Credit-Based Flow Control with Explicit Rate Feedback for Adaptive Video Multicast [†]

Célio Albuquerque, Brett J. Vickers and Tatsuya Suda
{celio,bvickers,suda}@ics.uci.edu
Dept. of Information and Computer Science
University of California, Irvine
Irvine, CA 92697-3425
714-824-4105 (phone) 714-824-2886 (facsimile)

Abstract

In an era of proliferating multimedia applications, support for video transmission is rapidly becoming a basic requirement of network architectures. Furthermore, since most video applications (e.g., teleconferencing, television broadcast, video surveillance, interactive video games) are inherently multicast in nature, network architectures that can efficiently transport high quality, multicast video are essential. The main problem complicating multicast video transport is variation in network bandwidth constraints. An attractive solution to this problem is to use an adaptive, multi-layered video encoding mechanism. In this paper, we consider a credit-based mechanism for the support of video multicast that relies on explicit rate congestion feedback from multicast destinations as well as hop-by-hop flow control. The responsiveness, bandwidth utilization, video quality and fairness of the mechanism are evaluated through simulations. Results suggest that the proposed mechanism is capable of providing a high quality video service in the presence of varying bandwidth constraints.

Sumário

Com a proliferação de aplicações multimídia, suporte para a transmissão de vídeo está se tornando um requisito básico de arquiteturas de redes de computadores. Além disso, como a maioria das aplicações de vídeo (teleconferência, distribuição de programas de televisão, vídeo vigilância, jogos de vídeo interativos) são multi-ponto por natureza, arquiteturas que transportem eficientemente vídeo multi-ponto de alta qualidade são essenciais. O principal problema para o transporte de vídeo multi-ponto é a variação da banda passante da árvore multi-ponto. Uma solução é o uso de um mecanismo de codificação de vídeo adaptativo e em multi-camadas. Este artigo apresenta um mecanismo baseado em créditos, para o suporte de vídeo multi-ponto. Este mecanismo se baseia em um controle de fluxo nó-a-nó e em informações de vazão retornadas pelos destinatários. O tempo de reação, a utilização da banda passante, a qualidade do vídeo e a imparcialidade do mecanismo são avaliadas através de simulações. Os resultados sugerem que o mecanismo proposto é capaz de prover um serviço de alta qualidade, mesmo na presença de variações na banda passante da árvore multi-ponto.

1 Introduction

Network architectures that can efficiently transport high quality, multicast video are rapidly becoming basic requirement of emerging multimedia applications. It has long been recognized that high speed networking technologies like ATM are capable of supporting the strict quality of service guarantees required by real-time traffic like video. Yet even in networks that have traditionally offered minimal or no quality of service guarantees, efforts

[†] This research is supported by the National Science Foundation through grant NCR-9628109. It has also been supported by grants from the University of California MICRO program, Hitachi America, Standard Microsystem Corp., Tokyo Electric Power Co., Nippon Telegraph and Telephone Corp. (NTT), Nippon Steel Information and Communication Systems Inc. (ENICOM), Matsushita Electric Industrial Co., and Fundação CAPES/Brazil.

are now underway to support real-time video applications. Quality of service support in the Internet, for instance, is the subject of a great deal of recent research attention [1].

Furthermore, since most video applications (e.g., teleconferencing, television broadcast, video surveillance, interactive video games) are inherently multicast in nature, support for point-to-point video communication is not sufficient. Unfortunately, multicast video transport is severely complicated by variation in the amount of bandwidth available throughout the network. See the example shown in Figure 1. The video source V attempts to transmit video to two destinations, D_1 and D_2, at a peak rate of 20 Mbps, but due to competing network traffic and varying link capacities, the path between V and D_1 can support 10 Mbps of video, while the path between V and D_2 can support only 4 Mbps. One potential solution to this problem of varying bandwidth constraints is to force the source to apply an adaptive video encoding technique and reduce its transmission rate to 4 Mbps, which is the highest rate that both paths can support. However, in a multicast connection with hundreds or even thousands of destinations, there is likely to be at least one very congested path. Limiting the video rate according to the most congested path penalizes the quality of video offered across all the other paths, regardless of how much bandwidth is available on them.

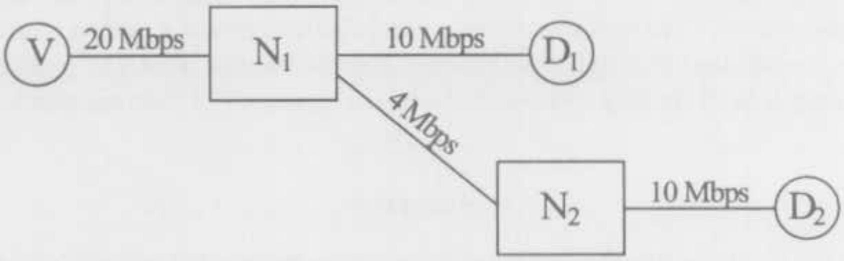
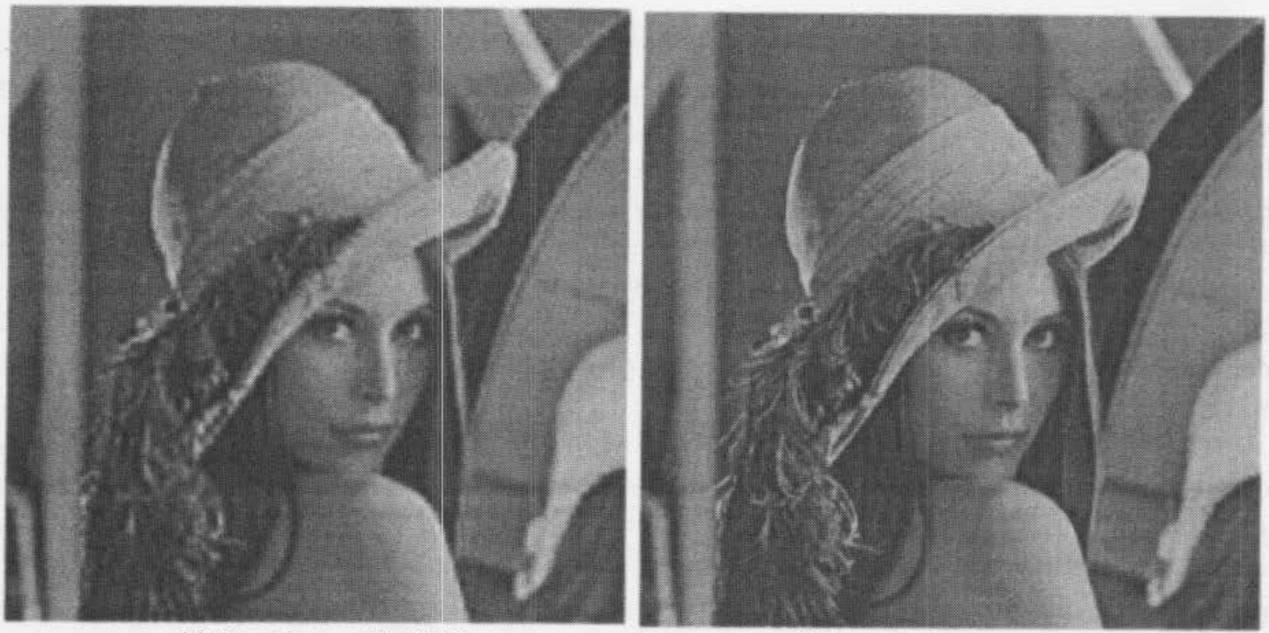


Figure 1: Example of multicast video session.

