RMT Working Group
INTERNET DRAFT
Expires May 2001

M. Luby/Digital Fountain
L. Vicisano/Cisco
A. Haken/Digital Fountain
November 2000

Reliable Multicast Transport Building Block: Layered Congestion Control <draft-ietf-rmt-bb-lcc-00.txt>

Status of this Memo

Copyright Notice

Abstract

This document describes LCC, a scalable layered congestion control building block for multicast. LCC is a combination of approaches that allow multiple receivers to concurrently receive packets from a single sender at varying rates depending on individual bandwidth connections and network conditions. Two basic goals of the approach are to allow each receiver to obtain the full benefit of the available bandwidth to the sender and to be fair to other flows in the network.

For all of the approaches described in this memo, a sender sends data for one or more objects to multiple multicast groups, potentially at different rates for each group, where an object is any well-defined

content or file. The set of multicast groups carrying data for an object or sequence of objects out of a single sender is called a session. LCC assumes that each receiver is free to join and leave at any time one or multiple groups carring the data for the session, as required by the congestion control algorithm. This implies that, in general, only a subset of the packets sent are actually received. Applications that use LCC must either be tolerant to this, or they must be able to compensate through some erasure-recovery mechanism (E.g. FEC techniques [FECBBSTAN] are an afficient way to achieve erasure-recovery if LCC is used in a reliable file transfer protocol). The number of groups and which groups in the session each receiver joins is dictated by the local bandwidth availability and network conditions experienced by the receiver. In particular, receivers reduce their reception rate as soon as they feel congestion as evidenced by measured packet loss.

LCC is receiver driven, i.e., each receiver adjusts its reception rate in response to measured packet loss at the receiver in a manner reminiscent of TCP congestion control. Thus, each receiver experiences a reception rate appropriate to that receiver independent of other receivers.

LCC has the following properties:

o To each receiver, it appears as if there is a dedicated unicast session from the sender to the receiver, where the reception rate adjusts to congestion along the path from sender to receiver similar to TCP.

o To the sender, there is no difference in load or outgoing rate if one receiver is joined to the session or a million (or any number of) receivers are joined to the session, independent of when the receivers join and leave.

o For each link in the network, the packet traffic from the session and its reaction to competing traffic is the same whether there is one receiver or a million receivers beyond the link, and this reaction is similar to how TCP reacts.

o Receivers adjust their reception rate by joining and leaving groups, i.e., with no feedback to the sender.

  Thus, LCC provides a massively scalable layered congestion control approach that is network friendly.

1. Introduction

This document describes a scalable layered congestion control (LCC) building block that can be used by applications built on top of IP

multicast [DEE88]. For example, LCC can be used as the congestion control scheme for LCT [LCT00]. Many congestion control schemes have been built on top of multicast. However, scalability is not a design goal for many of these congestion control schemes, in the sense that they require all receivers to be receiving at the same rate, and in general this rate is dictated by the available bandwidth along the most constrained path from the sender to a receiver. Thus, the reception rate of receivers is restricted to that of the worst case receiver. In contrast, a general design goal of LCC is that each receiver can be receiving at the rate that is dictated by the available bandwidth along the path from the sender to that receiver. Thus, the reception rate of receivers is independent of that of other receivers.

One of the key difficulties in scaling sender-based multicast congestion control schemes is dealing with the amount of data that flows from receivers back to the sender to adjust the sender rate. LCC avoids any such feedback, and thus is massively scalable.

An attractive feature of a scalable congestion control scheme is the ability for different receivers to join and leave the session asynchronously without adversely affecting the reception experience of other receivers and without affecting the scalability of the scheme. This is one of the features provided by LCC.

To transmit data about an object or sequence of objects using LCC, a sender sends the data concerning the object(s) to one or more multicast groups. The rate at which data is sent to different multicast groups may vary. The set of groups pertaining to an object or set of objects emanating from a single server over which congestion control is performed is called a session.

The original ideas for LCC are from [MCC96], [VIC98A], [VIC98B], [BYE00]. In all of these schemes, the sender places congestion control information into the header of each packet sent to the session. The set of groups a receiver joins is determined by the receiver based on signals placed into packets by the sender and by loss measured along the path from the sender to that receiver. Receivers that can receive packets at a rate higher than their current rate are allowed to periodically increase their reception rate, and receivers that are receiving packets at a higher rate than they have the capacity for (as evidenced by packet loss) MUST reduce their rate.

A primary goal for LCC session is to be fair to other LCC sessions as well as to sessions using other congestion control schemes within other protocols such as TCP. In particular, if several sessions are flowing through a bottleneck link, then it is desirable for the sessions to share the bandwidth capacity of the link fairly, and it is also desirable that the link not be overly congested.

This document describes two congestion control schemes, a static layer scheme called FLID-SL and a dynamic layer scheme called FLID-DL. FLID-SL is applicable to networks where joins to multicast groups and leaves from multicast groups are processed quickly, e.g., on the same time scale as the round trip time from the receiver to the sender.

