QoS Control using Adaptive Layerd Data Transmission*

Dorgham Sisalem, Frank Emanuel
GMD-Fokus, Berlin
{sisalem,emanuel}@fokus.gmd.de

December 3, 1997

Abstract

To improve the QoS of best-effort based networks and accommodate heterogeneous receivers different solutions based on applying hierarchical transmission of data streams have been introduced. These solutions are based on multicasting a base layer which is received by all interested end systems. In addition, enhancement layers are transmitted that improve the quality of the base layer. Each receiver can determine the number of enhancement layers to join and thereby the QoS he would like to receive based on its own capabilities and view of the network congestion state. With this approach, however, the quality of a received data stream is improved or reduced with the granularity of one layer. Depending on the encoding and transmission method used, the additional bandwidth required by each layer can vary greatly. To fine tune the QoS changes, we propose in this paper to enhance the hierarchical transmission schemes through sender adaptation using a scheme called adaptive layered transmission (ALT). With the ALT approach, the sender dynamically redistributes a data stream on the different layers based on the feedback of the receivers. In addition, ALT allows for dynamic discovery of the appropriate transmission rate for each layer. First results obtained through simulations suggest an overall improvement of the QoS perceived by heterogeneous receivers.
Keywords: QoS control; RLM; RSVP; layered transmission; adaptive applications; ALT

1 Introduction

Many distributed multimedia applications have the ability to adapt to fluctuations in the network conditions. By adjusting temporal and spatial quality to available bandwidth, or manipulating the playout time of continuous media in response to variations in delay, multimedia flows can keep an acceptable QoS level at the end systems. However, due to the heterogeneity of the Internet a single sender transmission rate cannot satisfy the conflicting bandwidth requirements at different sites. Therefore, the sender rate is usually adapted to the requirements of the worst positioned receiver, thereby reducing the quality of the data perceived at all receiving sites. This limitation can be overcome with a layered transmission scheme [10, 9, 7]. In this model, each source divides a data stream

*This work was funded in part by the BMBF (German Ministry of Education and Research) and the DFN (German Research Network).

across multiple network channels.That is, each layer is sent to a different multicast address. Based on their capabilities, the receivers can determine how many sessions to join and thereby adapting the quality of the received data to their own requirements. Depending on the details of the used approach, a receiver might dynamically join and leave the multicast sessions carrying the different layers. That is, the QoS of a receiver might change with the granularity of one layer. However, adding a new layer or removing one results in large changes in the perceived QoS and adds high loads to the links leading to the receiver. Hence, we propose a new scheme called adaptive layered transmission (ALT) to enhance the layered transmission schemes by dynamically adapting the size of the different layers based on the network congestion state. That is, in the case of congestion on one of the lower layers a receiver should not just leave the multicast session of that layer, but with ALT he would ask the sender to reduce the load on that layer. By shifting the load to a higher layer the overall transmission rate can be maintained and the receiver of the lower layer would have lower losses. Note, that in this paper, the terms layer and the multicast session carrying this layer are used interchangeably.

In addition to improving the granularity of QoS changes in hierarchical transmission schemes, ALT also enables a dynamic discovery of the bandwidth available for each data layer. Currently, data is divided among different layers based on encoding characteristics or possible network topologies. For example to support users connected through ISDN, T1 and ATM links to a multicast session, a sender might divide its data stream into three layers with the base layer having a data rate of 64 kbit/s, a first enhancement layer with a few hundreds kbits/s and the rest of the data to be sent is then placed on the second enhancement layer. However, with no a priori knowledge of the traversed links, such a static distribution might be disadvantageous as it does not consider the actual resources of the receivers. For example, consider the distribution presented above for accommodating receivers connected over ISDN, T1 and ATM links. For the case that no members are connected over ISDN links, and some members have enough resources to receive a data stream larger than the base layer but smaller than the first enhancement layer, they would still be restricted to the lower layer. With ALT, the sender adjusts its transmission rate on each layer in accordance with the worst receiver capacities of that layer. That is, the sender would increase the transmission rate of the base layer to a level than the initial one and lower than that used for the first enhancement layer.

In [1] the authors describe a mechanism called SCUBA that enhances hierarchical data transmission schemes by assignment of signal layers to network channels based on user preferences computed by the SCUBA scheme. Given a user preference, each source determines its signal layer mapping. While this approach accommodates receivers preferences in accordance with the sent data, the QoS of the participating receivers is still only changed with the granularity of a layer and provides no assistance in dynamically determining the appropriate transmission rate of each layer.

In Sec. 2, we describe different layered transmission proposals. Sec. 3 describes the details of the ALT scheme consisting of an adaptation scheme used for determining the rate to be used on the different layers and the interaction approach between the adaptation scheme and the layered transmission approach. Implementation of the ALT scheme in the PTOLEMY [2] simulation tool and simulation results are presented in Sec. 4 and Sec. 5 respectively. Finally, in Sec. 6 we conclude the paper and present some open questions.

2 Background and Related Work

In packet-switched networks like the Internet, packets traverse paths of widely ranging bandwidths capabilities and load factors. In such heterogeneous environments, a single sender transmission rate cannot satisfy the conflicting bandwidth requirements at different sites. Therefore, the sender rate is usually adapted to the requirements of the worst positioned receiver, thereby reducing the quality of the data perceived at all receiving sites.

A straightforward approach for accommodating the heterogeneity aspect of packet-switched networks is by transmitting a data stream on different layers. A base layer provides in this case the minimal amount of data needed for an acceptable representation of the original data stream. Each higher layer provides a QoS improvements.

Figure 1 depicts an example of a multi-layered transmission approach. The sender transmits its data split into three streams over three layers. As there is enough bandwidth available on the link between the upper router and receiver R1, R1 decides to subscribe to all three layers. The link between the two routers has a lower capacity. As a result, only the two lower layers of the three can be forwarded. The available bandwidth on the link to receiver R3 is even more restricted, so R3 decides only to receive the base layer. The link to R2 has a capacity of 10 Mb/s, but as the capacity between the two routers is restricted to 512 kbit/s, R2 can only subscribe to two layers, i.e., the base layer and enhancement layer 1.

