Network Working Group
INTERNET DRAFT

Dino Farinacci
Yakov Rekhter
cisco Systems
Peter Lothberg
Sprint
Hank Kilmer
Digex
Jeremy Hall
UUnet
June 25, 1998

Multicast Source Discovery Protocol (MSDP)
<draft-farinacci-msdp-00.txt>

Status of this Memo

Abstract

This proposal describes a mechanism to connect multiple PIM-SM domains together. Each PIM-SM domain uses it's own independent RP(s) and do not have to depend on RPs in other domains.

This proposal is being submitted as a method for the initial phase of Inter-Domain Multicast deployment in the Internet and may be upward compatible with the IDMR protocols being proposed for subsequent phases.

1.0 Introduction

This proposal describes a mechanism to connect multiple PIM-SM domains together. Each PIM-SM domain uses it's own independent RP(s) and do not have to depend on RPs in other domains.

Some advantages of this proposal:

- o PIM-SM domains can rely on their own RPs only.
- o Domains with only receivers get data without globally advertising group membership.
- o Global source state is not required.

2.0 Overview

An RP in a PIM-SM domain will have a MSDP peering relationship with an RP in another domain. The peering relationship will be made up of a TCP connection in which only control information is primarily exchanged. Each domain will have a connection to this virtual topology.

The purpose of this topology is to have domains discover multicast sources from other domains. If the multicast sources are of interest to a domain which has receivers, the normal source-tree building mechanism in PIM-SM will be used to deliver multicast data over an inter-domain distribution tree.

We envision this virtual topology will essentially be congruent to the existing BGP topology used in the unicast-based Internet today. That is the TCP connections between RPs can be realized by the underlying BGP routing system.

3.0 Procedure

A source in a PIM-SM domain originates traffic to a multicast group. The PIM DR which is directly connected to the source sends the data encapsulated in a PIM Register message to the RP in the domain.

The RP will construct a "Source-Active" (SA) message and send it to it's MSDP peers. The SA message contains the following fields:

- Source address of the data source.
- Group address the data source sends to.
- IP address of the RP.

Each MSDP peer receives and forwards the message away from the RP address in a "peer-RPF flooding" fashion. The notion of peer-RPF flooding is with respect to forwarding SA messages. The BGP routing table is examined to determine which peer is the next hop towards the originating RP of the SA message. Such a peer is called an "RPF peer".

If the MSDP peer receives the SA from a non-RPF peer towards the originating RP, it will drop the message. Otherwise, it forwards the message to all it's MSDP peers.

The flooding can be further constrained to children of the peer by interrogating BGP reachability information. That is, if a peer advertises a route (back to you) and you are the next to last AS in the AS-path, the peer is using you as the next-hop. In this case, you *should* forward an SA message (which was originated from the RP address covered by that route) to the peer. This is known in other circles as Split-Horizon with Poison Reverse.

When each MSDP peer (which are also RPs for their own domain) receive an SA message, they determine if they have any group members interested in the group the SA message describes. If the (*,G) entry exists with an non-empty outgoing interface list, the domain is interested in the group, and the RP triggers an (S,G) join towards the data source. This sets up a branch of the source-tree to this domain. Subsequent data packets arrive at the RP which are forwarded down the shared-tree inside the domain. If leaf routers choose to join the source-tree they have the option to do so according to existing PIM-SM conventions.

This procedure has been affectionately named flood-and-join because if any RP is not interested in the group, they can ignore the SA message. Otherwise, they join a distribution tree.

4.0 Controlling State

RPs which receive SA messages are not required to keep MSDP (S,G) state. However, if they do, newly formed MSDP peers can get MSDP (S,G) state sooner and therefore reduce join latency for new joiners.

RPs which originate SA messages do it periodically as long as there is data being sent by the source. RPs will not send more than 1 SA message for a given (S,G) within a 1 minute interval. Originating periodic SA messages are important so new receivers who join after a source has been active can get data quickly via the receiver's own RP when it is not caching SA state.

Intermediate RPs do not send periodic SA messages on behalf of sources in other domains. They only do for their own sources.

As the number of (source,group) pairs increases in the Internet, an RP may want to filter what sources it describes in SA messages. Also, filtering may be used as a matter of policy which at the same time can reduce state. Only the RP colocated in the same domain as the source can restrict SA messages. Other RPs should not filter or the flood-and-join model becomes broken.

If an MSDP peer decides to cache SA state, it may accept SA-Requests from other MSDP peers. When a MSDP peer receives an SA-Request for a group range, it will respond to the peer with a set of SA entries, in a SA-Response message, for all active sources sending to the group range requested in the SA-Request message. The peer that sends the request will not flood the responding SA-Response message to other peers.