2. Environmental Requirements

FLID-DL is applicable to networks where joins to multicast groups are process quickly but leaves from multicast groups may take several seconds. One of the current problems with many implementations of IP multicast routing protocols is that the IGMP protocol used between receiver and access routers to join and leave groups has the limitation that IGMP leave messages can take a significant amount of time to take effect, leading to a large amount of data flowing to a receiver on a given group long after the receiver has issued an IGMP leave message for that group in reaction to packet loss. This is a problem because the network remains congested for the period of time it takes to process the IGMP leave request, thus making it difficult to rely on IGMP leave requests to provide a network friendly congestion control scheme. The dynamic layer scheme uses groups where the transmission rates on the groups change over time in a way that makes it possible for receivers to quickly reduce their reception rate by taking no action. Instead, receivers that want to maintain their current reception rate must periodically issue IGMP join requests, and receivers that want to increase their reception rate must issue an additional IGMP join request.

One of the attractions of both congestion control schemes is that they are multicast routing independent and that they do not require multicast reverse connectivity, i.e. LCC receivers do not send multicast traffic or any other traffic for purposes of congestion control. In particular, LCC works with the original multicast model introduced in [DEE88], which we call Internet Standard Multicast (ISM) in this document, and with the Source Specific Multicast (SSM) model that is based on [HOL99]. The definition of a group that is used throughout this document is slightly different with ISM and with SSM. When using ISM, packets of a group are sent to a multicast group address G. When using SSM, packets of a group are sent to a channel address (S,G), where S is the IP address of the sender and G is a multicast group address.

SSM is more attractive to LCC than ISM for a few reasons. First, a session is made up of multiple groups, and LCC can be used to deliver a large number of objects over time using the same set of groups assigned to the session. With ISM, the multicast group address G that corresponds to a group must be allocated so that it is unique across the Internet. With SSM, the multicast group address G can be allocated locally by the sender with the only requirement that it is unique to the sender, because it is the (S,G) channel that corresponds to the group

that a receiver joins.

Second, LCC supports an unlimited number of receivers that are dynamically joining and leaving groups within a session. Changes in the multicast tree topology with SSM are light weight operations (a new branch from the receiver towards S grows when a receiver joins, and the branch is deleted when the receiver leaves), and with ISM changes can be heavier weight (involving transitions from a (*,G)-tree rooted at an RP to the tree rooted at S).

Third, LCC is scalable to an unlimited number of receivers that may span the global Internet. Thus, the light weight mechanisms that SSM uses to cross ISP boundaries (standard BGP+ routing tables) is a distinct advantage over the heavier weight mechanisms used by ISM (the MSDP and BGMP protocols, both of which are not needed by SSM).

Finally, a receiver joins a group by joining a channel (S,G) with SSM, and thus the receiver will only receive packets sent from the sender S. With ISM, the receiver joins a group by joining a multicast group G, and all packets sent to G, regardless of their origin sender, will be received by the receiver. With ISM, if another sender application is inadvertently sending packets to G at a high rate, these packets will be sent over portions of the network and delivered to the receiver even though they were not requested by the receiver, potentially resulting in a large amount of unintended network congestion. Thus, SSM has compelling advantages over ISM for prevention of unintended and wasteful network congestion.

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [R2119].

3. General Architecture

The Layered Congestion Control (LCC) schemes described in this document applies to a session emanating from a single sender. A session consists of data sent to one or more groups for some period of time that is determined by the application. For example, within LCT [LCT00], a session is defined as the data sent from a single sender about one object or a sequence of objects over which congestion control is to be performed.

A sender sends data to one or more groups within a session. The transmission rate of data to different groups may vary among groups, and the transmission rate to a particular group may vary over time. Congestion control is performed globally over all data sent to the session. Typically, the sender continues to send data to all groups in a session until the session is terminated. The session may be terminated when

some amount of time has expired, a certain amount of data has been sent, or some out-of-band signal (from a higher level protocol, perhaps) has indicated completion by a sufficient number of receivers.

It is possible that a receiver may concurrently join multiple sessions to receive data from multiple senders. For example, three different sessions could be transmitted from three different senders, each session consisting of four groups to which data is being sent at different rates, and a receiver may join and receive packets from all 12 groups concurrently. However, since the senders may be located at different points in the network that experience varying network conditions, a receiver MUST perform congestion control independently on each session it is receiving.

All packets sent to a session should be roughly of the same size in order for the congestion control scheme to work effectively. Larger packet sizes are generally desirable so that the fraction of bandwidth lost to packet headers is reduced. On the other hand, if the packet size is larger than the network's maximum transmission unit (MTU), packets would be fragmented, and the loss of any fragment would require the entire packet to be lost. Therefore a packet size close to, but not exceeding, the MTU is best, as this reduces packet header overhead and packet handling overhead in routers.

A receiver must first obtain a transmission session description before joining a session. This includes the information about the groups associated with a session that is needed in order to join those groups. Once a receiver has obtained this information the receiver may join one of the groups in the session. In order to be in compliance with LCC, receivers MUST join and leave groups within a session as described in detail in this document in order to vary their reception rate in the face of varying bandwidth capacity between the receivers and the sender.