A more scalable solution to the problem of available bandwidth variation is to use multi-layered video. A multi-layered video encoder encodes raw video data into one or more streams, or layers, of differing priority. The layer with the highest priority, called the *base layer*, contains the most important portions of the video stream. One or more *enhancement layers* with progressively lower priorities may then be encoded to further refine the quality of the base layer stream. For instance, in the example of Figure 1, the ideal deployment of multi-layered video results in a base layer stream transmitted at 4 Mbps and a single enhancement layer stream transmitted at 6 Mbps. Figure 2 provides a visual example of multi-layered encoding using wavelet subband coding [2]. Figure 2(a) is the base layer image and is composed of the first 10 subbands of the image. The remaining 9 subbands constitute the enhancement layer, and Figure 2(b) is the result of combining the base and enhancement layers.

There are two primary advantages to using multi-layered video encoding in multicast-capable networks. First is the ability to perform graceful degradation of video quality when loss occurs. Because each video layer is prioritized, a network experiencing congestion may discard packets from low priority layers, thereby protecting the important base layer and higher priority enhancement layers from corruption. The second advantage, which is related to the first, is the ability to support multiple destinations with different bandwidth constraints or end-system capabilities. For each source-to-destination path with a unique bandwidth constraint, an enhancement layer of video may be generated.

Multi-layered video is not by itself sufficient to provide ideal network bandwidth utilization or video quality, however. To improve the bandwidth utilization of the network and



(a) Base Layer, 28.52 dB (b) Base+Enh.Layers, 53.35 dB

Figure 2: Example of multi-layered picture encoded using wavelet subband coding.

optimize the quality of video received by each of the destinations, the source must respond to constantly changing network conditions by dynamically adjusting the number of video layers it generates as well as the rate at which each layer is transmitted. For the source to do this, it must have congestion feedback from the destinations and the network.

In this paper, we study a novel and promising feedback mechanism, which relies on adaptive, multi-layered video encoding. The proposed feedback mechanism is a *credit-based* mechanism that uses hop-by-hop flow control to reduce loss and optimize utilization. Intermediate nodes exchange feedback packets containing "credits," which reflect the amount of buffer space available at the next downstream node. Feedback packets also contain the rate at which each destination is receiving video from the destinations to the source. When the source receives a returning feedback packet, it adjusts its encoding behavior to generate the specified number of layers at the specified rates.

The remainder of this paper is organized as follows. Related research on the transport of video traffic in high speed networks is reviewed in section 2. The multicast, multi-layered feedback mechanism introduced by this paper is detailed in Section 3. The video quality, responsiveness, utilization, and fairness of the mechanism are evaluated through simulations in section 4, and concluding remarks are provided in section 5.

2 Related Work

A number of researchers have examined the use of congestion feedback for the adaptive control of the video encoding process [3, 4, 5, 6]. In [3], [4] and [5], information regarding the occupancies of internal network buffers is passed via network feedback packets to the video source. The encoding of the video sequence is then rate-controlled to avoid buffer overflow within the network. In [6], network switches implement an explicit rate control policy and inform the video source of the exact rate at which to encode video, thereby rapidly adjusting to changes in the network's available bandwidth due to transient congestion effects. However, in none of these works is the specific problem of transmitting multicast video across paths with varying bandwidth constraints taken into account.

In another work [7], a scenario in which a single end system transmits a single layer of video to several IP destinations is considered, and congestion feedback from the destinations is used to control the rate of the video stream. A form of probabilistic feedback used to prevent feedback implosion. Based on feedback responses from the destinations, the source adaptively modifies the video encoding rate to reduce network congestion when necessary and increase video quality where possible. While this scheme takes multicast connections into account, it uses only a single layer of video, and thus a few severely bandwidth-constrained paths can negatively impact the rate of video transmitted across paths that have more plentiful bandwidth.

The destination set grouping approach [8] attempts to satisfy the bandwidth constraints of multiple source-to-destination paths in the distribution of multicast video. The source maintains a small number of independent video streams, each encoded from the same raw video material but at different rates. The video streams are then targeted to destination groups with different bandwidth constraints. Feedback from the destinations is used to control the encoding rates of each offered video stream, and destinations are allowed to choose which stream to receive based on their current bandwidth constraints. Although this multicast approach is adaptive, transmitting several independently encoded video streams may result in an inefficient use of network bandwidth.

Another potential solution to the multicast of video to destinations with varying bandwidth constraints is transcoding [9]. In this approach, a single layer of video is encoded at a high rate by the source, and intermediate network nodes transcode (i.e., decode and re-encode) the video down to a lower rate whenever they become bottlenecked. While this approach solves the available bandwidth variation problem, it requires complex and computationally expensive video transcoders to be present throughout the network.

In the receiver-driven layered multicast (RLM) approach for IP networks [10], the source generates a fixed number of layers, each at a fixed rate, and the destinations "subscribe" to as many layers as they have the bandwidth to receive. This approach, while it improves the efficiency of video transport through multi-layered encoding, is not adaptive; it limits the destinations to choosing among the layers the source is willing to provide. Unfortunately, in some cases the provided selection may not be adequate enough to optimize network utilization and video quality.

The authors' previous research on adaptive multi-layered multicast [11] has investigated two congestion control mechanisms: an end-to-end, rate-based mechanism that relies on explicit rate congestion feedback; and a credit-based mechanism that uses hop-by-hop feedback. Simulation results suggested that both mechanisms are capable of providing a high quality video service in the presence of varying bandwidth constraints. However, the two mechanisms exhibit performance trade-offs, namely, the credit-based mechanism provided better network utilization and slightly better fairness, while the rate-based mechanism provided better responsiveness and slightly better goodput.

The adaptive approach described in this paper uses feedback from the network to optimize both the network utilization, responsiveness, fairness and the quality of video received by the destinations. This work is a significant extension of the authors' prior work [11]. This paper presents a novel credit-based feedback approach, which allows for an arbitrary number of encoded video layers.

3 Proposed Mechanism

To satisfy a large number of video multicast destinations with varying bandwidth constraints, a credit-based congestion control algorithm is introduced.

Credit-based mechanisms have been widely studied, especially in regard to the flow and congestion control of data traffic [12, 13, 14, 15]. The credit-based scheme proposed in this paper for multi-layered video is influenced largely by the *Quantum Flow Control* (QFC) mechanism [12] used for ABR data traffic in ATM networks [16, 17]. The primary advantage of QFC is its ability to achieve 100% network utilization while ensuring zero packet loss, regardless of the amount of network congestion.