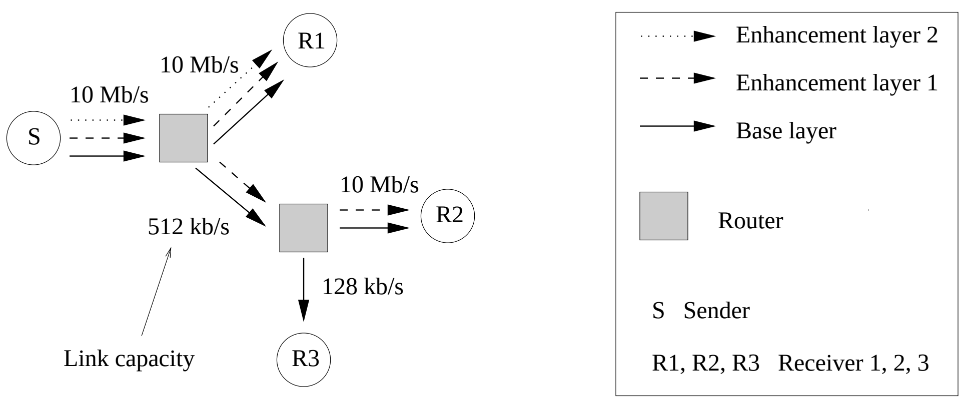

Figure 1: Example of multi-layer transmission

The different approaches introduced in the literature for utilizing layered data transmission can roughly be divided into two categories: cumulative and independent layering.

2.1 Cumulative Layered Data Transmission

In the cumulative case, each layer provides refinement information to the previous layers and the receiver must listen to all lower layers up to and including the highest one it wants to listen to, i.e., if a receiver wants to listen to layer 3 it also has to listen to layer 1 and layer 2.

For example, the MPEG-2 standard offers four different so-called “scalability modes” [5]: spatial, SNR, and temporal scalability, as well as data partitioning. MPEG-2 allows the simultaneous use of up to two different scalability modes (except from data partitioning) in any combination, hence resulting in a 3-level representation of the signal. In spatial scalability, the base and enhancement layers operate at different spatial resolutions. In SNR scalability, both adjacent layers have the same resolution and the enhancement refines the quantization process performed in the base layer. In temporal scalability, the enhancement layer increases the number of frames per second of the base layer. Data partitioning splits encoded data into two bit streams, one containing more and the other containing less important data. More important data are low frequency coefficients and motion vectors. Data partitioning and SNR scalability lead to a drift problem caused by motion compensated coding if only the lower bit rate layer can be received. Under this condition, different reference frames for motion compensation are used at coder and decoder [8]. Another problem with cumulative layering is that of resynchronization. As the different layers might actually take different routes to the receiver, they might suffer from different delays. To reconstruct a data segment consisting of data packets sent over different layers the receiver would need to wait until all the different parts are received. This incurs additional delay and thus might reduce the overall reception quality.

2.1.1 Hierarchical Data Transmission and Reservation

In [7] the authors describe a scheme that combines hierarchical data transmission and resource reservation. In their approach, the different data layers are transmitted over different multicast sessions. End systems wishing to receive a certain data layer, join the multicast group on which the data layer is being sent and issue a reservation request based on the characteristics of that data layer. The reservation is requested by using the RSVP protocol [16]. If the reservation request cannot be granted due to lack of resources, the receiver must leave the session. There are two strategies how the reservation requests can be done: with the first one, the receiver joins all sessions, tries to issue reservations for all layers at once, and leaves the groups for layers where the reservation fails. This method could lead to severe temporary congestion while the reservation errors propagate back to the receiver. Therefore, the authors prefer the second strategy: The layer streams are joined and reserved incrementally, starting with the base layer, until a reservation error occurs.

Reserving enough resources for supporting a certain Quality of Service (QoS) in advance guarantees this quality level and would be the most straightforward approach for handling the problem of QoS control. However, as it is usually difficult to know the exact characteristics of a certain stream in advance, there would be a tendency to over reserve resources in order to guarantee the requested QoS level. As a consequence, this would lead to under utilizing the network. Also, if a connection's request for a certain QoS level arrived during an overload situation, the network would only allow for lower quality, even though the network might become idle later. Finally, reservation can only be used if all the network nodes involved support it, which currently is not the case.

2.1.2 Receiver-Driven Layered Multicast

In [10] the authors describe a scheme called "Receiver-Driven Layered Multicast" (RLM). As the title indicates, it is a receiver-based rate adaptation technique which combines a layered source coding algorithm with layered transmission. It also uses one IP Multicast group for each coding layer. In contrast to [7], is does not require special support (RSVP) from the network. Within the RLM protocol, the data source does not take an active role, it simply transmits each layer of its signal on a separate multicast group. The protocol machinery is run at each receiver, where adaptation is done by joining and leaving multicast groups. Conceptually, each receiver runs the following control loop: on congestion it drops a layer; on spare capacity its adds a layer. If losses occur, the current level of subscription is too high. Otherwise, if there are no losses, this does not automatically mean that the level of subscription is too low. To find out if the level is too low, so-called “join-experiments” are performed: layers are added spontaneously at “well chosen” times. If this results in congestion, the new layer is dropped quickly and a timer is set for a new experiment. Otherwise the layer is kept. Scalability and interference problems introduced with the join experiments are solved by “shared learning”.

2.2 Independent Layered Data Transmission

Another approach for layered transmission is to simply transmit the same data stream encoded with different quality levels on different multicast sessions [9].This scheme is often called *simulcast* because the source transmits multiple copies of the same signal simultaneously at different rates resulting in different qualities. As the streams contain all the necessary information for decompression, the receivers need only to join one multicast session. This approach avoids the resynchronization and drift problems seen with cumulative approaches. However, this is reached at the cost of sending multiple replicated streams and thus possibly congesting the network.

3 The Adaptive Layered Transmission Scheme (ALT)

Using layered transmission approaches such as RLM results in QoS adjustments with the granularity of one layer. This might result in sudden and obvious changes in the perceived QoS at the receivers. With ALT we propose an intermediate solution as an enhancement to layered transmission schemes such as RLM. With ALT, instead of leaving a congested layer, the receiver notifies the sender of the congestion state upon which the sender reduces its transmission rate on the congested layer and increases it on a higher one if such a layer exists. Thereby, the overall transmission rate may stay constant and the end systems receiving all layers will not be affected by the rate changes.