The transmission session description is determined and agreed upon by the senders and communicated to the receivers out-of-band, or, in some cases, included or partially included in the header of each packet. The session description pertinent to LCC could include the packet format and length, the sender address and the multicast group address(es) used in the session. The session description could be in a form such as SDP [HAN98]. There must be an out-of-band mechanism for receivers to obtain the session description. The session description might be carried in a session announcement protocol such as SAP [HAN96], located on a Web page with scheduling information, or conveyed via E-mail or other out-of-band methods. Discussion of session description format, and distribution of session descriptions is beyond the scope of this document.

4. Overview of congestion control schemes

LCC performs congestion control by dedicating multiple groups to a session. Receivers joined to the session are subject to heterogeneous reception rates, obtained by having the receiver selectively join a subset of all the groups available from the sender.

The original ideas for both of the LCC congestion control schemes are from a combination of [MCC96, VIC98A, VIC98B, BYE00].

When a sender is instructed to start a session, the range of possible reception rates is specified by rmin and rmax, where rmin is the minimum available reception rate and rmax is the maximum available reception rate. It is recommended that rmin be small enough so that any receiver can receive at rate rmin without incurring significant packet loss.

If rmin = rmax is specified then there is only one reception rate specified for the session. If rmax > rmin is specified, then multiple reception rates should be made available in the session. Let X = 1.3. It is recommended that the number of available reception rates be set to A+1, where A is the largest value such that rmin * X^A <= rmax, and that for all i=0,...,A the reception rate R(i) be set to rmin * X^i.

Let A+1 be the number of different reception rates available for a session, and let R(0) < ... < R(A) be the set of available reception rates. As an example, set R(0) = 24 Kbps, set X = 1.3, set A=30, and for all i=1,...,30 set R(i) = R(0) * X^i. Then, R(30) = 62.9 Mbps, and thus the ratio of the maximum achievable rate to the minimum rate is a factor of 2,620 in this example. When a receiver is receiving packets at reception rate R(i), we say the receiver is at layer i. Let r(0) = R(0), and for all i = 1,..., A, let r(i) = R(i) - R(i-1) be the incremental rate needed to increase the rate from layer i to layer i+1. Suppose the current reception rate of a receiver is R(i), i.e., the receiver is at layer i. If the receiver is to increase its reception rate to layer i+1, it does so increasing its reception rate by r(i+1) to R(i+1). If the receiver is to decrease its reception rate to layer i-1, it does so decreasing its reception rate by r(i) to R(i-1). If the receiver is to maintain it reception rate at layer i, it does not change its reception rate from R(i).

5. Fair Layered Increase/Decrease + Static layer scheme (FLID-SL)

This section provides a detailed operational description of how to apply the LCC design principles to congestion control when IGMP leave latency is minimal. The next section described extends these principles to describe a second way to achieve congestion control that overcomes long IGMP leave latencies.

5.1. Sender operation

The sender uses A+1 groups numbered 0,..., A. For all i = 0,...,A, the ith group carries packets at rate r(i) at all times. Group 0 is referred to as the base group.

The sender partitions time into equal duration intervals called time slots. The time slot duration TSD determines the reaction time of receivers to changing network congestion conditions. It is recommended that the time slot duration TSD be set to one of either 0.5, 1.0, or 2.0 seconds. Associated with each time slot is the time slot index. The range of values for the time slot index is [0..G-1] for some value G. The time slot index increments by one modulo G between each consecutive time slot. For example, if G = 32 then the time slot index is 0, 1, 2, 3, ... , 30, 31, 0, 1, ... in consecutive time slots. It is recommended that G be set to 128. G must be set to at least 3.

In order to be able to measure loss within each group, the sender places consecutive sequence numbers in the packets sent to a group. The sequence numbers ignore time slot boundaries, i.e., the sequence numbers within the same group across the time slot boundary are still consecutive. Sequence numbers wrap around at 2^16, i.e., the consecutive sequence numbers are 0, 1, 2, ..., 2^16-2, 2^16-1, 0, 1, 2, ...

The sender places into each packet the index of the group within the set of groups used for that session.

The sender places into each packet the time slot index. Thus, all packets within the same time slot must have the same time slot index.

The sender places an increase signal trigger into each packet. The increase signal trigger is set to either 0 or 1 for each packet. The increase signal trigger must be the same for all packets sent to a group within the same time slot. An increase signal trigger of 0 indicates no increase allowed to receivers, and an increase signal trigger of 1 indicates increase is allowed to receivers.

