to those in (6), (7) and (8). We can therefore parallel the approach taken in Section 2.1. Define for $0 \leq k \leq K_{\max} - 1$.

$$r_{k+1} = \lambda r^+_{k+1} + k\mu r^-_{k+1.}$$

The Laplace transform of $\overline{S}_k(h)$ satisfies the following equations.

$$\begin{aligned} \overline{\mathbf{S}}_k(s) &= r_{k+1}\mathbf{F}^d_k(s) + \frac{\lambda}{\lambda + k\mu}\overline{\mathbf{S}}_{k+1}(s)\mathbf{F}_k(s) \\ &+ \frac{k\mu}{\lambda + k\mu}\overline{\mathbf{S}}_{k-1}(s)\mathbf{F}_k(s),\ 0 \leq k \leq K_{\max} - 2, \end{aligned} \tag{31}$$

$$\begin{aligned} \overline{\mathbf{S}}_k(s) &= r_{k+1}\mathbf{F}^d_k(s) + \frac{\lambda}{\lambda + k\mu}\overline{\mathbf{S}}_k(s)\mathbf{F}_k(s) \\ &+ \frac{k\mu}{\lambda + k\mu}\overline{\mathbf{S}}_{k-1}(s)\mathbf{F}_k(s),\ k = K_{\max} - 1. \end{aligned} \tag{32}$$

where $\mathbf{F}_k(s)$, $\mathbf{F}^d_k(s)$ are given by (13a) and (13b) respectively.

We observe that (31), (32) are essentially the same as (11), (12), the only difference being that b_k is replaced by r_k. Hence the analysis of the previous section holds in this case as well.

Note: If it is desired to count only the times when reduction of bandwidth occurs during the lifetime of an accepted connection, then we can simply set $r_k^- = 0$ for all values of k.

3 Algorithms and Numerical Results

In this section we present examples of algorithm design for rate adaptation, as well as numerical examples of system performance. In Section 3.1 we present a basic algorithm for adapting the connection rate in a fair manner while making maximum use of the available channel capacity, while in Section 3.2 we present the design of an algorithm that reduces the frequency of rate adaptation without adversely affecting the allocated rate.

3.1 ***An Algorithm for Rate Adaptation***

Let us assume that the connection rate can be adapted in a continuous manner within the range $[b_{\min},\ b_{\max}]$ and that the link capacity is C. Hence, if there are at most $K_{\min} = \lfloor C/b_{\max} \rfloor$ connections in the system, they can all receive the maximum rate $b_{\max}$, while if the number of connections is larger than $\lfloor C/b_{\max} \rfloor$, then the connection rate must be reduced. In the latter case, we make the "fair" choice to adapt the rate of all currently running connections to the same value. With this choice, the maximum number of connections that can be accepted by the system is $K_{\max} = \lfloor C/b_{\min} \rfloor$. Hence, we have the following values for b_k.

$$b_k = \begin{cases} b_{\max} \text{ if } & k \leq K_{\min} \\ C/k \text{ if } K_{\min} < k \leq K_{\max} \end{cases}.$$

With this choice, applications either receive the maximum rate for best quality, or they share the channel capacity equitably while still receiving rate larger than the minimum specified. Since the rate can be adapted to any value within the specified range, the whole channel capacity is used by the connections whenever possible.

We consider a channel with capacity $C = 2MBps$ and assume $b_{\max} = 0.5Mbps$, $b_{\min} = .2Mbps$, $\mu = 1$. Hence, $K_{\min} = 4$ and $K_{\max} = 10$. We vary the arrival rate (and hence the utilization ρ) from 0.2 to 7.0. When $\rho = 7.0$. the connection blocking probability is about 0.1, which is at the upper limit of acceptable values.

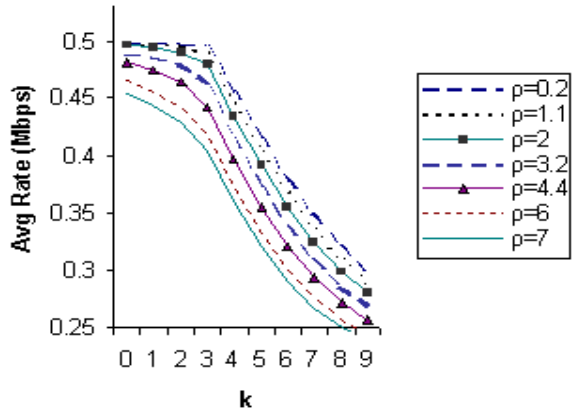

Fig. 3. Average allocated rate $\overline{B}_k$ as a function of k, for various values of ρ.

