Multicast Source Discovery Protocol (MSDP) Deployment Scenarios
<draft-ietf-mboned-msdp-deploy-01.txt>

Status of this Document

   This document is an Internet-Draft and is in full conformance with
   all provisions of Section 10 of RFC2026.

   Internet-Drafts are working documents of the Internet Engineering
   Task Force (IETF), its areas, and its working groups. Note that
   other groups may also distribute working documents as Internet-
   Drafts.

   Internet-Drafts are draft documents valid for a maximum of six months
   and may be updated, replaced, or obsoleted by other documents at any
   time. It is inappropriate to use Internet-Drafts as reference
   material or to cite them other than as "work in progress."

   The list of current Internet-Drafts can be accessed at
   http://www.ietf.org/ietf/1id-abstracts.txt

   The list of Internet-Draft Shadow Directories can be accessed at
   http://www.ietf.org/shadow.html.

   This document is a product of an individual. Comments are solicited
   and should be addressed to the author(s).

Abstract

This document describes best current practices for intra-domain and
inter-domain deployment of the Multicast Source Discovery Protocol
(MSDP) in conjunction with Protocol Independent Multicast Sparse Mode
(PIM-SM).

Table of Contents

1. Introduction

MSDP [MSDP] is used primarily in two deployment scenarios:

o Between PIM Domains

   MSDP can be used between Protocol Independent Multicast Sparse
   Mode (PIM-SM) [RFC2362] domains to convey information
   about active sources available in other domains. MSDP peering
   used in such cases is generally one to one peering, and
   utilizes the deterministic peer-RPF (Reverse Path Forwarding)
   rules described in the MSDP specification (i.e., does not use
   mesh-groups). Peerings can be aggregated on a single MSDP
   peer. Such a peer can typically have from one to hundreds of
   peerings, which is similar in scale to BGP peerings.

o Within a PIM Domain

   MSDP is often used between Anycast Rendezvous Points
   (Anycast-RPs) [RFC3446] within a PIM domain to synchronize
   information about the active sources being served by each
   Anycast-RP peer (by virtue of IGP reachability). MSDP peering
   used in this scenario is typically based on MSDP mesh groups,
   where anywhere from two to tens of peers can comprise a given
   mesh group, although more than ten is not typical. One or more
   of these mesh-group peers may then also have additional
   one-to-one peering with MSDP peers outside that PIM domain for
   discovery of external sources. MSDP for anycast-RP without
   external MSDP peering is a valid deployment option and common.

Current best practice for MSDP deployment utilizes PIM-SM and the
Border Gateway Protocol with multi-protocol extensions (MBGP)
[RFC1771, RFC2858]. This document outlines how these protocols work
together to provide an intra-domain and inter-domain Any Source
Multicast (ASM) service.

2. Inter-domain MSDP peering scenarios

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

2.1. Peering between PIM border routers

   In this case, the MSDP peers within the domain have their own RP
   located within a bounded PIM domain. In addition, a domain has it's
   own Autonomous System (AS) number MBGP speakers. The domain may also
   have multiple MSDP speakers. Each border router has an MSDP and MBGP
   peering with its peer routers. These external MSDP peering
   deployments typically configure the MBGP peering and MSDP peering
   using the same directly connected next hop peer IP address or other
   IP address from the same router. Typical deployments of this type are
   providers who have a direct peering with other providers, providers
   peering at an exchange, or 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 MBGP best path to the originating RP should be the
   same as the AS of the MSDP peer. As an example, consider the
   following topology:


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


   In this case, AS4 receives a Source Active (SA) message, originated
   by AS1, from AS2. AS2 also has an MBGP peering with AS4. The MBGP
   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. In this
   case, the peer-Reverse Path Forwarding check (peer-RPF check) passes
   and the SA message is forwarded.

   A peer-RPF failure would occur in this topology when the MBGP first
   hop AS, in the best path to the originating RP, is AS2 while the
   origin AS of the sending MSDP peer is AS3. This reliance upon BGP AS
   PATH information prevents endless looping of SA packets.

   Router code, which has adopted the latest rules in the MSDP draft,
   will relax the rules Between AS's a bit. In the following topology we
   have an MSDP peering between AS1<->AS3 and AS3<->AS4:


```
                    RP
  AS1----AS2----AS3----AS4
```

If the first AS in best path to the RP does not equal the MSDP peer,
MSDP peer RPF fails. So AS1 cannot MSDP peer with AS3 since AS2 is
the first AS in the MBGP best path to AS4 RP. With the latest MSDP
draft compliant code, AS 1 will choose the peer in the closest AS
along best AS path to the RP. AS1 will then accept SA's coming from
AS3. If there are multiple MSDP peers to routers within the same AS,
the peer with the highest IP address is chosen as the RPF peer.