The increase signal triggers are calculated as follows by the sender. For all i = 0,...,A-1, let p(i) = min {1.0, 20*packet size*TSD/R(i)}, and set p(A) = 0. Note that 1 >= p(0) >= p(1) >= ... >= p(A-1) >= p(A) = 0. Let B be an integer associated with each time slot that increases by one for each consecutive time slot, and thus t = (B mod G) is the time slot index. For each time slot, for each i = 0,...,A, for all packets sent to the group i that carries rate r(i) during the Bth time slot, the increase signal trigger is set to 1 if BB <= p(i), and the trigger is set to 0 otherwise. Here, BB is derived from B by writing B in reverse binary notation and considering it as a fraction that is between 0 and 1. For example, if B = 253 when written base ten, then when written in binary B = 11111101, and then BB = 0.10111111 when written as a binary fraction. As a decimal fraction, BB = 0.7421875. This

method of computing BB from B, where B increases by one at each time slot, guarantees that increase signal triggers equal to 1 for a given layer i are very well-spaced out over the time slots, and that on average the fraction of time slots with increase signal 1 for layer i is p(i). Note that increase signal trigger for the group carrying rate r(A) during a time slot is always set to 0, since p(A) = 0. This ensures that there is never an attempt by a receiver to increase the reception rate above R(A).

This method of setting the increase signal triggers implies the following monotonicity property: if the trigger is set to 1 for layer i during some time slot, then the trigger is also set to 1 for lower layers i' < i in the same time slot. This allows receivers at lower layers behind a bottleneck link to increase to a higher layer when receivers at higher layers increase to a higher layer. During other time slot periods, receivers at lower layers will be allowed to increase to a higher layer when receivers at higher layers aren't allowed to increase to a higher layer, thus giving receivers at lower layers a chance to catch up.

5.2. Receiver operation

When a receiver first joins a session, it must only join the base group and remain joined only to the base group for at least one complete time slot. The rate r(0) = R(0) of the base group must be small enough that when a receiver is joined to just this base group there is no significant packet loss. If there is significant packet loss over any significant period of time when the receiver is only joined to the base group then the receiver must leave the session by leaving the base group. If there is only one reception rate offered by the session, i.e., there is only a base group offered by the session and no other groups, then only this first paragraph is relevant to the receiver congestion control scheme. If there are two or more reception rates offered by the session, i.e. there is a base group and (at least three) other groups, then the rest of this subsection is relevant to the receiver congestion control scheme.

The receiver must keep a timer that tracks the maximum interarrival time between packets. Whenever there is an interarrival time that exceeds TSD, the receiver must leave the session by leaving all groups in the session immediately. The receiver may thereafter try to rejoin the session.

During a generic time slot t a receiver is joined to some number i, 0 <= i <= A, of the groups 0, ..., i that have rates r(0), r(1), r(2),...,r(i), respectively, within time slot t. Thus, during time slot t, the receiver is at layer i and has a reception rate R(i). The receiver does not join or leave any groups in the middle of time slot t. To simplify the following discussion, let t+1 indicate t+1 mod G. The

receiver only makes changes in group membership at the beginning (indicated by the first packet received) of the next time slot t+1.

If there is at least one packet loss measured in time slot t in any group then the receiver must leave group i at the beginning of the next time slot t+1. This will drop the reception rate for the receiver from layer i to layer i-1, i.e., the reception rate will drop from R(i) to R(i-1).

If there is no measured packet loss in time slot t then the action of the receiver depends on the increase signal trigger in group i in time slot t. If the increase signal trigger is 0 for group i in time slot t, indicating the receiver must not increase above layer i, the receiver does not join or leave any groups at the beginning of time slot t+1. If the increase signal trigger is 1 for group i in time slot t (in which case i < A, since the increase signal trigger is always 0 for group A), indicating the receiver can increase to layer i+1, then the receiver joins group i+1 at the beginning of time slot t+1.

5.3. General considerations

Generally, the multicast group addresses associated with the groups constitute a consecutive range of multicast address space. For example, the 21 groups [0..20] may be bound to the SSM channel addresses (192.35.134.26, 232.153.220.0) through (192.35.134.26, 232.153.220.20). However, it is not a requirement that these multicast group addresses be consecutive. There can be at most 256 groups associated with an LCC session using the static layer scheme, because there are 8 bits available for specifying groups in the abstract LCC packet header.

The number of groups associated with a session, and the addresses of the multicast groups or channels bound to these groups, must be part of the session description information communicated out-of-band.

5.4. Fairness

A crucial variable in determining the fair share of multiple TCP sessions flowing through a bottleneck link is the round trip time (RTT). In general, the smaller the RTT for TCP the more aggressive the session will be against other sessions, including other TCP sessions with larger RTTs. With FLID-SL it is desirable that receivers behind a common bottleneck link are joined to the same set of groups in the session at each point in time independent of their distance from the sender. Thus, the role that RTT plays in TCP with respect to fairness does not make sense for FLID-SL. In this document, the FLID-SL parameters are set so that a session shares approximately equally with other FLID-SL sessions, and a FLID-SL session shares approximately equally with a TCP session with a RTT of 200 ms. This is assuming all sessions use the same packet

length.

FLID-SL will not be fair to other sessions if the IGMP leave latency is high. Thus, FLID-SL should not be used with multicast routing protocols that do not support fast IGMP leaves.