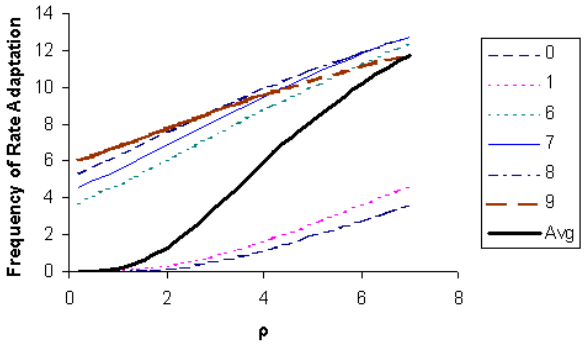

Fig. 4. Frequency of rate adaptation $\overline{R}_k$ as a function of ρ, for various values of k.

In Figure 3 we plot the values of average allocated rate, $\overline{B}_k$, as k varies, for various values of ρ. As expected, for given ρ the average bandwidth decreases as k increases. Moreover, for a given k, $\overline{B}_k$ fluctuates by at most $0.05Mbps$ as ρ varies from 0.2 to 7.0. Therefore, the average allocated bandwidth is not heavily dependent on the system utilization ρ. In figure 4 where we plot the frequency of rate adaptation, $\overline{R}_k$, as ρ varies, for various values of k. We see that the maximum value of frequency of rate adaptation is 12. In the same figure we also plot the average value of the frequency or rate adaptation, $\overline{R}$, as a function of ρ. We observe that the $\overline{R}$ is negligible $\rho < 1$ and increases sharply for $\rho > 2$.

Another interesting observation in Figure 4 is that $\overline{R}_9$ becomes smaller than $\overline{R}_8$, $\overline{R}_7$, $\overline{R}_6$, as ρ increases. This is explained as follows. A rate adaptation may occur either when a new arrival or a departure occurs during the lifetime of a connection. The higher k is, the higher the likelihood that a departure occurs. For small utilization where the likelihood of new arrivals is small, rate adaptation is due mainly to departures and hence $\overline{R}_9$ is larger than $\overline{R}_k$

for $k \leq 8$. As the utilization increases the likelihood of new arrival causing a rate adaptation is increased. However, when $k = 9$, the total number of connections in the system is 10, and new arrivals will be rejected thus causing no rate adaptation. Hence $\overline{R}_9$ is not affected by the increased likelihood of arrivals and eventually it becomes smaller than $\overline{R}_k$ for $k \leq 8$.

3.2 ***Reducing the Frequency of Rate Adaptation***

Consider now that we would like to reduce the frequency of rate adaptation observed in the previous setup. This could be achieved by setting $b_k = b_{k+1} = ..b_{k+n}$ for some values of k and n, and hence avoiding rate adaptation for certain state changes. Trivially, one could set $b_1 = ... = b_{K_{\max}} = b_{\min}$ so that no rate adaptation occurs. However, this way the connection rate will always be minimal. Hence, there is a trade-off between frequency of rate adaptation and average rate consumed by a connection. In addition, it is desirable to avoid abrupt rate adaptation, ensuring that the difference between two consecutive rate adaptations is always smaller than a given number a.

The previous considerations lead to the following design. We assume that the connection rate can again be adapted in a continuous manner. We also assume that $C/b_{\max}$ is integer

(1) Set $b_k = b_{\max}$ for $k \leq K_{\min}$.
(2) Set $K = K_{\min}$, $l = 0$
(3) Until $K \geq K_{\max}$ do
(4) For $k = K + 1, ..., \max\{K + l, K_{\max}\}$ set $b_k = \frac{C}{\max\{K+l, K_{\max}\}}$, where l is determined from the requirement that $b_K - b_{K+l} \leq a$, i.e.,

$$l = \left\lfloor \frac{aK^2}{C - aK} \right\rfloor .$$

(5) $K \leftarrow K + l$
(6) end /* Until do loop */

In order to ensure that $l \geq 1$ in the previous algorithm, the value of a must satisfy

$$a \geq \frac{C}{K_{\min} + K_{\min}^2}.$$

In Figure 5 we plot the values of b_k for various values of a, including the original algorithm in Section 3.1 (denoted as "cont" in the figure). We see that as expected, as a increases there are larger flat areas in the resulting