2.2. Peering between non border routers

When MSDP peering between border routers, intra-domain MSDP
scalability is restricted because it is necessary to also maintain
MBGP and MSDP peerings internally towards their border routers.
Within the intra-domain, the border router becomes the announcer of
the next hop towards the originating RP. This requires that all
intra-domain MSDP peerings must mirror the MBGP path back towards the
border router. External MSDP (eMSDP) peerings rely upon AS path for
peer rpf checking, while internal MSDP (iMSDP) peerings rely upon the
announcer of the next hop.

While the eMBGP peer is typically directly connected between border
routers, it is common for the eMSDP peer to be located deeper into
the transit providers AS. Providers, which desire more flexibility in
MSDP peering placement, 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 group and configured for Anycast RP.
All multicast routers in the providers AS should statically point to
the Anycast RP address. Static RP assignment is the most commonly
used method for group to RP mapping due to its deterministic nature.
Auto-RP [AUTORP] and/or the Bootstrap Router (BSR) [BSR] dynamic RP
mapping mechanisms could be also used to disseminate RP information
within the provider's network

For an SA message to be accepted in this (multi-hop peering)
environment, we rely upon the next (or closest, with latest MSDP
spec) AS in the best path towards originating RP for the rpf check.
The MSDP peer address should be in the same AS as the AS of the
border routers MBGP peer. The MSDP peer address should be advertised
via MBGP.

For example, using the diagram below, if customer R1 router is MBGP
peering with R2 router and if R1 is MSDP peering with R3 router, then
R2 and R3 must be in the same AS. The MSDP peer with the highest IP
address will be chosen as the MSDP RPF peer. R1 must also have the

MSDP peer address of R3 in its MBGP table.

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

From R3's perspective, AS1 (R1) is the MBGP next AS in the best path
towards the originating RP. As long as AS1 is the next AS (or
closest) in the best path towards the originating RP, RPF will
succeed on SAs arriving from R1.

In contrast, with the single hop scenario, with R2 (instead of R3)
border MSDP peering with R1 border, R2s MBGP address becomes the
announcer of the next hop for R3, towards the originating RP, and R3
must peer with that R2 address. And all AS2 intra-domain MSDP peers
need to follow iMBGP (or other IGP) peerings towards R2 since iMSDP
has a dependence upon peering with the address of the MBGP (or other
IGP) announcer of the next hop.

[2.3](). **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.
These cases are when there is only a single MSDP peer connection, a
default peer (default MSDP route) is configured, the originating RP
is directly connected, a mesh group is used, or an implementation is
is used which allows for an MSDP peer RPF check using an IGP.

An enterprise will typically configure a unicast default route from
their border router to the provider's border router and then MSDP
peer with the provider's MSDP 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. If only a single MSDP peering was used (no
internal MSDP peerings) towards the provider, then this stub site
will MSDP default peer towards the provider and skip the BGP RPF
check.

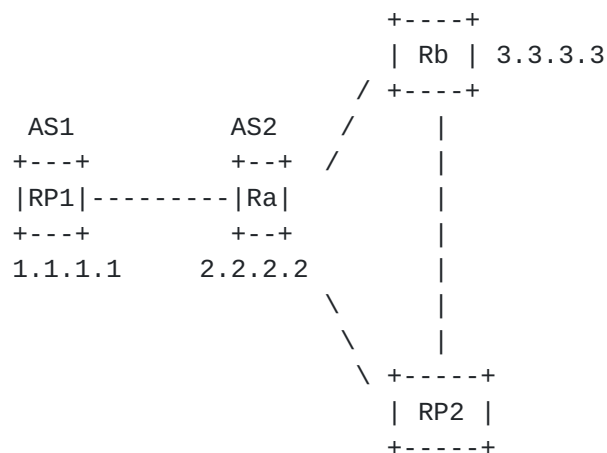## 2.4. MSDP peering at a Multicast Exchange

   Multicast exchanges allow multicast providers to peer at a common IP
   subnet (or by using point to point virtual LANs) and share MSDP SA
   updates. Each provider will MSDP and MBGP peer with each others
   directly connected exchange IP address. Each exchange router will
   send/receive SAs to/from their MSDP peers. They will then be able to
   forward SAs throughout their domain to their customers and any direct
   provider peerings.

