[RSchSt02] *Ivica Rimac, Jens Schmitt, Ralf Steinmetz;* **Is Dynamic Multi-Rate Multicast Worthwhile the Effort?;** 28th Euromicro Conference 2002, Dortmund, September 2002. S.233-239.

Is Dynamic Multi-Rate Multicast Worthwhile the Effort ?

Ivica Rimac, Jens Schmitt, and Ralf Steinmetz
Multimedia Communications—KOM
Department of Electrical Engineering and Information Technology
Darmstadt University of Technology
Merckstr. 25 • D-64283 Darmstadt • Germany
Email: {ivica.rimac, jens.schmitt, ralf.steinmetz}@kom.tu-darmstadt.de

Abstract

To accommodate heterogeneous transmission conditions in a streaming scenario several multi-rate multicast solutions have been proposed, based on simulcasting or hierarchical layering. At present, most of these schemes follow a receiver-driven layered multicast approach, where the receivers join or leave a subset of the session's fixed-rate layers in response to changing network conditions. Yet, recently fine-grained coding schemes are being developed, e.g., as proposed in MPEG-4. This will permit a sender to dynamically adapt the size of the layers according to the reported transmission conditions. In this paper, we briefly discuss a general multi-rate congestion control protocol based on dynamic layering and present the basic design challenges. Since adaptation involves the sender and makes dynamic layering fundamentally more complex than its static counterpart, we compare both approaches by extensive simulations in order to explore the theoretical benefit of dynamic over static layering. The main contribution of this paper is to quantitatively describe the performance of both approaches in different scenarios by means of an inter-receiver fairness measure that captures the collective satisfaction of the session receivers.

1. Introduction

With a progressing trend towards more continuous media distribution we are facing problems with the existing Internet, where end systems are expected to adopt the "social" rules and be cooperative by reacting to congestion signals and adapting their transmission rates properly and promptly. Interacting with proactive QoS (Quality of Service) mechanisms based on explicit reservation to ensure the availability of appropriate resources might be one solution to the problem. However, even in networks that support reservation streaming applications will probably rely on network QoS realized by reservation schemes that are based on aggregated flows, due to scalability considerations. Consequently, different sessions still compete for resources as in best-effort networks which demands for reactive congestion control mechanisms. While for unicast transmission several proposals have been made and quantitatively evaluated, the development of such mechanisms for multicast transmission is challenging, since feedback implosion poses a threat on scalability, among others.

Congestion control in single-rate sessions is usually performed by the sender adjusting its sending rate according to feedback from receivers or network nodes. Corresponding protocols as those proposed in [7] and [4] typically use the feedback of the limiting receiver.

However, to accommodate the heterogeneous transmission conditions of a set of receivers in a streaming scenario multi-rate multicast is a much more desirable transmission mode. Rubenstein et al. [5] showed that in theory, multi-rate sessions can achieve several desirable fairness properties that cannot be obtained in general networks using single-rate sessions. McCanne et al. describe in [2] a receiver-driven approach where the sender transmits the data stream in multiple cumulative layers, and the receivers join or leave the static layers according to experienced congestion losses.

The conditions and distribution of possible receiver rates are usually not known in advance and are quite likely to change during a session. Hence, the sending rates in a best effort environment are hard to predefine optimally and usually are determined by coding limitations with respect to scaling. A scheme with a reasonable number of static layers can support only coarse-grained adaptation, while it might be much more reasonable, e.g., to slightly reduce the rate of a layer in order to avoid collective leave actions. Deployment of recently proposed fine-grained coding schemes, such as in MPEG-4 [3], will enable the sender to adapt the layers of a session to the dynamic conditions, and thus improve network utilization and collective receiver satisfaction.

Sisalem et al. describe in [6] a general multi-rate framework for achieving TCP-friendly congestion control in heterogeneous multicast environments. While Jiang et al. in [1] propose the use of heuristics, Yang et al. [8] introduce an algorithm to find an optimal solution to the problem. The former work layers the data into a fixed base and only one enhancement layer. The latter calculates optimal rates for a given number of layers but relies heavily on intelligence in the network for rate computation and feedback aggregation.