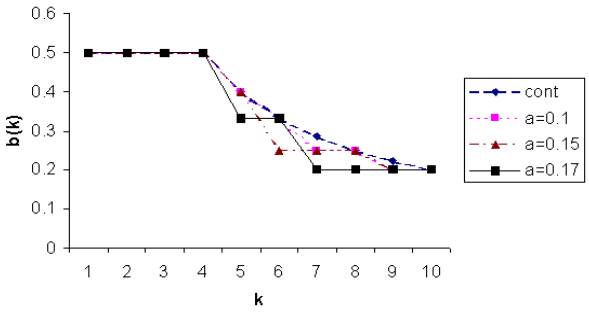

Fig. 5. Plots of b_k for various for various values of a.

curves. In figure 6 we plot the frequency of rate adaptation as k varies, for various values of ρ and for $a = .15$. We see that the maximum frequency of rate adaptation is 6.5 while for the algorithm of Section 3.1 the corresponding value is 12.

In Figures 7 and 8 we present the average allocated rate and frequency of rate adaptation as ρ varies, for various values of a. We see that for high utilization the frequency of rate adaptation can be reduced significantly at the expense of a moderate decrease in average allocated rate.

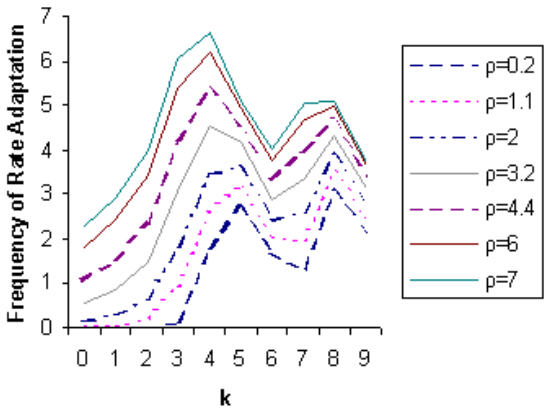

Fig. 6. Frequency of rate adaptation $\overline{R}_k$ as a function of k for various values of ρ.

4 Conclusions

We provided a model and analysis for a class of algorithms for channel sharing by rate adaptive applications. The developed formulas provide closed

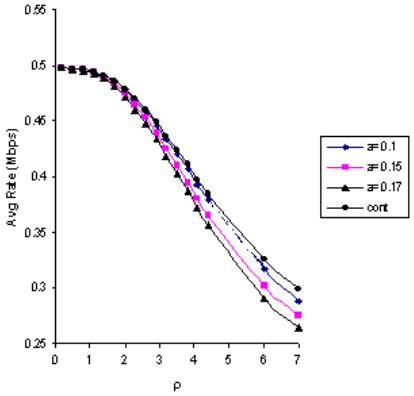

Fig. 7. Average allocated rate, $\overline{B}$, as a function of ρ, for various values of a.

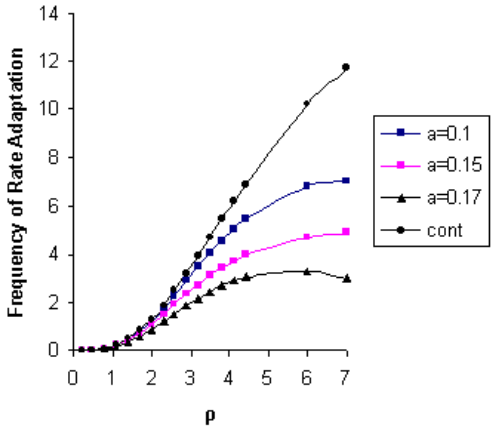

Fig. 8. Frequency of rate adaptation $\overline{R}$ as a function of ρ for various values of a.

form solutions for small systems, while for larger systems the measures of interest can be computed numerically. The formulas for the average allocated rate and the frequency of rate adaptation are very simple and hold for general service time distributions. We also showed how the results can be used in the design and evaluation of channel sharing algorithms.