The QFC mechanism maintains a separate control loop for each link of a connection by using *credits*. Credits reflect the amount of buffer space available at the next downstream node and give a node permission to transmit packets. Buffers are allocated on a per-connection basis and each time a node transmits a packet, it consumes one credit. If a node has no credits available, then it must wait for credits to arrive before transmitting a packet. To prevent the inefficient use of bandwidth by credit packets, several credits are collected by each node before being transmitted together to an upstream node. Packets are transmitted to downstream nodes only if there are credits available and there is no interfering traffic packets queued to be transmitted. Higher scheduling priority is given to interfering traffic, since the QFC mechanism is designed to exploit only the available, unutilized bandwidth. In this proposed mechanism, the credit-based flow control is designed to serve video packets at a guaranteed, minimum video rate (MVR) in addition to exploiting the available bandwidth on the network. Therefore, higher scheduling priority can occasionally be given to video traffic in order to guarantee a minimum rate for the video service.

For multicast connections, the original QFC algorithm is designed to reduce the source's transmission rate in response to the connection's most congested branch. For multi-layered video, this type of behavior is undesirable since full utilization of network bandwidth is one of the primary goals and losses to low priority video layers are tolerable. This paper introduces a modified credit-based mechanism that extends QFC to potentially achieve full utilization on all branches of a multicast connection. In the modified credit-based mechanism, losses are allowed to occur, but when buffers overflow, only the packets from the lowest priority layers are discarded. Destinations also supply feedback in order to inform the source the rate at which each destination is receiving video, and thereby adjust the number of layers as well as the rate of each layer. A detailed description of the credit-based mechanism for multicast video follows.

An intermediate node returns a feedback packet to its upstream neighbor whenever one of the following two conditions is satisfied:

1. Each of the multicast connection's output ports has transmitted at least N_t packets, or

2. At least one output port of a multicast connection has transmitted N_t packets, and the difference between the occupancies of any two video output queues in the same multicast connection is at least D_t packets.

The first condition guarantees that credits are periodically returned to an upstream node whenever each of the connection's adjacent downstream nodes is continually draining packets. The second condition allows credits to be returned to an upstream node even if one or more adjacent downstream nodes fails to drain packets rapidly enough. This condition is only verified if one of the video output queues has a possibility of imminent underflow, i.e., if one of the video output queue occupancies is less than 33%. This second condition prevents a node from becoming a bottleneck as long as at least one downstream path continues to accept packets. While this condition may result in packet losses on some links, the losses are isolated to low priority packets through a priority discard mechanism. In both conditions, feedback packets carry N_t credits to the node's upstream neighbor, which increments its credit counter by N_t.

Table 1 lists the information carried by the proposed credit-based mechanism's feedback packet. C is equal to the total number of credits that the downstream node has sent to the upstream node since call establishment, and L is the maximum number of video layers that can be generated by the source and transported by the network. The feedback packet also contains an explicit rate array field (r_i) and a counter array field (c_i) initially set by the multicast destinations.

Multicast destinations monitor the incoming video traffic over a *destination monitoring interval*, through the use of a moving window. Every time a destination receives N_t video packets, it generates a feedback packet containing N_t credits and sets the fields r_i and c_i. The field r_i is set with the destination's desired video rate. When it is time to send a feedback packet upstream, the destination indicates the desired video rate, by filling the first slot of the feedback packet's rate array (r_1) with the average video rate received over the past destination monitoring interval. It also sets the corresponding slot's counter (c_1) to one in order to indicate that one destination has requested rate r_1 so far.

Field	Description
L	Maximum number of layers allowed
C	Credit counter, indicating the total number of credits sent so far to the upstream node
r_i	An array (i =1,...,L) listing the cumulative rates of each video layer
c_i	An array (i =1,...,L) listing the number of destinations requesting each layer in the array r_i

Table 1: Contents of feedback packets used by the credit-based mechanism.

Feedback congestion information (r_i, c_i) is stored on the outputs of each intermediate node on a per-connection basis. When it is time to send a feedback packet upstream, the intermediate node collects the rate (r_i) and counter (c_i) entries from each output port of the multicast connection and stores them into a temporary local array, sorted by rate. Each rate entry corresponds to a video rate requested by one or more downstream destinations, while the counter values indicate how many downstream destinations have requested each rate. Ultimately, the rate values will be used by the source to determine the rates to transmit each video layer. If two or more packets contain identical rate values (or nearly identical values[1]), then their corresponding counter values are summed together and stored with the rate as a single local array entry.

After filling the local rate array, the number of entries in the array is compared to the maximum number of layers allowed for the connection (L). If the number of entries in the local rate array is less than or equal to the maximum number of layers allowed, then a new feedback packet is immediately generated, filled with the contents of the local rate array,

[1] Two rate values that are separated by less than 100 kbps are considered the same rate, and the lesser of the two rates is stored in the local rate array.

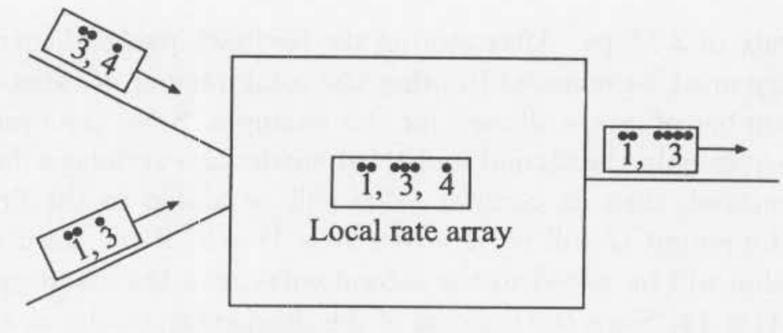


Figure 3: Example of feedback packet merging (L=2) at intermediate nodes.

and sent to the next hop. Otherwise, one (or more) of the entries must be discarded and its counter values added to the next lower entry. To determine which entry (or entries) to discard, the intermediate node attempts to estimate the impact of dropping each listed rate on the overall video quality. This is done through the use of a simple estimated video quality metric.

The estimated video quality metric attempts to measure the combined "goodput" of video traffic that will be received by all downstream destinations. The goodput for a single destination is defined as the total throughput of all video layers received by the destination *without loss*. For instance, suppose a source transmits three layers of video at 1 Mbps each. If a destination entirely receives the most important first two layers but only receives half of the third layer due to congestion, then its total received throughput is 2.5 Mbps, but its *goodput* is equal to the combined rate of the first two layers, namely 2 Mbps. The goodput is a relatively useful estimate of video quality because it measures the total combined rate of uncorrupted video traffic arriving at an end system.

As intermediate nodes merge feedback packets, they attempt to estimate the goodput that downstream destinations will receive. The combined goodput G is estimated from the values listed in a rate array calculated as follows:

$$G = \sum_{i=1}^{N} r_i \times c_i,$$

where N is the number of entries in the local rate array, and r_i and c_i are the rate and counter values for each entry. To determine which entry to remove from the local rate array, it is necessary to calculate the combined goodput that will result from each potential entry removal. The entry removal that results in the highest combined goodput is then removed from the rate array. This process is repeated until the number of entries in the local rate array is equal to the maximum number of layers allowed. The number of entries in the rate array is set to L, and a merged feedback packet is transmitted to the next hop.

