Internet Drat
<draft-lim-ip-reliable-multicast-01.txt>

Man Yeob Lim
Dae Young Kim
Chungnam Nat. Univ.
May 1998

IP Extension for Reliable Multicasting

Status of Memo

Abstract

This memo presents IP extension for recovering multicast packets from congestion. Dropped packets can be recovered far faster by IP routers with extension of this memo than by group member end-hosts. Because necessary interactions are limited among adjacent routers, this scheme substantially reduces overall signaling overhead among group members for packet recovery.

1. Introduction

Since the IP Multicast was proposed [1], there have been many research works on reliable multicast protocols. However the fact that the multicast itself is done in the IP layer but the solutions are sought in the transport or higher layer makes the search for solutions more difficult. The transport protocol sits on group members' end hosts which are spread over a large geographical area, and so if packet losses occur in the network, it not only takes long to detect in the transport layer but also makes group coordination very complicated. Even though many schemes were proposed to overcome the multicast communication losses[2-6], it is hard to devise a general solution without any attempt involving the IP protocol in the task.

There are two types of packet losses in Internet environments. One type is transparent to routers, while the other is not. The first type includes packet losses due to link error or router failure. Because these losses occur outside of the routers, in order to recover these packets an end-to-end ACK/NAK operation is required. The second type includes packet drops due to congestion. This type of packet loss is made by explicit router decision when the router encounters congestion. As the transmission quality improves, the first type of packet loss is diminishing and the congestion becomes major reason for packet loss. Because packet drops at congestion are done with routers' knowledge, we can think of a recovery scheme by explicit coordination among routers. If recovery of lost packets is done instantaneously and actively by the IP routers before later intervention by the higher protocol, not only the end-to-end multicast protocol can be significantly simplified but also the recovery can be done in a much faster fashion. A minimal requisite for the routers' capability at congestion in order to make the proposed scheme possible is that the router should be able to see the packet to collect necessary information before actually dropping one.

A study [7] shows that the loss on the links of the multicast network is observed to be only 2% or less of the whole packet loss and also that the rest congestion loss are again classified into two types, single and burst. Most of the congestion losses consist of isolated single losses, but a few of very long loss bursts, lasting from a few seconds up to 3 minutes(around 2000 consecutive packets) contribute heavily to the total packet loss. If retransmission request is made before congestion is relieved, the traffic will further become worse and congestion will be extended. The appropriate congestion control mechanism is required to cope with the burst loss.

The multicast packets are further categorized into two groups; real time packets and non real time packets. The real time multimedia

packets are time sensitive packets like audio and video that should be

recovered within a very short time if they are to be meaningful. These packets also have the best-effort characteristics, that is, the more packets are transferred the better quality is provided. Most upper layer protocols take long time to recover any loss depending on the location where the congestion occurs. If the congestion occurred upstream, the distance between the unreceived members and the received members are far apart. Even though the congestion occurred close to the receivers it takes time for the group members to coordinate to recover. Recovery by coordination among group members is not likely to meet real time requirement. Instead, the feasible way is to recover by coordination between adjacent routers.

The non real time multicast packets do not require critical timing requirement but typically are very critical to multicast coordination and reliability should be completely guaranteed. If these information is lost then the conference coordination collapses and too much endeavor to restore the control is required with time delay.

We propose extensions to IP and ICMP protocols for efficient recovery of both real time and non real time packets dropped from router congestion. We give multicast routers recovery cache or buffer so that lost packets can be recovered by coordination among routers. It is not suitable that all lost packets be recovered by routers. Recovery should be limited only to important multicast packets which are specially tagged, so that the cache/buffer size can be minimized and multicast routers are not required to do too much a processing overhead. In IP version 4, reliability bit in the type of service field can be used to identify multicast packets which require recovery. Low delay bit in the type of service can specify fast recovery using cache routers. In IP version 6, one entry in the priority field is suitable to specify reliable multicasting per packet or flow label can be used to specify reliable multicasting or fast recovery per data stream.

