INTERNET DRAFT
draft-sola-pim-static-multicast-00.txt

Manolo Sola (1)
Masataka Ohta (2)

(1) Waseda University
(2) Tokyo Institute of Technology

August 1998

Modifications to PIM-SM for Static Multicast

Status of this Memo

Abstract

The Protocol Independent Multicast - Sparse Mode (PIM-SM) is currently defined as an intra-domain multicast protocol. Although in PIM-SM more than one Candidate Rendezvous Point (C-RP) may exist, only one can be active at a given time, and this will be the one to which receivers will send Join messages or sources will send Register messages. The method used in PIM-SM to make public the set of C-RPs for a multicast group is to flood all over the domain packets with the list of C-RPs using the so-called Bootstrap method. This approach may scale in domains with few routers but does not scale if the protocol would have to be applied to provide multicast throughout the whole internet.

In this draft we specify the modifications needed in PIM-SM in order to support more than one active RP simultaneously, and also show how Static Multicast can be used to make public in an scalable way the information regarding RPs as well as the IP multicast address for the multicast group. In the presence of more than one source, having more than one RP and placing them according to sources and receivers'

distribution on the internet will reduce the average number of hops from sources to receivers, and will also improve the multicast protocol's fault tolerance.

1. Introduction

PIM-SM [PIM-SMv2] delivers multicast traffic to receivers of a multicast group by means of a tree rooted at the RP for that multicast group.

As PIM-SM is an intra-domain multicast protocol, it runs over a tree which can not cross domain boundaries.

To connect trees in different domains for the same multicast group, an inter-domain multicast protocol is needed and PIM-SM has to be extended to interoperate with such a protocol.

One motivation behind this multicast architecture was to be able to improve multicast scalability, in what respect to the size of the multicast routing table, by the aggregation of multicast addresses in a similar way unicast addresses were aggregated.

However, it has been shown that it is not possible to aggregate multicast routing table entries, even if multicast addresses are assigned aggregated [Sola98]. That is, aggregation of multicast address assignment is meaningless.

Other argument which is usually given to justify this multicast architecture, is that administrators of different domains may desire to run different multicast protocols. However, an intra/inter-domain multicast architecture does not allow administrators of different domains to run the same multicast protocol (at least 2 different protocols, intra-domain and inter-domain, would have to be used).

This way of delivering multicast lacks also of location independence (the protocol is bounded to a domain) and fault tolerance (only one C-RP can be active at a given moment).

In PIM-SM the way to discover the RP for a multicast group is based on flooding the RP list to all routers in the domain. However, it is necessary to find a method with better scalability properties to distribute that information.

In [static-multicast] there is no need for inter/intra domain splitting, multiple independent RPs for the same multicast group are allowed, and the tree rooted at each RP is no longer limited to a single domain. Furthermore, a scalable solution is presented to the problems of how to delegate and allocate multicast addresses, and how

to discover RPs for a multicast group.

In order to complete the Static Multicast architecture, a protocol to intercommunicate independent multicast trees for the same multicast group has to be defined.

In this draft we describe the modifications needed in PIM-SM in order to support that architecture.

When an administrator desires to run a multicast group, first of all, an IP multicast address must be obtained from an IP multicast address' delegation authority.

The proposed delegation mechanism for multicast address in [static-multicast] is summarized in the following paragraph:

> An administrator who has been delegated the set of 2^8 addresses {X.Y.Z.0 to X.Y.Z.255} is also delegated the set of 2^4 multicast addresses {M.X.Y.Z, with M=224,...,239} as:
>
> ```
> 224.X.Y.Z     CNAME     mcast0.Z.Y.X.in-addr.arpa.
>     .           .                .
>     .           .                .
>     .           .                .
> 239.X.Y.Z     CNAME     mcast15.Z.Y.X.in-addr.arpa.
> ```
>
> , and that administrator becomes an IP multicast address delegation authority, so he or she can further subdelegate any subset of the 2^4 multicast addresses in the same way he or she can further subdelegate any subset of the 2^8 unicast addresses.

