TCP-like congestion control for layered multicast data transfer

Lorenzo Vicisano[1] Luigi Rizzo[2]
Jon Crowcroft[1]

[1]Department of Computer Science, University College, London
Gower Street, London WC1E 6BT, UK
email: {L.Vicisano,J.Crowcroft}@cs.ucl.ac.uk
[2]Dip. di Ingegneria dell'Informazione, Università di Pisa
via Diotisalvi 2 – 56126 Pisa (Italy)
tel. +39-50-568533 – fax +39-50-568
email: l.rizzo@iet.unipi.it (corresponding author)

Abstract

We present a novel congestion control algorithm suitable for use with cumulative, layered data streams in the MBone. Our algorithm behaves similarly to TCP congestion control algorithms, and shares bandwidth fairly with other instances of the protocol and with TCP flows. It is entirely receiver driven and requires no per-receiver status at the sender, in order to scale to large numbers of receivers. It relies on standard functionalities of multicast routers, and is suitable for continuous stream and reliable bulk data transfer.

In the paper we illustrate the algorithm, characterize its response to losses both analytically and by simulations, and analyse its behaviour using simulations and experiments in real networks. We also show how error recovery can be dealt with independently from congestion control by using FEC techniques, so as to provide reliable bulk data transfer.

Keywords: congestion control, multicast.

1 Introduction

The design of one-to-many data transfer protocols which run on top of the Internet multicast service have to face the fundamental problem of *congestion control*. The network is shared by many competing users, and uncontrolled, aggressive flows can bring the network to a congestion collapse. To avoid this, protocol instances should behave in such a way to achieve a fair sharing of network resources with instances of other well-behaved protocols. At the same time, in multicast protocols, effective congestion control should not come at the expense of scalability, since we wish to address scenarios comprising thousands of receivers.

The problem is becoming more and more important with the increasing diffusion of of bandwidth-intensive multimedia tools for video and audioconferencing, which stimulate the use of the network for reaching large, heterogeneous groups of users. For these applications, techniques have been proposed [17, 18] for transmitting the same stream using multiple data rates (which map into different quality levels) based on a cumulative layered data organization. By using proper data arrangements [19], a similar approach can be used for the reliable transfer of bulk data.

Multicast routing provides the basic mechanism for distributing data with different rates to subsets of the multicast distribution trees. By joining/leaving multicast groups, receivers in a subtree can in principle adapt the data rate to the available network capacity. But this approach can only work if receivers are able to evaluate and adapt to the network capacity, act in coordination, and if the algorithms that are used can scale to large groups of receivers. The requirement of being a fair competitor with other well-behaved protocols such as TCP further complicates the problem of devising a suitable algorithm for letting receivers dynamically choose the appropriate data rate.

In this paper we will describe a receiver driven congestion control algorithm which is TCP-friendly and suitable for use in the Mbone. We support different bandwidths by using a layered organization of data, and letting receivers adapt to the available bandwidth by joining to one or more multicast groups. Each receiver takes decisions autonomously, but techniques are used to synchronise receivers behind the same bottleneck (and belonging to the same protocol instance), so that an effective use of the bandwidth is achieved. No explicit forms of group membership are used, so that the algorithm is fully scalable and simple to implement, both at the sender and the receiver side.

Previous works on this subject [2, 7, 11] have already used layered data organization to implement some form of multicast congestion control. The novel results of our work are the development of a congestion control algorithm which competes fairly with TCP flows, and the achievement of receiver synchronisation without recurring to the use of explicit group membership protocols. Other significant contributions of the paper lie in the evaluation of the algorithm and its interaction with TCP using analytical techniques, simulations, and experiments in real networks.

The paper is structured as follows. In Section 2 we briefly describe the basic mechanisms our congestion control algorithm is based upon, namely the relation between throughput and loss rate, multicast group membership in the Mbone, and layered data organization. Section 3 describes in detail our congestion control algorithm, while in Section 4 we analyse its behaviour. First, using a simplified steady state model, we show that the relation between loss rate and throughput is similar to that of TCP. Simulations are then used to validate the model, and, together with experiments on real networks, to show that the algorithm has the desired features in terms of receiver aggregation, adaptivity to varying network conditions, and intra- and inter-protocol fairness. Finally, Section 5 briefly discusses the application of the algorithm to continuous streams, and then shows techniques to achieve reliable bulk data transfer while preserving scalability and optimizing the usage of the network.

2 Basic mechanisms

In this Section we briefly review the three basic mechanism our congestion control algorithm is based upon.

2.1 Relation between throughput and loss rate

IP networks do not have an explicit congestion notification mechanism. The response to congestion is dropping packets, and the loss of one or more packets at the receiver is interpreted as a congestion signal. A congestion control algorithm must react to congestion signals by reducing the transmission rate. The relation between the packet loss rate, p, and the throughput, T, of a well behaved session is such that increasing values of p correspond to decreasing values of T. For the congestion control algorithms used in TCP [8, 16], various authors have derived the following approximate relation between throughput and loss rate:

$$T = \frac{s \cdot c}{RTT\sqrt{p}} \qquad (1)$$

where s is the packet size, RTT is the round trip time for the session, and the constant c is in the range 0.9...1.5 [6, 10, 13]. The relation holds for small values of p; at high loss rates, TCP tends to lose its ACK-based self clock, the session assumes a chaotic behaviour, and the actual throughput has a sharp decrease and becomes highly dependent on the length of the retransmission timeouts.

Equation 1 does not capture completely the behaviour of the control system. Another important parameter of any control algorithm is the speed of response to variations in the controlled system – in our case, in network conditions. For TCP congestion control, this response time is in the order of the RTT of the session, which is the time required for the sender to notice a congestion signal and react (by either shrinking or inflating the congestion window).

Equation 1 shows that, in presence of competing receivers behind the same bottleneck (thus experiencing a similar loss rate), the actual share of the bandwidth for each of them strongly depends on s and RTT. Apart from this variance, competing TCP sessions all reduce their throughput similarly in response to losses. Fairness among different protocols requires that all of them have a similar behaviour in response to losses, with comparable functional dependencies, absolute throughput values, and response times.

2.2 Multicast group membership

Forwarding of multicast data in the Mbone is implemented by multicast routers, which communicate with hosts using a dedicated group membership protocol, IGMP [4, 5]. Packets for a given multicast group are forwarded into a network segment only if the segment leads to active receivers for that group (be them end nodes or other multicast routers serving active receivers). In the *join* phase, a receiver joining a new group sends an IGMP message to the router, which will in turn enable forwarding for that group, possibly propagating the information to the upstream router (*graft* message). The join phase is fast, taking roughly one RTT (computed between the joining node and the first router towards the source which is blocking the group). In the *leave* phase, a receiver informs its local router that it is not interested in the group anymore. This triggers a polling phase in which the multicast router checks if other receivers in the local subnet are still interested in the group. If no active receivers exist, the router sends a *prune* message for the group to the upstream multicast router, so that forwarding for that group is blocked. Since the absence of active receivers can only be determined after a timeout, the leave phase can take a considerable time (a few seconds) which we call the *leave delay.*

By joining and leaving a group, a set of receivers acting in coordination can exercise a simple on/off control the flow data in a subtree, although the minimum duration of the on period is constrained by the leave delay.