6. Fair Layered Increase/Decrease + Dynamic layer scheme (FLID-DL)

Several ideas are used to circumvent the problems associated with long leave latency in the design of FLID-DL. As with FLID-SL, let A+1 be the number of different reception rates available for a session, and let R(0) < ... < R(A) be the set of available reception rates. One idea is to allocate G+1 groups to the session, numbered from 0 to G, where G > A. **Group 0 is called the base group, and this group always carries** packets at rate r(0) = R(0). Groups 1 thru G are called dynamic groups, because over time they carry packets at different rates. At each point t in time, for all i = 1, ..., A, there is exactly one dynamic group that carries the rate r(i), and these groups are called active groups at time t. The remaining Q = G-A dynamic groups carry zero rate at time t, and these groups are called quiescent groups at time t. Which dynamic groups are active and what rate they carry and which groups are quiescent varies over time in such a way that the problem of long leave latency is overcome.

6.1. Sender operation

As for FLID-SL, the sender partitions time into equal duration intervals called time slots. The time slot duration TSD determines the reaction time of receivers to changing network congestion conditions. It is recommended that the time slot duration TSD be set to one of either 0.5, 1.0, or 2.0 seconds. Associated with each time slot is the time slot index. The range of values for the time slot index is [0..G-1], i.e., the number of possible time slot indices is equal to the number of dynamic groups. The time slot index increments by one modulo G between each consecutive time slot. For example, if G = 32 then the time slot index is 0, 1, 2, 3, ... , 30, 31, 0, 1, ... in consecutive time slots.

Given the definition of a time slot and time slot index, we can now define how the rates on the dynamic groups vary over time. For i = A+1,...,G, define r(i) = 0. The idea is that dynamic group j in time slot t carries packets flowing at rate r(((j-t-1) mod G)+1). Thus, in the time slots with indices 0, 1, 2, ..., G-A-1, G-A, G-A+1, ..., G-1, dynamic group G carries packets at rate r(G) = 0, r(G-1) = 0, r(G-2) = 0,...,r(A+1) = 0, r(A), r(A-1),...r(1), respectively. Thus, dynamic group G is quiescent for Q = G-A time slots, then carries rate r(A), r(A-1), r(A-2), ..., r(1) over the subsequent A time slots. This same pattern is then cyclically repeated thereafter. Each of the other G-1 dynamic groups goes through the same cycle, where the cycle for dynamic

group G-1 is shifted one time slot forward from dynamic group G, and in general the cycle for dynamic group j is shifted one time slot forward from dynamic group j+1. Thus, during each time slot t, for each i = 1,...,G, dynamic group ((i+t-1) mod G) + 1 carries packets at rate r(i). Thus, during each time slot A groups are active and Q are quiescent.

The reason for this organization of the dynamic groups is that it allows receivers to quickly decrease their reception rate within one time slot without requiring small leave latencies. For 1 <= i <= A, suppose a receiver is joined to dynamic group j = ((i+t-1) mod G)+1 at time t in order to receive at rate r(i). Then, over the i-1 subsequent time slots t+1, t+2,..., t+i-1 the reception rate of the receiver from group j is r(i-1), r(i-2),...,r(1). Then, in the next time slot t+i, the rate of group j drops to zero and remains there for a total of Q time slots. The idea is that once a receiver is joined to a dynamic group, the receiver remains joined to the group independent of all other factors such as packet loss until the group becomes quiescent, and at this point in time the receiver immediately leaves the group. Let LL be the upper bound on the leave latency, i.e., LL is an upper bound on the time between when a receiver issues a leave to a group and the time when the leave takes effect. Then, in order for the leave to take effect before the group becomes active again in Q time slots, it must be the case that LL <= (Q-1) * TSD. For example, if LL = 10 seconds and TSD is set to 1 second, then Q must be at least 11. Once a receiver joins a dynamic group it remains joined to the dynamic group until it becomes quiescent, at which time the group is left and this takes effect before the group becomes active again. As we describe below in the receiver congestion control scheme, with this property long leave latencies are no longer a problem, as the reaction time to network congestion is at most one time slot.

Suppose there is only one reception rate available, i.e., rmin = rmax. Then the base group is used to transmit at rate rmin and no dynamic groups are used in the session, and A = 0, Q = 0, and thus G = 0. When only the base group is used, the group number in each packet is set to zero, the time slot index in each packet is set to zero and the increase signal trigger in each packet is set to zero. The sequence numbers are used as normal within the base group, as described below.

The rest of this subsection concerns the case when rmax > rmin, i.e., when multiple reception rates are available in the session. Since there are at least two different reception rates, A >= 1. Let LL be the maximum leave latency. Based on TSD and LL, Q must be at least LL/TSD + 1. For example, if LL is 9.3 seconds and TSD is 1.0 second then the minimum possible value for Q is 11. Based on this, Q >= 2, and thus the number G = Q+A of dynamic groups must be at least 3.

In order to be able to measure loss within each group, the sender places

consecutive sequence numbers in the packets sent to a group. The sequence numbers ignore time slot boundaries, e.g., even though the rate of packets sent to a dynamic group changes when the time slot changes, the sequence numbers within the group across the time slot boundary are still consecutive. Sequence numbers wrap around at 2^16, i.e., the consecutive sequence numbers are 0, 1, 2, ..., 2^16-2, 2^16-1, 0, 1, 2, ...