This paper is concerned with the possible benefit of an dynamic multi-rate multicast solution with respect to inter-receiver fairness, i.e., the collective satisfaction of the receivers of a session. The objective is to contribute to answering the question whether and when the gain of dynamic multi-rate multicast compensates for the higher implementation costs associated with the coding scheme and dynamic partitioning. So far, to the best of our knowledge this has not been adequately addressed in current literature.

In Section 2, we introduce a metric to capture collective receiver satisfaction of a multi-rate multicast session, the inter-receiver fairness. In Section 3, we give a brief overview of the issues inherent to the development of an adaptive multi-rate protocol. In Section 4 we then present the experiments conducted and interpretation of the results. Finally, in Section 5 we conclude and present possible future work items.

2. Inter-Receiver Fairness

The "satisfaction" of a receiver is represented by its utility function u_i, which is a function of the actual rate g_j of the receiver—the cumulative rate of all the multicast groups it is subscribed to—and its theoretical fair allocation r_i.[1] In the rest of the paper, we use a logarithmic receiver utility function as presented in Figure 1, which takes the optimal value of $u_{opt} = 1$ when the actual rate g_j matches the determined fair rate of the receiver r_i, and zero for $g_j > r_i$. The latter takes into consideration, that if the actual rate exceeds the fair rate, the receiver and all receivers of other flows sharing the same bottleneck link will experience loss.

The goal of a dynamic multi-rate scheme in an environment like the current Internet is to maximize collective satisfaction of the receivers of a multicast session, i.e., the sum of the utility values of the receivers in the session, while maintaining TCP-compatibility. We choose a concave, wide-sense increasing utility function to represent receiver satisfaction (see Figure 1), in order to preferably increase satisfaction of receivers with smaller utility. For comparison reason, we consider a linear utility function as well, which underlies the same restrictions regarding maximal value and range .

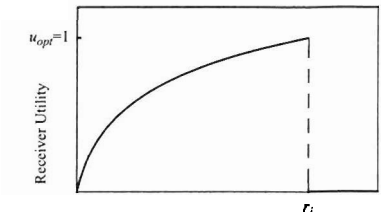

Figure 1. Receiver utility function.

To determine the optimal receiver partition we implemented and modified the algorithm originally described in [8]. For the class of receiver utility functions we consider wide-sense increasing functions as depicted in Figure 1. The utility U_j of the set of receivers G_j subscribed to the same layers is maximized when g_j equals r_i of the worst receiver of G_j. Thus, the complexity of the algorithm to compute U_j is reduced to a search over the discrete set of fair receiver rates.

In the rest of the paper, we use the following variables:

- L the number of layers (groups) in a session,
- N the number of receivers in a session,
- R_i the ith receiver,
- r_i the theoretical fair allocation for R_i ($i \in 1..N$),
- l_j the data rate of the jth layer ($j \in 1..L$),
- G_j the set of receivers subscribed to layers 1 to j,
- n_j the number of receivers in G_j,
- g_j the cumulative data rate in G_j, where

$$g_j = \sum_{k \in 1...j} l_k ,$$

- u_i the actual utility of R_i,

$$u_i = u(r_i, g_j), R_i \in G_j, \quad (2)$$

- u_{opt} the maximal utility of a receiver ($u_{opt} = 1$),
- U_j the utility of G_j, where

$$U_j = \sum_{R_i \in G_j} u_i , \quad (3)$$

1. The interested reader can find the formalized definition in [1]

U_S the session utility,

$$U_S = \sum_j^L U_j \quad (4)$$

3. Protocol Issues

With the recent development of fine-grained coding schemes [3], interest in dynamic multi-rate protocols for streaming applications is increasing. Currently, there are few solutions proposed (e.g., [1][6][8]) which unavoidably have limitations. In this section, we will briefly describe some general issues which we have to deal with when developing a protocol.

3.1 Optimal Rate Estimation

In an end-to-end approach, the task of determining the optimal rate of receivers is usually distributed. We will assume in the rest of this paper, that the receivers decide when to send feedback, which avoids loss path multiplicity and extensive sender-side computation. The former phenomenon is put down to the fact that if n receivers have an individual packet loss probability p_i (and losses are independent), the source would perceive a loss probability $p_n = 1 - (1 - p_i)^n$.