The ALT scheme consists of three major parts: the feedback information, an adaptation scheme and the mechanisms for the dynamical redistribution of the bandwidth among the different layers.

3.1 The Real Time Transport Protocol

The adaptation scheme we have investigated is based on the real time protocol (RTP) [11] designed within the Internet Engineering Task Force(IETF). RTP is an end-to-end protocol that is often used together with other transport protocols, in particular UDP for transporting multimedia data. RTP has no notion of a connection; it may operate over either connection-oriented (say, ATM AAL5) or connectionless lower-layer protocols (typically, UDP/IP). It does not depend on particular address formats and only requires that framing and segmentation is taken care of by lower layers. RTP offers no reliability mechanisms. It is typically implemented as part of the application or as a library rather than integrated into the operating system kernel.

RTP sessions consist of two lower-layer data streams, namely a data stream for, say, audio or video and a stream of control packets (using the sub-protocol called RTCP). Each session member periodically multicasts RTCP control reports to all other session members. Senders transmit reports describing the amount of data they have sent. The receivers send for each incoming stream a receiver report indicating the fraction of lost data, timing information and various identification and QoS information.

3.2 The Adaptive Bandwidth Adjustment Scheme

As an adaptation scheme we use the algorithm proposed in [4]. With this scheme the sender reduces its transmission rate by a multiplicative decrease factor after receiving feedback from the receiver indicating losses above a certain threshold called the upper loss threshold (λ_c). With losses below a second threshold called the lower loss threshold (λ_u) the sender can increase its transmission rate additively. For the case that the feedback information indicates losses in between the two thresholds the sender can maintain its current transmission level, see Fig. 2. Reducing the rate multiplicatively allows for a fairer reaction to congestion. That is, connections utilizing a disproportionately large bandwidth share are forced to reduce their transmission rate by a larger amount.

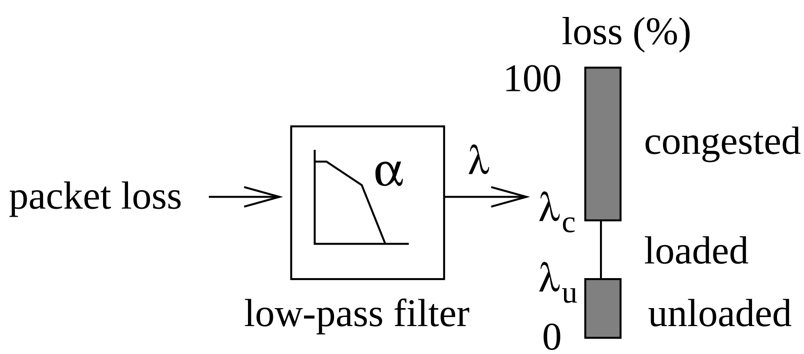

Figure 2: Additive increase/multiplicative decrease algorithm

To smoothen the effects of sudden changes in the network congestion state the scheme uses a smoothed loss (λ) for determining the congestion state and not the actual values reported in the RTCP messages.

$$\lambda_i = (1 - \alpha)\lambda_i + \alpha \, \text{loss}_i$$

with loss_i as the loss indicated in the RTCP reports from receiver i and λ_i as the smoothed loss value for that member. For the case of multiple receivers, λ is set to $\max(\lambda_i)$. Finally, the adaptation is further limited by setting a minimum transmission rate and a maximum one. That is, the sender can not increase its transmission rate above a certain limit and should not decrease it below the minimum rate.

We have chosen this scheme based on its similarity to a large number of other adaptation schemes proposed in the literature such as [3, 12]. Thereby, we believe that the results obtained with other schemes would not differ substantially from the here reported results.

3.3 The Dynamical Bandwidth Arbitrator (DBA)

At each layer, an instance running the adaptation scheme presented in Sec. 3.2 estimates the transmission rate to be used on that layer. The dynamical bandwidth arbitrator distributes the overall transmitted rate among the different layers based on the transmission rates for each layer.

1. In the first step, the sender starts transmitting data with an initial rate R_i° on each layer L_i. This rate can be specified based on the details of the used data encoding or network topologies supported.

2. Receivers join the first layer, and if no losses were noticed they join the second layer and do so on.

3. If losses were noticed on a layer (L_i), then on the contrary to the approaches presented in Sec. 2 the receiver does not need to leave the congested layer but should inform the sender about the loss. This notification is done using the periodic reports sent to the senders with RTCP.

4. Upon receiving a loss notification at the sender site the adaptation instance of layer L_i calculates a new reduced transmission rate (R_i) and informs the arbitrator about the new transmission rate.

5. To maintain an overall constant transmission rate, the DBA tries to shift the reduced transmission rate of layer L_i up to layer L_{i+1}.

6. If the layer L_{i+1} was transmitting data at a rate (R_{i+1}) less than its initial rate R_{i+1}° which indicates congestion on that layer, then the rate should be shifted directly from layer L_i up to L_{i+2} or up the first higher layer (L_x) transmitting at a rate ($R_x \geq R_x^\circ$).

7. In the case that a layer (L_i) is underutilized, the adaptation instance on that layer would signal the DBA to increase the transmission rate (R_i). This is done either by increasing the overall transmission rate by the requested amount or by reducing the transmission rate of the highest layer. The first approach would lead to a QoS improvement at all members. The second approach is suitable for the cases, where the sent data stream is constant. In this case, only members listening up to layer (L_i) would benefit from the rate shifting.

8. To avoid reducing the rate on a certain layer down too much, a receiver should leave a layer if he is facing losses higher than a certain threshold.

9. In the common case, a low loss value over a layer (L_i) would result in an increased transmission rate (R_i). However, as the sender adjusts its transmission rate down to the worst receiver level, the only way for a receiver listening to this layer and having no losses to improve its QoS is by joining a higher layer. In this, the receiver should follow a scheme such as RLM or the reservation based scheme presented in Sec. 2. The same would be valid for the case that the sender is sending with R_i equal to the maximum transmission rate set for layer L_i.