(There is one important caveat when removing an entry from the local rate array: the first entry can never be removed. Even minor losses in the base layer can cause precipitous drops in video quality, so the base layer should ultimately reflect the amount of bandwidth available on the most congested path. Hence, the array entry with the lowest rate can never be removed, because it may ultimately determine the rate of the base layer.)

For an example of the feedback merging process, consider Figure 3. Two feedback packets are shown arriving at an intermediate node, both with two rate entries (r_1 and r_2) in units of Mbps stored in their rate arrays. The counter values (c_i) are indicated by the number of dots over each listed rate. Since both packets contain a rate entry of 3 Mbps, these entries are merged into a single entry in the local array, and their counter values of 1 and 2 are added together, as shown, in order to indicate that three downstream destinations have requested a rate of 3 Mbps. After storing the feedback packets' entries into the local rate array, one entry must be removed to bring the total number of rates down to 2, which is the maximum number of layers allowed for this example. Since the first entry can never be removed, this leaves only the second and third entries as candidates for removal. If the second entry is removed, then its counter value will be added to the first entry and the resulting combined goodput G will be $(1 \times 5) + (4 \times 1) = 9$. If the third entry is removed, then its counter value will be added to the second entry, and the resulting goodput will be $G = (1 \times 2) + (3 \times 4) = 14$. Since the removal of the third entry results in a higher combined goodput than the removal of the second entry, the third entry is removed. The resulting feedback packet contains two rate entries and is forwarded to the next hop.

By the time a feedback packet arrives at the source, it contains the number of video layers to encode and a list of cumulative rates at which to encode each layer. The base layer is always transmitted at the minimum video rate guaranteed by the network. Intermediate layers are transmitted at 90% of the rate reported by feedback packets. The reason for this is to allow 10% of the available bandwidth to be filled by lower priority packets, and in case of fluctuations in the available bandwidth, low priority packets would be dropped first, allowing some time for the video sources to adjust their transmission rates before higher priority layers are corrupted. The overall cumulative rate can be less than the available bandwidth on the path to the least congested destination, though. In order to fully utilize this bandwidth, the source monitors its buffer occupancy and increments the rate of the lowest priority video layer if the source buffer occupancy falls below a threshold of 33%.

The effect of this credit-based mechanism with explicit rate feedback is to dynamically establish the number of video layers to encode nearly optimal rates for each of the layers. The rates are optimal in the sense that they are selected by the network in a manner that optimizes the combined goodput. Under this mechanism, bandwidth in the network is almost fully utilized, and the quality of video received by most of the destinations is determined not solely by the source, but also by the current state of congestion in the network.

4 Performance

This section presents the results of several simulations designed to evaluate the performance of the proposed multicast, multi-layered feedback mechanism. Various network topologies are used to evaluate several performance metrics including the responsiveness, utilization, fairness and video quality. All simulations assume the use of ATM cell-sized packets. Unless otherwise specified, all link capacities are equal to 100 Mbps, propagation delays between end systems and intermediate nodes are 5 μs (1 km), and propagation delays between intermediate nodes are 100 μs (20 km). Feedback packets are generated once for every 16 packets transmitted or when the difference between the occupancies of any two video buffers for the same multicast connection is 16 ($N_t = D_t = 16$). A minimum video rate (MVR) of 1 Mbps is reserved throughout the multicast tree and a destination monitoring interval of 20 ms is used.

4.1 Video Quality

Providing better video quality is the ultimate reason for exploiting the unused, available bandwidth on the network. This experiment illustrates how the proposed mechanism enhances the video quality delivered to destinations with varying available bandwidths on the path from the video source. In this experiment, a network model based on a tree topology is used. As shown in Figure 4, it consists of one video source V, two destinations D_1 and D_2, and three intermediate nodes $\{N_1,\ldots,N_3\}$. Persistent interfering traffic is applied with a constant rate of 96 Mbps on link L_1 and with a constant rate of 98 Mbps on link L_2.

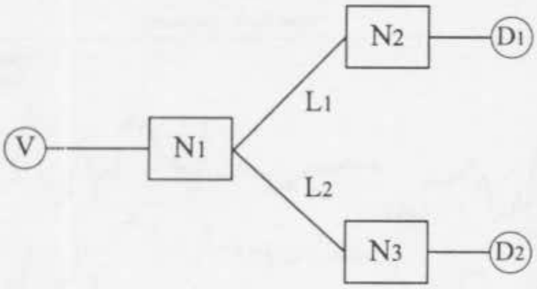
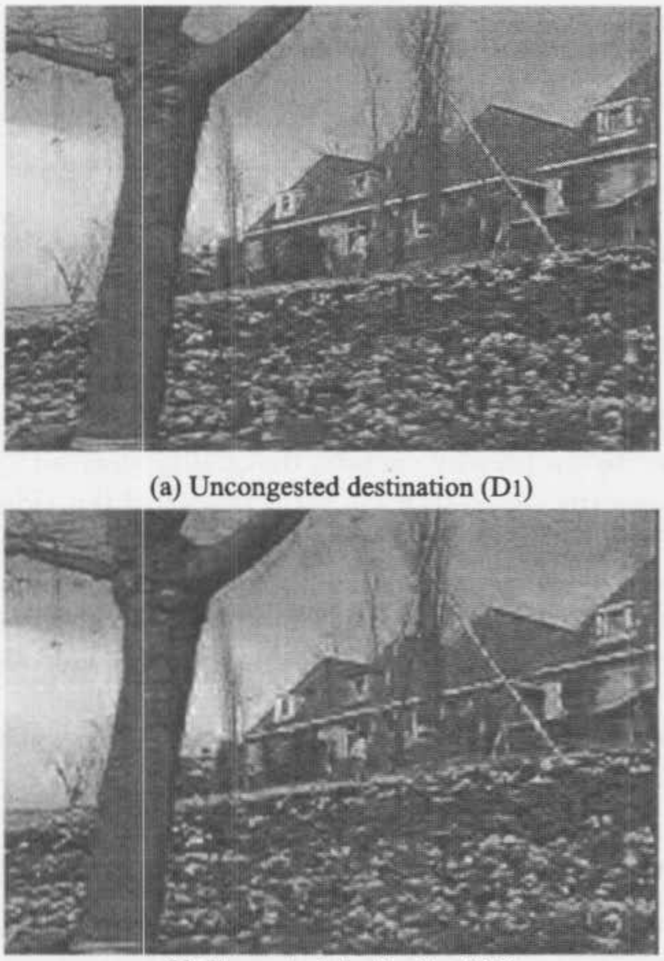


Figure 4: Simulation model for evaluating video quality enhancement.



(a) Uncongested destination (D1)

(b) Congested destination (D2)

Figure 5: Sample frame of video received by the destinations.

In this experiment, the source encodes raw video sequences from the movie *Star Wars: Return of the Jedi* and from the sequence *Flower Garden*. The video encoder performs a block-based multi-layered wavelet subband coding and adaptively adjusts the rates and the number of video layers sent to the simulator. The simulator receives each encoded video block, segments it into packets and sends the packets to the video source output queue, to be transmitted to the network. Packets may be dropped due to congestion in the network. Destinations receive packets, and reconstruct each video block. In this process, if a packet is missing, the whole subband (or subbands) associated with the lost packet is discarded.