While there is still no agreement in the research community on the definition of fairness for multicast flows, TCP-compatibility of protocols is desired in the context of the current TCP-dominated Internet. This led to development of several congestion control mechanisms, among others, TFMCC as proposed in [7]. Currently, we are following the basic idea of deriving the optimal rate of a receiver from an equation modeling long-term TCP throughput. This equation-based approach has been originally designed and evaluated for unicast traffic, and recently extended to single-rate multicast. In order to extend this approach to multi-rate multicast, the following problems have to be solved, among others:

- Measuring the lost event rate.
  In the multi-rate case, data of the different layers may travel different paths which makes loss event estimation more complex than in the single-rate case. Furthermore, in the case where a non-bottle-necked receiver has less allocated resources than its allowable share, no loss might be experienced which would lead to an over-estimation of the optimal rate.
- Measuring the round trip time.
  The simplest approach for estimating the round trip time by sending a request from the sender to the receivers and having the receivers acknowledging the request right away, does not scale well for multicast communication. Thus, a method like the one presented in [6] seems promising, combining one way measurement with clock skew estimation.

3.2 Feedback Suppression

To achieve optimal partitioning, Jiang et al. propose in [1] a protocol, where the sender asks the receivers for feedback of their optimal rates. With the information of all receivers the sender computes the optimal receiver partitioning. But the computation is quite expensive and the protocol demands for receiver feedback periodically. Originally, the scalability problem was solved by feedback aggregation performed at the routers. Since we are interested in an end-to-end solution, router-support cannot be assumed, and other mechanisms have to be applied.

It is obvious that in a large multicast session mechanisms to keep feedback bounded to avoid feedback implosion are necessary for scalability reasons. Intuitively, in periods with no or little changes where utility gain is negligible and doesn't justify the cost for feedback and repartitioning, there is no need for sender action, while heavy changes should cause immediate reaction. Thus, in the approach we are currently developing, a receiver is allowed to send feedback once its utility degradation ($u_{opt} - u_i$) exceeds a certain degradation threshold:

$$\Delta u_j = u_{opt} \times (1 - \alpha_j) \quad (5)$$

To derive α_j, the following has to be considered:

1. There is solely one receiver in G_j.
   In this case, the receiver could send feedback more frequently. An increase or decrease of l_j does not effect the actual utility of any other receiver, but might increase U_S. Consequently, no other receiver might be triggered to leave the group.
2. G_j getting populated.
   The higher G_j gets populated, the lesser weights the utility of a single receiver. Consequently, keeping α_j and Δu_j independent of the population might cause more feedback of relative unimportant changes.
3. Relative effect of rate change in different layers.
   As depicted in Figure 2, the receivers in G_k are already better served than those in G_j, where $k > j$. Thus, if the difference $(r_{i1} - g_j)$ equals $(r_{i2} - g_k)$, the receiver in G_j might be allowed to send feedback while its counterpart in G_k might not, which intuitively seems reasonable.

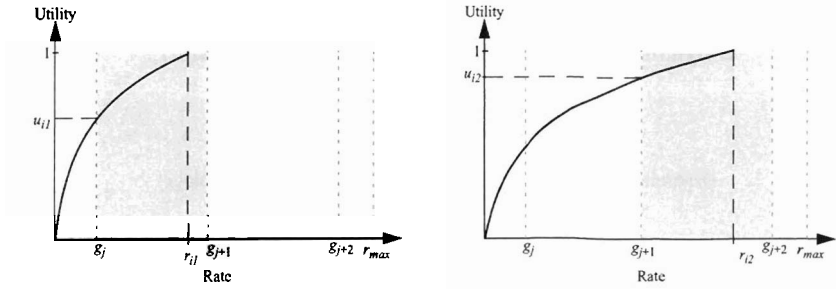

Figure 2. Effect of a rate increase in (a) a lower layer, (b) and a higher layer.