10. Finally, note, that in the IP multicast model underlying our work, the multicasted data would most probably traverse the same paths until reaching a splitting point. If congestion were to happen on a link traversed by all the layers, the ALT scheme would try to readjust the bandwidth distribution, by sending more on the highest layer and less on the lower layers. However, in this case the highest layer is congested as well. With the adaptation applied on the highest layer as well, the sender would reduce its transmission rate. If the transmission rate was reduced below the minimum transmission rate of highest layer, the sender should

just quit this layer and only restart it when the lower layers seem to be satisfied. This applies then recursively to all lower layers as well.

As an example, consider the case of a sender wishing to transmit a stream with an average rate of 3000 kbit/s in three layers with each layer having an initial rate of 1000 kbit/s. So, the sender starts transmission with 1000 kbit/s at each layer. Due to congestion, the sender might have to reduce its transmission rate down to 600 kbit/s on the first (base) layer. The remaining 400 kbit/s should then be transmitted on the second layer, resulting in 1400 kbit/s on that layer. However, if the capacity of some member listening to the second layer limits the transmission rate on this layer to 1200 kbit/s, the remaining 200 kbit/s should be transmitted on the third layer. In this case, the transmission rates of the three layers would be 600 kbit/s, 1200 kbit/s, and 1200 kbit/s, respectively. The total average of 3000 kbit/s is still maintained. Now, due to an underload situation on the first layer, the sending rate on that layer could be increased up to 900 kbit/s. To allow the members of the first two layers to have the best possible quality, the rate of the first layer is increased by decreasing the rate of the third layer simultaneously. That is, the final rate distribution looks as follows (layer one to layer three): 900 kbit/s, 1200 kbit/s, and 900 kbit/s. As another option, the sender might just increase the overall transmission rate. Thus, the bandwidth distribution would be: 900 kbit/s, 1200 kbit/s and 1200 kbit/s, with an overall transmission rate of 3300 kbit/sec. This option would result in a better reception quality at receivers at the cost of sending more data.

4 Implementation of the ALT Scheme

Figure 3 depicts the implementation concept of the ALT scheme based on a three-layer model. The main components in the diagram are the adaptive sender and the controller (DBA). The sender consists of three instances of the adaptation scheme presented in Sec. 3.2 with each instance controlling the bandwidth of one layer. The sender starts transmitting data with the initial rate for each layer. The receivers listening to a certain layer send back RTCP packets containing receiver reports (RR) about the data they received over that layer. According to the reported loss rates, the adaptation instance of that layer makes an adjustment decision. The amount of bandwidth needed to be increased or decreased on a certain layer is communicated to the controller. According to the bandwidth calculated on layer 1 and 2, the controller recalculates the bandwidth distributions. This is done in the following way: The controller waits for a message from the adaptation instance on layer 1 or layer 2 asking for a rate increase or decrease. The layer of the adaptation instance asking for a rate change is called the “actual layer”. If the bandwidth of the actual layer is to be decreased, the amount of the desired decrease is subtracted from the current rate of the actual layer and added to the rate of the next higher layer. The controller then redistributes the overall used bandwidth rate among the three layers. On the other hand, if the bandwidth on the actual layer is to be increased, the controller decreases the rate of the highest layer (layer 3). After redistributing the bandwidth the actual layer is set to the next higher layer which may ask for decrease/increase. That is, the controller waits for a message from the adaptation instance on the new actual layer. Thereby, before a layer can change its rate again, the influence of the previous change on the other layers will be considered.

At the highest layer, the rate changes need not be conveyed to the DBA as there are not any layers to which rate can be shifted to.

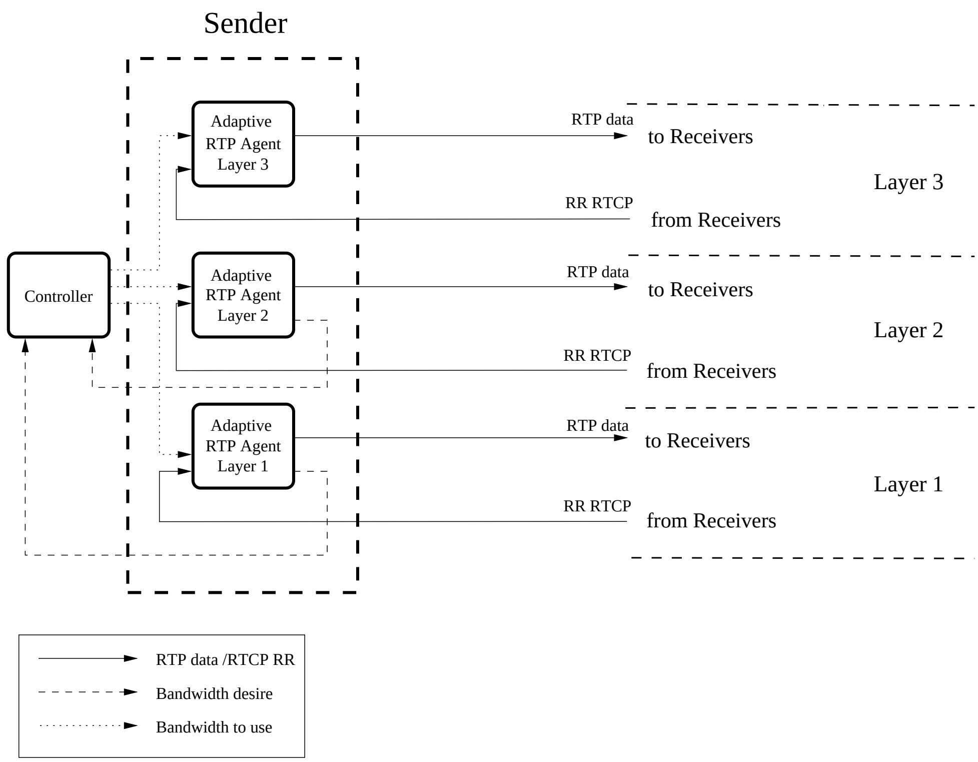

Figure 3: Diagram of the control scheme with three layers

5 Experimental Results of Utilizing the ALT Scheme

Fig. 5 depicts the network topology used for the simulation test of the ALT scheme. A sender is transmitting data over three layers to there receivers. One of the receivers is only receiving one layer, another is receiving two and the final one is receiving all of the layers. As background traffic we used an active-idle source as was described in [15], see Fig. 4. The number of packets generated during an active state is geometrically distributed with mean N_p. The pause between two packets during the active state is drawn from a negative exponential distribution with mean T_{off}. The idle states can have any general distribution with mean T_{idle}. Based on the traffic characteristics described in [15], we used an active-idle source with N_p set to 50, T_{off} set to 2 msec and T_{idle} set to 5 seconds.

