ELSEVIER

Computer Communications 23 (2000) 1336–1348

computer communications

www.elsevier.com/locate/comcom

Traffic shaping for MPEG video transmission over the next generation internet

M.F. Alam[a], M. Atiquzzaman[b,*], M.A. Karim[c]

[a]*Electro-Optics Program, University of Dayton, Dayton, OH 45469-0245, USA*
[b]*Department of Electrical and Computer Engineering, University of Dayton, Dayton, OH 45469-0226, USA*
[c]*Department of Electrical and Computer Engineering, University of Tennessee, Knoxville, TN 37996-2100, USA*

Abstract

The Internet Engineering Task Force (IETF) has proposed the Guaranteed Service (GS) in the Integrated Services model with firm delay and bandwidth guarantees. In this paper, we study the effects of a token bucket traffic shaper at the source on the transmission characteristics of Motion Picture Experts Group (MPEG) compressed video streams over the GS. We develop an analytical model of the traffic shaper, and also carry out numerical simulation of the transmission performance using MPEG trace data from several different video sequences. The analytical model and the numerical simulation results are in excellent agreement. Our study provides a technique to determine the token bucket parameters that have to be specified while setting up a GS flow. Our study reveals that the traffic shaping process gives an end application a wide range of flexibility in controlling the delay. By choosing appropriate token bucket traffic shaper parameters, delay can be reduced significantly without the need to reserve costly network resources for less delay-sensitive applications, while time-critical applications may specify the required token bucket parameters for minimum delay.

Keywords: Quality of service; Token bucket traffic shaping; Internet video; MPEG; Integrated services

1. Introduction

Present-day Internet Protocol (IP) provides a service that can be characterized as a best-effort service. Current IP network elements treat all traffic equally without any mechanism for providing priority to the packets carrying delay-sensitive applications like real-time video. The best-effort model is sufficient for non-real-time data communication applications like web browsing, file transfer, remote login and electronic mail. However, within the next few years, an exponential growth of real-time applications such as video and multimedia over data communication networks and the Internet is expected. In order to provide quality of service (QoS) guarantee to real-time applications, the Internet Engineering Task Force (IETF) has defined two new services on IP networks which are collectively called *Integrated Services*: the Guaranteed Service (GS) and the Controlled Load (CL) Service. The GS [1] provides guarantees of bandwidth as well as end-to-end-delay, and is designed to support real-time applications like video-conferencing, which is based on variable-bit-rate (VBR) compressed video transmission. The CL Service [2] provides a service which closely approximates the behavior visible to applications receiving best-effort service under lightly loaded conditions of the network. There is no guarantee of delay in the CL service, and it is best suited for real-time applications that can adapt themselves to the changing conditions of the network.

The MPEG standard for video coding [3] has received worldwide acceptance for storage and transmission of compressed video. MPEG transmission on Asynchronous Transfer Mode (ATM) networks is already a topic of extensive research [4–13]. Owing to the highly bursty nature of compressed video streams, traffic shaping and traffic smoothing are required for efficient utilization of bandwidth and network resources at various points in a network. While setting up a video or multimedia flow, network resources have to be reserved. The reservation process depends heavily on the characteristics of the video traffic to be generated by the end application and the required level of service guarantees. A traffic shaper at the source is required so that the traffic generated by the source is conformant to the pre-negotiated traffic specification. In addition, a traffic shaper can be employed to reshape the traffic so that the

* Corresponding author. Tel.: + 1-937-229-3183; fax: + 1-937-229-4529.

E-mail addresses: alammdfe@flyernet.udayton.edu (M.F. Alam), atiq@ieee.org (M. Atiquzzaman), karim@utk.edu (M.A. Karim).

0140-3664/00/$ - see front matter
PII: S0140-3664(00)00180-8

packet stream sent out to the network is less bursty than the original MPEG stream. This smoother traffic will require fewer resources from the network (e.g. buffer size, burst capacity etc.) and hence may be less costly than a packet stream without any traffic shaping. The *aim* of this paper is to study the design requirements of a traffic shaper for efficient digital video transmission over the GS category of the Integrated Services.

Previous studies on video traffic shaping primarily focussed on dynamically controlling the data rate of the source [4–6], dynamic adjustment of the characteristics (e.g. bandwidth) of a connection [7–11], and reduction of resource requirements by statistical multiplexing of multiple sources [12,13]. Most of these studies are based on MPEG video transmission over ATM networks, although some studies on MPEG transmission using the Transmission Control Protocol (TCP) over IP networks have been carried out [14]. However, the IETF has specified the User Datagram Protocol (UDP) for transmitting MPEG streams in the Integrated Services model [15]. The connectionless UDP protocol has been chosen as the transport layer protocol for fast transport of real-time streams on IP networks instead of the highly reliable but slow connection-oriented TCP protocol [15]. As far as the authors are aware, *there has not been any significant study on traffic shaping for MPEG video transmission over the Integrated Services IP networks*. Thus, the objective of this paper is to analyze the requirements of a traffic shaper for MPEG video transmission using the UDP-over-IP transport mechanism, and the effects of the shaper parameters on the transmission characteristics.

Transmission of MPEG-compressed video streams over the emerging services of the Internet, like the Integrated Services, is of significant importance due to the large installed base of the IP networks and the possibility of implementation of these services in near future, specially in corporate networks. Recent studies on end-to-end delay of video transmission over GS have assumed either a constant-bit-rate traffic [16] or a simple leaky-bucket traffic shaping [17] at the transmitting end. Depending on the shaper parameters chosen, a leaky-bucket or constant-bit-rate traffic shaper either fails to utilize the full burst-handling capability of the GS, or cannot ensure that the IP packets sent out to the network will be conformant to the traffic specifications (TSpec) that are specified when a GS flow is set up. In addition, a leaky-bucket traffic shaper reduces or removes the burstiness of the MPEG data stream, and may introduce large amount of delay which can result in a large buffer requirement at the traffic shaper. Such shaping delays contribute to considerable increase in end-to-end delays in MPEG video transmission, specially if the traffic shaper fails to utilize the allowed bursts in the traffic that can be supported by a GS flow. Thus, a proper traffic shaping mechanism is required at the transmitting end for MPEG video transmission over the GS so that the sending application can transmit traffic bursts up to its allowed limit which was negotiated at flow setup time. The shaper should also ensure that the traffic is not too bursty, and the shaper output is conformant to the negotiated traffic specification.

Since the token bucket algorithm (with peak rate control) has been specified as the traffic policing mechanism in the GS [1], a token bucket (followed by a leaky bucket for peak rate control) can be used also at the source to shape the traffic [18] so that the traffic profile is conformant to the negotiated traffic specification. Although it is possible to adopt a traffic shaping algorithm [19] which ensures that the transmitted traffic is conformant to a token bucket traffic policing element in the network, the token bucket itself can also be used at the source to shape MPEG video traffic.