2. Recovery Operation

2.1. Overview

When a router receives more data than it can handle, congestion occurs. Once a congestion occurs the router receives no more data until the congestion is removed. Because the input queue is full, the incoming packets are not accepted in the queue but just ignored. We propose to allocate small size of extra queue to store header of dropped packets. The IP header is 20 bytes long and this information is enough to search the same packet from duplicated packet storage. The header includes source address, destination group address, identity and offset fields. The identity field identifies packets out of a packet stream which is going from one source to one destination. This field was defined for

packet segmentation and reassembly. When a packet is travelling through

a network the packet is segmented into pieces if the underlying network does not support long packet size. The segmented packets are reassembled in the destination host based on the identity and the offset field. If copies of the dropped packets are stored in a place it is possible to identify the same packet with the identity and the offset field. In order to retransmit the dropped packet, the congested router requests retransmission sending a request packet which includes header of the dropped packet.

Retransmission can be made by the source host whose IP address is in the header. As packets are dropped farther from the source host, it takes longer to recover by the sending host. We propose multicast routers have internal buffer to hold duplicate copies of the multicast packets as long as the packets can reach the next multicast routers. The multicast routers are not all equipped with buffer, but routers in every several hops in the routing tree are selected to have buffer. As a packet travels along the routing tree, it passes buffering routers once in several. If congestion occurs and a packet is lost, the packet is repaired by the buffering router which the packet passed at last. For this purpose we add an option field to the IP data packet to store the last buffering router's IP address. The address field is updated by every buffering router when the packet passes through the router.

If the buffering router receives repair request, the router searches the duplicate packet in the buffer and encapsulates the packet into a unicast one. The unicast packet is transferred to the requesting router. Then the requesting router restores multicast packet and resume multicasting.

When a host is located close to a multicast router and is available as a buffering device instead of the buffering router itself, the host can be used to store copies of multicast packets. If a host registers itself as a buffering device to a multicast router, the router sends all duplicated multicast packets which pass through the router to the buffering host and updates the buffering router address in the option field with the IP address of the buffering host.

2.2. Delay before retransmission

When congestion occurs there arises a question that how soon the router will be recovered from congestion and become ready to receive packets. If the congestion extends for a long period of time fast retransmission is useless or makes problem even worse. In reliable unicast protocol, congestion control is provision of TCP, which seems working acceptable because there are enough delay before the congestion is detected by the end host even though the response is not optimally fast. In multicast congestion makes reliable transport protocol harder and getting an

optimum congestion control even worse. The optimum scenario against

congestion is to detect a congestion as soon as possible, to reduce transmission rate and to wait until the congestion is removed before any retransmission is attempted. This IP extension detects congestion at the same time as congestion occurs because recovery operation is initiated by the router itself which is congested. We propose to make repair request after the congestion is removed. Because the header size is only 20 bytes long there is no big impact to the memory usage even though thousand headers are saved before requesting requests.

2.3. Recovering burst loss

When a burst loss occurs and thousand packets are lost, thousand repair requests should be sent. In order to reduce the number of repair request packets we can combine the requests into one or more combined repair request. Because the routers are several hops in between, and the routing table is not updated in small amount of time the IP packets are travelling in order as they were produced, i.e. in sequence same as the identity number. When congested router requests a repair, it can analyze the headers and combine as many headers in series into one, having same source and destination addresses. The repair request can be to transmit packets from source S to destination D with identity between M and N. Even though there is one or several holes in the burst, the retransmission can include them only to be ignored in the destination host. Using this expression the requesting router can reduce the memory holding headers of the lost burst packets.