The sender places into each packet the index of the group within the set of groups used for that session.

The sender places into each packet the time slot index. Thus, all packets within the same time slot must have the same time slot index.

The sender places an increase signal trigger into each packet. The increase signal trigger is set to either 0 or 1 for each packet. The increase signal trigger must be the same for all packets sent to a dynamic group within the same time slot. An increase signal trigger of **0 indicates no increase allowed to receivers, and an increase signal** trigger of 1 indicates increase is allowed to receivers, in a manner described in detail in the description of the receiver congestion control scheme.

The increase signal triggers that indicate to receivers when to increase from a given layer to the next layer are calculated as follows by the sender. For all i = 0,...,A-1, let p(i) = min {1.0, 20*packet size*TSD/R(i)}, and set p(A) = 0. Note that 1 >= p(0) >= p(1) >= ... >= p(A-1) >= p(A) = 0. Let B be an integer associated with each time slot that increases by one for each consecutive time slot, and thus t = (B mod G) is the time slot index. For the time slot, for each i = 0,...,A, for all packets sent to group j = ((i+B-1) mod G)+1 that carries rate r(i) during the Bth time slot, the increase signal trigger is set to 1 if BB <= p(i), and the trigger is set to 0 otherwise. Here, BB is derived from B by writing B in reverse binary notation and considering it as a fraction that is between 0 and 1. For example, if B = 253 when written base ten, then when written in binary B = 11111101, and then BB = 0.10111111 when written as a binary fraction. As a decimal fraction, BB = 0.7421875. This method of computing BB from B, where B increases by one at each time slot, guarantees that increase signal triggers equal to 1 for a given layer i are very well-spaced out over the time slots, and that on average the fraction of time slots with increase signal 1 for layer i is p(i). Note that increase signal trigger for the dynamic group carrying rate r(A) during a time slot is always set to 0, since p(A) = 0. This ensures that there is never an attempt by a receiver to increase the reception rate above R(A).

This method of setting the increase signal triggers implies the following monotonicity property: if the trigger is set to 1 for layer i during

some time slot, then the trigger is also set to 1 for lower layers i' < i in the same time slot. This allows receivers at lower layers behind a bottleneck link to increase to a higher layer when receivers at higher layers increase to a higher layer. During other time slot periods, receivers at lower layers will be allowed to increase to a higher layer when receivers at higher layers aren't allowed to increase to a higher layer, thus giving receivers at lower layers a chance to catch up.

6.2. Receiver operation

When a receiver first joins a session, it must only join the base group and remain joined only to the base group for at least one complete time slot. The rate r(0) = R(0) of the base group must be small enough that when a receiver is joined to just this base group there is no significant packet loss. If there is significant packet loss over any significant period of time when the receiver is only joined to the base group then the receiver must leave the session by leaving the base group. If there is only one reception rate offered by the session, i.e., there is only a base group offered by the session and no dynamic groups, then only this first paragraph is relevant to the receiver congestion control scheme. If there are two or more reception rates offered by the session, i.e. there is a base layer and (at least three) dynamic layers, then the rest of this subsection is relevant to the receiver congestion control scheme.

The receiver must keep a timer that tracks the maximum interarrival time between packets. Whenever there is an interarrival time that exceeds TSD, the receiver must leave the session by leaving all groups in the session immediately. The receiver may thereafter try to rejoin the session.

During a generic time slot t a receiver is joined to the base group at rate r(0) and some number i, 0 <= i <= A, of dynamic groups that have rates r(1), r(2),...,r(i), respectively, within time slot t. Thus, during time slot t, the receiver is at layer i and has a reception rate R(i). The receiver does not join or leave any groups in the middle of time slot t. The receiver only makes changes in group membership at the beginning (indicated by the first packet received) of the next time slot after t.

If the receiver is joined to one or more dynamic groups in addition to the base group (i >= 1) in time slot t then at the beginning of the next time slot after t the receiver must leave the dynamic group that was carrying packets at rate r(1) during time slot t, i.e. the receiver must leave dynamic group j = (t mod G) + 1. Dynamic group j will be quiescent for Q time slots after the end of time slot t, and thus by the time group j becomes active again, the leave request will have been processed.

If there is at least one packet loss measured in time slot t then the receiver must not join any groups at the beginning of the next time slot after t. If the receiver is joined to at least one dynamic group in addition to the base group (i >= 1) then the effect of this inaction will be to drop from layer i to layer i-1, i.e., the reception rate will drop from R(i) to R(i-1). This is because, for all 2 <= i' <=i, the dynamic group that carries rate r(i') in time slot t will carry rate r(i'-1) in the time slot after t, and the dynamic layer that carries the rate r(1) in time slot t drops to zero and remains at zero for Q time slots at the end of time slot t. The net effect is that the reception rate drops by r(i) at the beginning of the time slot after t, and thus the reception rate drops from R(i) to R(i-1).