In this paper, *we analyze a token bucket (with leaky bucket rate control) traffic shaper* of IP packets at the source end of an MPEG video source. The token bucket serves as a controlling mechanism for the traffic profile generated by the source. We develop an *analytical model* for studying the effects of the traffic shaper parameters on delay and buffer requirement. Methods for *choosing parameters* for the traffic shaper are discussed, and then *the effect of the shaper parameters on delay, jitter, and buffer requirement* are studied. In order to *verify the validity of the proposed model* under real MPEG traffic flow, we used *simulation techniques* to calculate delay and buffer requirements using trace data from several different MPEG video sequences. The simulation results show *excellent agreement* with the analytical model. The significance of the study carried out in this paper is that there has not been any previously known study on traffic shaping of MPEG video sequences using the token bucket algorithm, and the results give us considerable insight into the behavior of an MPEG video stream when subjected to token bucket traffic shaping. In addition, a *major contribution* of this paper is that it discusses a *clear methodology for quantitatively specifying the traffic parameters* that need to be specified for a certain level of QoS (e.g. delay, jitter, etc.) when setting up a GS flow.

The rest of the paper is organized as follows. In Section 2, we discuss MPEG video transmission using the GS, and the need for traffic shaping at the source. In Section 3, we develop a model to analyze the effects of various traffic shaper parameters on the delay and buffer size of the token bucket traffic shaper. In Section 4, we discuss the simulation procedure for calculating delay and buffer requirement using MPEG trace data. In Section 5, we present and compare results from both the analytical model and numerical simulation. The results include delay, jitter and buffer requirement as a function of various shaper parameters. Conclusions from our study are presented in Section 6.

2. Guaranteed service and MPEG transmission

In this section, we discuss the transport of MPEG streams

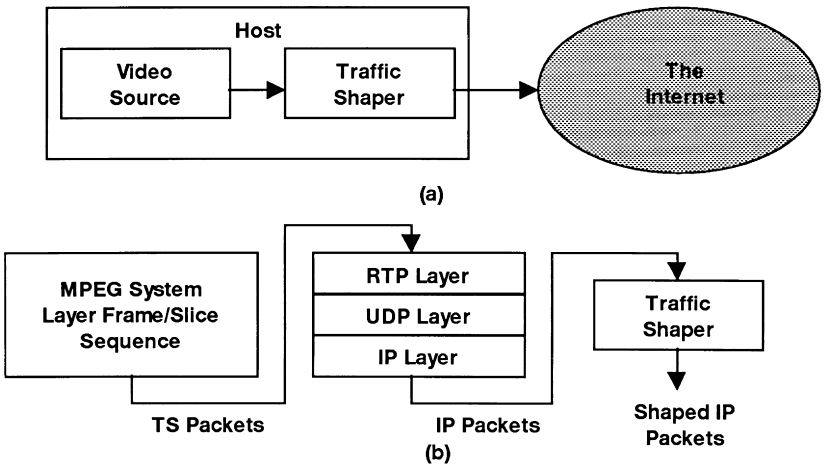

Fig. 1. (a) The system block diagram for transmission of an MPEG video stream; and (b) The network protocol layers for MPEG transmission over the GS IP network.

over UDP, the TSpec required for setting up an MPEG flow, and the traffic shaping of MPEG streams.

2.1. Transport of MPEG streams

A real-time stream using any of the Integrated Services on the Internet is established when an end application reserves bandwidth from the network using the Resource ReSerVation Protocol (RSVP) [20]. RSVP is a control protocol, which is used by network elements to exchange information regarding bandwidth and delay guarantees that can be supported by each element. IETF has defined the Real-time Transport Protocol (RTP) for delivering real-time traffic on IP networks, which includes timing information for synchronization during reconstruction as well as feedback on reception quality. The encapsulation mechanism of MPEG-1 and MPEG-2 video streams in IP packets using the RTP protocol has already been specified [15] by IETF. Fig. 1(a) shows the system block diagram for MPEG video transmission, while Fig. 1(b) shows the protocol stack for transporting MPEG video Transport Stream (TS) packets over an IP network. Fig. 2 shows the structure of a typical IP packet carrying an RTP payload encapsulated in UDP.

2.2. Traffic specifications

Setting up a flow over the GS requires the TSpec for the flow to be specified in advance in terms of a number of token bucket traffic descriptors. The traffic descriptors include: bucket size b, the average (or token generation) rate r and the peak bucket rate p. A pure token bucket does not have any peak rate control. In order to ensure that the peak rate does not exceed p, a leaky bucket with bucket rate p and bucket size b is required following the token bucket [21]. Fig. 3 shows the *token bucket with leaky bucket rate control* traffic policing mechanism adopted in the Integrated Services. The network elements providing delay and bandwidth guarantees in the Integrated Services execute some type of fair queuing algorithm like *weighted fair queuing* [22–24], and an upper bound on the delay for a flow can be computed by the network elements if each of the flows arriving at the router is shaped by a token bucket with leaky bucket peak rate control [24].

The TSpec parameters give the end application a way to specify the burstiness present in its generated traffic. In the GS, the end user also specifies a number of reservation specifications (RSpec) including a reservation level R, which is a rate anywhere between the average rate r and the peak rate p. The reservation level gives the end user a way to control the queuing delay suffered by the IP packets while traversing the network. For a given TSpec and RSpec, the network delay and bandwidth are guaranteed after a flow is set up.

2.3. Traffic shaping of MPEG streams

MPEG video sequences are arranged into Group of

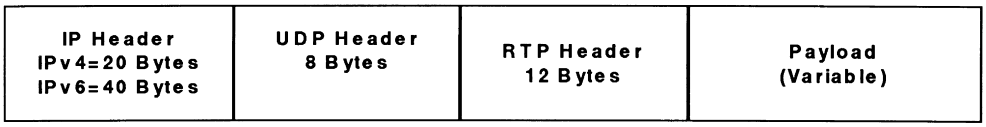

Fig. 2. An IP packet structure for real-time MPEG stream transmission using the RTP protocol.

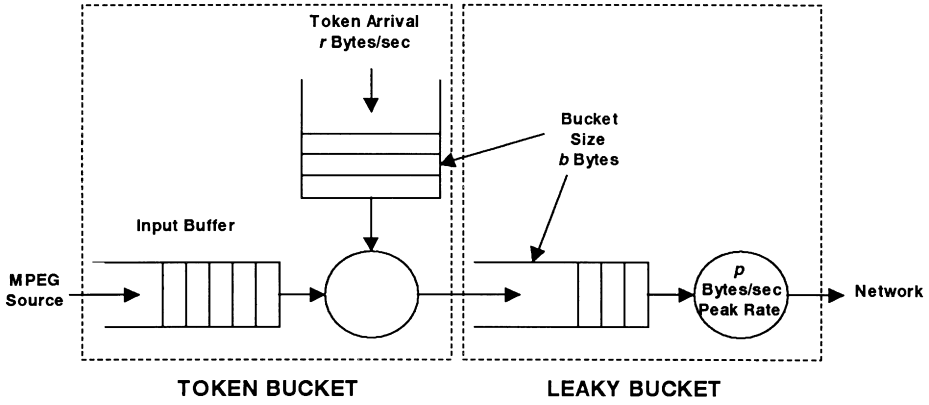

Fig. 3. Schematic diagram of the *token bucket with leaky bucket rate control* traffic policing arrangement and TSpec parameters (*r*, *b*, *p*) specified in the GS.