5.0 SA Encapsulated Data Packets

For bursty sources, the SA message may contain multicast data from the source. Interested RPs can decapsulate the SA message and forward the original data packet down the shared-tree inside of a domain. We recommend this not be the default setting.

6.0 Auto-configuration versus Manual-configuration of MSDP Peers

MSDP peers can be configured manually or can be learned automatically. The two automatic mechanisms can be achieved by:

- PIM Query/Hello messages
- BGP capability parameter negotiation

In either case, each side of the peering relationship will indicate it's desire to participate in the MSDP protocol. If so, the TCP peer relationship is set up.

7.0 Other Scenarios

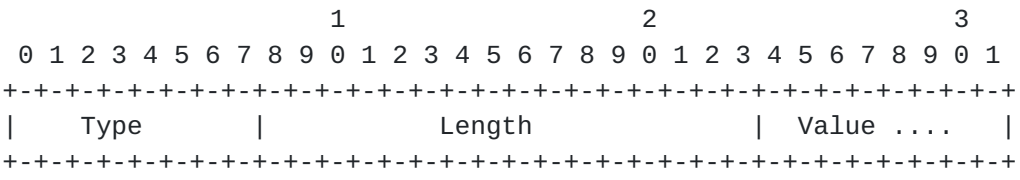
MSDP is not limited to deployment across different routing domains. It can be used within a routing domain when it is desired to deploy multiple RPs for different group ranges. As long as all RPs have a interconnected MSDP topology, each can learn about active sources as well as RPs in other domains.

MSDP can be used in domains that operate a dense-mode multicast routing protocol. However, in some cases SA messages with encapsulated source data may be required.

8.0 Packet Formats

MSDP messages will be encapsulated in a TCP connection using well-known port 639. The one side of the MSDP peering relationship will listen on the well-known port and the other side will do an active connect on the well-known port. The side with the higher IP address will do the listen. This connection establishment algorithm avoids call collision. Therefore, there is no need for a call collision procedure.

MSDP messages will be encoded in TLV format:



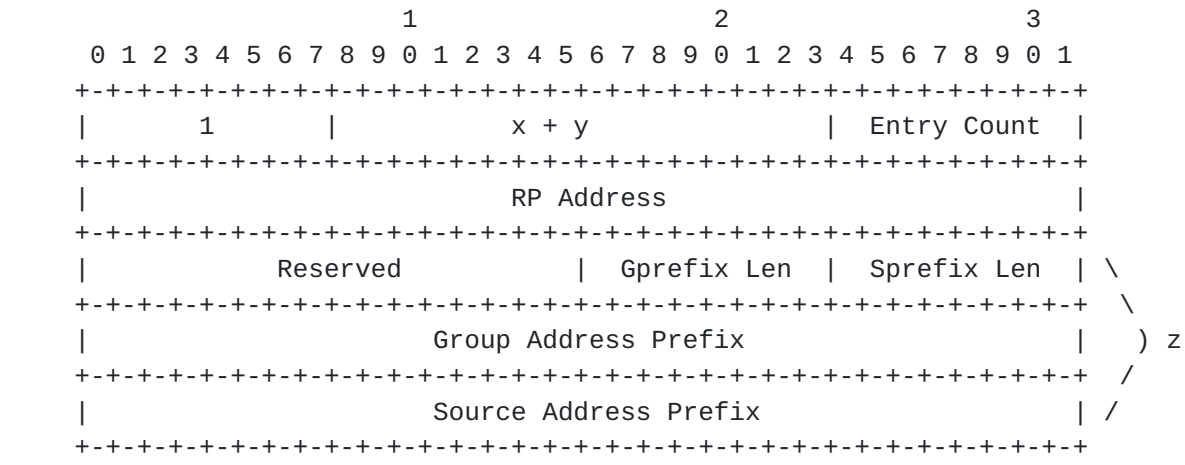
Type (8 bits)
Describes the format of the Value field.

Length (16 bits)
Length of Type, Length, and Value fields in octets. Minimum length required is 3 octets.

Value (variable length)
Format is based on the Type value. See below. The length of the value field is Length field minus 3.

Documented Types:

IPv4 Source-Active TLV



Type
IPv4 Source-Active TLV is type 1.

Length x
Is the length of the control information in the message. x is 8 octets (for the first two 32-bit quantities) plus 12 times Entry Count octets.

Length y
If 0, then there is no data encapsulated. Otherwise an IPv4 packet follows and y is the length of the total length field of the IPv4 header encapsulated. If there are multiple SA TLVs in a message, and data is also included, y must be 0 in all SA TLVs except the last one. And the last SA TLV must reflect the source and destination addresses in the IP header of the encapsulated data.

Entry Count
Is the count of z entries (note above) which follow the RP address field. This is so multiple (S,G)s from the same domain can be encoded efficiently for the same RP address.