As some of the multicast address space has already been assigned, some addresses can not be used. To ensure that these addresses will never be used for static multicast, or to reserve some multicast address space for other purposes, the set of valid values for M may be reduced.

2. Summary of the new feautures introduced in this draft

We call Multi-RP PIM-SM (MRP PIM-SM) to the protocol resulting from the modifications to PIM-SM in order to accomplish with the Static Multicast architecture. The current version of MRP PIM-SM is based on PIM-SM version 2 [PIM-SMv2]. For brevity, we will refer to PIM-SM version 2 as PIM-SM.

In this draft Static Multicast is used for distributing RP and multicast group address information.

Also, this draft proposes that RP placement and migration strategy must be decided by the administrator of the multicast group, and that a router must be configured by its administrator in order to behave as an RP.

As a Designated Router (DR) may not have enough routing information to decide which is its nearest RP, the format of the Join/Prune message is modified and a list of RPs is included in the message instead of a unique RP. This message travels hop by hop, and a router forwards it towards the nearest RP in a way which will be commented later. As a Join/Prune is sent with a list of RPs instead of only one RP, another new message (Join Ack message) will inform the router about the RP that has been finally joined.

Other new message, the State message, is defined to inform RPs about other operative RPs for the same multicast group. The use of these messages will be explained later in this document.

For the case in which the DR doesn't need to join a RP but to send Registered messages to him, first, the nearest RP is discovered and then the traffic is forwarded to that RP.

Four new kind of control messages are proposed:

2.1. Join/Prune message

In order to make possible for non-member sources to send packets to their nearest RP, an additional field in the current Join/Prune message is defined. That field, the Control-Only-Join field, has no arguments. With this flag set, a Join/Prune message indicates that it is a join to receive only Join Ack messages with no intention to receive any data stream. With this flag unset, a Join/Prune message indicates that it is a join to receive Join Ack messages and any data stream.

An RP is said to be an RP with members for a multicast group if it has state instantiated due to the reception of Join/Prune messages for that multicast group having the Control-Only-Join field unset.

2.2. Join Ack message

For each multicast group, each RP must multicast periodically this message. It must include the IP address of the RP so that routers in that RP's tree for a multicast group will be able to know which RP they have joined. This message must also include a list containing the complete list of RPs obtained through DNS, and a RP-unreachable flag associated to each RP.

2.3. State message

It is sent periodically from a RP to each one of the RPs of the multicast group. A RP knows about other reachable RPs through this message.

2.4 Join Probe

It is used to detect RPs marked as unreachable that becomes reachable.

3. RP behaviour

3.1 Number of RPs and distribution of RP's information

The administrator of a multicast group G must decide how many RPs will exist for G and where they will be located and, following the rules in [static-multicast], DNS must be used to make worldwide available the information about RPs and the IP multicast address for the multicast group.

For PIM-SM, a new RR should be defined in DNS. We illustrate it with an example:

```
bbc.com          RVP      england.bbc.com
                 RVP      scotland.bbc.com
                 RVP      wales.bbc.com
                 RVP      north-ireland.bbc.com
```

The DNS packet format is identical to that of PTR [RFC1035], and the QTYPE value for a RVP query is <to be assigned by IANA>. The administrator of the multicast group must contact the administrator of a router that will behave as an RP so that the router can be configured properly. A router must know whether it is an RP by means of a configuration paramenter at the router which must be set by the administrator of that router.

When a router configured as an RP for multicast group G starts to run, it must query DNS to get the list of RPs for G.

3.2 How to discover the nearest RP

Lets suppose that a router with no state for multicast group G receives a Join message. This message must include:

a. A complete list of RPs for G as seen by the DR or RP that originated this Join message. The DR or RP used DNS to obtain that list. In a particular case which will be explained later, an

intermediate router can also modify this list.

b. A Control-Only-Join flag. A DR with no members that need to forward multicast data traffic from non-member senders must set this flag when sending a Join message. This will allow the DR to receive Join Ack messages to discover the nearest RP without receiving multicast data traffic which is not needed because there is no members.