Pictures (GoP). Each GoP contains three different types of frames: Intra (I), Bi-directional (B) and Predictive (P). At the beginning of a GoP, an I-frame is transmitted. After the I-frame, a number of B-frames are transmitted with P-frames inserted between the B-frames. A typical sequence of frames, for example, is IBB–PBB–PBB–PBB. The GoP structure is usually described as *MmNn* where *n* is the total number of frames in a GoP, and *m* is the I–P or P–P frame interval. For example, IBB–PBB–PBB–PBB is an *M*3*N*12 sequence. During the transmission of the I-frame, a complete image is transmitted which makes the I-frame much larger in data content than other frames. The P-frames usually require fewer bits than the I-frame, and the B-frames usually require the least number of bits. Due to the presence of variable number of bits in different frames, MPEG video streams are highly bursty in nature with a large burst usually being present at the beginning of each GoP when the I-frame is transmitted. Thus, in order to reduce delay and buffer size requirement, it is required that the traffic shaper can transmit a burst of high data rate at the beginning of every GoP period.

Fig. 4 shows a typical plot of data output rate as a function of time. If a token bucket traffic shaper is lossless, then the token bucket is full of tokens at the beginning of each GoP period. Under these conditions, the traffic shaper initially transmits a burst at the peak data rate *p* for a period determined by the bucket size *b*. When the excess tokens are exhausted (due to the transmission of a large I-frame, for example) then a period of data transmission at the average rate *r* continues. Finally, when the input buffer becomes empty, the output of the traffic shaper matches the input and tokens are accumulated in the token bucket until another GoP starts. The period during which data is transmitted at peak rate *p* is the *burst length*, and the area under the burst length (shaded area in Fig. 4) is the *burst capacity* (or *burst volume*) of the token bucket.

During the initial period of higher data rate, the I-frame is transmitted (completely or partially), while the remaining period is utilized for transmitting the remaining portion of the I-frame if necessary, and then other frames. However, IP packet transmission with burst characteristics exactly matching the burst characteristics of the MPEG transport stream packets may require too much network resources (buffer size and bandwidth in each network element) due to the burstiness of the MPEG traffic. An IP stream, which specifies such high burstiness in its TSpec, may be subjected

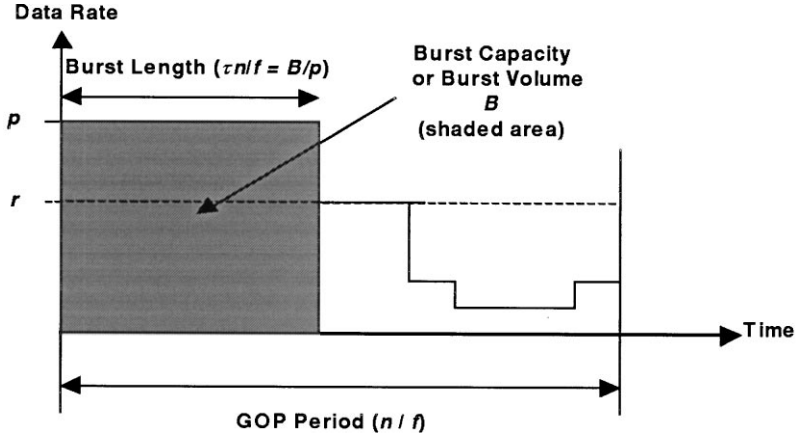

Fig. 4. A typical time versus data-rate plot of the *token bucket with leaky bucket peak rate control* traffic shaper.

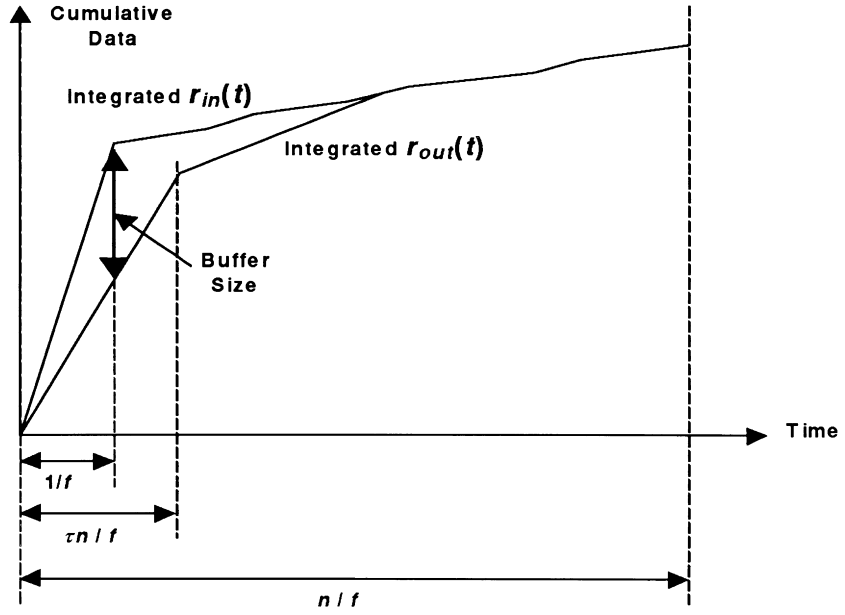

Fig. 5. Input data rate $r_{in}(t)$ and output data rate $r_{out}(t)$ integrated to find the maximum difference between the input cumulative data and output cumulative data for calculating the buffer size requirement of the traffic shaper.

to higher tariff. Thus, the traffic shaper can be employed to reshape the traffic so that the IP packet stream sent out to the network is less bursty than the original MPEG stream, but still conformant to a less costly TSpec. Also, by reshaping the traffic entering the network, it is possible to control the delay experienced by the MPEG stream according to the requirements of the end application, although the buffer size required is dependent on the chosen delay. Thus, a token bucket traffic shaper fully utilizes the traffic burst transmission guarantees provided by the GS on an IP network. As far as the authors are aware, *token-bucket traffic shaping for MPEG video transmission that fully utilizes the burst-handling capability of the GS has not been reported previously*. In a multi-hop transmission of GS stream, traffic shaping has been shown not to introduce any additional end-to-end delay since a properly shaped traffic sent from the transmitting end suffers less queuing delays in intermediate network elements, and the shaping process also has the beneficial effect of reducing jitter [25].

3. Analytical model

To calculate the delay and buffer characteristics of the traffic shaper in our analytical model, we assume that the I-frames require a data rate R_I, the B-frames require a data rate R_B, and the P-frames require a data rate R_P. The average data rate, r, for an *MmNn* GoP sequence can be written as

$$r = \frac{R_I + (n - n/m)R_B + (n/m - 1)R_P}{n} \tag{1}$$

Let us assume that the number of frames per second is f, the peak rate of transmission is p, and the burst volume, which is the area under the peak-rate transmission (at rate p) in the shaper, is B. Also, we define a *normalized burst length* τ as:

$$\tau = \frac{\text{Burst length}}{\text{GOP period}} = \frac{B/p}{n/f} \tag{2}$$

which is the fraction of time the traffic shaper transmits at the peak rate compared to the time taken to transmit one GoP. In addition, we also define a *normalized burst volume* β as:

$$\beta = \frac{\text{Burst volume}}{\text{I-frame data size}} = \frac{B}{R_I/f} \tag{3}$$

which is the ratio of the burst volume to the I-frame data size in the MPEG video sequence.