As a result, we can summarize that the more G_j is populated, the higher the utility degradation a receiver has to experience until it is allowed to send feedback. Thus, the value of α_j is a function of the number of receivers n_j in G_j, thus:

$$\alpha_j \propto n_j \quad (6)$$

An exact function for α_j has to be thoroughly determined, e.g., through extensive simulations, since it has a substantial impact on scalability and accuracy. Finally, when a large set of receivers in G_j exceeds the threshold, feedback implosion might still occur.

3.3 Avoiding Leave Action

If a receiver calculates the theoretical rate to be less than the current receiving rate, it may leave the highest layer immediately [6]. In the cases where a receiver's estimated rate is falling slightly below its current receiving rate, this might cause avoidable coarse-grained quality degradation. Adapting the layer in response to feedback might prevent some receivers to perform leave actions. We propose the following approach:

- Over a time interval T, the sender is collecting receiver feedback for each layer.
- Each receiver calculates its r_i. If $r_i < g_j$, a report is sent to the sender and the receiver waits for the next announcement of the sending rates.
- Only if the new g_j has not been lowered to accommodate a receivers reported rate, then the receiver is forced to leave the group.

For the receivers which still have to leave a group, this introduces a higher leave latency, but might keep overall satisfaction on a much higher level.

3.4 Summary

In this section, we briefly described the issues and complexity inherent to the design of an adaptive end-to-end multi-rate multicast approach and our initial ideas for a protocol. But what is the quantitative gain? When does the gain justify the additional effort? To our knowledge, these questions have not been addressed in existing work. Hence, in the next section we take a step back and investigate the impact of different receiver rate distributions on session utility.

4. Experiments

For each of the following experiments we generated 500 rates according to the corresponding distributions in each of 100 runs, and set the minimal rate r_{min} and maximal rate r_{max} to 64 kbps and 2,560 kbps, respectively. The inter-receiver fairness of a session is maximized when all receivers are served optimally, i.e., the number of layers of a session equal the number of different rates. In this case, the session utility becomes

$$U_{S_{opt}} = \sum_{R_i | i \in N} u_{opt} = N$$

To quantify the inter-receiver fairness of a session, we define the goodness of session as

$$\text{goodness of session} = \frac{U_S}{U_{S_{opt}}} \quad (8)$$

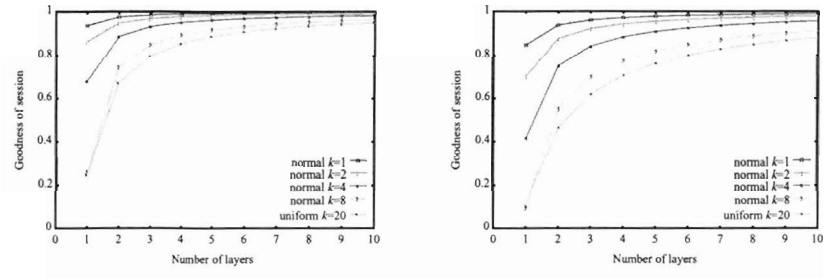

Figure 3. Impact of number of layers for varying normal distributed receiver rates. (a) Logarithmic utility function; (b) linear utility function.

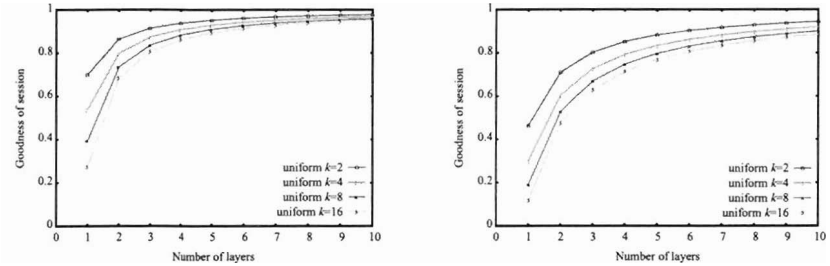

Figure 4. Impact of number of layers for varying uniform distributed receiver rates. (a) Logarithmic utility function; (b) linear utility function.

With Equation 8, the goodness metric is bound to the interval [0, 1].

Since the standard deviation in all the runs was lower than 2 percent of the average value calculated from the 100 runs, we just show the average values.