2.3 Layered organisation of data

Our congestion control protocol is aimed to the distribution of the same data (possibly, with different quality) to a set of receivers with different, increasing bandwidths B_i's, $i = 0 \ldots l-1$. The use of l different multicast groups, each with bandwidth B_i, would make a very poor use of the network capacity. The problem can be solved in a more efficient way by using a cumulative, layered data organisation (*layered data organisation* for brevity) based on l multicast groups or *layers*. Each layer carries data with bandwidth L_i, in such a way that

$$B_i = \bigcup_{j=0}^{j=i} L_j$$

Each receiver can tune its receive bandwidth by joining the appropriate number of layers, and there is no duplication of data if receivers with different requirements share the same links.

A layered data organisation is only possible if the data to be transferred supports it, i.e. $\forall i$, the data carried with bandwidth B_i is a subset of the data carried with bandwidth B_{i+1}. This is usually feasible for multimedia data streams; for bulk data such as a file, suitable data arrangements which minimize replications of data are shown in [19]. In both cases, however, the problem of finding a suitable data organisation becomes more and more complex as the number of layers grows. This, and scalability reasons, suggest that only a small number of values for B_i are made available from the transmitter, e.g. exponentially spaced; receivers can then choose the *subscription level* which best matches the available bandwidth.

3 Congestion control for multicast layered data

Our congestion control mechanism is designed for a transmitter sending data to many receivers in the Mbone. The communication might involve a continuous stream (e.g. video or audio data), or a reliable, bulk data transfer (e.g. a file). In this Section we will ignore the fact that lost packets will need to be retransmitted in order to complete a reliable file transfer; this problem will be tackled in Section 5.

In unicast communications, the sender takes part to congestion control by changing its sending rate according to the congestion signals that it receives. In multicast communications, this approach would be problematic, since different groups of receivers with different requirements may exist, and adapting to the needs of one set of receivers would penalize others with contrasting requirements.

As in other proposals appeared in the literature [11], we achieve the desired effect by using the multicast router driving the bottleneck to modulate the sending rate for the subtree. This is possible thanks to the layered data organisation, which gives receivers the possibility to modulate the receive rate by joining/leaving layers. Adaptation to heterogeneous requirements becomes possible (and simple) because it can be done independently on each subtree served by a multicast router. As a consequence of this approach, the transmitter does not even need to take part in congestion control (except for using a layered data organisation).

In our algorithm, receivers try to infer the status of the network using congestion signals (typically, packet losses), and control the flow of incoming data using the pruning capability of IP multicast routing. Each receiver joins or leaves layers depending on the received congestion signals. The strategy to join or leave layers is chosen in such a way to emulate the behaviour of TCP and generate a similar relation between throughput and loss rate. Intuitively, a receiver with subscription level i will behave as follows:

> if a loss is experienced, decrease the subscription level; otherwise, if no losses are experienced for a time $t_p(i)$, increase the subscription level.

The above rule mimics the behaviour of TCP. With a proper choice of the bandwidth B_i corresponding to each subscription level, and the delay $t_p(i)$ before moving to the next level, the above algorithm yields a relation between throughput and loss rate similar to that of TCP, namely $T \propto 1/\sqrt{p}$. Parameters can also be tuned to yield a response time comparable to that of TCP congestion control.

Unfortunately, we also have to deal with some additional problems, not present in unicast communications, for which the simple algorithm presented above is not adequate. First, congestion control cannot be effective if receivers behind the same router act in uncoordinated way. Second, the length of the leave delay makes *failed* joint attempts (i.e. joins to a layer which cause the bandwidth to exceed the bottleneck bandwidth) have long lasting effects. Both phenomena can severely compromise the effectiveness of the algorithm, so we need to develop appropriate countermeasures.

3.1 Synchronisation points

With more than one receiver behind a bottleneck, synchronization among receivers is fundamental in order to make the pruning mechanism work. In fact:

(a) the action of a receiver dropping a layer (leaving a group) has no effect unless all the receivers sharing the bottleneck drop that layer;

(b) if a receiver causes congestion on the bottleneck by adding a new layer, another receiver might interpret the resulting losses as a consequence of its too high level of subscription and and drop a layer;

(c) if two receivers behind the same bottleneck have different subscription levels, the bottleneck bandwidth allocated to that protocol instance is not fully exploited;

All these problems can be minimised if join attempts from different receivers are coordinated. Unfortunately, receivers tend to diverge because of different start times, propagation delays, loss patterns, and the history of events experienced by each of them. We have avoided the introduction of an explicit protocol for receiver coordination, since it could pose scalability problems, and could be slow to converge. Rather, we have used an implicit coordination protocol based on the use of *synchronisation points* (SP's).

SP's correspond to specially flagged packets in the data stream. A receiver can only make a join attempt immediately after an SP, and can only base its decisions on the history of events since the last SP. Both strategies help in keeping receivers in synch, especially for nearby nodes which will be less affected by skews in propagation delays and losses caused by background traffic. The deletion of history across SP's serves to avoid that random events experienced by each receiver accumulate their effects and cause receivers to diverge.

The distance between SP's on each layer is related to the time $t_p(i)$ that a receiver must spend at a given level before doing a join attempt. The phenomenon illustrated in point *(b)* above suggests that when a receiver at level i does a join attempt, all receivers at the same or lower levels might experience the losses related to a failure. To overcome this problem, SP's at each level are always a subset of the SP's at the previous level. This has two implications. First, when a receiver does a join attempt, all receivers at the same or lower levels will do an attempt as well; as a consequence, any congestion signal will not cause further divergent behaviours. Second, receiver with a lower subscription level have more chances to increase the level and aggregate to other receivers at a higher level.

3.2 Sender-initiated probes

The problems described in the previous section are further aggravated by the length of the leave delay, which makes a join attempt (and especially, a failed one) have consequences on the network for a long time. Join attempts (and consequently, SP's) should then be sufficiently far apart to let the network settle in between. Furthermore, as an optimisation of the algorithm, we would like to be able to increase the subscription level only when there is confidence that the attempt can be successful.

Our algorithm involves a form of sender-initiated probe for bandwidth, consisting in the periodic generation of short bursts of packets, followed by an equally long relaxation period during which no packets are sent. The bursts have the effect of a join attempt. For the duration of the burst, the bandwidth is effectively doubled; if the bottleneck bandwidth is not sufficient, queues build up and possibly cause packet losses, or an increase in the packet interarrival times.

Such congestion signals are not interpreted as a signal to lower the subscription level, but rather as hints for *not* increasing the subscription level. The advantage of these sender-initiated probes over real join attempts is that the surge of packets that would follow the failed join attempt will not last for the whole leave delay, but just for the (much shorter) duration of the burst, and recovery will be possible during the subsequent relaxation period. We note that sender initiated probes also mimic the behaviour of TCP sessions, which, in a number of occasions (e.g. when opening windows, or receiving an ACK for more than one MSS) transmit two or more back to back packets.