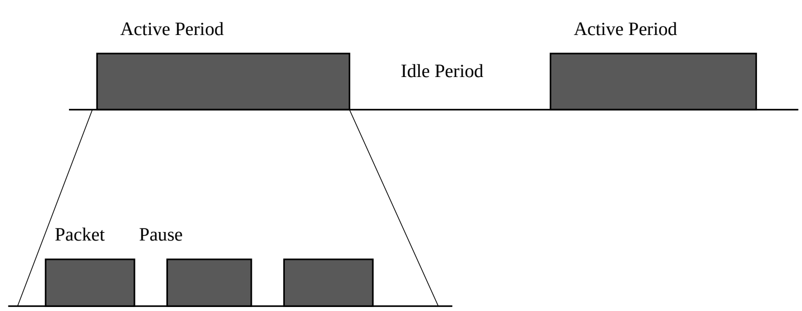

Figure 4: A three state active-idle source

The sender starts transmitting data on the three layers with the initial rates of 64 kbit/s, 500 kbit/s and 1.5 Mbit/s respectively. In addition, background traffic is turned on at router 3, thus affecting the traffic of the first layer. To increase the congestion state, additional background traffic is added at router 2 starting from time 500 seconds and thereby affecting both layer 1 and layer 2. Note, that in the tests described here, the source aims at maintaining an overall constant transmission rate. That is, when the transmission rate of a certain layer can be increased, the transmission rate of the highest layer is reduced.

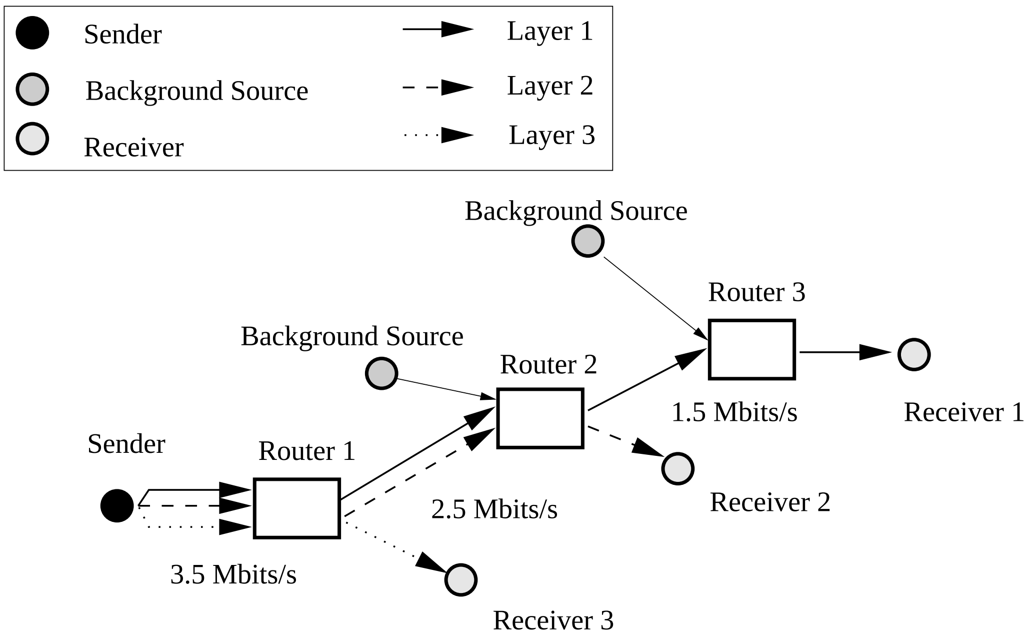

Figure 5: Experiment configuration

Parameter	**Value**
Initial Bandwidth	64 kbit/s, 500 kbit/s, 1.5 Mbit/s
Lower Loss Threshold λ_u	5%
Upper Loss Threshold λ_c	10%
Additive Increase ν	50 kbit/s
Multiplicative Decrease μ	0.875
Minimum transmission rate b_{min}	50 kbit/s
Maximum transmission rate b_{max}	10 Mbit/s

Table 1: Parameter values for the adaptation instances

The results depicted in Fig. 8 reveal that while the sender starts transmitting data on the first layer with an initial rate of 64 kbit/s, it finally oscillates around 500 kbit/s. Fig. 7 depicts a similar result for the second layer, that is the source starts with 500 kbits/s and reaches 1.5 Mbits/s on the average. For the third layer, the sender actually decreases the transmission rate. As the rate gets reduced down to the minimum transmission rate the user could actually drop this layer. While this would slightly reduce the overall data amount arriving at the receivers, it might improve the QoS at the end systems as the receivers would only need to consider two layers which results in easier resynchronization of the data and possibly lower delays.

Figures 9 through 11 show the total amount of RTP bandwidth measured at receiver 3 that listens to all 3 layers, receiver 2 listening to layers 1 and 2, and receiver 1 listening only to layer 1, respectively. We notice that receivers listening to all layers, receive a constant data stream equal to the sum of the initial bandwidth rates of the three layers of around 2 Mbit/s. Receivers of the first two layers have a somewhat oscillatory data stream of around 1.8 Mbit/s which is considerably higher than the sum of the initial rates of the first and second layer (564 kbit/s). The oscillations are a result of the adaptation to the bursty background traffic and are typical to the additive increase/multiplicative decrease kind of adaptation scheme similar to the one we are using here [6].

Figures 12 through 14 show the data losses occurring on link 3, link 2 and link 1, respectively.