If there is no measured packet loss in time slot t then how many groups the receiver joins at the beginning of the time slot after t depends on the increase signal trigger for layer i in time slot t. The increase signal trigger for layer i in time slot t is carried in packets in group ((i+t-1) mod G)+1, since this is the group that is carrying packets at rate r(i) in time slot t. If the increase signal trigger is 0 for layer i in time slot t, indicating the receiver must not increase above layer i, the receiver joins dynamic group j = ((i+t) mod G)+1 at the beginning of the time slot after t. This is because j is the group carrying packets at rate r(i) during the time slot after t and this will maintain the receiver reception rate at R(i) = R(i-1) + r(i). If the increase signal trigger is 1 for layer i in time slot t (in which case i < A, since the increase signal trigger is always 0 for the group carrying rate r(A) in time slot t), indicating the receiver can increase to layer i+1, then the receiver joins dynamic groups j = (i+t) mod G + 1 and j' = (i+1+t) mod G + 1. This is because j is the group carrying packets at rate r(i) during the time slot after t, and j' is the group carrying packets at rate r(i+1) during the time slot after t, and this will increase the receiver reception rate to R(i+1) = R(i-1) + r(i) + r(i+1).

Thus, at the beginning of each time slot, the receiver does at most one leave and at most two joins.

6.3. General considerations

Generally, the multicast group addresses associated with the groups constitute a consecutive range of multicast address space. For example, the 21 groups [0..20] may be bound to the SSM channel addresses (192.35.134.26, 232.153.220.0) through (192.35.134.26, 232.153.220.20). However, it is not a requirement that these multicast group addresses be consecutive. Besides the LCC base group 0, there can be at most 128 dynamic groups associated with an LCC session using a dynamic layer scheme, or 129 groups in total including the base group. This is because there are 7 bits allocated to time slot indices in the abstract LCC packet header, and thus there are at most 128 time slot indices

possible, and there is a one-to-one correspondence between time slot indices and dynamic groups.

The number of groups associated with a session, and the addresses of the multicast groups or channels bound to these groups, must be part of the session description information communicated out-of-band.

6.4. Fairness

A primary goal for LCC session is to be fair to other LCC sessions as well as to sessions using other congestion control schemes within other protocols such as TCP. In particular, if several sessions are flowing through a bottleneck link, then it is desirable for the sessions to share the bandwidth capacity of the link fairly, and it is also desirable that the link not be overly congested. A crucial variable in determining the fair share of multiple TCP sessions flowing through a bottleneck link is the round trip time (RTT). In general, the smaller the RTT for TCP the more aggressive the session will be against other sessions, including other TCP sessions with larger RTTs. With FLID-DL it is desirable that receivers behind a common bottleneck link are joined to the same set of groups in the session at each point in time independent of their distance from the sender. Thus, the role that RTT plays in TCP with respect to fairness does not make sense for FLID-DL. In this document, the FLID-DL parameters are set so that a session shares approximately equally with other FLID-DL sessions, and a FLID-DL session shares approximately equally with a TCP session with a RTT of **200 ms.** This is assuming all sessions use the same packet length.

7. Abstract LCC packet header

LCC defines the congestion control information that must be carried in each packet header. This information is of the same format for both the static layer scheme and for the dynamic layer scheme. Other information may be required in the packet header to support the scheme that is using LCC to implement congestion control, but this is outside the scope of this document. Thus, although we refer to packets as LCC packets, other protocols that embed LCC header information into packets may consider the packets to be of the type of the overall protocol. For example, the LCT protocol instantiation [LCT00] may use LCC for congestion control, and the packets that LCT uses are considered to be LCT packets, even though these packets contain LCC header information.

The LCC packets that contain the required congestion control information are sent by the sender(s) to a multicast IP destination address. In the LCC header, all integer fields are carried in "big-endian" or "network order", that is, most significant byte (octet) first. Unless otherwise noted, numeric constants are in decimal (base 10).

A LCC packet header contains the following congestion control information that is placed into each packet by a sender:

o Increase signal trigger (T): 1 bit

o Time slot index (TSI): 7 bits

o Group number (GN): 8 bits

o Sequence number (SEQNO): 16 bits

The required congestion control information required in a LCC packet is depicted in Figure 1 below.

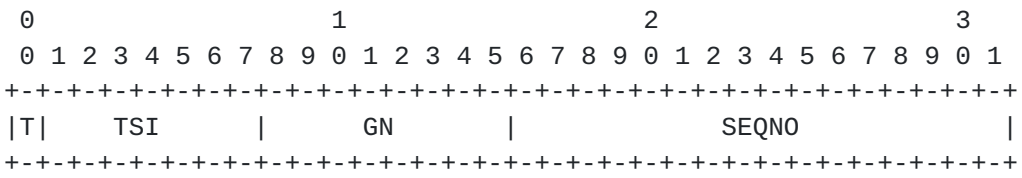

Figure 1 - LCC packet header layout

8. Applications

LCC is a good choice for a congestion control scheme to use with LCT [LCT00]. With LCT, FEC codes [FECBBINFO00, FECBBSTAN00] are used to provide reliability for content download applications. When using LCC, LCT sends coded data about one or more objects to each group that is part of a session. With appropriate use of FEC codes, the data sent to the different groups in the session, or in some cases even multiple sessions from different senders, can be effectively used by receivers to recover the original object(s). With the LCT approach to reliable multicast, the reliability scheme based on FEC codes can be made to be completely separate and independent of the congestion control scheme based on LCC. Furthermore, both the FEC codes described in [FECBBINFO00, FECBBSTAN00] and LCC can be used so that in the overall LCT protocol receivers do not send packets to the sender except perhaps to request extension of an ongoing session or to confirm complete receipt of an object. Thus, using LCC for congestion control and FEC codes for reliability within LCT yields a massively scalable network friendly content distribution protocol.