3.3 The deaf period

The leave phase takes a long time to complete, even when all receivers in a subtree drop the layer which caused congestion. A naive application of the algorithm presented in Section 3 would probably result in decreasing the subscription level multiple times in response to a failed join attempt. To overcome this problem, we have introduced a *deaf period* t_D . After a loss, and the subsequent decrease in the subscription level, a receiver does not react to further losses for a time t_D, in order to account for the slow response of the multicast router to leave requests. The deaf period is set slightly larger than the leave delay, so as to leave the multicast router sufficient time to respond to the leave request.

3.4 The congestion control algorithm

In this Section we provide the full description of our congestion control algorithm, with the addition of the mechanisms discussed so far.

We assume that l different bandwidths are offered, $B_i, i = 0 \ldots l-1$, with increasing values (e.g. exponentially spaced by a factor of 2). Data is transmitted over l multicast groups with bandwidths L_i, where $L_0 = B_0$ and $L_i = B_i - B_{i-1}, i = 1 \ldots l-1$.

For each bandwidth B_i packets are normally transmitted at a constant[1] rate, appropriately placed onto the different layers; we call τ_i the inter-packet time at each layer i.

During bursts, which have a duration τ_0, two back-to-back packets are sent at each transmission. Following the burst, there is an interval τ_0 during which the transmission is suspended. Finally, an interval $(W-2)\tau_0$ follows during which transmissions occur at the nominal rate.

Synchronisation points are spaced proportionally to the bandwidth corresponding to the subscription level, and are located at the end of a burst (which corresponds to the last packet

[1] depending on the actual values of B_i, this might only apply to the average rate

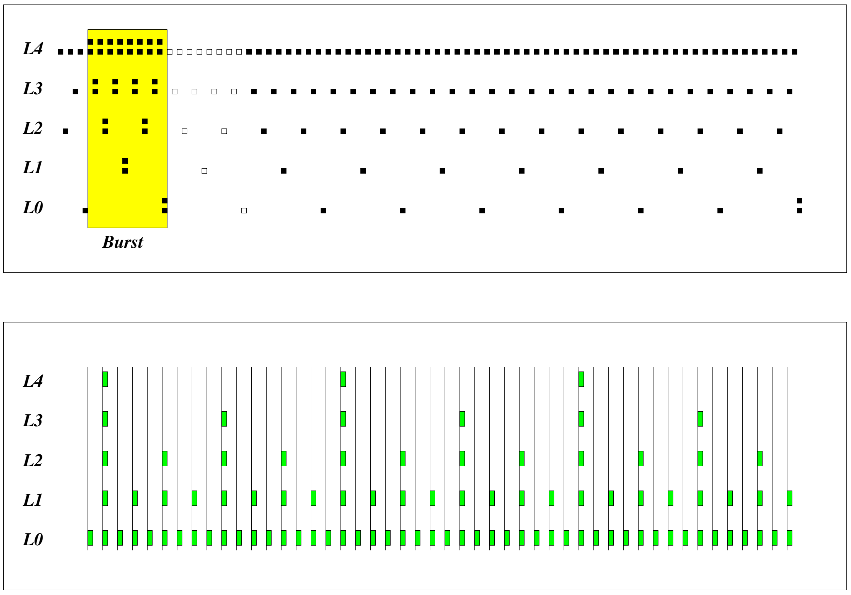

Figure 1: Above: the sequence of packet transmissions; below: the location of synchronisation points at the various layers. $P = 1, W = 8$ and five layers.

---

of the burst in layer L_0). As a reference, Figure 1 shows the sequence of packet transmissions for $P = 1, W = 8$ and five layers, and the location of synchronisation points.

The distance between SP's at subscription level i is $(B_i/B_0)PW\tau_0$. The constants τ_0, P and W appear in the relation between throughput and loss rate, and will be discussed in Section 4.2.

Subscription levels can increase only at SP's, and decrease at any time, using the following rules:

- *decrease* if a loss is experienced during normal transmissions (except for a time t_D after a previous decrease);
- *increase* at a SP if no losses are experienced during the burst preceding that SP;
- *unchanged* otherwise (this includes the case of losses experienced during a burst, or during the deaf period).

The deaf period t_D after a decrease serves to avoid cascaded losses while the leave can effectively complete.

It should be noted that the protocol does not depend heavily on the bandwidths being exponentially distributed, or on the use of probes (although probes do show a performance

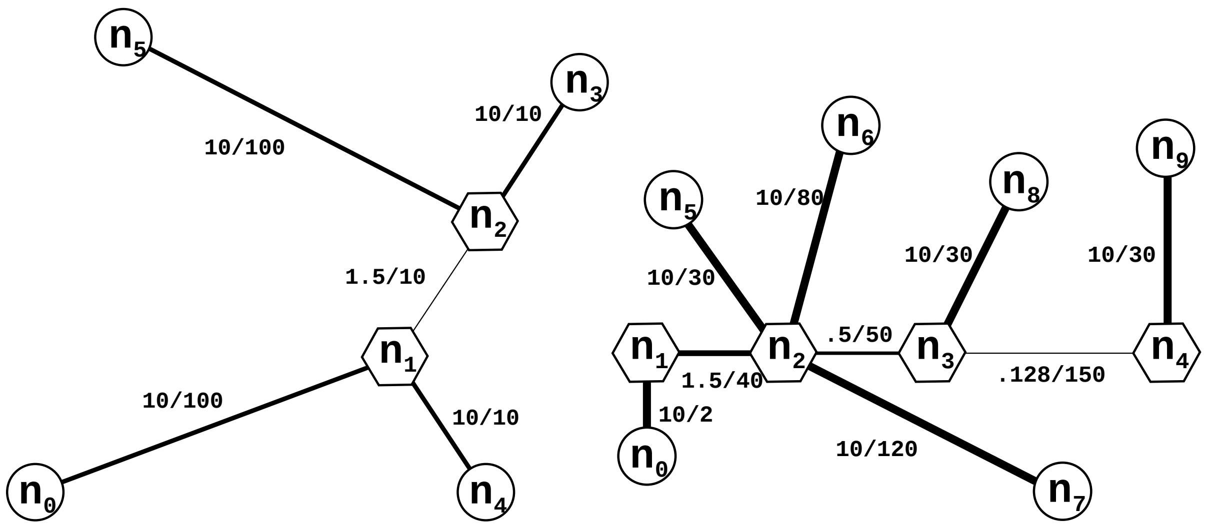

Figure 2: The network topologies used in the simulations.

improvement in our simulations). The main, and novel, elements of the algorithm are the use of synchronisation points and the rules for changing subscription levels.