The steps to discover the nearest RP are as follows:

Step 1.

In order to determine the next hop for the Join message:

(1) The entries in the routing table allowing to arrive to the RPs in the list of RPs which are not marked as unreachable are selected, excluding those associated to the incoming interface by which the Join message arrived;

(2) From this set, all entries with the same best metric are selected;

(3) From these entries, those with the same more specific route are selected;

(4) From this set of routing entries, one is selected randomly and becomes the next hop.

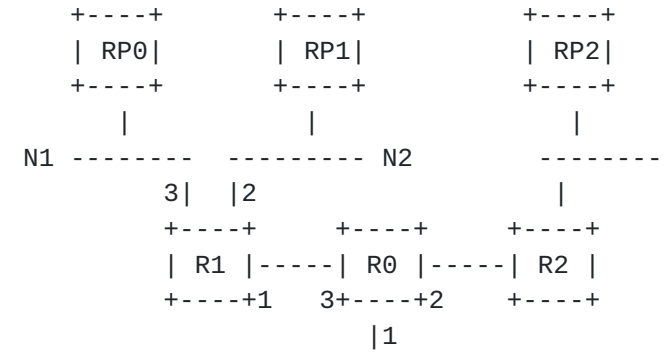

Fig.1: Example to ilustrate how Join messages are forwarded.

Lets illustrate these points with an example.

In Fig.1 there are three RPs (RP0, RP1, RP2). RP0 belongs to subnet N1=131.112.0.128/25, and RP1 to subnet N2=131.112.0.0/25. The IP addresses of RP0, RP1, and RP2 are 131.112.0.129, 131.112.0.1, and

131.112.1.1, respectively.

Lets suppose that there is no state at any router for multicast group G, that a Join message for G arrives to R0 through interface 1, and that the list of RPs in the message is {RP0, RP1, RP2}.

Lets also suppose that the routing entries at R0 allowing to arrive to {RP0, RP1, RP2} are:

destination	interface	cost	next hop
131.112.0/24	3	1	R1
default	2	1	R2

In order to determine the next hop to forward the Join message, previous points (1) to (4) has to be followed. In point (1) these two entries are selected. In point (2) again these (2) entries are selected because they have the same cost. In point (3) the entry associated to interface 3 is selected because the prefix 131.112.0/24 is more specific than the default entry.

R0 uses this entry to forward the Join message to R1.

Lets also suppose that at R1 the routing entries allowing to arrive to {RP0, RP1, RP2} according to (1) are

destination	interface	cost	next hop
131.112.0.128/25	3	0	RP0
131.112.0.0/25	2	0	RP1

In this case, both entries are selected in points (2) and (3). In point (4), and to solve the tie, one is selected randomly, and the Join message is forwarded to the corresponding RP.

Step 2.

The router must forward the Join message, and wait for a Join Ack or a time-out. The time-out for the forwarded Join message should be set to T/R, where T is the time-out of the incoming Join message at the router sending the message, and R is the number of entries in the set of routing entries.

If the forwarded Join message times-out, the RPs associated to entry in the routing table are marked as unreachable and the process must be restarted at (2).

If the join process succeeds, the router will start to receive Join

Ack for G from the RP which has been joined. This message includes the RP which has been joined.

3.3. DNS look-up

As a general rule for DRs or RPs involved in DNS look-ups, DNS has only to be queried if the TTL of the records in the last look-up is about to expire.

3.4. Forwarding tree

When a router configured as an RP for multicast group G starts to run, it must query DNS to get a list of RPs for G. Also an RP with receivers for G must send a (S,G) Join message towards each of the RPs in the list placing in S the address of the RP to join. RPs are the only routers that can initiate source specific joins.

