MBONED Working Group | Mike McBride
Internet Draft | John Meylor
Cisco Systems
David Meyer
Sprint
Category: Best Current Practice
draft-ietf-mboned-msdp-deploy-00.txt
February, 2002

Multicast Source Discovery Protocol Deployment Scenarios

<draft-ietf-mboned-msdp-deploy-00.txt>

1. Status of this Memo

2. Abstract

This document describes best current practices for intra-domain and inter-domain MSDP deployment.

3. Copyright Notice

4. Introduction

The Multicast Source Discovery Protocol [MSDP] is a mechanism to connect multiple PIM-SM [RFC2117] domains together. Each PIM-SM domain uses its own independent Rendezvous Point, or RP, and does not have to depend on RPs in other domains. Current best practice for MSDP deployment utilizes Protocol Independent Multicast (Sparse Mode) and the Border Gateway Protocol With multi-protocol extensions [RFC2858,NICKLESS]. This document outlines how these protocols work together to provide Intra-domain and Inter-domain Any Source multicast (ASM) service. In addition, this document describes how MSDP can provide a PIM-SM domain with RP redundancy and load balancing using the Anycast RP mechanism [ANYCAST-RP].

5. Inter-domain MSDP peering scenarios

The following sections describe the different inter-domain MSDP peering possibilities and their deployment options.

5.1. Peering between PIM border routers (Single hop peering)

In this case, the MSDP peers within the domain each have their own RP located within a bounded PIM domain. In addition, a domain has it's own Autonomous Number (AS) and BGP speakers. The domain may also have multiple MSDP speakers. Each router has an MSDP and BGP peering with its peer routers. These deployments typically configure the BGP peering and MSDP peering using the same directly connected next hop peer IP address or another IP address from the same router. Typical deployments of this type are providers who have a direct peering with other providers or with providers who use their edge router to MSDP/MBGP peer with customers.

For a direct peering inter-domain environment to be successful, the

first AS in the BGP best path to the originating RP must be the same as the AS of the MSDP peer [MSDP]. As an example, consider the following topology:

```
AS1----AS2----AS4
        |    /
        |   /
        |  /
       AS3
```

In this case, AS4 receives an Source Active SA Message (SA), originated by AS1 via AS2, which also has an BGP peering with AS4. The BGP first hop AS from AS4, in the best path to the originating RP, is AS2. The origin AS of the sending MSDP peer is also AS2. The peer-RPF (Reverse Path Forwarding) check passes and the SA message is forwarded.

A peer-RPF failure will occur in this topology when the BGP first-hop AS in the best path to the originating RP is AS2 while the origin AS of the sending MSDP peer is AS3. An MSDP peering between AS2 and AS4 would prevent this failure from occurring.

5.2. Peering between non border routers (Multi-hop peering)

While the eBGP peer is typically directly connected between border routers, it is common for the MSDP peer to be located deeper into the transit providers AS. However, MSDP scalability is sacrificed if a provider must maintain BGP and MSDP peerings with all their edge routers so that they can BGP and MSDP peer with customer routers. Alternatively, providers commonly choose a few dedicated routers within their core network for the inter-domain MSDP peerings to their customers. These core MSDP routers will also typically be in the providers intra-domain MSDP mesh [MSDP] group and configured for Anycast RP. All multicast routers in the providers AS should statically point to the Anycast RP address. AutoRP and BSR mechanisms could be used to disseminate RP information within the provider's network.

For an SA message to be accepted in this (multi-hop peering) environment, the MSDP peer address must be in the same AS as the AS of the MBGP peer and must be advertised via MBGP. For example, using the diagram below, if customer R1 router is MBGP peering with AS2 provider's R2 router and if R1 is MSDP peering with R3 router, then R2 and R3 must be in the same AS. R1 also must have the MSDP peer

address of R3 in its BGP table.

```
     +--+     +--+     +--+
     |R1|----|R2|----|R3|
     +--+     +--+     +--+
      AS1      AS2      AS2
```

5.3. MSDP peering without BGP

In this case, an enterprise maintains its own RP and has an MSDP peering with their service provider, but does not BGP peer with them. MSDP relies upon BGP path information to learn the MSDP topology for the SA peer-RPF check. MSDP can be deployed without BGP, however, and as a result there are some special cases where the requirement to perform an peer-RPF check on the BGP path information is suspended. In this case (when there is only a single MSDP peer connection) a default peer (default MSDP route) is configured and either the originating RP is directly connected or a mesh group is used. An enterprise will also typically configure a unicast default route from their border router to the provider's border router and then MSDP peer with the provider's border router. If internal MSDP peerings are also used within the enterprise, then an MSDP default peer will need to be configured on the border router pointing to the provider. In this way, all external multicast sources will be learned and internal sources can be advertised.