4 Performance analysis

In this section we evaluate the performance of our congestion control algorithm. We first derive an analytical relation between throughput and loss rate basing on an approximate steady state model. The analytical model is then validated through simulations and measurement on the Internet, which show a very close match to the analytical value. Subsequently, a number of experiments are conducted to determine the behaviour of the algorithm in different conditions. We investigate the start up behaviour, steady state behaviour, adaptation to changing network conditions, fairness among protocol instances and with TCP instances. Measurements on the real network have been performed using a prototype implementation of the algorithm, while simulations have been done using the `ns` simulator [12], both using tail-drop and RED routers. The network topologies used in the simulations are shown in Figure 2, with transmitters and receivers spread on the various nodes. Experiments (both simulation and real measurements) generally include multiple instances of our multicast congestion control algorithms, possibly competing with TCP protocol instances, and background traffic. The protocol we evaluated used exponentially distributed bandwidths and parameters $P = 1, W = 8$.

4.1 Throughput versus loss rate

The relation between throughput and loss rate can be derived using an approximate steady state model of the system, in which a receiver oscillates among two subscription levels, i and $i+1$, which surround the "correct" bandwidth. In the approximate model, the throughput depends on the fraction α of time which is spent at level $i+1$, which in turn depends on the loss rate experienced by the connection. In Appendix A we derive the following approximation for the throughput:

$$T \approx \frac{s}{\tau_0} \frac{c'}{\sqrt{p}\sqrt{PW}} \tag{2}$$

where W is the number of packets between two bursts, P is the distance (in bursts) between synchronisation points at the lower level, and the parameter $1 \leq c' \leq 2$ is a function of α. The parameter p represents the burst loss rate, computed accounting closely spaced (i.e. closer than t_D) losses as a single one. Note that $s/\tau_0 = B_0$, but we have left s and τ_0 in the equation to remark its similarity to the same relation for TCP. Apart for a multiplying constant, Equations 1 and 2 reflect the same behaviour in response to losses.

4.2 Parameter tuning

Our algorithm uses several parameters (P, W, B_0) whose values influence the actual response (both steady state and dynamic) to losses. Our goal is to tune parameters to achieve the same response as TCP.

TCP sessions have a wide range of RTT's, with common values for non-local communications in the 0.1-1 s range. For the reasons we will discuss in the following paragraphs, it is difficult for us to achieve very short response times, so we can only aim for the high end of the range. i.e. a response time of about 1 s.

W is the distance between bursts, measured in packets transmitted at the lowest level. W cannot be less than 2, but in practice it needs to be larger in order to make bursts sufficiently far apart. On the other hand, W should not be too large or it would make synchronisation points too distant, hence making the algorithm slower to converge to the correct operating point. In our experiment we have used $W = 8$

P represents the distance between synchronisation points at the lowest level, measured in bursts (the same distance, measured in packets, is PW). The contrasting requirements of making W large and synchronisation points not too distant almost unavoidably forces us to use $P = 1$, unless we work with very high bandwidths.

$B_0 = s/\tau_0$ is the lowest bandwidth that we can use on our session. There are some practical lower bounds on its acceptable values. The first one comes from the application, in that some data streams (e.g. audio or video) cannot be transferred with an acceptable quality below a certain bandwidth. A second lower bound is related to the speed of response of the congestion control protocol. Since the distance between synchronisation points is at least PW packets, lowering the bandwidth also means increasing the distance between synchronisation points. Reducing the packet size is only possible to a certain extent, above which the overhead for packet headers becomes unacceptably large.

In our experiments, we have used a packet size $s = 256$ bytes. Our target of 1 second between synchronisation points at the lowest subscription level, and the minimum values for P and W of 1 and 8 respectively, translate into a minimum bandwidth of 2 KB/s.

4.3 Differences with TCP

Although our algorithm tries to mimic the behaviour of TCP congestion control, and Equations 1 and 2 only differ by a multiplying constant, there are some differences between the behavior of the two algorithms.

First of all, with our algorithm, the relative throughputs for competing sessions experiencing the same loss rate does not depend on the "distance" (RTT) between the sender and the receiver. In this respect, TCP congestion control has what can appear as a degree of unfairness, in that the throughput is inversely proportional to the RTT. In fact, both approaches are appropriate having different targets:

- In TCP, congestion control is completely controlled by the sender, and the only way the receiver has to reduce throughput is to delay the generation of acknowledgements, making the RTT appear larger than it really is. Furthermore, connections with a larger RTT are less responsive than those with a shorter RTT. For both reasons, it is more than acceptable that bandwidth is not allocated evenly to sessions with a different RTT, and this should not be seen as a sign of unfairness.

- Our congestion control algorithm lets receivers control throughput via the join/leave mechanism, so there is no need for an additional mechanism such as the dependency on the RTT. Besides, the response time of the algorithm does not depend on the RTT but rather on (usually longer) parameters such as τ_0 and the leave delay, so it is reasonable that all protocol instances get the same share of bandwidth irrespective of the RTT.

A second difference is that our protocol works by controlling the data rate, while TCP congestion control is based on the 'conservation of packets' principle [8]: a new packet is not put into the network until an old packet leaves; this allows TCP to fine-tune its traffic generation and to react quickly in the face of short-term network load changes. On the other hand it heavily relies on the acknowledgements to work, making TCP break at high loss rate, as discussed below.

Another important difference is that our algorithm has a limited set of throughputs allowed (exponentially distributed), while TCP is able to adjust its throughput which much finer granularity. This has two major consequences: one is that we obtain a worse bandwidth utilization, the other is that the minimum allowed throughput can be higher than what the network can sustain; this happens in presence of high losses. Both Eq. 1 and 2 are only valid at low loss levels, corresponding to a TCP session able to recover through fast retransmit, and our algorithm used with a bandwidth $B \geq B_0$. If these conditions do not hold, the source becomes not controllable, since we do not have a way to reduce the sending rate to an acceptable level.

The effect of high losses on TCP is that a connection tends to lose its ACK-based self clock, resulting in timeouts and consequent hiccups in the flow of data. The actual throughput quickly decreases (perhaps by an order of magnitude or more) and becomes heavily dependent on the value of the retransmission timeouts, rather than on the actual loss rate. In the same conditions, our algorithm behaves differently since the sender continues to transmit at a fixed rate, the congestion control mechanism ceases to be effective, and the actual throughput becomes dependent on the number of uncontrolled sources and the queue management policy at the router.

It could be possible, in principle, to implement a behaviour similar to TCP even in our algorithm, so that when excessive losses are experienced at subscription level 0, a receiver disconnects even from layer 0 to resume reception at a later time. Of course, all receivers behind the same bottleneck should act synchronously, for the reasons explained in Section 3.1. However, we believe such a behaviour to be implemented more properly at the application level. In fact, the time constants involved in this mechanism must be extremely long (several seconds at the very best). This can make the resulting quality unacceptable for some applications (e.g. audio or video) suggesting a permanent, rather than temporary, disconnection. Furthermore, experience with the use of Web browser on congested networks suggests that the user will always try to override timeouts when they become too long (e.g. pushing the 'Reload' button, or opening new connections), thus defeating the mechanism.

As a consequence, we believe that the proper approach to deal with high losses lies in the use of appropriately low values of B_0, and in the use of queue management mechanisms in the routers to deal with uncontrollable flows [6].