3.1. Calculation of buffer size

In this section, we calculate the buffer size required at the shaper to achieve zero packet loss. Let us assume that $r_{in}(t)$ is the instantaneous input data rate to the traffic shaper, and $r_{out}(t)$ is the instantaneous output data rate from the traffic shaper at time t where $t = 0$ represents the beginning of a GOP period. We assume that the input buffer is empty at the beginning of each GoP period. The differences between the integrated values of $r_{in}(t)$ and $r_{out}(t)$ represent the instantaneous buffer queue length (see Fig. 5). This integral reaches a maximum value in each GoP period. The buffer requirement S for a single GoP is given by that maximum value of the integral as

$$S = \max\left[\int_0^t [r_{in}(t) - r_{out}(t)]\,dt\right]. \tag{4}$$

As explained in Appendix A, the buffer size S for a sequence of GoPs can be evaluated from Eq. (4) for the case when the burst length is at least equal to the I-frame transmission time, (i.e. $\tau > 1/n$) as

$$S = K\frac{R_I}{f}\left[1 - \frac{\beta}{\tau n}\right] \tag{5}$$

where we take into consideration the variation of the data sizes of I-frames from GoP to GoP by introducing K which represents the ratio of maximum to average buffer size. In order to compare different video sequences, we normalize the buffer size S to the average I-frame data size R_I/f, and define the *normalized buffer size* as $S/(R_I/f)$. This normalized buffer size is used in the discussions in Section 5 where we present results from analysis and simulation.

3.2. Calculation of delay

The delays experienced at the shaper buffer by the frames within the same GoP are different for different frames, but is maximum for the I-frame due to its large size. Assuming that the transmission of the I-frame begins exactly at the same instant when the shaper starts its transmission at the peak rate p, the maximum delay experienced by the frames in a GoP is given by the delay experienced by the I-frame. We define this maximum delay as the *GoP delay*. The GoP delay is calculated by the time difference between the end of the arrival of an I-frame and the end of the departure of an I-frame from the traffic shaper. As explained in Appendix A, this time difference is the maximum delay experienced by all the frames and is defined as the delay through the traffic shaper and is given by

$$t_{\text{delay}} = \begin{cases} \dfrac{1}{f}\left[\tau n - 1 + R_I\dfrac{(1-\beta)}{r}\right] & \beta < 1 \\ \dfrac{1}{f}\left[\dfrac{\tau n}{\beta} - 1\right] & \beta \geq 1 \end{cases}. \tag{6}$$

Eq. (6) represents the delay experienced by a single GoP through the token bucket traffic shaper when R_I is the I-frame data rate. When R_I represents the average I-frame data rate for a video sequence comprising of a number of GoP sequences, t_{delay} represents the *average GoP delay*. In order to compare different video sequences, we normalize the average GoP delay t_{delay} to the GoP period $1/f$, and define the *normalized average GoP delay* as $t_{\text{delay}}/(1/f)$. We use this normalized average GoP delay also in Section 5 where we present results from both analysis and simulation.

3.3. Token bucket burst parameters

The token bucket parameters are the peak data rate p, the average data rate r, and the bucket size b. These parameters (b, r, p) can be related to the burst capacity B and the burst length $(B/p = \tau n/f)$ as follows:

$$\frac{b}{p-r} = \tau n/f = B/p \tag{7}$$

or its similar relationship,

$$\frac{pb}{p-r} = B. \tag{8}$$

4. Simulation

The analytical model presented in Section 3 allows us to determine the buffer requirement and delay at the traffic shaper under study. Since the analytical model takes into account the average bit rates of each of the frame types, we verify the applicability of the analytical model using simulation. MPEG trace data from several video sequences have been used as input to a simulation program to find the statistical properties of delay and buffer size. Real MPEG video sequences contain streams of I-, B-, and P-frames, the bit rates of which vary statistically from frame to frame. Thus the delay and buffer requirements vary for each GoP, and statistical properties of the delay and buffer size need to be studied by simulation in order to design a realistic traffic shaper, and to compare the simulation results with the analytical model. For an MPEG video sequence, all GoPs do not require the same buffer size. We use the maximum buffer requirement of all the GoPs as the required buffer size because a buffer of smaller size will cause data loss at the shaper. For delay, the average delay is a very good indicator of the overall delay performance of the traffic shaper.

A simulation program was written to simulate the behavior of a first-in-first-out (FIFO) queue with the output data transmission rate controlled according to the characteristics of the token bucket traffic shaper with leaky bucket peak rate control. MPEG trace data from several different movie clips [26] were used as input to the simulator. The program takes the burst volume and the burst length of the traffic shaper as its input. The shaper parameters are specified in terms of β (burst-volume to average I-frame-data-size ratio) and τ (burst-length to GoP-period ratio). The simulation program calculates the delay experienced by each frame, and also keeps track of instantaneous buffer occupancy as it runs. Statistics on delay, jitter and buffer queue length are gathered and saved. The simulation is performed for different ranges of values of β and τ.

The parameters for the traffic shaper need to be chosen in a way so that the shaper output does not limit the performance of the MPEG video transmission over an IP network. For example, the choice of the average data rate r should be such that it represents the "worst-case" traffic scenario as specified by IETF [1] and there is no data loss when a GoP of large size is transmitted. Also, the burst volume should be

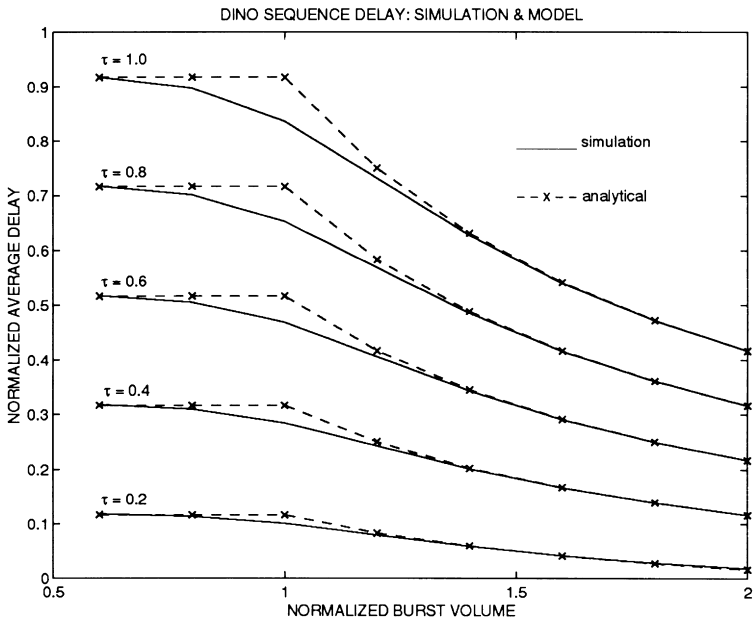

Fig. 6. Normalized average delay for a traffic shaper for the DINO sequence as a function of the normalized burst volume specified in the traffic shaper. The solid lines represent simulation results while the broken lines (with × mark) represent analytical results for comparison. Different curves represent different values of τ. Five different curves are plotted for values of τ in the range $0.2 < \tau < 1.0$.

comparable to the average I-frame data size of the MPEG video sequence for efficient bandwidth utilization. For these reasons, the simulation program first collects some statistics like the average I-frame data size, the maximum GoP data size etc. to suggest suitable average data rate r for the traffic shaper before running the actual simulation.