Since each video block contains 13 subbands, if one or a few subbands are dropped, the block can still be decoded. Also, since losses occur preferentially at low priority subbands, in case of congestion, a graceful degradation of the video quality is observed.

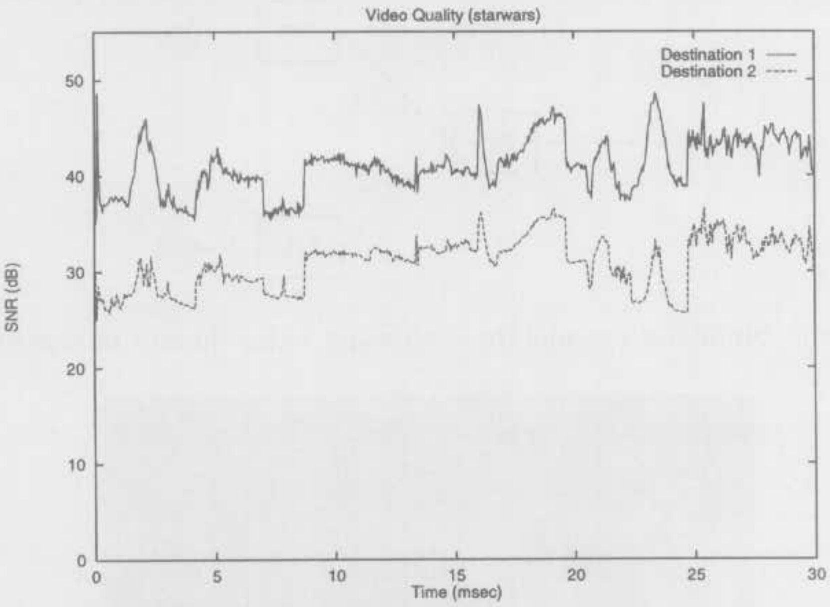


Figure 6: SNR of video received at destinations.

Figure 5 shows a sample video frame of the sequence *Flower Garden* received by both destinations. The frame received at destination D_1 suffers no losses. Almost half of the packets transmitted by the source are dropped before reaching destination D_2. However, since losses are restricted to low priority packets, the quality received by D_2 is only gracefully degraded. Figure 6 shows the signal-to-noise ratio (SNR) of the video sequence *Starwars: Return of the Jedi* received by the destinations versus time. Again, it is important to notice that although destination D_2 loses half of the packets, no large drops in the video quality is observed throughout the whole sequence. It's interesting to observe how the SNR curves follow each other, clearly displaying that the base quality of the video is preserved througout the sequence.

A set of video sequences resulted from various simulations is available for demonstration of the performance of the mechanism.

4.2 Responsiveness

In order to be effective, feedback-based traffic control mechanisms must react in a timely fashion to changes in the network's congestion status. The proposed mechanism attempts to react rapidly to changes in the network's available bandwidth by adjusting the number of video layers the source generates as well as the rate of each layer.

A tree topology network model is used to evaluate responsiveness. As shown in Figure 7, it consists of eight video sources $\{V_1,\ldots,V_8\}$, two destinations D_1 and D_2, and three intermediate nodes $\{N_1,\ldots,N_3\}$. Interfering traffic is applied on the links connecting intermediate nodes, and three responsiveness experiments are conducted. The first experiment is organized so that the sources are required to create and delete video layers in response to changes in the available bandwidth in the network. The second experiment is designed to require the sources to adjust the rate of one of its video layers in response to changes in network congestion. And the third experiment evaluates how the responsiveness is affected

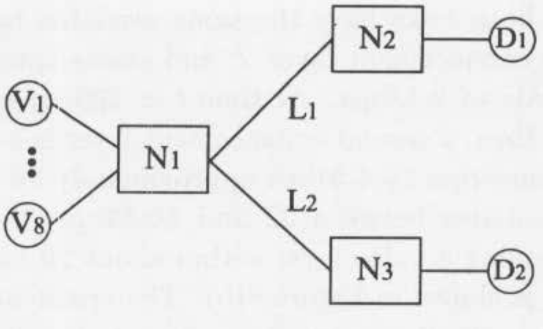
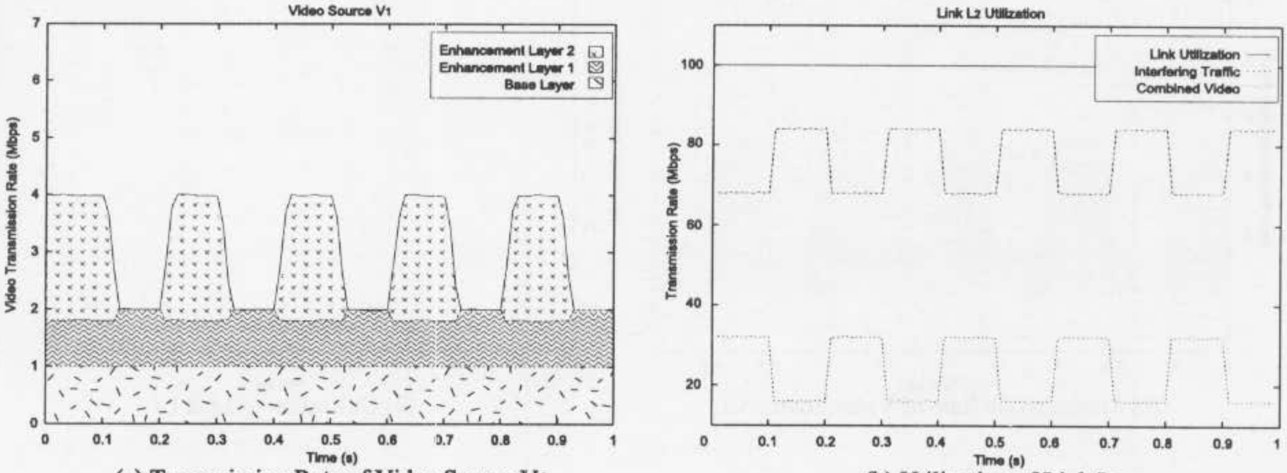


Figure 7: Simulation model for evaluating responsiveness and utilization.

by varying the size of the transitions on the available bandwidth as well as the size of network buffer allocated for the video service.

In the first experiment, a persistent stream of constant rate interfering traffic is applied to link L_1. The transmission rate of this interfering stream is 84 Mbps, leaving 2 Mbps of available bandwidth for use by each video connection. On link L_2, square-wave interfering traffic that oscillates with a period of oscillation of two hundred milliseconds between constant rates of 68 and 84 Mbps is applied in order to test the responsiveness of the source to rapid changes in the network's available bandwidth.