LCC is potentially also applicable to other applications. For example, it is possible that LCC could be used as the congestion control scheme for a layered media streaming application with LCT [LCT00].

9. LCC and Generic Router Assist

Router filtering of packets to assist in congestion control is described in [LUB99], [CAI99]. The addresses of the multicast groups can be communicated to the routers and they can do filtering of groups based on congestion. This is one of the reasons why it is good to have the congestion control information portion of the packet header in a fixed place at the beginning of the packet, so that the routers can probe this field if necessary (although it may not be). A full exploration of this approach is outside the scope of this document.

10. Intellectual Property Issues

Digital Fountain has patents pending for congestion control schemes that may be needed to use some of the congestion control schemes described in this document in a commercial product or service. Digital Fountain is willing to provide a blanket royalty free license to the rights it holds that are needed to use the congestion control schemes described in this document if and when these congestion control schemes become part of the IETF standards.

11. References

[AFZ95] Acharya, S., Franklin, M., and Zdonik, S., "Dissemination-Based Data Delivery Using Broadcast Disks", IEEE Personal Communications, pp.50-60, Dec 1995.

[R2119] Bradner, S., Key words for use in RFCs to Indicate Requirement Levels (IETF RFC 2119) http://www.rfc-editor.org/rfc/rfc2119.txt

[BYE00] Byers, J.W., Frumin, M., Horn, G., Luby, M., Mitzenmacher, M., Roetter, A., Shaver, W., "FLID-DL Congestion Control for Layered Multicast", Int'l Workshop on Networked Group Communication, Vol. 2, pp. 71-81, Palo Alto, CA, Nov. 2000.

[CAI99] Cain, B., Speakman, T., and Towsley, D., "Generic Router Assist (GRA) Building Block, Motivation and Architecture", Internet Draft draft-ietf-rmt-gra-arch-00.txt, a work in progress.

[DEE88] Deering, S., "Host Extensions for IP Multicasting", RFC 1058, Stanford University, Stanford, CA, 1988.

[LCT00] Luby, M., Gemmell, J., Vicisano, L., Rizzo, Handley, M., L., Crowcroft, J., "Layered Coding Transport: A massively scalalable multicast protocol", Internet draft draft-ietf-rmt-lct-00, November 2000.

[FECBBINFO00] Luby, M., Vicisano, L., Gemmell, J., Rizzo, L., Handley, M., Crowcroft, J., "The use of Forward Error Correction in Reliable Multicast", Internet Draft draft-ietf-rmt-info-fec-00.txt, November 2000.

[FECBBSTAN00] Luby, M., Gemmell, J., Vicisano, L., Rizzo, L., Handley, M., Crowcroft, J., "RMT BB: Forward Error Correction Codes", Internet Draft draft-ietf-rmt-bb-fec-01.txt, November 2000.

[HAN96] Handley, M., "SAP: Session Announcement Protocol", Internet Draft, IETF MMUSIC Working Group, Nov 1996.

[HAN98] Handley, M., and Jacobson, V., "SDP: Session Description Protocol", RFC 2327, April 1998.

[HOL99] Holbrook, H., Cheriton, D., "IP Multicast Channels: Experss Support for Large-scale Single-source Applications", ACM SIGCOMM'99

[LUB99] Luby, M., Vicisano, L., Speakman, T. "Heterogeneous multicast congestion control based on router packet filtering", presented at RMT meeting in Pisa, March 1999.

[MCC96] McCanne, S., Jacobson, V., Vetterli, M.,, "Receiver-driven Layered Multicast", Sigcomm'96, Stanford, CA, August 19996.

[VIC98A] L.Vicisano, L.Rizzo, J.Crowcroft, "TCP-like Congestion Control for Layered Multicast Data Transfer", IEEE Infocom'98, San Francisco, CA, Mar.28-Apr.1 1998.

[VIC98B] Vicisano, L., "Notes On a Cumulative Layered Organization of Data Packets Across Multiple Streams with Different Rates", University College London Computer Science Research Note RN/98/25, Work in Progress (May 1998).

Authors' Addresses

Michael Luby
luby@digitalfountain.com
Digital Fountain
600 Alabama Street
San Francisco, CA, USA, 94110

Lorenzo Vicisano
lorenzo@cisco.com
cisco Systems, Inc.
170 West Tasman Dr.,
San Jose, CA, USA, 95134

Armin Haken
armin@digitalfountain.com
Digital Fountain
600 Alabama Street
San Francisco, CA, USA, 94110