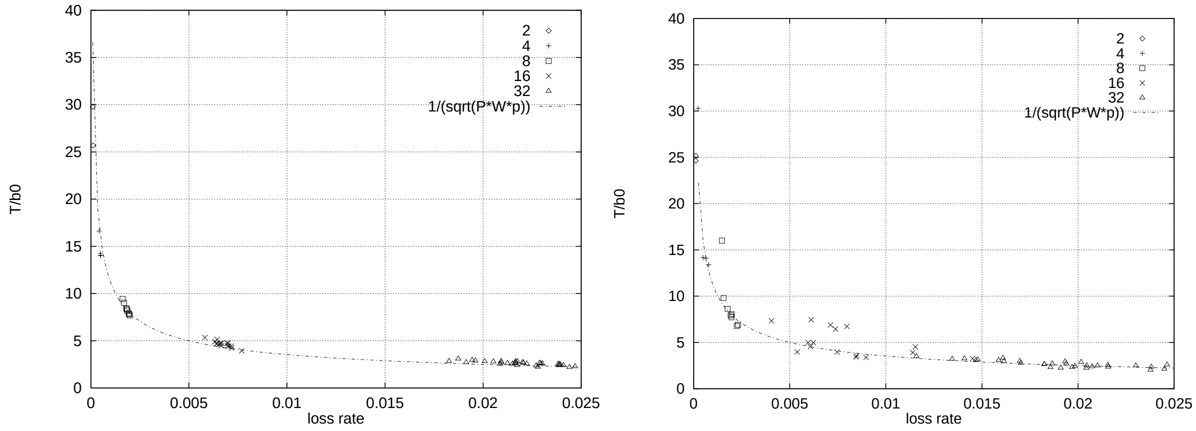

Figure 3: The relation between throughput and (burst) loss rate in simulations, compared to the analytical results from Eq. 2 (solid line). Left: RED routers; right: tail-drop routers.

4.4 Validation of the model – inter-protocol fairness

To validate our analytical model, we have both simulated the behaviour of several instances of the protocol sharing a bottleneck link, and run an implementation of our protocol on a real Internet path.

For the simulations we have used the topology of Figure 2,a with both RED and tail-drop routers. Routers' parameters setting was the default with `ns` (queue size of 50 packets and, for RED routers, thresholds of 5 and 15 packets and queue weight of 0.002). In this experiment, we spread N pairs of sender/receiver on the two paths with different delays (n_0—n_5 and n_4—n_3). Senders were started at random times over an interval of $N/2$ seconds, and receivers had been kept active for 300 seconds.

In Figure 3 we compare Eq. 2 with simulation results: as we can see there is a very close match between the model and simulations, in particular when using RED routers.

A similar experiment has been done using an experimental implementation of the algorithm on a real Internet path that connect two subnets, one in UCL (London) and the other in DEIT (Pisa). The path crosses the 'TEN-34' network[2], the Italian research network (GARR) and the English one (JANET). It is characterized by an unloaded average RTT of about 70 ms and a loaded RTT of about 300 ms; it traverses 15 nodes in both directions, the route having been

[2]TEN-34 is a high speed pan-European interconnect facility between the national research networks. TEN-34's connection between UK and Italy is made of two links (IT-DE and DE-UK) with 20Mb/s and 22Mb/s respectively, for further information see `http://www.dante.net/ten-34.html` .

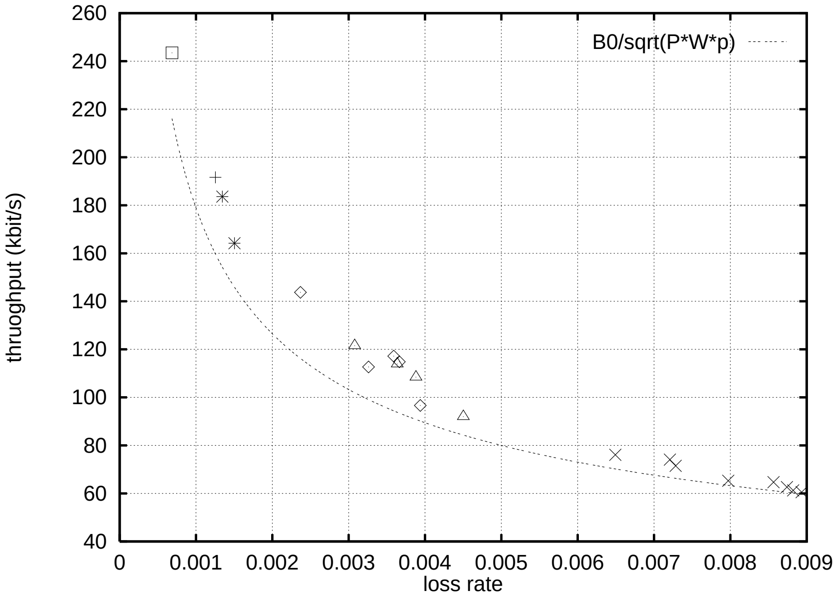

Figure 4: The relation between throughput and (burst) loss rate in real measurement, compared to the analytical results from Eq. 2 (solid line). Protocol instances run concurrently are represented by the same symbol.

---

stable for the whole duration of the experiment. The maximum throughput experienced on the path is a little greater than 1Mb/s, probably due to a 2Mb/s link in the Italian part of the path. RTT and nodes traversed have been monitored periodically using `ping` and `traceroute` programs respectively. For the experiments we have setup a multicast tunnel (without throughput limitation) between two multicast routers feeding the two subnets. We have run senders on four UNIX machines in the subnet at one site, while receivers have been run on two machine at the other site. The protocol parameters are the same as those used in the simulations. We have run from one to eight concurrent instances of the protocol, sharing the links with the normal Internet traffic —although it was quite low due to the non busy time chosen for the experiments.

Figure 4 shows the experiment results, which are in accord with the model prediction and with simulations. The actual throughput, slightly higher than the analytical value, can be motivated by the (probable) use of tail-drop routers in the network, which cause more bursty losses.

The experiments presented above have also been used to evaluate the fairness in sharing bottleneck bandwidth among several instances of the same protocol. Both Figure 3 and Figure 4 show that protocol instances running concurrently present a similar throughput. To evaluate

protocol instances	2	4	8	16	32
RED					
fairness index	0.998	0.978	0.992	0.993	0.994
router loss rate	0.0054	0.0125	0.0250	0.0354	0.0737
perceived loss rate	0.0030	0.0070	0.0143	0.0231	0.0495
burst loss rate	0.000077	0.000510	0.001752	0.007092	0.021459
bandwidth exploitation	0.563	0.624	0.696	0.776	0.907
tail-drop					
fairness index	0.986	0.779	0.876	0.897	0.978
router loss rate	0.0034	0.0180	0.0283	0.0486	0.0825
perceived loss rate	0.0029	0.0117	0.0135	0.0267	0.0504
burst loss rate	0.000150	0.001458	0.001894	0.005198	0.018356
bandwidth exploitation	0.579	0.773	0.766	0.857	0.942

Table 1: Various Protocol Instances sharing the same bottleneck: Fairness Index, Loss Rate at the router, Perceived Loss Rate, burst Loss Rate and Bandwidth Exploitation. Average values on all the instances. Top: RED router, bottom: Drop-Tail router.

better this feature, we have computed the fairness index[3] [3] from the results of the simulations.