4.1 Single-Rate vs. Multi-Rate

The first question we examined is the reasonable number of layers for a session. It is obvious that when the number of layers L approaches the number of receivers N—or, more exactly, the number of receivers with different fair rates—session utility will become optimal. However, the higher the number of layers the more overhead is incurred, e.g., in multicast address allocation, routing tables, synchronization of the layers at receiver side, etc.

In the first experiment we studied the effect of increasing the number of layers for (a) a logarithmic utility function, and (b) for a linear utility function. We generated the receiver rates according to the following distributions:

1. Uniform distribution with varying range $[r_{min}, 2k \times r_{min}]$, $k \in 1..20$.
   The results of the experiment are depicted in Figure 4. It demonstrates, that for a single-rate approach with an expected distribution of $3 \times r_{min}$, the goodness of session is less than 50 percent. Even for such a narrow range overall satisfaction can be increased by approx. 20 percent (a) and 30 percent (b), respectively, by providing 3 layers instead of a single one. For the extreme case where the rates are expected to cover the range of 64 kbps up to 2.56 Mbps uniformly, the gain approaches 50 percent.
2. Normal distribution with mean μ = 1,248 kbps and varying standard deviation $\sigma = 2^k$, $k \in 1..9$.
   As Figure 3 demonstrates, for receiver rates following a normal distribution, when 97 percent of all rates are in the interval [μ–32 kbps, μ–32 kbps] the session goodness calculated for the single-rate scheme is approx. 0.7 and 0.4 for (a) and (b), respectively, while with 3 layers it increases to approx. 0.9 and 0.8, respectively.

The preceding experiment demonstrates that even with the introduction of a relatively low number of 3 layers, a remarkable gain in session goodness can be expected. Since usually the quality of inelastic data is not acceptable beneath a certain threshold, it might make sense to use one base layer and 3 enhancement layers as a reasonable trade-off between overhead regarding group management and session goodness.

4.2 Static Layers vs. Dynamic Layers

While in Section 4.1 we tried to provide simulation data for comparison of single- and multi-rate approaches, in this section we are interested in the impact of changing rate distributions during a session. The objective is to quantitatively describe the session degradation of predermined layers compared to dynamically adapting layers.

In the first experiment we generated uniformly distributed rates in the interval $[r_{min}, r_{max}]$, and calculated the rates of the 3 and 4 static layers according to the optimal partitioning algorithm. In the experiment, these values serve as the predefined rates for the static layering, and the distribution range of the rates serves as the variable, i.e., $[r_{max}-2k \times r_{min}, r_{max}]$ with $k \in 1..19$.

In the second experiment, we first assumed a trimodal distribution to determine the rates for the static layers. Then, we simulated the effect of receivers drifting from the last mode to an additional one. The rates are distributed as follows:

- 20 percent uniform distributed $[r_{min}, 2 \times r_{min}]$
- 30 percent uniform distributed $[10 \times r_{min}, 12 \times r_{min}]$

w percent uniform distributed
$[r_{max}-12 \times r_{min}, r_{max}-8 \times r_{min})$
and (50–w) percent uniform distributed
$[r_{max}-8 \times r_{min}, r_{max}]$, $w = k \times 5\%$, $k \in 0..10$

The results of the first experiment are depicted in Figure 5, which demonstrates the relative session degradation, i.e., the degradation of a static session with predefined rates compared to a session where rates are recalculated to adapt to the dynamics of the distribution. It is obvious that if the actual distribution approaches the expected distribution $[r_{min}, r_{max}]$, session degradation will be minimized.

If we consider logarithmic utility functions, a static session degrades by 15 percent for a 3-layer session, and 10 percent for a 4-layer session, as a result to halving the distribution range. In the linear case, these degradations amount to 18.5 percent and 21 percent, respectively.

The results of the second experiment are depicted in Figure 6, which demonstrates that while in the static case degradation is roughly linearly increasing, in the dynamic case it is kept almost constant, due to the adaptation to the changing conditions. In the extreme situation where all the receivers drift to the new mode, degradation reaches 19 percent when a logarithmic function is chosen to represent receiver satisfaction, and 32 percent when it is represented by linear function.