2.4. Recovering real time packets

Real time packets are recovered by the so called cache router, which is located at just one previous hop from where congestion occurs. Cache routers continuously copy all multicast packets with service type of low delay in its ring type cache. When a cache router forwards a multicast packet, it updates the cache router address in the option. While this multicast packet travels along the network toward destination, the cache router address is updated every time the packet passes through cache routers. When a router has to drop a packet due to congestion, it sends a repair request to the cache router whose IP address is specified in the option field of the packet. Upon reception of the request, the cache router looks for the same packet in the cache. If it is still there, the cache router retransmits the packet. If repair request is made delayed the packet may not be recovered because the cache is overwritten with new incoming packets. Knowing that the real time packets are recovered by the just previous routers from congestion point, all routers should be equipped with cache in order for a network provides reliable multicast on real time packets.

2.5. Congestion control

When a congestion occurs, there need a mechanism to control congestion. Even though recovery mechanism against burst loss is prepared, if data rate from source is not suppressed the congestion lasts longer and the recovery operation might finally be overwhelmed. Because congestion is both detected and handled in the IP layer congestion control should be incorporated by IP protocol also. When a router detects internal queue usage reaches close to the congestion state, for example 90%, the router sends an Explicit Congestion Notice (ECN) to the source host, so that the source host reduces data rate. Because it takes certain amount of time for ECN travels to the source host and reduced packet stream reaches the congestion point, the pending packets may cause congestion. In order to reduce this possibility the ECN is sent earlier reserving larger queue free.

We propose additional congestion control scheme which distributes the congestion status to adjacent routers and share multiple queues against congestion. If adjacent routers are notified that congestion occurred on the next router, then the adjacent routers queue the packets which are to go to the congested router as much as the queue size allows. If the queue becomes full and another congestion happens or queue usage is over the predetermined threshold then congestion is notified to the adjacent routers again thus the congestion statues is propagated outward from the congestion point. If the congestion at the initial congestion point is relieved then this is notified to the adjacent routers and packet forwarding is resumed. This congestion control scheme can distribute congestion at specific router to many routers in wide area dynamically using queues of those routers combined efficiently. Disadvantage is that congestion may happen in many routers simultaneously, thus overall network goes down instead one single router. Instead of using this scheme separately, using together with the recovery operation and congestion control of source host may accomplish better performance.

3. Protocol Data Unit(PDU) description

3.1. Extension to IP datagram adding cache/buffering router address

An option is defined in IP datagram to store cache or buffering router IP address. Figure 1 shows packet format of the option in the IP datagram.

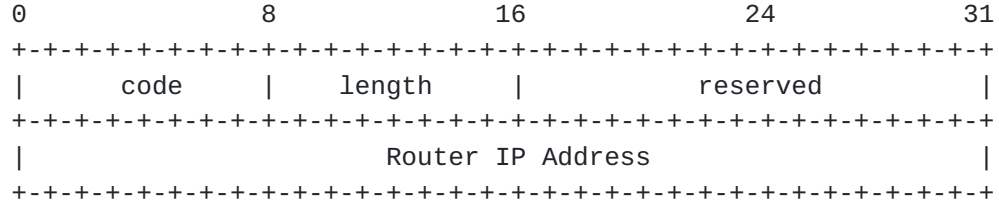


Figure 1. The format of the router option in an IP datagram

The source host initializes the option field as the IP address of the source host. Cache routers or buffering routers updates the field as its IP address, selectively based upon the packet type which is identified by the low delay priority bit; real time packet vs. non real time packet. If a host is used as a buffering device, this field is updated as the IP address of the buffering host. This makes it possible that recovery is implemented by the nearest router from the congested router.

3.2. Repair request ICMP Message

This message is sent to the cache/buffer router or buffering host when a drop from congestion occurs. Receiving the request the router searches the requested packet in the cache/buffer. The router sends the packet to the congested router. The buffering router converts the multicast packet into a unicast packet and sends to the congested router. One message can request single or multiple packets. The no_of_identity field specifies the number of packets with different identity number. Because there are multiple packets with same identity, this number is not same as the number of packets. Figure 2 shows the format of the drop ICMP message.