3. Intra-domain MSDP peering scenarios

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

3.1. Peering between MSDP and MBGP configured routers

   The next hop IP address of the iBGP peer is typically used for the
   MSDP peer-RPF check (IGP can also be used). 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 MBGP
   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 MSDP peer 2.2.2.2 which is also the correct MBGP next hop for
1.1.1.1.

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

 was MBGP peering to an address other than 2.2.2.2 on
 Ra. Intra-domain deployments must have MSDP and MBGP (or other
 IGP) peering addresses which match, unless a method to skip the
 peer rpf check is deployed.

3.2. 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 MBGP or where the domain is not running MBGP.
The problem here is that the MSDP peer address needs to be the same
as the MBGP peer address. To get around this requirement, the intra-
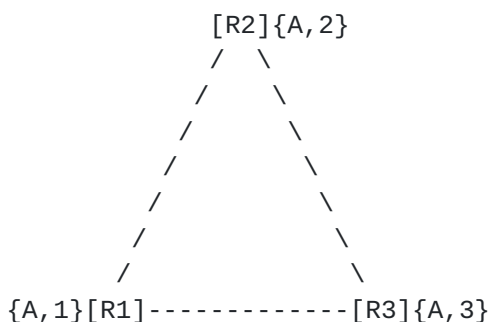domain MSDP RPF rules have been relaxed in the following topologies:

 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

 o By peering with the originating RP.

 o By relying upon an IGP for MSDP peer RPF

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 were developed to reduce the amount of SA traffic in
the network since SAs, which arrive from a mesh group peer, are not
flooded to peers within that same mesh group. Mesh groups must be
fully meshed.

If recent (but not currently widely deployed) router code is running
which is fully complaint with the latest MSDP draft, another option,
to work around not having BGP to MSDP RPF peer, is to RPF using an
IGP like OSPF, IS-IS, RIP, etc. This new capability will allow for
Enterprise customers, who are not running BGP and who don't want to
run mesh groups, to use their existing IGP to satisfy the MSDP peer
RPF rules.

3.3. Hierarchical Mesh Groups

Hierarchal Mesh Groups are occasionally 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 (due to the full mesh requirement) 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 leaf (or
second tier) 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. 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
messages with each other or an endless SA loop will occur.

Redundancy, between mesh groups, will also cause a loop and is
subsequently not available with Hierarchical mesh groups. For
instance, assume R3 had two routers connecting it's leaf mesh group 3
with the core mesh group A. A loop would be created between mesh
group 3 and mesh group A because each mesh group must be fully meshed

between peers.

3.4. MSDP and Route Reflectors

   or confederation members, be fully meshed to prevent loops. In
   the route reflector environment, MSDP requires that the route
   reflector clients peer with the route reflector since the RR is
   the BGP announcer of the next hop towards the originating
   RP. The RR is not the BGP next hop, but is the announcer of the
   BGP next hop. The announcer of the next hop is the address
   typically used for MSDP peer RPF checks. For example, consider
   the following case:


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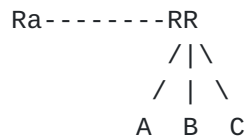

   Ra is forwarding MSDP SAs to the route reflector RR. Routers A, B,
   and C also MSDP peer with RR. When RR forwards the SA to A, B, and C,
   these RR clients will accept the SA because RR is the announcer of
   the next hop to the originating RP address.

   An SA will peer-RPF fail, if Ra MSDP peers directly with Routers A,
   B, or C, because the announcer of the next hop is RR, but the SA
   update came from Ra. Proper deployment is to have RR clients MSDP
   peer with the RR. MSDP mesh groups may be used to work around this
   requirement. External MSDP peerings will also prevent this
   requirement since the next AS is compared between MBGP and MSDP
   peerings, rather than the IP address of the announcer of the next
   hop.

   Some recent MSDP implementations conform to the latest MSDP draft
   which relaxes the requirement of peering with the Advertiser of the
   Next Hop (the Route Reflector). This new rule allows for peering with
   the Next-Hop, in addition to the Advertiser of the next hop. In the
   example above, for instance, if Ra is the Next-Hop (perhaps due to
   using BGP's Next hop self attribute) and if routers A,B,C are peering
   with Ra, the SA's received from Ra will now succeed.