The results show that with the ALT scheme, the sender adapts the transmission rate for each layer in accordance with the available network resources. Thereby, the QoS of the different receivers improves considerably. Receiver 1, for example, does not only get the 64 kbits/s initial rate chosen for the base layer, but can actually benefit from around 500 kbits/s. Also, while receivers 1 and 2 suffer from losses due to the background traffic, receiver 3 which listens to all layers gets a constant data stream and is not affected by the rate shifting procedure.

6 Summary and Open Issues

In this paper, we described a QoS control scheme (ALT) that integrates rate adaptation and hierarchical data transmission. Hierarchical transmission is used to accommodate for the heterogeneity found in multicast Internet environments. With hierarchical transmission, the data is split into several streams at the data source and each stream is transmitted as a separate multicast session. There are two kinds of relationships between the information sent over the single layers: cumulative or independent. Receivers that want to receive data of a certain layer have to join the multicast group for that layer. In order to drop a layer, receivers leave the corresponding multicast group.

With ALT, receivers notify the sender of their view of the network congestion state using the RTP protocol. The sender reacts to the reported loss rates by shifting bandwidth between the layers and keeping the overall sending rate constant. Thereby, the scheme tries to provide all connected receivers on all layers with the best possible quality.

As a future task we are planning to implement the scheme into a video tool, e.g., into the NEVIT video tool [14]. For our shifting process, we have to use a layered coding mechanism where the bit rate of the single streams can be controlled.

We still need to investigate issues of scalability of the scheme and its performance under more dynamic environments with heterogeneous receivers joining and leaving the multicast sessions frequently. In this paper, we only considered receivers listening to a static number of layers. An important aspect, which we did not address yet, are the effects of adding a new layer at a receiver or leaving one on the overall distribution mechanisms. In this context, the integration of ALT and RLM needs to be closer defined and tested.

The scheme proposed here shows promising results in terms of adapting the layer sizes to the different receiver groups and improving the overall reception quality. However, we believe that there is still lots of work to be done in improving the adaptation mechanisms as well as the details of the DBA. We are currently working on testing various other adaptation schemes especially with