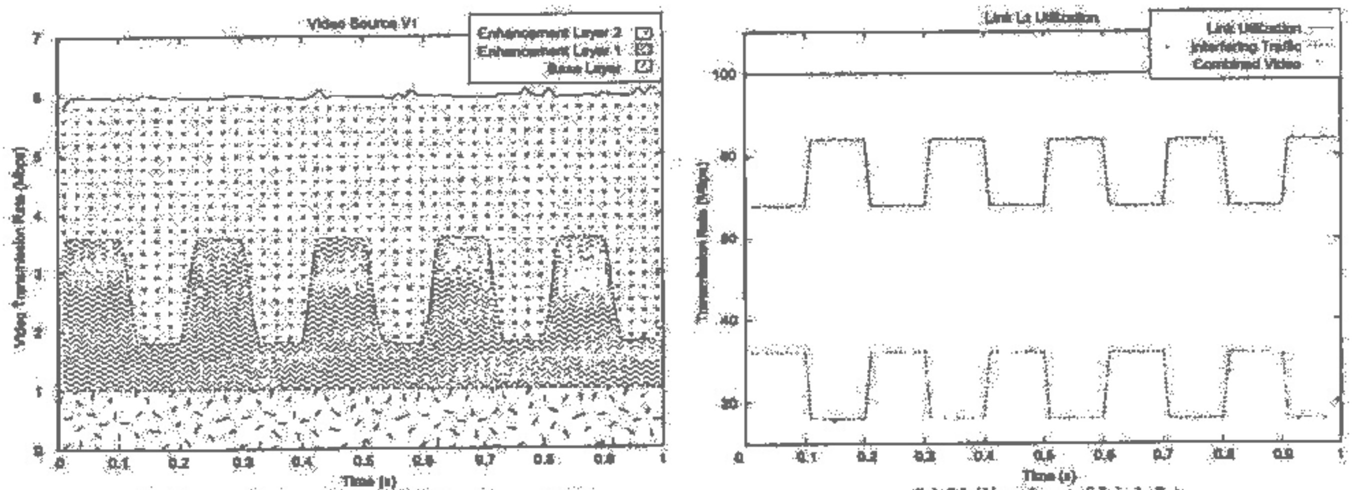
(a) Transmission Rate of Video Source V1

(b) Utilization of Link L2

Figure 8: Creation and deletion of video layers.

Figure 8(a) displays the rates of the video traffic layers generated by the video source V_1 (all other sources exhibited similar behavior). For the first 100 msec of the simulation, 4 Mbps is available for each video traffic on link L_2, while only 2 Mbps is available for each video traffic on link L_1. It then requires approximately 20 msec for the video sources to adapt to the available bandwidth. Since the destination monitoring interval is 20 ms, it takes about this amount of time for the destinations to start reporting to the source the new status of the network. The result is three layers of video, the base layer transmitted at the minimum video rate of 1 Mbps, the enhancement layer 1 transmitted at a cumulative rate of 1.8 Mbps, and the enhancement layer 2 transmitted at a cumulative rate of 4 Mbps. Note that the intermediate enhancement layer is transmitted at 90% of the available bandwidth on link L_1. Therefore, 10% of the available bandwidth on link L_1 is utilized by lower priority, layer 2 packets. At time $t = 100$ msec, the available bandwidth on link L_2 drops from 32 Mbps to 16 Mbps, and again the mechanism requires about 20 msec to react. During the next 100 msec, since both links have the same available bandwidth of 2 Mbps, each video source removes the enhancement layer 2, and starts transmitting the enhancement layer 1 at a cumulative rate of 2 Mbps. At time $t = 200$ msec, the available bandwidth on link L_2 returns to 32 Mbps, a second enhancement layer is added by each video source, and its cumulative rate converges to 4 Mbps approximately 20 ms later. As the available bandwidth on link L_2 oscillates between 32 and 16 Mbps, the mechanism responds by cyclically adding and removing a video layer within about 20 ms, as shown in Figure 8(a). The utilization of link L_2 is shown in Figure 8(b). The credit-based algorithm was able to utilize 100% of the link throughout the entire experiment. It is important to observe that as soon as bandwith becomes available, the rate of the combined video traffic is increased, and when the available bandwidth is reduced, the video rate is reduced accordingly. Throughout the whole experiment no video packets were lost on link L_2, and only low priority, layer 2 packets were dropped on link L_1 during periods when the sources were transmitting three layers of video.

The results of this first responsiveness experiment illustrate how the the proposed mechanism is able to respond to changes in network congestion by adding or removing an enhancement layers of video.



(a) Transmission Rate of Video Source V1

(b) Utilization of Link L2

Figure 9: Increasing and decreasing the rates of video layers.

In the second experiment, persistent interfering traffic is applied at a rate of 52 Mbps on link L_1. The square-wave interfering traffic applied to link L_2 in the first experiment is also applied in the second experiment. With 48 Mbps available on link L_1 and between 32 and 16 Mbps available on link L_2, it is expected that the feedback mechanism will generate three layers of video at all times, but with an oscillating rate for the enhancement layer 1. Figure 9(a) displays the transmission rates of each video layer generated by the video source V_1 (again, all other video sources produce similar behavior).

Since the available bandwidth on link L_1 (48 Mbps) always exceeds the available bandwidth on link L_2, no layers are added or deleted once the three layers have been established. A base layer is generated at the MVR of 1 Mbps. The persistent interfering traffic on link L_1 results in a cumulative rate of 6 Mbps per video source. The oscillating interfering traffic on link L_2 results in an enhancement layer 1 generated by each video source, with a cumulative rate fluctuating in concordance with the oscillations in available bandwidth on link L_2. Again, responses to changes in available bandwidth on link L_2 require approximately 20 msec to be reflected at each source. The cumulative rate of enhancement layer

1 oscillates between 1.8 and 3.6 Mbps. Again observe that 10% of the available bandwidth on link L_2 is utilized by low priority, layer 2 packets. In this experiment, 100% of link utilization is achieved at all times and no losses are experienced on link L_1, while losses on link L_2 are isolated to low priority, layer 2 packets.

The results from the second experiment illustrate the ability of the mechanism to adapt the transmission rate of a layer of video when bandwidth availability in the network changes.

In order to further investigate how the mechanism respond to oscillations in the available bandwidth in the network, a third set of experiments was performed. In these experiments a *responsiveness metric* is defined and measured as the time between a change in the available bandwidth and the time at which the source rate converges to a target rate. A sliding window of length 20 msec is used to detect when the source rate is within 0.5% of the target rate. The responsiveness metric is then equal to the time between the left side of the sliding window and the time of the bandwidth change. The sliding window jumps at intervals of 10 μsec. In this experiment, the buffer allocated in the network to each video connection is set at values of 50, 100 and 200 packets. Since larger buffers imply larger queueing delays, it is expected that they may also imply slower responsiveness. The size of the transitions might also have a direct effect on the responsiveness time. Larger changes in the available bandwidth may require longer time for the sources to converge to the new rate. In this experiment, separate responsiveness metrics are obtained for increases and decreases in the available bandwidth. Table 2 summarizes the average responsiveness metric, obtained from a sequence of 300 transitions in the available bandwidth.