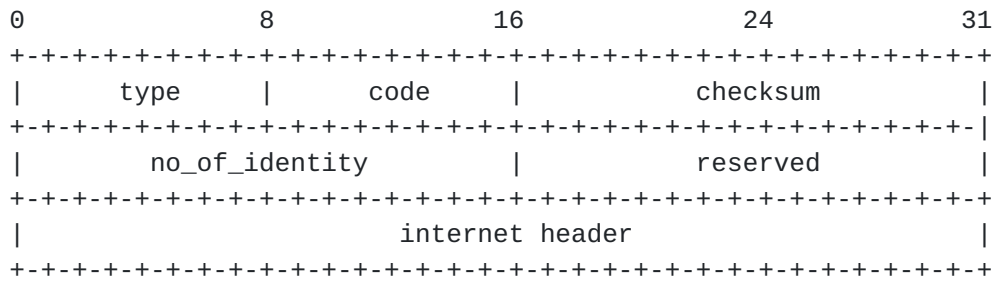


Figure 2. Repair request ICMP message format

3.3. Repair packet

When a buffering router receives a repair request then the router searches the packet in the buffer. If it succeeds finding the packet it converts the packet into a unicast packet saving the multicast address in the option field and changing the destination address to the congested router address. This repair packet is tunneling routers which received original packet without congestion so that packet is recovered where congestion occurred and no duplicate retransmissions happen. Receiving router should convert the unicast packet to a multicast packet and continue multicast routing. The figure 3 shows the format of the multicast address option. Retransmission packet by cache routers is forwarded in multicast format because the cache router is located adjacent to the congested router.

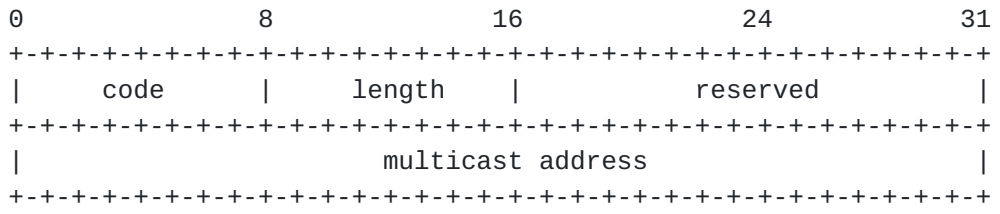


Figure 3. The format of the multicast address option in repair packet

3.4. Register buffering host ICMP message

This message is used for a host to register itself to a multicast router as a buffering host. When a multicast router receives this message the router forwards all multicast packets with reliability service type to the buffering host. Upon receiving repair request from a congested router the buffering host transmits to the congested router in unicast format.

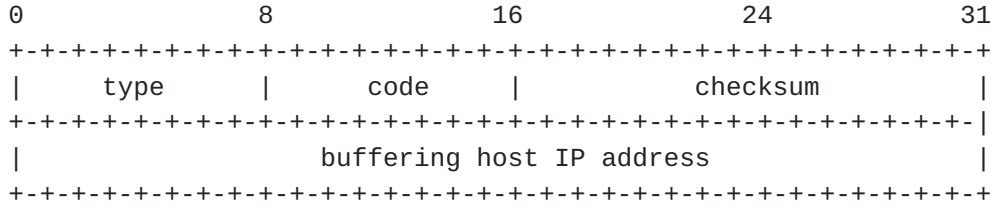


Figure 4. Register buffering host ICMP message format

3.5. Flow control ICMP message

This message is used to control flow rate from source. The flow control figure means the router state how fast the router can process packets. The figure is expressed from 0 to 255. Figure 0 means completely free and figure 255 means that router is completely blocked.