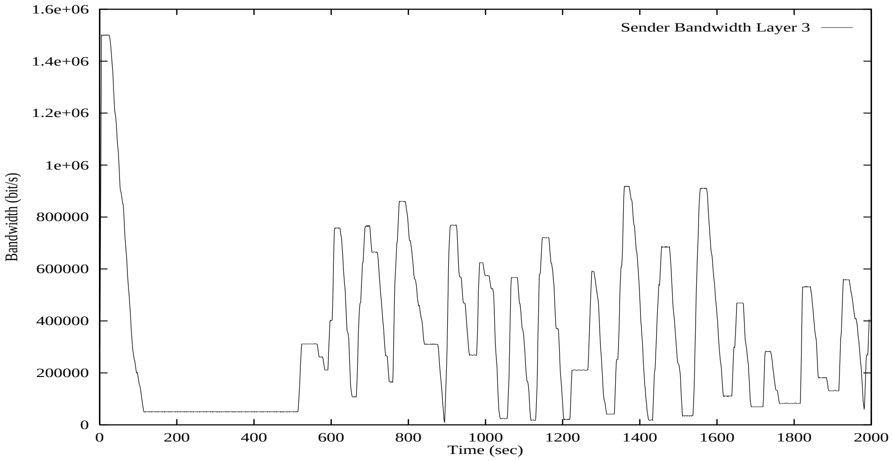

Figure 6: Sender bandwidth layer 3

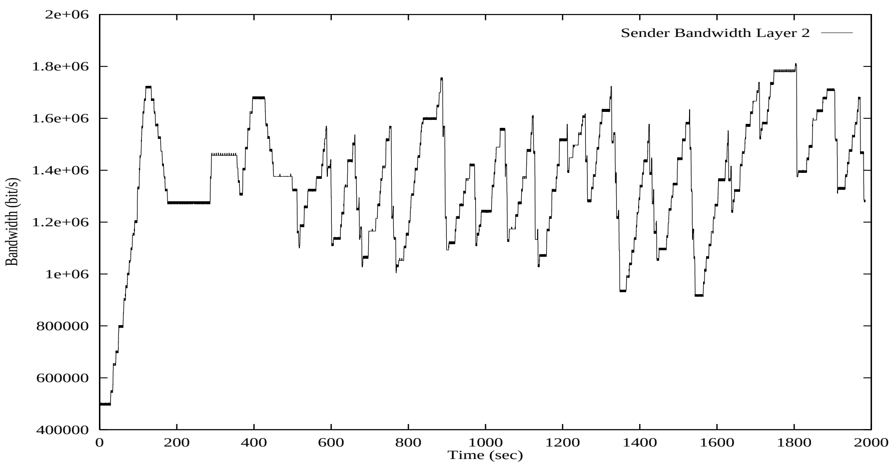

Figure 7: Sender bandwidth layer 2

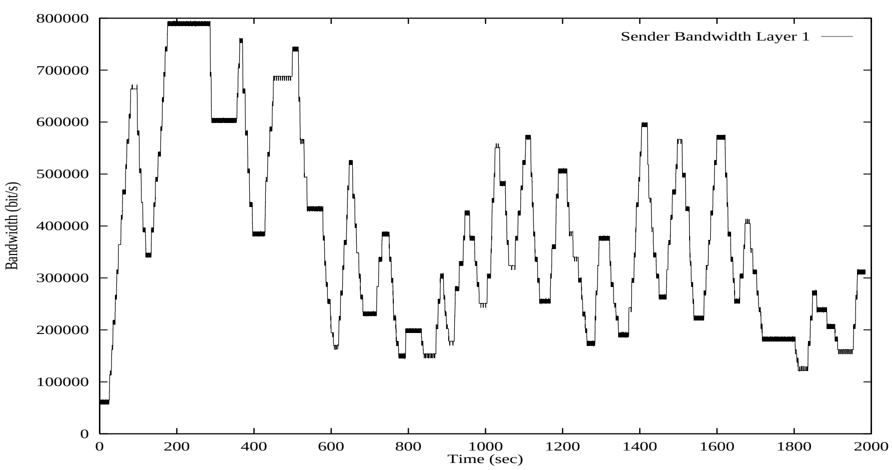

Figure 8: Sender bandwidth layer 1

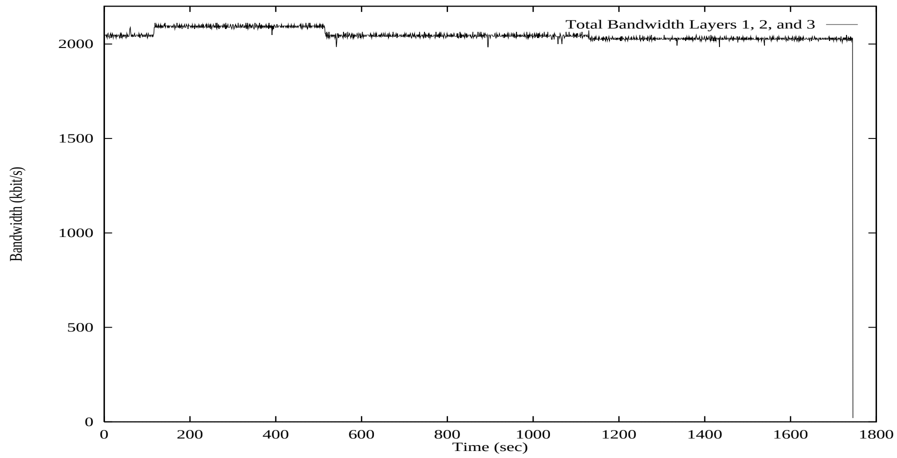

Figure 9: Total data bandwidth delivered to receivers on all 3 layers

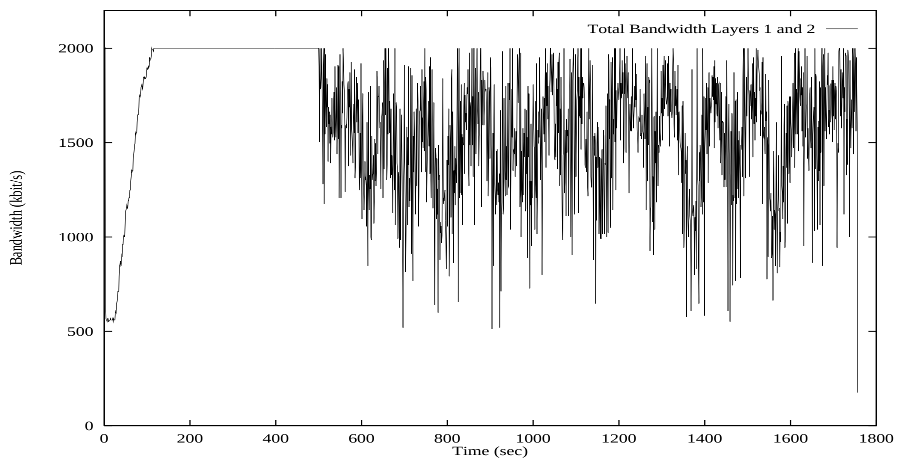

Figure 10: Total data bandwidth delivered to receivers on layers 1 & 2

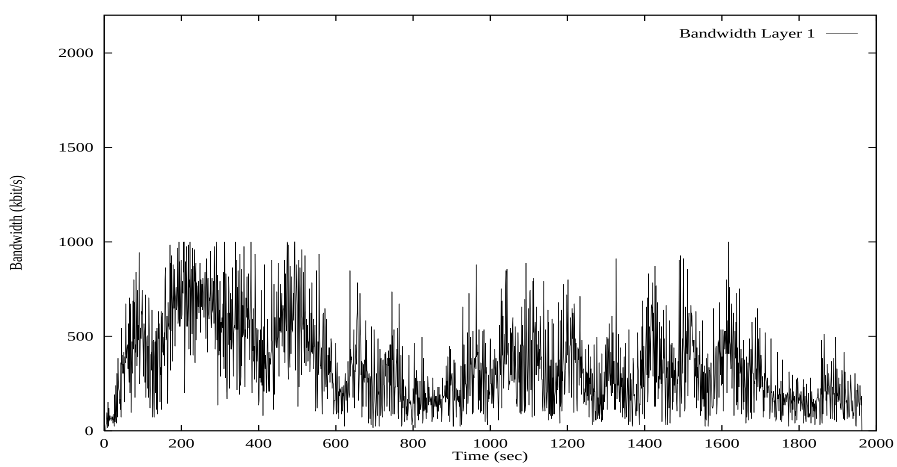

Figure 11: Data bandwidth delivered to receivers on layer 1

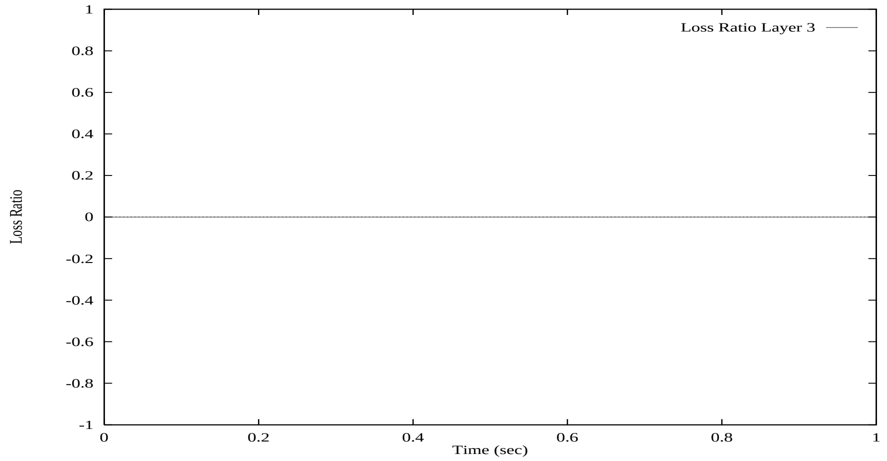

Figure 12: Loss ratio layer 3

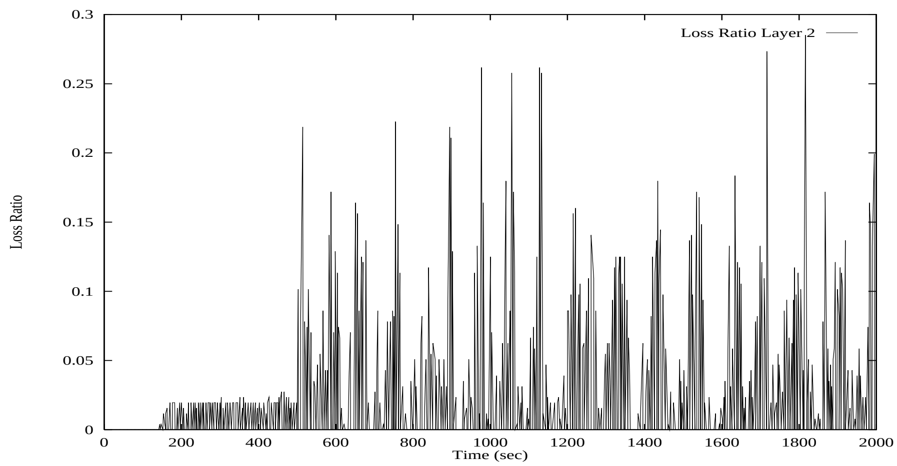

Figure 13: Loss ratio layer 2

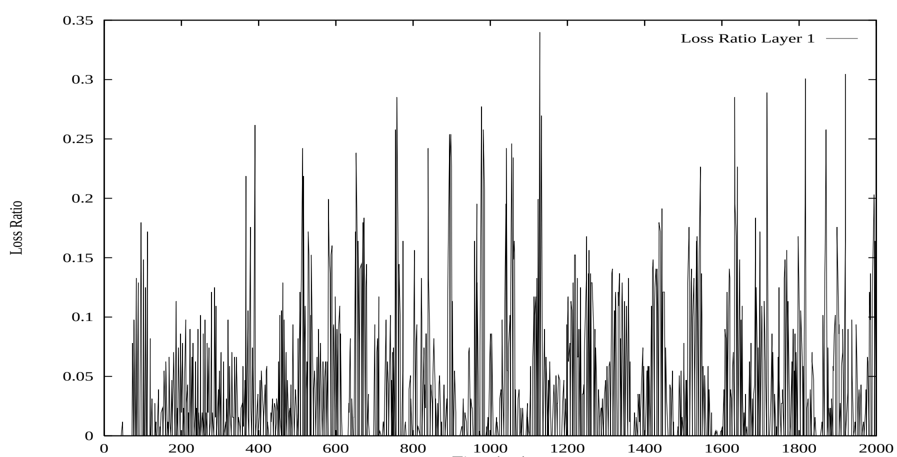

Figure 14: Loss ratio layer 1

regard to their fairness and scalability [13, 12]. Also, the details of the arbitrator need further refinement and more detailed testing.

References

[1] E. Amir, S. McCanne, and R. Katz. Receiver-driven bandwdith adaptation for light-weight sessions. In *ACM Multimedia'97*, Seattle, Nov. 1997.

[2] S. Bhattacharyya, J. T. Buck, et al. *The Almagest: A manual for Ptolemy*. Berkeley, California, 1994.

[3] J.-C. Bolot, T. Turletti, and I. Wakeman. Scalable feedback control for multicast video distribution in the internet. In *SIGCOMM Symposium on Communications Architectures and Protocols*, pages 58–67, London, England, Aug. 1994. ACM.

[4] I. Busse, B. Deffner, and H. Schulzrinne. Dynamic QoS control of multimedia applications based on RTP. *Computer Communications*, 19(1):49–58, Jan. 1996.

[5] A. Campbell, D. Hutchinson, and C. Aurrecoechea. Dynamic QoS management for scalable video flows. In *Proc. International Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV)*, Lecture Notes in Computer Science, pages 107–118, Durham, New Hampshire, Apr. 1995. Springer.

[6] D. Chiu and R. Jain. Analysis of the increase/decrease algorithms for congestion avoidance in computer networks. *Journal of Computer Networks and ISDN*, 17(1):1–14, June 1989.