5.4. MSDP peering between mesh groups

Mesh groups which are within different PIM domains can MSDP peer with one another to exchange information about active sources. An RP within AS1's mesh group may MSDP peer with an RP which is within AS2's mesh group. However, there should be no mesh group in common between PIM domains. It is important to note however, that mesh groups that span PIM domains is not recommended, as SA forwarding loops can develop. As an example, consider the following topology:

```
              AS2
              / \
             /   \
            /     \
           /       \
          /         \
       AS1 -------- AS3
```


If each AS had their own intra-domain MSDP mesh group, and if there was an inter-domain MSDP mesh group between AS1-AS2, AS1-AS3, and AS2-AS3 then an SA loop would be created. Since there is no RPF check between mesh groups, the SAs would loop around from one PIM domain to another.

5.5. MSDP peering at a Multicast Exchange

Multicast exchanges allow multicast providers to peer at a common IP subnet and share MSDP SA updates. Each provider will MSDP and BGP peer with each others directly connected exchange IP address. Each exchange router will send/receive SAs over the exchange fabric. They will then be able to forward SAs throughout their domain to their customers and any direct provider peerings.

6. Intra-domain MSDP peering scenarios

The following sections describe the different intra-domain MSDP peering possibilities and their deployment options.

6.1. Peering between routers configured for both MSDP and MBGP

The next hop IP address of the iBGP peer (that is MSDP is advertising as the next hop toward the originating RP) is used for the peer-RPF check. This is different from the inter-domain BGP/MSDP case, where AS path information is used for the peer-RPF check. For this reason, it is necessary for the IP address of the MSDP peer connection be the same as the internal BGP peer connection whether or not the MSDP/MBGP peers are directly connected. A successful deployment would be similar to the following:


```
                                 +----+
                                 | Rb | 3.3.3.3
                               / +----+
         AS1          AS2     /     |
        +---+        +--+   /       |
        |RP1|---------|Ra|          |
        +---+        +--+           |
        1.1.1.1     2.2.2.2         |
                             \      |
                              \     |
                               \ +-----+
                                 | RP2 |
                                 +-----+
```

Where RP2 MSDP and MBGP peers with Ra using 2.2.2.2 and with Rb using 3.3.3.3. When the MSDP SA update arrives on RP2 from Ra, the MSDP RPF check for 1.1.1.1 passes because RP2 receives the SA update from 2.2.2.2 which is the correct BGP next hop for 1.1.1.1.

When RP2 receives the same SA update from MSDP peer 3.3.3.3, the BGP lookup for 1.1.1.1 shows a next hop of 2.2.2.2 so RPF correctly fails, preventing a loop.

This deployment would also fail on an update from Ra to RP2 if RP2 was BGP peering to an address other than 2.2.2.2 on Ra. Intra-domain deployments should have MSDP and MBGP peering addresses which match.

6.2. MSDP peer is not BGP peer (or no BGP peer)

This is a common MSDP intra-domain deployment in environments where few routers are running BGP or where the domain is not running BGP. The problem here is that the MSDP peer address needs to be the same as the BGP peer address. To get around this requirement, the intra-domain MSDP RPF rules have been relaxed in certain as follows:

- o By configuring the MSDP peer as a mesh group peer,
- o By having the MSDP peer be the only MSDP peer,
- o By configuring a default MSDP peer, or
- o By peering with the originating RP.

The common choice around the intra-domain BGP peering requirement, when more than one MSDP peer is configured, is to deploy MSDP mesh groups. When a MSDP mesh group is deployed, there is no RPF check on arriving SA messages when received from a mesh group peer. Subsequently, SA messages are always accepted from mesh group peers.

MSDP mesh groups are helpful in reducing the amount of SA traffic in the network since SAs are not flooded to other mesh group peers.

7. MSDP and Anycast RPs

A network with can achieve RP load sharing and redundancy by using the Anycast RP mechanism in conjunction with MSDP mesh groups [ANYCAST-RP]. This mechanism is a common deployment technique used by service providers, who commonly deploy several RPs within their domain. These RPs will all have the same IP address configured on a Loopback interface (making this the anycast addresses). These RPs will MSDP peer with each other using a separate loopback interface and are part of the same MSDP mesh group. This second Loopback interface will typically also be used for the MBGP peering. All routers within the provider's domain will learn of the Anycast RP address either through AutoRP, BSR, or a static RP assignment. Each designated router in the domain will send source registers and group joins to the Anycast RP address. Unicast routing will direct those registers and joins to the nearest Anycast RP. If a particular Anycast RP router fails, unicast routing will direct subsequent registers and joins to the nearest Anycast RP. That RP will then forward an MSDP update to all peers within the global MSDP mesh group. Each RP will then forward (or receive) the SAs to (from) external customers and providers.