		Changes in the Available Bandwidth					
		8 Mbps ABW=[16,24]Mbps		32 Mbps ABW=[16,48]Mbps		72 Mbps ABW=[16,88]Mbps	
Buffer Size	ABW	Adjust Rate	Add/Rem Layer	Adjust Rate	Add/Rem Layer	Adjust Rate	Add/Rem Layer
50	Up	21.0214	21.1417	21.0461	21.0461	20.5829	20.5829
	Dn	22.9554	22.4952	16.7575	17.0146	17.8480	17.8307
100	Up	21.1585	21.2624	20.5308	20.5308	20.6289	20.6289
	Dn	22.0598	22.1422	16.6132	16.9184	18.0631	18.0343
200	Up	21.5726	21.6780	20.7632	20.7631	20.1297	20.1297
	Dn	21.2580	21.6557	16.6899	17.0735	18.7253	18.6847

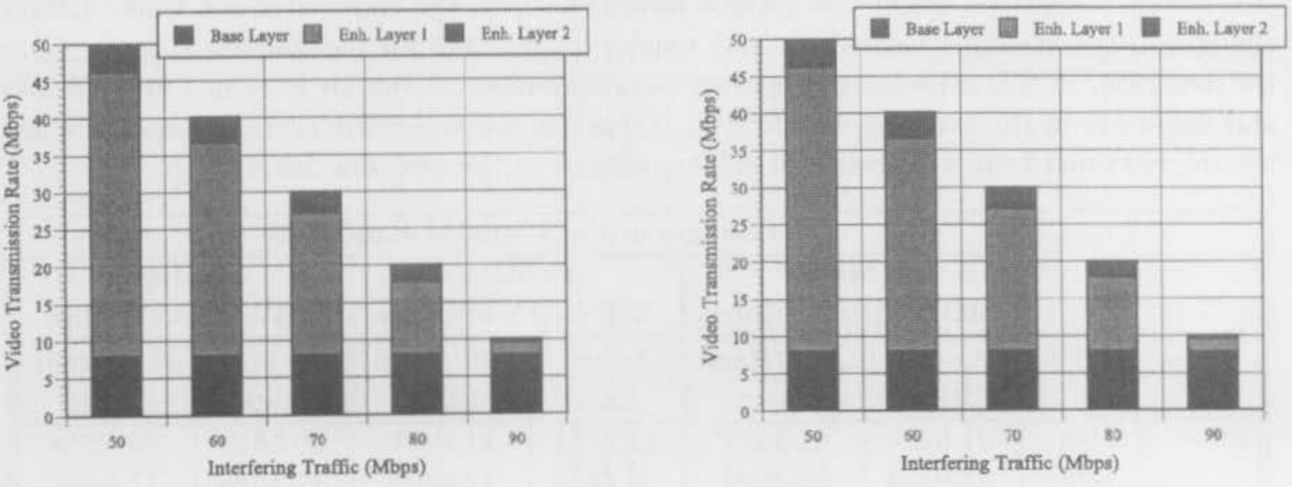
Table 2: Responsiveness Metrics.

Surprisingly, the results show little or no correlation between the responsiveness of the mechanism and the size of the network buffers allocated for the video service, or the size of the changes in the available bandwidth on the network. The conclusion we can draw from this experiment is that the major component determining how fast the mechanism can adapt to changes in the network is the destination monitoring interval of 20 ms. Another factor contributing to the invariance of the responsiveness metrics is the abscence of interfering traffic in the backward path. Changes in the available bandwidth on the network are reflected on the video rate received by the destinations, which report the average video rate received over the past *destination monitoring interval* on feedback packets. Since there's no backward interfering traffic, after the link propagation delays from the destinations back to the sources, the sources start adjusting their transmission rates to the new status of the network.

A few extra observations can be extracted from Table 2. First, the responsiveness metrics for merely adjusting rates or for adding and removing layer and adjusting rates were similar throughout all the test cases. Also, for medium and large changes in the available bandwidth, decreasing the rate converged faster to the target rate than increasing it. This is most likely due to the fact that the proposed mechanism increases its transmission rates incrementally whenever the source buffer is below a threshold of 33%, whereas decreases in rate occur immediately in response to explicit rate indications provided by the destinations.

4.3 Utilization

One of the goals of adaptive congestion control techniques is to optimize utilization of network bandwidth. In a multi-layered multicast service, the combined throughput is bounded by the utilization of the least congested source-to-destination path. The results of the experiments to evaluate responsiveness showed that 100% of utilization is achieved when oscillating, square-wave interfering traffic is applied. In order to better evaluate the utilization of the mechanism, Poisson interfering traffic is applied on both links L_1 and L_2. The load of the interfering traffic (ρ) is the same on both links.



(a) Link L1

(b) Link L2

Figure 10: Video Utilization on Links L_1 and L_2.

In this experiment, it is expected that at a given time each source generates at most three layers of video. Since both links contain the same load of interfering traffic, losses are equally likely to occur on both links.

On this experiment, the average rate of the interfering traffic is varied between 50, 60, 70, 80 and 90 Mbps. In all experiments, 100% utilization is observed on links L_1 and L_2. Figure 10 shows the combined average video transmission rate versus the interfering traffic load on links L_1 (a) and L_2 (b). The histograms also show the average rate of each video layer. Packet losses observed in all experiments are isolated to the low priority enhancement layers 1 and 2. Table 3 shows the average loss ratio observed in both links L_1 and L_2. The loss ratio of each enhancement layer increases exponentially with the load of the interfering traffic (ρ). The loss ratio of the enhancement layer 2 was 1.64% for ρ equals to 0.5 and reached 7.74% when ρ was 0.90. The loss ratio of the enhancement layer 1 was 0.069% for ρ equals to 0.5 and reached 3.867% when ρ was 0.90.

Loss Ratio	Interfering Traffic Load				
	ρ=0.5	ρ=0.6	ρ=0.7	ρ=0.8	ρ=0.9
Base Layer	0	0	0	0	0
Enh. Layer 1	0.069%	0.088%	0.093%	0.193%	3.867%
Enh. Layer 2	1.643%	2.290%	2.640%	4.070%	7.740%

Table 3: Loss Ratio versus Interfering Load (ρ)

4.4 Fairness

An important factor in the evaluation of any traffic control mechanism is its fairness. If the mechanism fails to divide bandwidth equally among competing connections, then some connections may unfairly receive better service than others. This set of simulation experiments evaluates how fairly the proposed feedback mechanism allocates bandwidth to competing video connections.

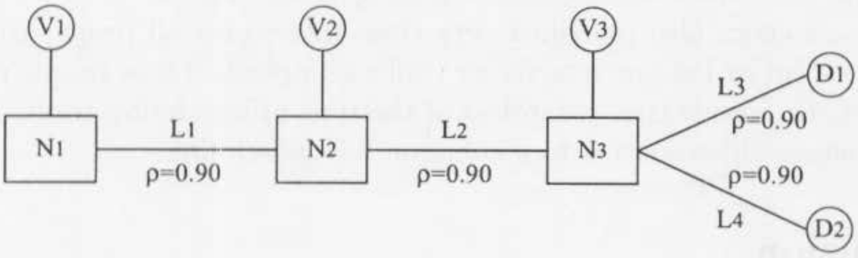


Figure 11: Simulation model for evaluating fairness.