5. Results

In this section we present results obtained from the analytical model in Section 3 and validate the accuracy of the analytical model by comparing the analytical results with simulation results obtained from MPEG trace data from

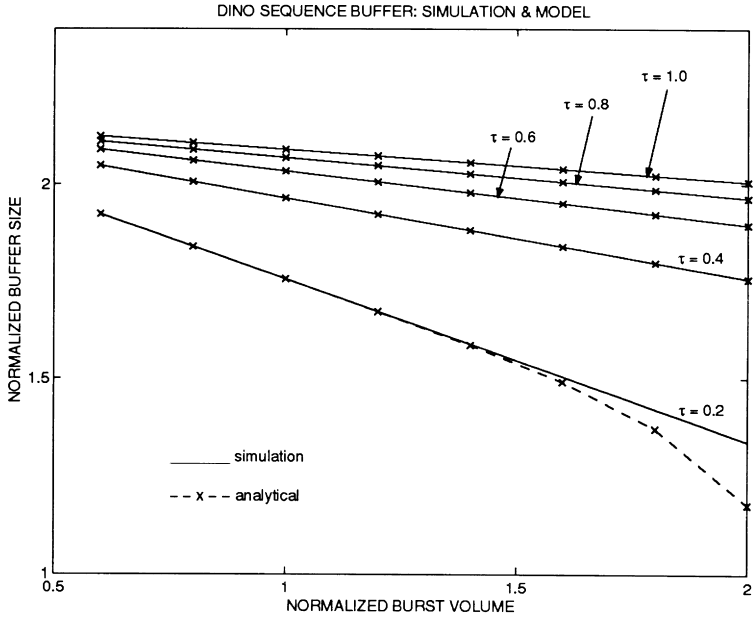

Fig. 7. Normalized buffer size for a traffic shaper for the DINO sequence as a function of the normalized burst volume specified in the traffic shaper. The solid lines represent simulation results while the broken lines (with × mark) represent analytical results for comparison. Different curves represent different values of τ. Five different curves are plotted for values of τ in the range $0.2 < \tau < 1.0$. The solid and broken lines overlap completely for $\tau = 0.4$, 0.6, 0.8 and 1.0.

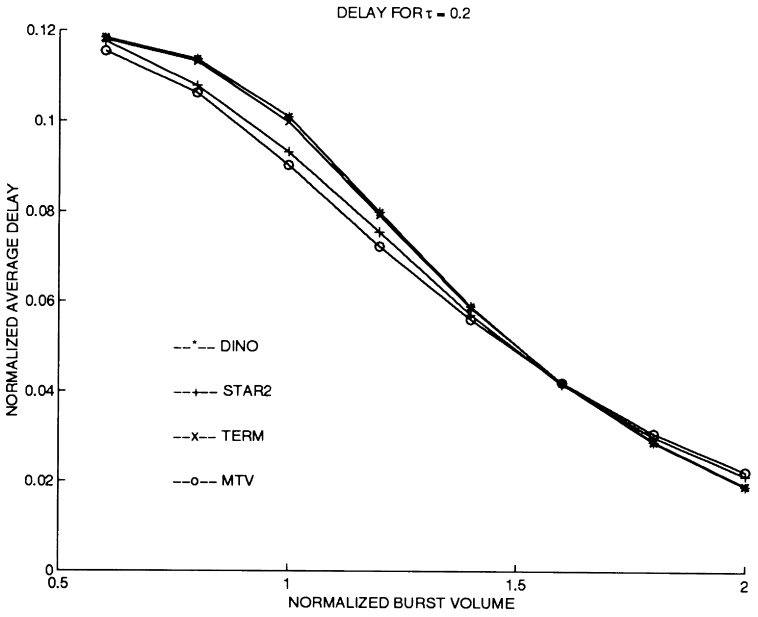

Fig. 8. Normalized average delay of four different video sequences as a function of normalized burst volume for the case $\tau = 0.2$. The sequences are DINO(*), STAR2(+), TERM(×), and MTV(W).

several video sequences. For conciseness, we define the following terms for the discussion to follow:

- *delay* to refer to *normalized average GoP delay* defined in Section 3.2;
- *buffer size* to refer to *normalized buffer size* defined in Section 3.1; and
- burst volume to refer to normalized burst volume β.

Fig. 6 shows the normalized average GoP delay t_{delay} as a function of the normalized burst volume β as given by Eq. (6). The broken lines represent the delay from analytical results for the DINO sequence, a sequence from the movie Jurassic Park. The delay is plotted for τ between 0.2 (short burst-length) and 1.0 (flat, or no burst). For a particular value of τ, the delay is reduced for increased burst volume. For a constant burst volume, increasing τ also increases the delay considerably.

$\tau = 1.0$ represents the case where the traffic shaper transmits at a constant data rate that is equal to the average data rate of the video. The delay for $\tau = 1.0$ is also identical to the delay experienced by a constant-bit-rate transmission as well as a leaky-bucket traffic shaper, since the delay is calculated based on the delay of the I-frame. The traffic shaping delay has been neglected in Ref. [17] where end-to-end delay for a network has been calculated based on a leaky-bucket traffic shaper, although the delay experienced by an MPEG stream through a leaky-bucket traffic shaper is comparable to the end-to-end delay. For example, the curve for $\tau = 1.0$ in Fig. 6 shows significant delay through the traffic shaper which is comparable to end-to-end-delays calculated in Ref. [17]. Fig. 6 also shows that significant reduction in delay can be obtained by choosing a lower value of τ while keeping the burst capacity B constant, i.e. by increasing the peak rate of transmission in the traffic.

Simulation results for the normalized average delay is shown for the DINO sequence in Fig. 6 in solid lines. Comparing the analytical model results with the simulation results, we observe that there is excellent agreement between the model and the simulation results for delay. The analytical model results for delay represents the delay experienced by a GoP of average GoP size. Thus, the statistical variations of the GoP sizes and frame sizes cause the simulation result curves to become smoother than the analytical results. This can explain the deviation of the analytical results from simulation results around $\beta = 1$ in Fig. 6.

Fig. 7 shows the normalized buffer size requirement as a function of the normalized burst volume for various values of τ as given by Eq. (5) for the same DINO sequence as in Fig. 6. The broken lines represent analytical results. In Fig. 7, for a constant burst-length, the buffering requirement is reduced for increased burst volume of the shaper due to the fact that the buffer size requirement is determined by the difference between the data rates at the input and the output of the traffic shaper. For a constant burst volume, decreasing the value of τ also decreases buffer size requirement because of the higher initial data rate with smaller values of τ.

Simulation results for buffer size are shown in Fig. 7 in solid lines for the same DINO sequence. Excellent agreement between analytical results and numerical simulation is observed for the buffer size.