7.1. Hierarchical Mesh Groups

Hierarchial Mesh Groups are typically deployed in intra-domain environments where there are a large number of MSDP peers. Allowing multiple mesh groups to forward to one another can reduce the number of MSDP peerings per router and hence reduce router load. A good hierarchical mesh group implementation (one which prevents looping) contains a core mesh group in the backbone and these core routers serve as mesh group aggregation routers:

```
                   [R2]{A,2}
                   /  \
                  /    \
                 /      \
                /        \
               /          \
              /            \
             /              \
     {A,1}[R1]-------------[R3]{A,3}
```


In this example, R1, R2, R3 are in MSDP mesh group A (the core mesh group) and each serves as MSDP aggregation routers for their mesh groups 1, 2, and 3. Since SA messages received from a mesh group peer are not forwarded to peers within that same mesh group, SA messages will not loop. In particular, do not create topologies which connect mesh-groups in a loop. In the above example for instance, "second tier" mesh-groups 1, 2, and 3 must not directly exchange SA message.

7.2. MSDP and Route Reflectors

BGP requires all iBGP speakers that are not route-reflector clients or confederation members be fully meshed. This requirement does not scale when there are large number of iBGP speakers. In the route-reflector environment, MSDP requires that the route reflector clients peer with the route reflector. For example, consider the following case:

```
    Ra---------RR
              /|\
             / | \
            A  B  C
```

Ra is forwarding MSDP SAs to the route reflector RR. Routers A, B, and C also MSDP peer with RR. When RR forwards the SA to A, B, and C, these RR clients will accept the SA because RR is the iBGP next hop for the originating RP address.

An SA will peer-RPF fail if Ra MSDP peers directly with Routers A, B, and C because the iBGP next hop for RR's clients is RR, but the SA update came from Ra. Proper deployment is to have RR's clients MSDP peer with RR.

7.3. MSDP Filtering

Typically there is a fair amount of (S,G) state in a PIM-SM domain that is local to the domain. However, without proper filtering, SA-messages containing these local (S,G) announcements may be advertised to the global MSDP infrastructure. Examples of this includes domain local applications that use global IP multicast addresses and sources that use RFC 1918 addresses [RFC1918]. To improve on the scalability of MSDP and to avoid global visibility of domain local (S,G) information, the following external SA filter list is recommended to help prevent unnecessary creation, forwarding, and caching of some of these well-known ³domain local³ sources [IANA].

```
224.0.0.0/4      Local application packets
   (packets from any application which are intended to stay
    adminstratively scoped, but use global addressing. The
    current list of applications which could be filtered
    is dynamic and subject to individual policy.  See WG
    mail group for latest recommendations)
224.0.1.39       AutoRP Announce
224.0.1.40       AutoRP Discovery
239.0.0.0/8      Admin. Scoped
10.0.0.0/8       private addresses [RFC1918]
127.0.0.0/8      private addresses [RFC1918]
172.16.0.0/12    private addresses [RFC1918]
192.168.0.0/16   private addresses [RFC1918]
232.0.0.0/8      Default SSM-range
```

8. Author's Addresses

Mike McBride
Cisco Systems
mcbride@cisco.com

John Meylor
Cisco Systems
jmeylor@cisco.com

David Meyer
Sprint
Email: dmm@sprint.net

9. REFERENCES

[ANYCAST-RP] D. Meyer et. al, "Anycast RP mechanism using PIM and MSDP", draft-ietf-mboned-anycast-rp-08.txt, May, 2001.

[NICKLESS] Bill Nickless, "IPv4 Multicast Best Current Practice", draft-nickless-ipv4-mcast-bcp-01.txt, February 2002.

[IANA] http://www.iana.org

[MSDP] D. Meyer and Bill Fenner (Editors), "The Multicast Source Discovery Protocol (MSDP)", draft-ietf-msdp-spec-13.txt, November 2001.

[RFC1771] Rekhter, Y., and T. Li, "A Border Gateway Protocol 4 (BGP-4)", RFC 1771, March 1995.

[RFC1918] Y. Rekhter, R. Moskowitz, D. Karrenberg, G. de Groot, E. Lear, "Address Allocation for Private Internets", 02/29/1996.

[RFC2117] D. Estrin et. al, "Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification", RFC 2117, June, 1997.

[RFC2362] D. Estrin, et. al., "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification", RFC 2362, June, 1998.

[RFC2858] T. Bates, Y. Rekhter, R. Chandra, D. Katz, "Multiprotocol Extensions for BGP-4", RFC 2858, June 2000.

10. Full Copyright Statement