Table 1 shows fi and some other parameters computed for 2, 4, 8, 16 and 32 protocol instances. From the table we can see that our protocol achieves a good fairness both with RED routers and with tail-drop routers, although, with the latter, we measured an appreciably worse fairness with 4 and 8 protocol instances.

Table 1 also shows three different loss rate indexes: the loss rate seen at the router, the loss rate seen at receivers, and the burst loss rate. The first two values differ in that a receiver, once a layer is left, does not consider packets (and losses) from that layer anymore. The last value is computed by considering closely spaced losses (namely less than t_D seconds apart) as a single loss.

The last parameter shown in Table 1 is the average bandwidth utilisation of the aggregate flow. The coarse quantisation of allowed throughputs sometimes leads to a non optimal utilisation of the available bandwidth. This is more evident with fewer protocol instances which stabilise on high throughput levels, characterised by larger rate steps between one level and the adjacent ones.

4.5 Fairness with TCP

A second set of simulations and experiments has been done to investigate the behaviour of our protocol when competing with other TCP instances. Using the simulator, we have run 8 multicast protocol instances as in the previous experiments, together with 8 TCP connections on the path n_0—n_5 characterised by a 420 ms RTT. Figure 5 shows the aggregate throughput of TCP connections and that of our protocol instances. The graph on the left refers to TCP with 256 bytes packet size, while the one on the right refers to 1024 bytes packet size. According to Eq 1 and 2, TCP with 256 bytes packet size and 420ms RTT should be as aggressive as our protocol (with the usual parameter setting). However Figure 5 shows that our multicast protocol is slightly more aggressive; that is due to the fact that our algorithm tends to aggregate burst losses more than TCP does; in other words, given the same loss rate, our protocol sees a lower congestion-event rate than TCP does.

Figure 6 shows the same experiment performed in the real network. Using the testbed

[3]The fairness index is defined as

$$fi = \frac{\left(\sum_{x=0}^{N-1} T(x)\right)^2}{N \sum_{x=0}^{N-1} T(x)^2}$$

where $T(x)$ is the throughput of the x-th protocol instance. fi always lies between 1 (indicating that all instances get the same share) and $1/N$ (when one of the instances gets all the bandwidth).

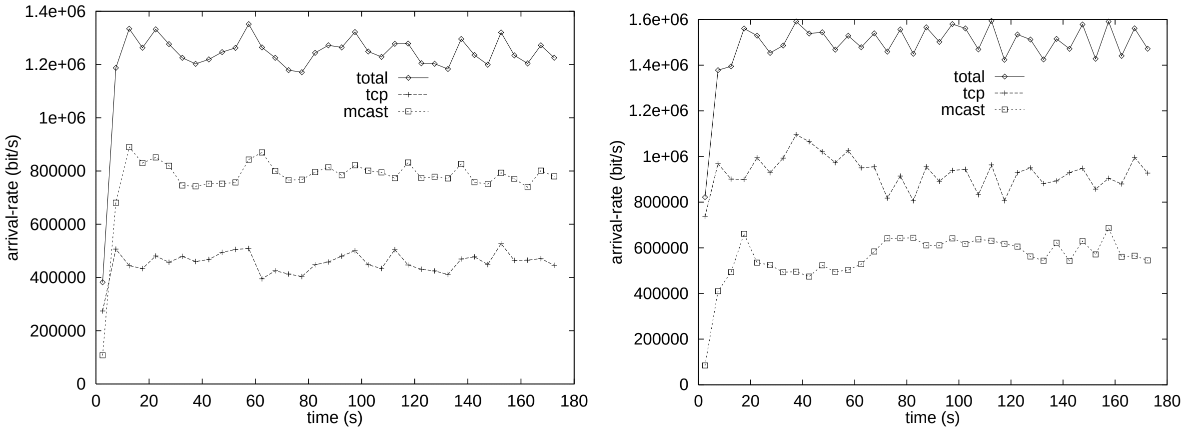

Figure 5: Aggregate throughput of TCP connections and that of our protocol instances. Left: TCP with 256 bytes packet size; right: 1024 bytes packet size.

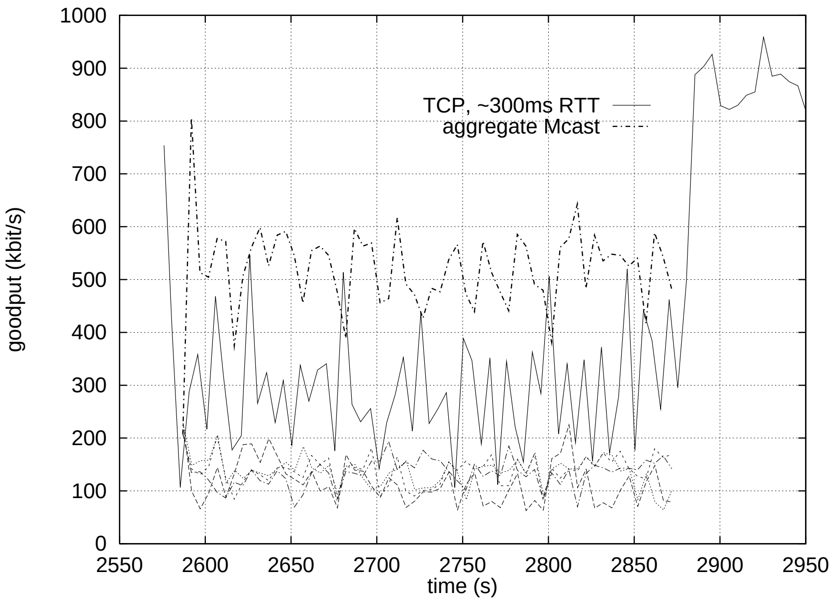

Figure 6: Bandwidth sharing among one TCP connection (top) and 4 protocol instances (bottom) on a real network. $RTT \simeq 300$ms, and TCP packet size of 1440 bytes.

described above, we have run 4 instances of our protocol and one TCP connection on our network path. Both in TCP case and in our case data rate has been evaluated at the application level averaging over non-overlapping windows 5 seconds wide.

As we can see in Figure 5, network load was almost entirely determined by our traffic: before we started the multicast transfers, the TCP connection was using all the available bottleneck bandwidth (excluding that used by non-responsive protocols); after our traffic started, the TCP throughput decreased, accounting for the bandwidth used by the multicast traffic. As we can see and according with Equations 1 and 2, with the parameters used (full-sized ethernet packets for TCP) and with with the RTT we had, our protocol was less aggressive than TCP.

4.6 Startup and steady state dynamics

A final set of experiments has been run to determine the dynamic behaviour of our algorithm. We were particularly concerned about the startup behaviour, to see how quickly a new receiver would synchronise with other receivers belonging to the same protocol instance and located behind the same bottleneck, and the steady state dynamics, to verify that receivers remain in synch and only move between adjacent layers.