In Fig. 8, we compare the simulation results for delay as a function of the burst volume for four different video sequences: DINO (Jurassic Park), STAR2 (Star Wars), TERM (Terminator 2) and MTV. The burst length for all

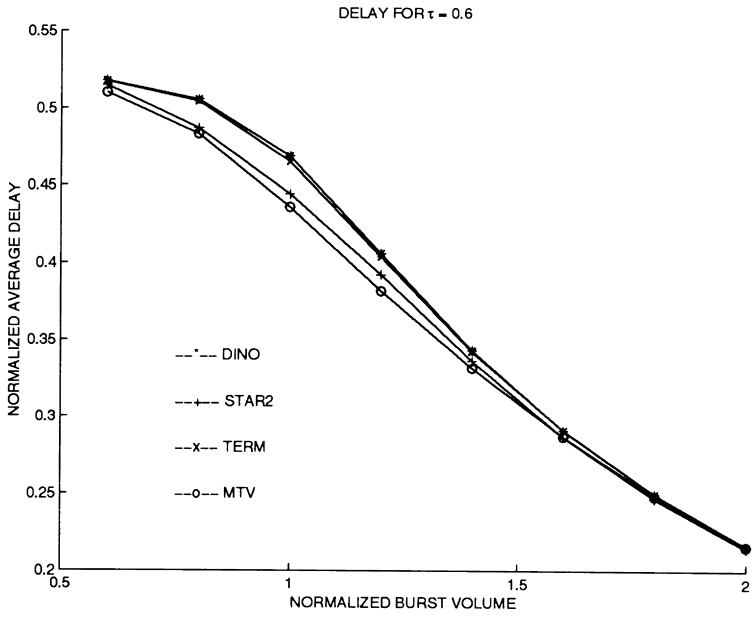

Fig. 9. Normalized average delay of four different video sequences as a function of normalized burst volume for the case $\tau = 0.6$. The sequences are DINO(*), STAR2(+), TERM(×), and MTV(W).

the sequences has been set for $\tau = 0.2$. Average delay in each sequence is normalized to unit GoP period while the burst volume is normalized to the average I-frame data size of each of the respective sequences. The close match between four different types of video sequences in Fig. 8 suggests that normalized delay and normalized burst volume are related by a curve which is almost identical for a wide range of video sequences. This general characteristic of the shaper parameters can be effectively utilized for specifying the shaper parameters for a given delay and vice versa. Fig. 9 compares the delay for $\tau = 0.6$ for the same four video sequences as in Fig. 8. Close match between different video sequences is also observed in Fig. 8. Although both Figs. 8 and 9 show similar characteristics, the delay ranges are different for the two cases. In Fig. 9, delay is longer due to the higher value of τ chosen compared to Fig. 8.

The simulation program also gathers information about delay variation (jitter). Since the standard deviation of delay is a good measure of jitter, we denote the standard deviation

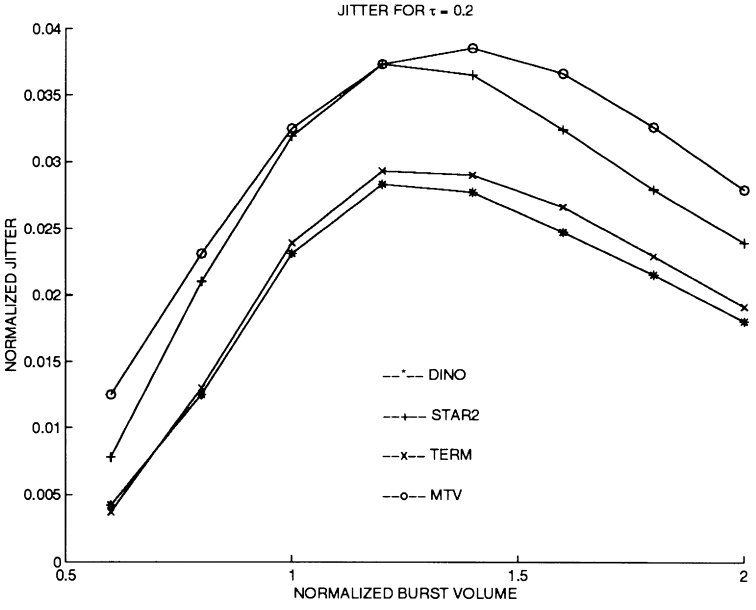

Fig. 10. Normalized jitter of four different video sequences as a function of normalized burst volume for the case $\tau = 0.2$. The sequences are DINO(*), STAR2(+), TERM(×), and MTV(W).

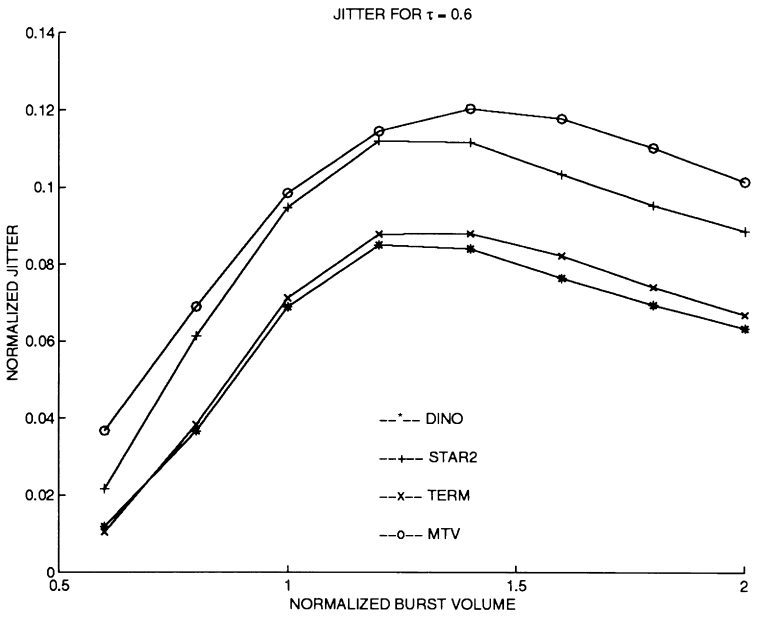

Fig. 11. Normalized jitter of four different video sequences as a function of normalized burst volume for the case $\tau = 0.6$. The sequences are DINO(*), STAR2(+), TERM(×), and MTV(W).

of normalized average GoP delay as *jitter* for brevity. In Fig. 10, jitter is plotted as a function of the burst volume for $\tau = 0.2$ for the same four video sequences as in Figs. 8 and 9. The jitter can be seen to increase, reach a maximum, and then decrease with increasing normalized burst volume. In Fig. 11, jitter is plotted for $\tau = 0.6$ for the same four video sequences as in Fig. 10. Fig. 11 shows the same characteristics as in Fig. 10 although the range of the jitter scale is different in the two figures. In Fig. 12, for $\tau = 0.2$, we compare the simulation results for the normalized buffer requirement for the same four video sequences as a function of the normalized burst volume. Wide variation in the normalized buffer requirement is observed for different types of video sequences. In Fig. 13, the buffer requirements for the four video sequences are compared where τ is chosen as 0.6. Again, a wide variation in the normalized buffer size is observed for different video sequences. These variations in buffer size suggest that careful attention is required for choosing the proper buffer size for a particular video sequence.