There are several directions in which the current work can be extended. Regarding the generality of the results, it is desirable to understand the effect of more general service time distributions on system performance, for the rest of the measures studied in this paper. As far as the system model is concerned, we addressed in the current work the case of a single service class where all applications have the same holding times and performance requirements. It is interesting to study in this context the case of multiple service classes with different QoS requirements, as well as other bandwidth allocation algorithms. The inclusion of VBR connections which may need to alter their rate in addition to adjusting to system request is also an important issue. Also another

issue is the extension of the current approach to multi-hop networks.

References

[1] C. Adjih, N. Argiriou, M. Chaudier, E. Deberdt, F. Dumontet, L. Georgiadis and P. Jacquet, "An Architecture for IP Quality of Service Provisioning over CATV Networks," *EMMSEC 1999,* Sweden, June 1999.

[2] J.C. Bolot and T. Turletti, "A Rate Control Mechanism for Packet Video in the Internet," *IEEE INFOCOM,* November 1994.

[3] N. G. Duffield, K. K. Ramakrishnan and A. R. Reibman, "SAVE: An Algorithms for Smoothed Adaptive Video over Explicit Rate Networks," *IEEE INFOCOM,* April 1998.

[4] A. Eleftheriadis and D. Anastasiou, "Optimal Data Partitioning of MPEG-2 Coded Video," *First IEEE International Conf. on Image Processing,* November 1994.

[5] M. Grossglauser, S. Keshav and D. Tse, "RCBR: A Simple and Efficient Service for Multiple Time-Scale Traffic," *ACM SIGMCOMM'95,* September 1995.

[6] ITU-T, "Objective quality measurement of telephone band and (300-3400 Hz) speech codecs", Series P: Telephone transmission quality, telephone installation, local line networks, P.861, February 1998.

[7] ITU-500-R, "ITU-500-R Recommendation BT.500-8 - Methodology for the subjective accessment of the quality of television pictures", 1998.

[8] H. Kanakia, P.P. Mishra and A. Reibmand, "An Adaptive Congestion Control Scheme for Real-Time Video Transport" *IEEE/ACM Transactions on Networking,* Vol.3, No. 6 (Dec. 1995), pp. 671-682.

[9] J. Kimura, F. A. Tobagi, J-M. Pulido and P. J. Emstad, "Perceived Quality and Bandwidth Characterization of Layered MPEG-2 Video Encoding," Proceedings of the SPIE International Symposium on Voice, Video and Data Communications, Boston, Mass., September 1999.

[10] I. N. Sneddon, *The Use of Integral Transforms,* Tata McGraw Hill Co, 1974.

[11] T. V. Lakshman, P.P. Mishra and K. K. Ramakrishnan, "Transporting Compressed Video over ATM Networks with Explicit Rate Feedback Control," *IEEE INFOCOM,* Kobe, Japan, April 1997.

[12] P.P. Mishra, "Fair Bandwidth Sharing for Video Sources Using Distributed Feedback Control," *IEEE GLOBECOM*, Singapore, November 1995.

[13] P. Pancha and M. Zarki, "Prioritized Transmission of Variable Bit Rate MPEG Video," *IEEE GLOBECOM,* pp 1135-1138, December 1992.

[14] A. R. Prasad, R. Esmailzadeh, S. Winkler, T. Ihara, B. Rohani, B. Pinguet and M. Capel, "Perceptual Quality Measurement and Control: Definition, Application and Performance," *Proc. 4th International Symposium on Wireless Personal Multimedia Communications*, pp. 547-552, Aarborg, Denmark, September 9-12, 2001

[15] D. Taubman and 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, August 1996.

[16] B. Vandalore, W. Feng, R. Jain and S. Fahmy, " A Survey of Application Layer Techniques for Adaptive Streaming of Multimedia," *Journal of Real Time Imaging*, volume 7, issue 3, pp. 221-235, June 2001.

[17] S. Winkler, "A perceptual Distortion Metric for Digital Color Video," *Proceedings of SPIE Human Vision and Electronic Imaging*, vol 3644, pp 175-184, Jan. 1999.

[18] R. W. Wolff, *Stochastic Modeling and the Theory of Queues*, Prentice Hall, 1989

5 APPENDIX A

5.1 Proof of Theorem 1

We restate the theorem here.