On the topology of Figure 2,b, we have run an instance of our protocol placing a sender on node n_0 and five receivers on nodes n_5–n_9. Receivers on node n_5–n_7 have a bottleneck bandwidth of 1.5Mb/s while receivers on nodes n_8 and n_9 have a bottleneck bandwidth of .5Mb/s and .128Mb/s respectively. Moreover each bottleneck link is crossed by a random uncorrelated traffic which accounts for about half of the available bandwidth.

Figure 7 (left) shows the result of the experiment: receivers on node n_5–n_7 (r_2–r_4) experience the same network conditions and converge reasonably fast to the same subscription level even if started at different times. They also keep synchronized until the end of the run. The other receivers, which stabilize on lower subscription levels, are not subject to any appreciable interference caused by the formers.

Figure 7 (right) shows the throughput seen at the application level in a similar experiment. This time we have run a single receiver after the first bottleneck, but we have run a TCP connection on the path n_0—n_7 with 256 bytes packet-size. We can notice that TCP competes for the available bandwidth on the first bottleneck with receiver r_2 interfering very little with the other receivers and achieving a good bandwidth sharing.

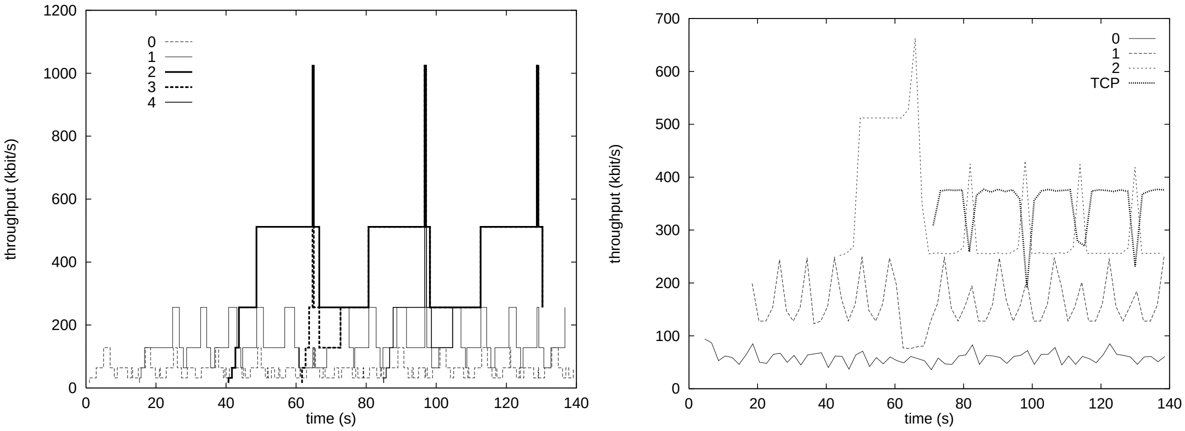

Figure 7: Different receivers of the same protocol instance, joining at different time and with different bandwidths. On the left is shown throughput pertinent to the subscription level while on the right is shown the throughput as seen at the application level.

5 Applications

The congestion control algorithm we have presented is applicable to both continuous data streams and to reliable bulk data transfer. In general, receivers will not receive all packets transmitted by the sender on all layers, either because their subscription level is too low, or because of losses due to congestion or to the hopping between layers. It should be noted, though, that our congestion control protocol aims at keeping the loss rate at very low values. Still, the effect of missing packets is obviously different for continuous data streams and for reliable bulk file transfers.

5.1 Continuous streams

In the transfer of continuous data streams, delay constraints generally do not allow to give hard guarantees on the integrity of the received data, and the recovery of lost packets is generally useless. The goal, for a continuous data stream, is to transfer data with the maximum achievable quality for the available bandwidth. This is simply achieved by using the layered data organisation, where each subscription level corresponds to the same data transmitted with a different quality.

It is desirable to limit the latency in transferring data, and to make burst losses reasonably short. These are contrasting requirements since burst losses can be reduced by using techniques such as interleaving, at the price of an increase in the latency. Regarding latency, we should also note that the use of the sender-initiated probe produces a degree of burstiness in the traffic which requires at least τ_0 seconds of buffering at the receiving end in order to produce a uniform flow of data.

5.2 Reliable bulk data transfer

We come now to the problem of reliable bulk data transfer, which requires to find a proper layered data arrangement for a fixed-size data object, and an effective way for recovering from any packet that may have been lost because of congestion or during the jumps between different layers. We assume that the transmitter starts sending packets as soon as it receives a request from the first receiver, and continues transmission until there are active receivers. Transmission is always done on all layers, which are then selectively forwarded by the multicast routers. We also assume the existence of an upper level session management protocol which possibly collects feedback to decide how long to continue transmissions and select which packets to send.

The approach we follow is based on the use of FEC techniques (see [1, 14]). Intuitively, the problems of finding a suitable transmission order for the data, and achieving an effective recovery from losses, can be solved by passing all the k packets which make up the object to an encoder which *can* produce a large number $n \gg k$ of packets. The encoding is such that any subset of k encoded packets suffice to reconstruct the original data.

When the transmitter has to send a packet, it takes a new one produced by the encoder. Provided that each packets is not transmitted twice during the time a receiver needs to complete reception (the condition $n \gg k$ insures this), each receiver only has to collect k different packets, no matter which ones or which layer they come from, in order to complete reception. This approach has already been used by the authors in the RMDP protocol [15], a reliable multicast protocol which can be easily extended with the congestion control algorithm described in this paper.

The use of a FEC encoding with high redundancy makes all packets equivalent for data reconstruction, and solves many problems at once. First, there are no scalability problems due to the presence of uncorrelated losses at different receivers, since each receiver can conduct recovery autonomously, by simply waiting for more packets to come. Second, there is no need for detailed feedback from the receivers, since all received packets are equally good for data reconstruction purposes. Third, there is no need to devise a special data arrangement for the different layers in order to avoid that a receiver sees the same packet more than once.

The use of FEC has a drawback, though, which consists in the potentially high encoding and decoding costs. However, this overhead can be limited to acceptable values.

The principle of operation of conventional FEC encoders and decoders is shown in [1], while [14] presents an implementation suitable for use in computer communication protocols. Such encoders/decoders require $O(k)$ operations per packet produced, although the multiplying constants are small and encoding speeds in the Mbit/s (or higher) can be achieved on cheap PCs with values of k in the range 32..64 (see [14] for such an implementation). The use of FFT techniques (see [1]) can reduce the complexity of FEC encoding to $O(\log k)$.

The encoding costs can be bounded, at the price of a slight decrease in efficiency (i.e. number of packets which are necessary to reconstruct the original data), by limiting the number of packets used to produce each encoded packet. A first approach consists in associating the k packets making up the file to be transferred to points of a $t-$dimensional cube of side k_0, i.e. $k = k_0^t$, with k_0 sufficiently small to make the computation feasible at the desired speed. For each group of k_0 packets differing in one coordinate, the encoder can produce n_0 packets. The encoding costs are then bounded by the value of k_0; the receiver efficiency decreases slightly, in that the condition of having a minimum number of packets must hold on each group of n_0 packets rather than on the whole file.