RP Address
The address of the RP in the domain the source has become active in.

Gprefix Len and Sprefix Len
The route prefix length associated with the group address prefix and source address prefix, respectively.

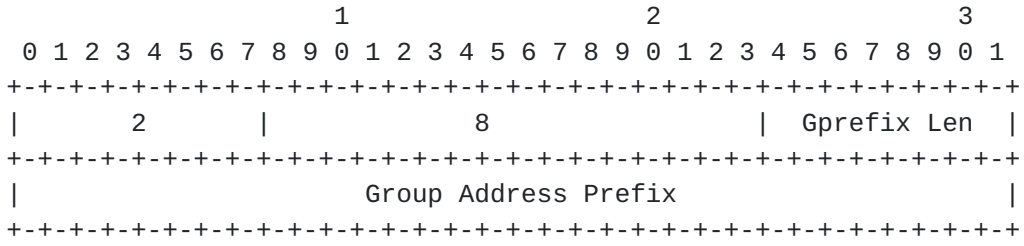
Group Address Prefix
The group address the active source has sent data to.

Source Address Prefix
The route prefix associated with the active source.

Multiple SA TLVs can appear in the same message and can be batched for efficiency at the expense of data latency. This would typically occur on intermediate forwarding of SA messages.

IPv4 Source-Active Request TLV

Used to request SA-state from a caching MSDP peer. If an RP in a domain receives a PIM Join message for a group, creates (*,G) state and wants to know all active sources for group G, and it has been configured to peer with an SA-state caching peer, it may send an SA-Request message for the group.



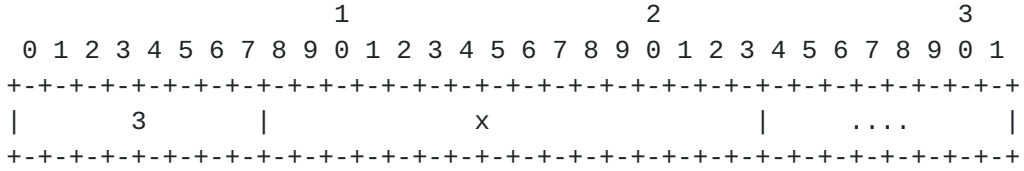
Type
IPv4 Source-Active Request TLV is type 2.

Gprefix Len
The route prefix length associated with the group address prefix.

Group Address Prefix
The group address prefix the MSDP peer is requesting.

IPv4 Source-Active Response TLV

Sent in response to a Source-Active Request message. The Source-Active Response message has the same format as a Source-Active message but does not allow encapsulation of multicast data.



Type
IPv4 Source-Active Response TLV is type 3.

Length x
Is the length of the control information in the message. x is 8 octets (for the first two 32-bit quantities) plus 12 times Entry Count octets.

9.0 Acknowledgements

The authors would like to thank David Meyer, John Meylor, Liming Wei, Manoj Leelanivas, Mark Turner, and John Zwiebel for their design feedback and comments.

10.0 Author's Address:

Dino Farinacci
Cisco Systems, Inc.
170 Tasman Drive
San Jose, CA, 95134
Email: dino@cisco.com

Yakov Rehkter
Cisco Systems, Inc.
170 Tasman Drive
San Jose, CA, 95134
Email: yakov@cisco.com

Peter Lothberg
Sprint
VARESA0104
12502 Sunrise Valley Drive
Reston VA, 20196
Email: roll@sprint.net

Hank Kilmer
Digex Inc.
One DIGEX Plaza
Beltsville, Maryland 20705
Email: hank@rem.com

Jeremy Hall
UUnet Technologies
3060 Williams Drive
Fairfax, VA 22031
Email: jhall@uu.net

11.0 References

[1] Estrin D., Farinacci, D., Helmy, A., Thaler, D., Deering, S., Handley M., Jacobson, V., Liu C., Sharma, P., Wei, L., "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification", draft-ietf-idmr-pim-sm-specv2-00.txt, September 9, 1997.

[2] Thaler, D., Estrin, D., Meyer, D., "Border Gateway Multicast Protocol (BGMP): Protocol Specification", draft-ietf-idmr-gum-01.txt, October 30, 1997.

[3] Rekhter, Y., and T. Li, "A Border Gateway Protocol 4 (BGP-4)", RFC 1771, March 1995.

[4] Bates, T., Chandra, R., Katz, D., and Y. Rekhter., "Multiprotocol Extensions for BGP-4", RFC 2283, February 1998.

[5] Deering, S., "Multicast Routing in a Datagram Internetwork", PhD thesis, Electric Engineering Dept., Stanford University, December 1991.

[6] Pusateri, T., "Distance Vector Multicast Routing Protocol", draft-ietf-idmr-dvmrp-v3-05.txt, October 1997.