3.5. **MSDP and Anycast RPs**

   A network, with multiple RPs, can achieve RP load sharing and
   redundancy by using the Anycast RP mechanism in conjunction with MSDP
   mesh groups [RFC3446]. This mechanism is a common deployment
   technique used within a domain by service providers and Enterprises
   which deploy several RPs within their domain. These RPs will each
   have the same IP address configured on a Loopback interface (making
   this the Anycast address). These RPs will MSDP peer with each other
   using a separate loopback interface and are part of the same fully
   meshed MSDP mesh group. This loopback interface, used for MSDP
   peering, 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 Auto-RP, 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 Anycast MSDP mesh
   group. Each RP will then forward (or receive) the SAs to (from)
   external customers and providers.

4. **Intellectual Property**

   The IETF takes no position regarding the validity or scope of any
   intellectual property or other rights that might be claimed to
   pertain to the implementation or use of the technology described in
   this document or the extent to which any license under such rights
   might or might not be available; neither does it represent that it
   has made any effort to identify any such rights. Information on the
   IETF's procedures with respect to rights in standards-track and
   standards-related documentation can be found in BCP-11. Copies of
   claims of rights made available for publication and any assurances of
   licenses to be made available, or the result of an attempt made to
   obtain a general license or permission for the use of such
   proprietary rights by implementors or users of this specification can
   be obtained from the IETF Secretariat.

   The IETF invites any interested party to bring to its attention any
   copyrights, patents or patent applications, or other proprietary
   rights which may cover technology that may be required to practice
   this standard. Please address the information to the IETF Executive
   Director.

5. Acknowledgments

   The authors would like to thank John Zwiebel and Swapna Yelamanchi
   for their feedback on earlier versions of this document.

6. Security Considerations

   An MSDP service should be secured by explicitly controlling the state
   which is created by, and passed within, the MSDP service. As with
   unicast routing state, MSDP state should be controlled locally, at
   the edge origination points. Selective filtering at the multicast
   service edge helps ensure that only intended sources result in sa-
   message creation, and this control helps to reduce the likelihood of
   state-aggregation related problems in the core. There are a variety
   of points where local policy should be applied to the MSDP service.

6.1. Filtering SA messages

   The process of originating sa-messages should be filtered to ensure
   only intended local sources are resulting in sa-message origination.
   In addition, MSDP speakers should filter on which sa-messages get
   received and forwarded.

   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 domain local sources.

   224.0.0.0/4   Specific local application packets   [IANA]
   224.0.1.39   Auto-RP Announce   [AUTORP]
   224.0.1.40   Auto-RP Discovery   [AUTORP]
   239.0.0.0/8   Administratively Scoped IP Multicast [RFC2365]
   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                         [SSM]
```

6.2. SA message state limits

   Proper filtering on sa-message origination, receipt, and forwarding
   will significantly reduce the likelihood of unintended and unexpected
   spikes in MSDP state However, a sa-cache state limit SHOULD BE be
   configured as a final safeguard to state spikes.

7. IANA Considerations

   This document creates a no new requirements on IANA namespaces
   [RFC2434].

8. References

8.1. Normative References

[MSDP]   Meyer, D. and W. Fenner (Editors), "The Multicast
                Source Discovery Protocol (MSDP)", draft-ietf-msdp-
spec-19.txt,
                May 2003. Work in Progress.

[SSM]   Holbrook, H., and B. Cain, "Source-Specific
                Multicast for IP", draft-ietf-ssm-arch-03.txt,
                May, 2003. Work in Progress.

[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", RFC 1918, Feburary, 1996.

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

[RFC2365]   Meyer, D. "Administratively Scoped IP Multicast",
                RFC 2365, July, 1998.

[RFC2434]   Narten, T., and H. Alvestrand, "Guidelines for
                Writing an IANA Considerations Section in
                RFCs", RFC 2434/BCP 0026, October, 1998.

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

[RFC3446]   Kim, D., et. al., "Anycast Rendezvous Point (RP)
                Mechanism using Protocol Independent Multicast
                (PIM) and Multicast Source Discovery Protocol
                (MSDP)", RFC 3446, January, 2003.

8.2. Informative References

[AUTORP]   Fenner, W., et. al., " Protocol Independent
                Multicast - Sparse Mode (PIM-SM): Protocol
                Specification (Revised)", draft-ietf-pim-sm-v2-new-07.txt,
                March, 2003. Work in Progress.

   [BSR]   Fenner, W., et. al., "Bootstrap Router (BSR)
                   Mechanism for PIM Sparse Mode", draft-ietf-pim-sm-
bsr-03.txt,
                   February, 2003. Work in Progress.

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

9. Author's Addresses

   Mike McBride
   Isac Systems
   Email: mcbride@cisco.com

   John Meylor
   Cisco Systems
   Email: jmeylor@cisco.com

   David Meyer
   Email: dmm@maoz.com

10. Full Copyright Statement