Theorem 1: *For* $0 \leq k \leq K_{\max} - 1$, $h > 0$ *and for general service time distributions the following relations hold.*

$$\lim_{h\to\infty} \frac{\overline{I}_k(h)}{h} = \overline{B} = \overline{B}(h) = \sum_{k=0}^{K_{\max}-1} b_{k+1} q_k.$$

We will need the following general theorem for M/G/c Loss Systems [18]. Consider an M/G/c system where the service (holding) time distribution is $G(x) = P(S \leq x)$ with average service rate μ. Let N be the steady-state number of customers (connections) in the system and S_{r_i} the steady-state remaining service times of each of these customers. Let also π_k be the probability distribution given by (3) and G_e the equilibrium distribution of G, that is,

$$G_e = \mu \int_0^x (1 - G(x))dx.$$

Theorem (Erlang's Loss Formula). For an M/G/c loss system with arrival rate λ and service rate μ, $0 < \mu < \infty$, the joint limiting and stationary

distribution of $(N, S_{r1}, ..., S_{rN})$ is

$$P(N = n, S_{r_i} \leq x_i,\ i = 1, ..., n) = \pi_n \prod_{i=1}^{n} G_e(x_i).$$

Proof. Consider first

$$\lim_{h\to\infty} \frac{\overline{I}_k(h)}{h} = \sum_{k=0}^{K_{\max}-1} b_{k+1} q_k. \tag{33}$$

Once an arriving connection is accepted by the system, then, throughout the connection's holding time, the system behaves as a loss system Q that can accept $K_{\max} - 1$ connections. The newly arriving connection receives a reward rate of b_{k+1} when the system Q is at state k (since there are in total $k + 1$ connection in the original system). From the general theory of regenerative processes [18] it is known that the long term time-average reward (left hand side of (33) is equal to the steady-state average reward (right hand side of (33)) -recall from (4) that the q_k is the steady-state distribution of the number of connections in system Q.

We now prove that

$$\overline{B}(h) = \sum_{k=0}^{K_{\max}-1} b_{k+1} q_k.$$

Let an arriving connection be accepted, i.e., it finds $K = k \leq K_{\max} - 1$ connections in the system and let $\widehat{S}_{r_i}$, $i = 1, ..., k$ be the remaining service time of the connections in the system found upon arrival. We have for $0 \leq k \leq K_{\max} - 1$.

$$P(\, K = k, \widehat{S}_{r_i} \leq x_i,\ i = 1, ..., k \Big|\, K \leq K_{\max}{-}1) = \frac{P(K = k, \widehat{S}_{r_i} \leq x_i,\ i = 1, ..., k)}{P(K \leq K_{\max} - 1)}.$$

Applying now PASTA and Erlang's Loss Formula we have

$$\begin{aligned}
\frac{P(K = k, \widehat{S}_{r_i} \leq x_i,\ i = 1, ..., k)}{P(K \leq K_{\max} - 1)} &= \frac{P(N = k, S_{r_i} \leq x_i,\ i = 1, ..., k)}{P(N \leq K_{\max} - 1)} \\
&= \frac{\pi_k}{\sum_{i=1}^{K_{\max}-1} \pi_i} \prod_{i=1}^{k} G_e(x_i) \\
&= q_k \prod_{i=1}^{k} G_e(x_i).
\end{aligned}$$

That is,

$$P(\, K = k, \widehat{S}_{r_i} \leq x_i,\ i = 1, ..., k \Big|\, K \leq K_{\max} - 1) = q_k \prod_{i=1}^{k} G_e(x_i).$$

From the last formula we see that after a given connection is admitted in the system and throughout its holding time, the system behaves as a $M/G/\left(K_{\max}-1\right)$ system *in steady state.* Therefore, if the connection holding time is h, then for any time $t \leq h$ we have

$$E\left\{B_K(t)\middle|\, H=h\right\} = \sum_{k=0}^{K_{\max}-1} b_{k+1}q_k.$$

It follows that

$$\begin{aligned} E\left\{I_K(H)\middle|\, H=h\right\} &= E\left\{\int_0^h B_k(t)dt\middle|\, H=h\right\} \\ &= \int_0^h E\left\{B_k(t)dt\middle|\, H=h\right\}dt \\ &= h\sum_{k=0}^{K_{\max}-1} b_{k+1}q_k, \end{aligned}$$

and

$$\overline{B}(h) = E\left\{\frac{I_K(H)}{H}\middle|\, H=h\right\} = \sum_{k=0}^{K_{\max}-1} b_{k+1}q_k.$$

Finally, $\overline{B} = E\left\{\overline{B}(H)\right\} = \sum_{k=0}^{K_{\max}-1} b_{k+1}q_k$. ■