MRP-PIM-SM always uses RP-trees and does not allow to switch from RP-trees to source's shortest path trees (SP-trees). The reason for not allowing SP-trees is due to the multicast architecture being used. The administrator of the multicast group decides the number of RPs and where they will be located. As is commented in [Sola98], the administrator should place RPs looking for an acceptable average number of hops between sources and receivers. Within that scheme, the scalability drawbacks of source-specific trees overcome their benefits.

When a Register message from a source arrives to an RP, the RP decapsulate the packet, encapsulate the packet again using as the source address its own address, and forwards the packet as in PIM-SM.

When a receiver receives a packet sent to the multicast group G, it decapsulates the packet and process it as in PIM-SM.

4. Designated Router (DR) and intermediate router's behaviour

4.1. First member for G

If a DR has no state for G and a join request for G arrives, the DR must query DNS to get a list of RPs for G. DRs join the nearest RP sending, according to the rules given previously, (*,G) messages with the Control-Only-Join falg unset and the complete list of RPs.

4.2. Sources and DR with no members

A Designated Router (DR) always sends packets from sources in Register messages towards the nearest RP. If the nearest RP is unknown, it is discovered sending a Join message with the Control-

Only-Join flag set and the complete list of RPs. The Join Ack to this message will include that RP.

4.3. Join messages with different lists of RPs

If a router with Join state for G receives a Join message for multicast group G, and the list of RPs in the message is not equal to the list of RPs at the router, the router must query DNS and build a new list of RPs. If the new list is not equal to the one at the router, the router must prune the current up-branch and send a new Join message based on the new list.

4.4. Unreachable RPs becoming reachable

A new message is defined, the Join Probe message, to detect when an RP marked as down becomes reachable.

Routers should send Join Probes to next hops (excluding the one through which the normal Join is sent) towards RPs nearer than the currently joined RP.

The Join Probe message is similar to the Join message except that it does not affect up data path and it includes only the RPs nearer than the currently joined RP.

If Join Ack is returned to Join Probe, the acked RP is removed from the list of unreachable RPs, and will be considered again for the next Join to be sent.

In this case, a Join Ack is sent selectively, that is, only to those routers which requested a Join Probe for that RP

5. Eliminate intra/inter-domain procedures

As the new multicast architecture does not make any intra/inter-domain assumptions, the procedures, timers, and data structures introduced in PIM-SM to interoperate with an inter-domain protocol should be eliminated.

6. Security Considerations

It should be noted that PIM-SM offers a floor control capability as good as that of unicast communication. That is, with a floor control at the RP based on source IP addresses, attacking senders, even with a forged source address, can not inject disturbing packets on the multicast tree unless they were located between an RP and receivers, even in which case only receivers downstream of the attacker are affected.

References

[Sola98] M. Sola, M. Ohta, T. Maeno, "Scalability of Internet Multicast Protocols," INET'9

[static-multicast] M. Ohta, J. Crowcroft, "Static Multicast," Work in progress, ftp://ftp.ietf.org/internet-drafts/draft-ohta-static-multicast-00.txt

[PIM-SMv2] D. Estrin et al. "Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification," RFC 2362, ftp://ftp.isi.edu/in-notes/rfc-2362.txt

[RFC1035] P. Mokapertis, "Domain Names - Implementation and Specification," RFC 1035, ftp://ftp.isi.edu/in-notes/rfc-1035.txt

Author's Address

Manolo Sola
Waseda University
School of Science and Engineering
Department of Information and Computer Science
3-4-1 Okubo, Shinjuku
Tokyo 169-8555, JAPAN

Phone: +81-3-5734-3299
Fax: +81-3-5734-3415
EMail: sola@jet.es

Masataka Ohta
Tokyo Institute of Technology
Computer Center
12-12-1, O-okayama, Meguro-ku
Tokyo 152, JAPAN

Phone: +81-3-5734-3299
Fax: +81-3-5734-3415
EMail: mohta@necom830.hpcl.titech.ac.jp