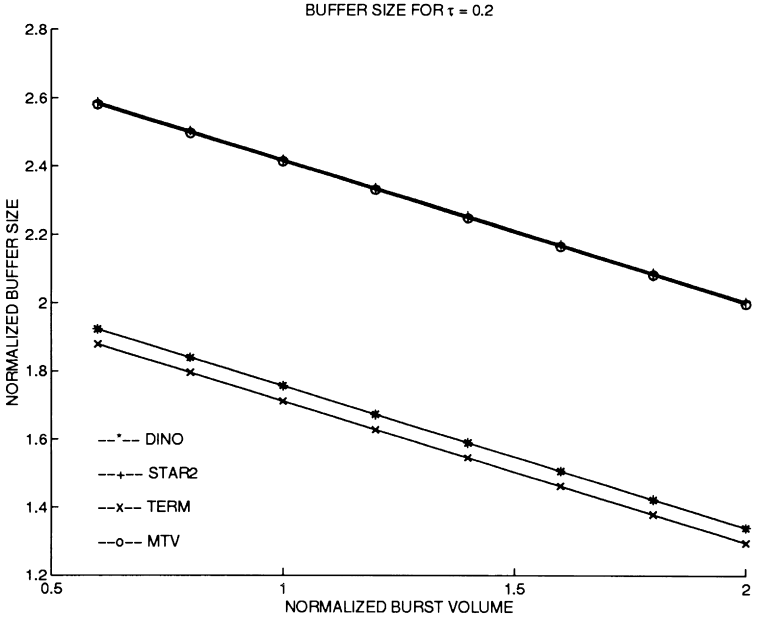

Fig. 12. Normalized buffer size of four different video sequences as a function of normalized burst volume for the case $\tau = 0.2$. The sequences are DINO(*), STAR2(+), TERM(×), and MTV(W).

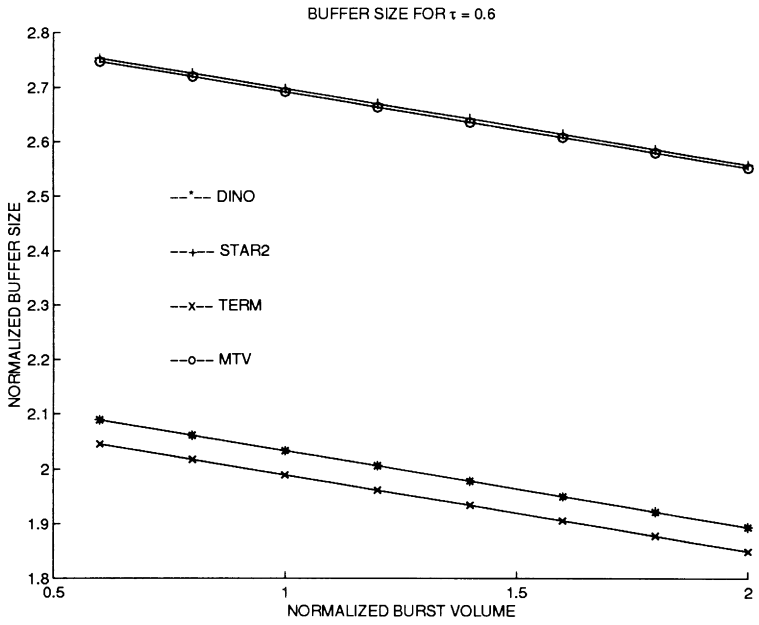

Fig. 13. Normalized buffer size of four different video sequences as a function of normalized burst volume for the case $\tau = 0.6$. The sequences are DINO(*), STAR2(+), TERM(×), and MTV(W).

Real-world applications of the results may include finding the proper token bucket parameters (b, r, p) when a delay is specified. For a given delay, we can choose a value for the burst length τ from a narrow range of values. For example, delay varies between 0.02 and 0.12 for $\tau = 0.2$ in Fig. 8, whereas delay is in the range 0.2 and 0.5 for $\tau = 0.6$ in Fig. 9. After specifying τ, the burst volume B can be calculated from Figs. 12 and 13. Then, the token bucket size b and the peak rate p can be calculated using Eqs. (7) and (8). The token generation rate is calculated from the maximum GoP data size. The maximum GoP data size and the average I-frame data sizes for the four sequences are listed in Table 1.

6. Conclusion

In this paper, we have analyzed the delay, jitter and buffer requirement of a token bucket traffic shaper (with a leaky bucket peak rate control) for efficient MPEG video transmission over the GS, which is a proposed service for the next generation IP networks. The results show that there is excellent agreement between the analytical model and numerical simulation. The significance of this study is that it shows how the traffic shaper can effectively control the delay of the traffic shaping process so that it is conformant to the negotiated traffic specification between the user and the network while fully utilizing the allowed burst-handling capability of the GS. While a leaky bucket type traffic shaper introduces large amount of delay, the token bucket traffic shaper can reduce the shaping delay considerably, thereby reducing the end-to-end delay.

Table 1
Normalization factors for the four different video sequences

Sequence	Largest GoP data size (bits)	Average I-frame data size (bits)
DINO	10 166 368	55 076
TERM	8 068 200	37 388
MTV	18 231 552	69 862
STAR2	7 052 696	44 012

The results also indicate that for larger burst volumes specified during flow setup, the average delay and buffer size are reduced. Reducing the burst length while keeping the burst volume constant (i.e. increasing peak data rate while keeping the burst volume constant by decreasing the burst length) reduces the delay and buffer size requirements at the shaper.

Our study also shows that the delay as a function of the burst volume of the token bucket traffic shaper is nearly identical for different video sequences. This characteristic of the traffic shaper can be utilized for choosing appropriate traffic shaper parameters. Jitter has been found to first increase, reach a maximum and then decrease with increase in burst volume for different types of video sequences. Our study further reveals that the buffer size requirement varies widely for different types of video sequences, and careful attention is needed while choosing the buffer size for a particular video sequence.

The major contribution of this paper is that these results can be utilized for specifying suitable TSpec parameters that are required while setting up a flow for transmission of stored MPEG video sequences over the GS. The TSpec parameters are functions of the statistical properties of the MPEG video sequence as well as the required QoS (delay, jitter etc.) guarantees. Specifying low delay requires

specifying higher burst handling capacity from the network during flow set up, which may be costlier than setting up a flow where the traffic is smoother but delay is longer. Optimum shaper parameters can be chosen to find a trade-off between short delay at high cost, and low cost for long delay.

Appendix A

Eliminating B from Eqs. (2) and (3), we get

$$p = \frac{\beta R_I}{\tau n} \tag{A1}$$

From Fig. 5, the buffer requirement reaches its maximum value at time $t = 1/f$, provided that $R_I > p > R_P > R_B$. Under this assumption, which is usually true for a GoP, the required buffer size can be written as

$$S = (R_I - p)/f = (R_I/f)\left[1 - \frac{\beta}{\tau n}\right] \tag{A2}$$

where Eq. (A1) is used to arrive at the last step. Here, the buffer size represents the average buffer size. However, we need to multiply this buffer size by a constant K to arrive at Eq. (5) such that K represents the ratio of the maximum-to-average buffer size. The value of K depends on statistical properties of the frame data sizes, and it is computed from MPEG trace data.