[7] D. Hoffman and M. Speer. Hierarchical video distribution over internet-style networks. In *ICIP'96*, Lausanne, Switzerland, Sept. 1996.

[8] U. Horn and B. Girod. Scalable video coding for the internet,. In *8th Joint European Networking Conference*, Edinburgh, England, May 1997.

[9] X. Li and M. H. Ammar. Bandwidth control for replicated-stream multicast video. In *HPDC Focus Workshop on Multimedia and Collaborative Environments (Fifth IEEE International Symposium on High Performance Distributed Computing)*, Syracuse, New York, Aug. 1996. IEEE Computer Society.

[10] S. McCanne, V. Jacobson, and M. Vetterli. Receiver-driven layered multicast. In *SIGCOMM Symposium on Communications Architectures and Protocols*, Palo Alto, California, Aug. 1996.

[11] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson. RTP: a transport protocol for real-time applications. RFC 1889, Internet Engineering Task Force, Jan. 1996.

[12] D. Sisalem. End-to-end quality of service control using adaptive applications. In *IFIP Fifth International Workshop on Quality of Service (IWQOS '97)*, New York, May 1997.

[13] D. Sisalem, H. Schulzrinne, and F. Emanuel. The direct adjustment algorithm: A tcp-friendly adaptation scheme. Technical report, Aug. 1997. Available from http://www.fokus.gmd.de/usr/sisalem.

[14] D. Sisalem, H. Schulzrinne, and C. Sieckmeyer. The network video terminal. In *HPDC Focus Workshop on Multimedia and Collaborative Environments (Fifth IEEE International Symposium on High Performance Distributed Computing)*, Syracuse, New York, Aug. 1996. IEEE Computer Society.

[15] L. Wojnaroski. Baseline text for traffic management sub-working group. Technical Report 94-0394r5, ATM Forum, Oct. 1994.

[16] L. Zhang, S. Deering, D. Estrin, S. Shenker, and D. Zappala. RSVP: A New Resource ReSerVation Protocol. Technical report, Xerox PARC and University of Southern California, Mar. 1993.