The so-called "parking lot" model depicted in Figure 11 is used to test fairness. This network topology consists of three video sources $\{V_1, \ldots, V_3\}$, each located at a different point in the network and transmitting video across intermediate nodes $\{N_1, \ldots, N_3\}$ to two common destinations D_1 and D_2. Links L_1, L_2, L_3 and L_4 are congested with independent interfering traffic loads of $\rho = 0.90$. This leaves, on average, 10 Mbps of available bandwidth on each of the bottleneck links. In this experiment, two sets of simulations were performed. In the first set of simulations, the bottleneck links are congested by a persistent stream of interfering traffic generated at a constant rate. In the second set of simulations, Poisson interfering traffic is used. In order to measure the effect of the round trip time on the fairness of the feedback mechanisms, propagation delays between intermediate nodes are varied between 5 μs and 500 μs, representing distances of 1 km and 100 km, respectively.

The allocation of bandwidth to competing video traffic streams is said to be optimal if it is *max-min fair*. A max-min fair allocation of bandwidth occurs when all active connections not bottlenecked at an upstream node are allocated an equal share of the available bandwidth at every downstream node [18, 19]. In the model shown in Figure 11, a max-min fair allocation of bandwidth occurs if all three sources are told to transmit at the same rate. To measure fairness, we calculate the standard deviation σ of the rates that each source transmits across the bottleneck links L_3 and L_4. An optimally fair allocation results in a standard deviation of zero.

Table 4 summarizes the results of both sets of simulations. It presents the average bit rate in Mbps used by each video traffic stream on bottleneck link L_3 (similar behavior is observed on link L_4). Since the average available bandwidth on link L_3 is 10 Mbps, the optimal fair share is 3.333 Mbps for each of the three video streams. The proposed mechanism proved to be fair in the sense that it equally divides the available bandwidth of

Interfering	Links $L_{\{1,2,3,4\}}$	Rate (Mbps)			Utilization	Fairness
Traffic	Prop. Delay	V_1	V_2	V_3		σ
Constant	5 μs	3.333	3.333	3.333	100%	0.0001
	50 μs	3.333	3.333	3.334	100%	0.0003
	500 μs	3.330	3.332	3.337	100%	0.0024
Poisson	5 μs	3.339	3.340	3.346	100%	0.0025
	50 μs	3.229	3.231	3.238	100%	0.0033
	500 μs	3.292	3.295	3.335	100%	0.0167

Table 4: Video transmission rates and fairness metric with Constant and Poisson interfering traffic.

link L_3. Optimal fair share was achieved in most cases. Slight unfairness can be observed as the round-trip delay increases, however for practical purposes the mechanism proved to fairly divide the bandwidth among the competing video sources.

The fairness metrics also remained very close to zero for all propagation delay values when both persistent or Poisson interfering traffic is applied. These results demonstrate the fair behavior of the mechanism, regardless of the type of interfering traffic or the distances from the competing video sources to a common bottleneck link.

5 Conclusion

A multi-layered, feedback-based mechanism for the transport of multicast video has been presented and investigated in this paper. In this mechanism, the source uses network feedback to dynamically adjust both the number of video layers it generates and the rate at which each layer is generated. By doing so, it optimizes bandwidth utilization and the quality of video received by each destination.

The proposed mechanism's performace was evaluated in terms of utilization, video quality, responsiveness and fairness. The mechanism's ability to enhance the video quality when bandwidth is available was illustrated. In terms of responsiveness, the most important factor determining how fast the mechanism adapts to changes in the network was the destination monitoring interval of 20 msec. Optimal utilization of 100% was observed in all experiments. The mechanism also proved to fairly share the available bandwidth among competing video sources, regardless of the distances to a common bottleneck link.

In future work, we intend to explore the impact of the mechanism described in this paper on an actual network, through implementation on a modified IP network testbed.

References

[1] P.P. White. PSVP and Integrated Services in the Internet: A Tutorial. *IEEE Communications Magazine*, May 1997.

[2] Martin Vetterli and Jelena Kovacevic. *Wavelets and Subband Coding*, chapter Subband and Wavelet Coding of Images and Video. Prentice Hall, 1995.

[3] Y. Omori, T. Suda, G. Lin and Yasuhiro Kosugi. Feedback-based Congestion Control for VBP Video in ATM Networks. In *Proc. of the 6th Int'l. Workshop on Packet Video*, 1994.

[4] C.M. Sharon, M. Devetsikiotis, I. Lambadaris, and A.P. Kaye. Pate Control of VBP H.261 Video on Frame Pelay Networks. In *Proc. of the International Conference on Communications (ICC)*, pages 1443–1447, 1995.

[5] H. Kanakia, P.P. Mishra, and A. Peibman. An Adaptive Congestion Control Scheme for Peal-Time Packet Video Transport. *IEEE/ACM Transactions on Networking*, 3(6):671–682, Dec. 1995.

[6] T.V. Lakshman, P.P. Mishra, and K.K. Pamakrishnan. Transporting Compressed Video over ATM Networks with Explicit Pate Feedback Control. In *Proc. of IEEE Infocom*, 1997.

[7] J.C. Bolot, T. Turletti, and I. Wakeman. Scalable Feedback Control for Multicast Video Distribution in the Internet. In *Proc. of ACM SIGCOMM*, pages 58–67, August 1994.

[8] S.Y. Cheung, M.H. Ammar, and X. Li. On the Use of Destination Set Grouping to Improve Fairness in Multicast Video Distribution. In *Proc. of IEEE Infocom*, 1996.

[9] P.A.A. Assunção and M. Ghanbari. Multi-Casting of MPEG-2 Video with Multiple Bandwidth Constraints. In *Proc. of the 7th Int'l. Workshop on Packet Video*, pages 235–238, March 1996.

[10] S. McCanne, V. Jacobson, and M. Vetterli. Peceiver-Driven Layered Multicast. In *Proc. of ACM SIGCOMM*, pages 117–130, August 1996.

[11] B. J. Vickers, C. V. N. Albuquerque and T. Suda. Adaptive Multicast of Multi-Layered Video: Pate-Based and Credit-Based Approaches. *to appear at INFOCOM*, March 1998.

[12] The Flow Control Consortium c/o Ascom Nexion Inc. Quantum Flow Control, Version 2.0, July 1995.

[13] H. T. Kung, T. Blackwell and A. Chapman. Credit-Based Flow Control for ATM Networks: Credit Update Protocol, Adaptive Credit Allocation, and Statistical Multiplexing. In *Proc. of ACM SIGCOMM*, 1994.

[14] H.T. Kung and K. Chang. Peceiver-Oriented Adaptive Buffer Allocation in Credit-Based Flow Control for ATM Networks. In *Proc. of IEEE Infocom*, 1995.

[15] K.K. Pamakrishnan and P. Newman. Integration of Pate and Credit Schemes for ATM Flow Control. *IEEE Network Magazine*, 1995.

[16] P. Morris and H.T. Kung. Impact of ATM switching and flow control on TCP Performance: Measurements on an experimental switch. In *Proc. of IEEE Globecom*, November 1995.

[17] P. Chandra, A. Fisher, C. Kosak and P. Steenkiste. Experimental Evaluation of ATM Congestion Control Mechanisms. In *Proc. of IEEE Infocom*, 1997.

[18] D. Bartsekas and P. Gallagher. *Data Networks, second edition.* Prentice Hall, 1987.

[19] F. Bonomi and K.W. Kendrick. The Pate-Based Flow Control Framework for the Available Bit Pate ATM Service. *IEEE Network Magazine*, pages 25–39, March/April 1995.