For $\beta \geq 1$, an I-frame is completely contained inside the burst volume of the shaper. A complete I-frame arrives at the input buffer of the shaper at time $t = 1/f$, and leaves the shaper at time $t = R_I/fP$. Hence, the delay (for the case $\beta \geq 1$), which is the time difference between the arrival and the departure of the end of the I-frame, is given by

$$t_{\text{delay}} = \frac{R_I}{fp} - \frac{1}{f} = \frac{1}{f}\left[\frac{\tau n}{\beta} - 1\right], \qquad \beta \geq 1 \tag{A3}$$

where Eq. (A1) is used for eliminating R_I. When $\beta < 1$, a complete I-frame is not contained inside the burst volume of the shaper. The burst length (B/p) can also be written as ($\tau n/f$) from definition of τ (Eq. (2)). Let us assume that the complete I-frame is transmitted at rate p for the burst length, and also transmission at the average rate r is required for an additional duration of time t'. Thus, the total amount of data transmitted is equal to the data size of the I-frame, and can be written as the sum of the data transmitted at rate p and at rate r as

$$p\frac{\tau n}{f} + t'r = \frac{R_I}{f} \tag{A4}$$

from which we get

$$t' = \frac{R_I}{f}\frac{(1-\beta)}{r} \tag{A5}$$

where b and p are eliminated using Eqs. (A1) and (3). The total delay for the case $\beta < 1$ can thus be written as the burst length plus t' minus the time of arrival of the I-frame as

$$t_{\text{delay}} = \left(\frac{\tau n}{f} + t'\right) - \frac{1}{f}, \qquad \beta < 1 \tag{A6}$$

which can be rewritten using Eq. (A5) as

$$t_{\text{delay}} = \frac{1}{f}\left[\tau n - 1 + R_I\frac{1-\beta}{r}\right], \qquad \beta < 1. \tag{A7}$$

References

[1] S. Shenker, C. Partridge, R. Guerin, Specification of guaranteed quality of service, RFC 2212, Internet Engineering Task Force, September 1997.

[2] J. Wroclawski, Specification of the controlled-load network element service, RFC 2211, Internet Engineering Task Force, September 1997.

[3] The MPEG family of standards includes MPEG-1, MPEG-2 and upcoming MPEG-4, formally known as ISO/IEC-11172, ISO/IEC-13818 and ISO/IEC-14496.

[4] Y.S. Saw, P.M. Grant, J.M. Hannah, B. Mulgrew, Video rate control using a radial basis function estimator for constant bit-rate MPEG coders, Signal Processing: Image Communication 13 (3) (1998) 183–199.

[5] K.T. Choi, S.C. Chan, T.S. Ng, Perceptual based rate control scheme for MPEG-2, 1998 IEEE International Symposium on Circuits and Systems, Monterey, CA,USA, May 31–June 3, 1998, vol. 5, pp. V546–V548.

[6] S. Lee, M.S. Pattichis, A.C. Bovik, Rate control for foveated MPEG/H.263 video, 1998 IEEE International Conference on Image Processing, Chicago, IL, USA, October 4–7, 1998, vol. 2, pp. 365–368.

[7] J.S. Kim, J.K. Kim, Adaptive traffic smoothing for live VBR MPEG video service, Computer Communications 21 (7) (1998) 644–653.

[8] W. Zhu, Y. Wang, Y.Q. Zhang, Jitter smoothing and traffic modeling for MPEG-2 video transport over ATM networks, International Journal of Imaging Systems and Technology 9 (5) (1998) 332–339.

[9] D.H.K. Tsang, B. Bensaou, Sh. T.C. Lam, Fuzzy-based rate control for real-time MPEG video, IEEE Transactions on Fuzzy Systems 6 (4) (1998) 504–516.

[10] J.D. Salehi, Z.L. Zhang, J. Kurose, D. Towsley, Supporting stored video: Reducing rate variability and end-to-end resource requirements through optimal smoothing, IEEE/ACM Transactions on Networking 6 (4) (1998) 397–410.

[11] X. Wang, S. Jung, J.S. Meditch, Dynamic bandwidth allocation for VBR video traffic using adaptive wavelet prediction, IEEE International Conference on Communication, Atlanta, GA, USA, June 7–11, 1998, vol. 1, pp. 549–553.

[12] B.N. Bashforth, C.L. Williamson, Statistical multiplexing of self-similar video streams: simulation study and performance results, 1998 Sixth International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunications Systems, pp. 119–126, Montreal, Canada, July 19–24, 1998.

[13] P. Cuenca, B. Caminero, A. Garrido, F. Quiles, L. Orozco-Barbosa, QoS and statistical multiplexing performance of VBR MPEG-2 video source over ATM networks, 1998 11th Canadian Conference on Electrical and Computer Engineering, Toronto, Canada, May 24–28, 1998, vol. 1, pp. 33–36.

[14] S. Jacobs, A. Eleftheriadis, Streaming video using dynamic rate shaping and TCP congestion control, Journal of Visual Communication and Image Representation 9 (3) (1998) 211–222.

[15] D. Hoffman, G. Fernando, V. Goyal, M. Civanlar, RTP payload format for MPEG1/MPEG2 video, RFC 2250, Internet Engineering Task Force, January 1998.

[16] K. Van der Wal, M. Mandjes, H. Bastiaansen, Delay performance

analysis of the new internet services with guaranteed QoS, Proceedings of IEEE 85 (12) (1997) 1947–1957.

[17] H. Naser, A. Leon-Gracia, Performance evaluation of MPEG2 video using guaranteed service over IP-ATM networks, Multimedia Computing and Systems Conference, Austin, TX, USA, June 28–July 1, 1998.

[18] M.F. Alam, M. Atiquzzaman, M.A. Karim, Effects of source traffic shaping on MPEG video transmission over next generation IP networks, 1999 International Conference on Computer Communications and Networks, Boston, MA, USA, October 11–13, 1999, pp. 514–519.

[19] M.F. Alam, M. Atiquzzaman, M.A. Karim, Traffic shaping for MPEG video transmission over Guaranteed Service on the Internet, 1999 IEEE Global Communications Conference, Rio de Janeiro, Brazil, December 5–9, 1999, pp. 364–368.

[20] R. Braden, L. Zhang, S. Berson, S. Herzog, S. Jamin, Resource ReSerVation Protocol (RSVP)—Version 1 functional specification, RFC 2205, Internet Engineering Task Force, September 1997.

[21] C. Partridge, Token bucket with leaky bucket rate control, Gigabit Networking, Addison-Wesley, Reading, MA, 1994 (pp. 262–263).

[22] A. Demers, S. Keshav, S. Shenker, Analysis and simulation of a fair queuing algorithm, Internetworking: Research and Experience 1 (1) (1990) 3–26.

[23] A.K. Parekh, R.G. Gallager, A generalized processor sharing approach to flow control in integrated services networks: the single-node case, IEEE/ACM Transactions on Networking 1 (3) (1993) 344–357.

[24] A.K. Parekh, R.G. Gallager, A generalized processor sharing approach to flow control in integrated services networks: the multiple node case, IEEE/ACM Transactions on Networking 2 (2) (1994) 137–150.

[25] L. Georgiadis, R. Guerin, V. Peris, R. Rajan, Efficient support of delay and rate guarantees in an internet, Computer Communications Review 26 (4) (1996) 106–116.

[26] O. Rose, Statistical properties of MPEG video traffic and their impact on traffic modeling in ATM systems, Report No. 101, Institute of Computer Science, University of Würzberg, February 1995.