As an alternative, a recent paper [9] presents a different encoding algorithm, based on probabilistic techniques, which is very efficient for large values of k, requiring constant time per packet produced.

6 Conclusions and future work

We have presented a congestion control algorithm for multicast data transfer on the MBone, evaluated its performance, and shown its applicability to continuous stream and reliable bulk transfer. The algorithm does not require router support, or per-receiver status at the sender; it is completely receiver driver, and does not rely on any explicit protocol for receiver synchronisation or managing group membership. Analytical results, simulations and experiments on real networks have shown the algorithm to achieve a fair sharing of bottleneck bandwidth with other protocol instances and with TCP instances.

Future work will include a more extensive set of simulations, and measurements using larger network topologies and more complex scenarios.

References

[1] R.E.Blahut, “Theory and Practice of Error Control Codes” Addison Wesley, MA, 1984

[2] T.Brown, S.Sazzad, C.Schroeder, P.Cantrell, J,Gibson, "Packet video for heterogeneous networks using CU-SeeMe", Proc. of IEEE Intl. Conf. on Image Processing, Lausanne, Sept.96

[3] D-M.Chiu, R.Jain "Analysis of the Increase and Decrease Algorithm hms for Congestion Avoidance in Computer Networks", *computer Networks and ISDN Systems*, V.17, pp.1-14, 1989.

[4] S.Deering, "Multicast Routing in a Datagram Internetwork", PhD Thesis, Stanford University, Dec.1991.

[5] W. Fenner, "Internet Group Management Protocol, Version 2", INTERNET-DRAFT (working draft), Jan. 20, 1997, `http://ds.internic.net/internet-drafts/-draft-ietf-idmr-igmp-v2-06.txt`

[6] S.Floyd, K.Fall, "Router Mechanisms to Support End-to-End Congestion Control", Technical report, `ftp://ftp.ee.lbl.gov/papers/collapse.ps`.

[7] D.Hoffman, M.Speer, "Hierarchical video distribution over Internet-style networks", Proc. of IEEE Intl. Conf. on Image Processing, Lausanne, Sept.1996

[8] V.Jacobson, "Congestion Avoidance and Control", *ACM SIGCOMM'88*, August 1988, Stanford, CA, pp.314-329.

[9] M.Luby, M.Mitzenmacher, A.Shokrollahi, D.Spielman, and V.Stemann. "Practical loss-resilient codes", Proceedings of the Twenty-Ninth Annual ACM Symposium on Theory of Computing, page 150, El Paso, Texas, 4-6 May 1997.

[10] M.Mathis, J.Semke, J.Mahdavi, T.Ott, "The Macroscopic Behavior of the TCP Congestion Aviodance Algorithm", CCR, Vol.27 N.3, July 1997.

[11] S. McCanne, V. Jacobson, and M. Vetterli, "Receiver-driven Layered Multicast", SIGCOMM'96, August 1996, Stanford, CA, pp.1-14.
Available as `ftp://ftp.ee.lbl.gov/papers/mccanne-sigcomm96.ps.gz`

[12] UCB/LBNL Network Simulator - ns - version 2, 1997, `http://www-mash.cs.berkeley.edu/ns/`.

[13] T.Ott, J.Kemperman, M.Mhathis, "The stationary distribution of ideal TCP congestion avoidance", Technical Report, Aug. 1996.

[14] L. Rizzo, "Effective erasure codes for reliable computer communication protocols", CCR, V.27 N.2, April 1997, pp.24-36.
Source code available as `http://www.iet.unipi.it/~luigi/vdm.tgz`

[15] L.Rizzo, L.Vicisano, "A Reliable Multicast data Distribution Protocol based on software FEC techniques", *The Fourth IEEE Workshop on the Architecture and Implementation of High Performance Communication Systems (HPCS'97), Sani Beach, Chalkidiki, Greece June 23-25, 1997*

[16] W. Stevens., "TCP Slow Start, Congestion Avoidance, Fast Retransmit, and Fast Recovery Algorithms.", RFC2001, January 1997.

[17] N.Shacham, "Multipoint communication by hierarchically encoded data", Proc. of IEEE Infocom'92, (1992), pp.2107-2114.

[18] D.Taubman, A.Zakhor, "Multi-rate 3-D subband coding of video", IEEE Trans. on Image Processing 3, 5 (Sept.1994), pp.572-588.

[19] L.Vicisano, "Notes on a cumulative layered organisation of data packets across multiple streams with variable-rate", `http://www.cs.ucl.ac.uk/staff/L.Vicisano/layers.ps` .

A Simplified Model of the Receiver

To compute the dependency of the throughput on the loss rate, we use a simplified model of our congestion control algorithm, assuming that receivers reach a steady state, where they oscillate between two subscription levels: i and $i+1$. Experimental observations have shown this to be a realistic assumption, and we will see that the results we obtain from this model conform to the simulation outcome.

Let the subscription level become $i+1$ after an SP at level i, and assume that the first loss is encountered after $t_H = \alpha 2^{i+1} PW\tau_0$. After the loss, the receiver will decrease the subscription level and remain at level i until the next SP, which will occur after a time t_L ($t_L = (1-2\alpha)2^i PW$ if $0 \leq \alpha < 0.5$, $t_L = (2-2\alpha)2^i PW$ otherwise). Let N_H and N_L the number of packets received during t_H and t_L, with $N_H = 4\alpha PW(2^i)^2$, and $N_L = (1-2\alpha)PW(2^i)^2$ or $N_L = (2-2\alpha)PW(2^i)^2$ depending on α. The throughput is then

$$T = s\frac{N_H + N_L}{t_H + t_L} \tag{3}$$

Assuming (approximately) one loss per cycle, we can express the reciprocal of the loss rate as

$$1/p = N_H + N_L \tag{4}$$

By substituting the values of t_H, t_L, N_H and N_L in Eqs. 3 and 4, and writing Eq.4 as

$$2^i = \frac{1}{\sqrt{(1+2\alpha)pPW}} \tag{5}$$

(for $\alpha < 0.5$), and plugging Eq.5 into Eq.3 we obtain

$$T = \frac{s}{\tau_0}\frac{c'}{\sqrt{p}\sqrt{PW}}$$

where c' varies between 1 and 2 depending on α (using $c' = 1$ gives a conservative estimate of the throughput).

The model suffers from several approximations, the main one being not taking into account the presence of the leave delay. Its presence has the effect of increasing the actual throughput for a given loss rate. In fact, after the loss which triggers the leave of layer $i+1$, further loss might occur at all layers for the duration of the leave delay. As a consequence, when matching experimental data (from simulations and measurements in real networks) to the analytical model, we have used the burst loss rate as a parameter, computed by considering closely spaced losses as a single one. With this correction, experimental results fit well our model.