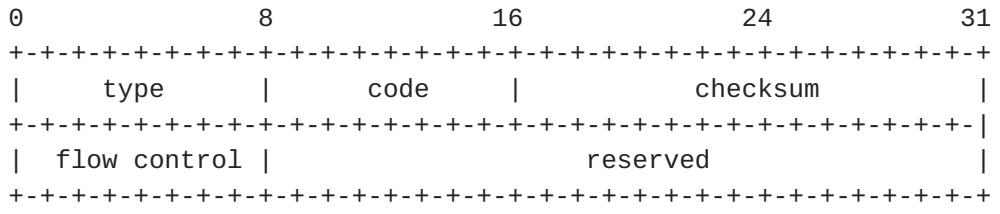
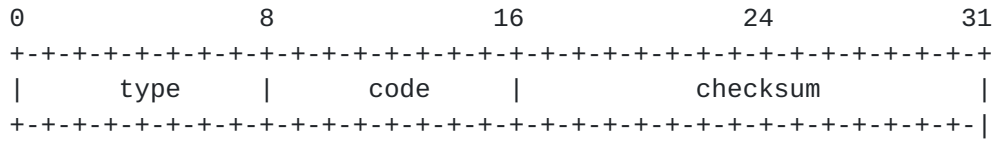


Figure 5. Flow control ICMP message format

3.6. Congestion propagation ICMP message

This message is used to notify adjacent routers that congestion occurred so that forwarding is suppressed. This message is broadcasted with TTL set to 1, i.e. only the surrounding routers can receive this

packet. The same message is also used when the congestion is removed with different code number. Therefore all routers have information whether the adjacent routers are working or blocked. If one router is blocked no packet is forwarded to the router. Instead the packets are buffered until it becomes unblocked. If this causes the router itself becomes congested then congestion is again propagated to the next surrounding routers and there is no more packets incoming.



code c1: congestion occurred
code c2: congestion removed

Figure 6. Congestion propagation ICMP message format

4. Implementation issues

Suppose cache routers are used in a gigabit network and routers are separated by 100 kilometers apart. Packet travel time is 0.3 millisecond for one way. If a congestion occurs, the cache router drops a received multicast packet and sends a repair request. If we assume the time to process a received packet and to generate a request is 0.4 millisecond, the cache router should store a duplicate copy of a multicast packet by 1 millisecond. This results in 1 Mbit cache required for each channel.

Considering buffer size of a buffering router, suppose round travel time between source and destination host is 10 seconds. If we give router's recovery time to recover from congestion 10 seconds the total time to store packets in the buffer will be 20 seconds. Supposing 1 percent of the total 1 gigabit traffic is multicasting traffic requiring retransmission, the buffer size will be 25 Mbyte. If we increase the recovery time to 3 minutes the buffer size becomes around **250 Mbyte, and we feel this figure is not difficult to implement.**

Authors:

Man Yeob Lim, Mr
InfoCom Eng. Dept.
Chungnam National University
Daejeon 305-764
Korea

Phone: +82 42 821 3544
Fax: +82 42 821 2225
Email: mylim@sunam.kreonet.re.kr

Dae Young Kim, Prof.
InfoCom Eng. Dept.
Chungnam National University
Daejeon 305-764
Korea

Phone: +82 42 821 6862
Fax: +82 42 823 5586
Email: dykim@ccl.chungnam.ac.kr

REFERENCES

[1] S. Deering, Host Extensions for IP Multicasting, RFC 1112, Jan. 1989.
[2] S. Kasera, J. Kurose, and D. Towsley, Scalable reliable multicast using multiple multicast groups, Proc. ACM Sigmetrics Conference, 1997
[3] S.Floyd, V. Jacobson, C. Liu, S. McCanne, and L. Zhang, A reliable multicast framework for light-weight sessions and application level framing, ACM SIGCOMM 95.
[4] S. Armstrong, A. Freier, K. Marzullo, Multicast Transport Protocol, RFC 1301, Feb. 1992.
[5] B. Whetten, T. Montgomery, S. Kaplan, A high performance totally ordered multicast protocol, Theory and Practice in Distributed Systems, Springer Verlag, LCNS 938.
[6] C. Papadopoulos, G. Paruklar, G. Varghese, An error control scheme for large-scale multicast applications, Washington University, St. Louis.
[7] M. Yajnik, J. Kurose, and D. Towsley, Packet loss correlation in the Mbone multicast network, University of Massachusetts at Amherst.