The experiments show that degradation in a session might become quite high due to unpredictive distribution of receiver rates, which advocates for dynamic layering approach. But depending on the ratio of $r_{min}/(r_{max}-r_{min})$,

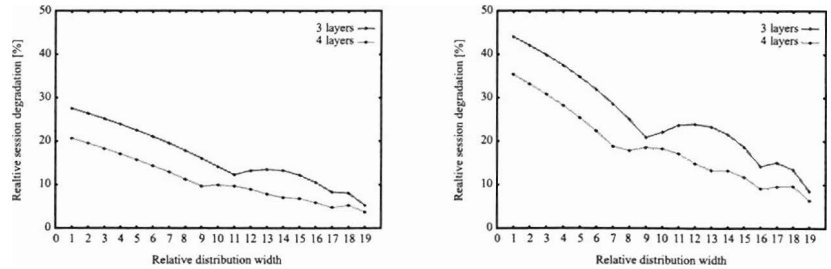

Figure 5. Comparison of a static multi-rate scheme to an adaptive one. (a) Logaritmic utility function; (b) linear utility function.

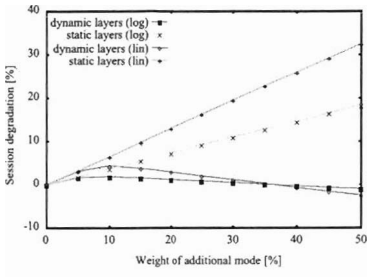

Figure 6. Impact of an additional mode.

session degradation in the static case might be acceptable when compared to protocol complexity of dynamic approaches.

5. Conclusion and Future Work

In this paper, we studied the effect of changing receiver rate distributions on the session goodness, that might theoretically be achieved in a multicast session. By means of experiments we showed that for environments where the distribution cannot be predicted or changes during the session are frequent, adaptation of the transmission rates in a multi-rate multicast mode may increase the overall satisfaction significantly with only a few layers. We also showed, that in some cases the degradation implied by a static multi-rate approach might be acceptable when compared to protocol complexity of its dynamic counterpart.

In future work, we plan to study different solutions to address the issues we identified for several problems inherent to host-based solutions for dynamic multi-rate multicast transmissions. We will further investigate and substantiate the mechanism proposed for feedback suppression, and simulations should help to determine the parameters, as well as to evaluate the short-term and transient behavior.

6. References

[1] Tianji Jiang, Ellen W. Zegura, and Mostafa Ammar. Inter-receiver fair multicast communication over the internet. In *Proccedings of the 9th International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV'99)*, Basking Ridge, NJ, June 1999.

[2] Steven McCanne, Martin Vetterli, and Van Jacobson. Receiver-driven layered multicast. In *Proceedings of ACM SIGCOMM'96*, Palo Alto, CA, August 1996.

[3] H. Radha, M. van der Schaar, and Y. Chen. The MPEG-4 fine-grained scalable video coding method for multimedia streaming over IP. *IEEE Transactions on Multimedia*, 3(1):53–68, March 2001.

[4] Luigi Rizzo. pgmcc: A tcp-friendly single-rate multicast congestion control scheme. In *Proceedings of ACM SIGCOMM 2000*, Stockholm, Sweden, August 2000.

[5] Dan Rubenstein, Jim Kurose, and Don Towsley. The impact of multicast layering on network fairness. In *Proceeding of ACM SIGCOMM'99*, Cambridge, MA, August 1999.

[6] Dorgham Sisalem and Adam Wolisz. MLDA: A TCP-friendly congestion control framework for heterogeneous multicast environments. In *Proceedings of the Eight International Workshop ob Quality of Service (IWQoS 2000)*, Pittsburgh, PA, June 2000.

[7] Jörg Widmer and Mark Handley. Extending equation-based congestion control to multicast applications. In *Proceedings of ACM SIGCOMM 2001*, SanDiego, CA, August 2001.

[8] Yang Richard Yang, Min Sik Kim, and Simon S. Lam. Optimal partitioning of multicast receivers. In *Proceeding of the 8th Conference on Network Protocols*, Osaka, Japan, November 2000.