Network Working Group   M. McBride
Request for Comments: 4611   J. Meylor
BCP: 121   D. Meyer
Category: Best Current Practice   August 2006

        Multicast Source Discovery Protocol (MSDP) Deployment Scenarios

Status of This Memo

   This document specifies an Internet Best Current Practices for the
   Internet Community, and requests discussion and suggestions for
   improvements. Distribution of this memo is unlimited.

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 [RFC3618] is used primarily in two deployment scenarios:

    o Between PIM Domains

       MSDP can be used between Protocol Independent Multicast Sparse
       Mode (PIM-SM) [RFC4601] 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., it 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
       also have additional one-to-one peerings 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)
    [RFC4271, RFC2858]. This document outlines how these protocols work
    together to provide an intra-domain and inter-domain Any Source
    Multicast (ASM) service.

    The PIM-SM specification assumes that SM operates only in one PIM
    domain. MSDP is used to enable the use of multiple PIM domains by
    distributing the required information about active multicast sources
    to other PIM domains. Due to breaking the Internet multicast
    infrastructure down to multiple PIM domains, MSDP also enables the
    possibility of setting policy on the visibility of the groups and
    sources.

    Transit IP providers typically deploy MSDP to be part of the global
    multicast infrastructure by connecting to their upstream and peer
    multicast networks using MSDP.

Edge multicast networks typically have two choices: to use their
Internet providers' RP, or to have their own RP and connect it to
their ISP using MSDP. By deploying their own RP and MSDP, they can
use internal multicast groups that are not visible to the provider's
RP. This helps internal multicast be able to continue to work in the
event that there is a problem with connectivity to the provider or
that the provider's RP/MSDP is experiencing difficulties. In the
simplest cases, where no internal multicast groups are necessary,
there is often no need to deploy MSDP.

1.1. BCP, Experimental Protocols, and Normative References

   This document describes the best current practice for a widely
   deployed Experimental protocol, MSDP. There is no plan to advance
   the MSDP's status (for example, to Proposed Standard). The reasons
   for this include:

   o MSDP was originally envisioned as a temporary protocol to be
      supplanted by whatever the IDMR working group produced as an
      inter-domain protocol. However, the IDMR WG (or subsequently, the
      BGMP WG) never produced a protocol that could be deployed to
      replace MSDP.

   o One of the primary reasons given for MSDP to be classified as
      Experimental was that the MSDP Working Group came up with
      modifications to the protocol that the WG thought made it better
      but that implementors didn't see any reasons to deploy. Without
      these modifications (e.g., UDP or GRE encapsulation), MSDP can
      have negative consequences to initial packets in datagram streams.

   o Scalability: Although we don't know what the hard limits might be,
      readvertising everything you know every 60 seconds clearly limits
      the amount of state you can advertise.

   o MSDP reached nearly ubiquitous deployment as the de facto standard
      inter-domain multicast protocol in the IPv4 Internet.

   o No consensus could be reached regarding the reworking of MSDP to
      address the many concerns of various constituencies within the
      IETF. As a result, a decision was taken to document what is
      (ubiquitously) deployed and to move that document to Experimental.
      While advancement of MSDP to Proposed Standard was considered, for
      the reasons mentioned above, it was immediately discarded.

   o The advent of protocols such as source-specific multicast and bi-
      directional PIM, as well as embedded RP techniques for IPv6, have
      further reduced consensus that a replacement protocol for MSDP for
      the IPv4 Internet is required.

The RFC Editor's policy regarding references is that they be split
into two categories known as "normative" and "informative".
Normative references specify those documents that must be read for
one to understand or implement the technology in an RFC (or whose
technology must be present for the technology in the new RFC to work)
[RFCED]. In order to understand this document, one must also
understand both the PIM and MSDP documents. As a result, references
to these documents are normative.

The IETF has adopted the policy that BCPs must not have normative
references to Experimental protocols. However, this document is a
special case in that the underlying Experimental document (MSDP) is
not planned to be advanced to Proposed Standard.

The MBONED Working Group has requested approval under the Variance
Procedure as documented in RFC 2026 [RFC2026]. The IESG followed the
Variance Procedure and, after an additional 4 week IETF Last Call,
evaluated the comments and status, and has approved this document.

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT",
"SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this
document are to be interpreted as described in RFC 2119 [RFC2119].

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, the domain will
typically have its own Autonomous System (AS) number and one or more
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
```
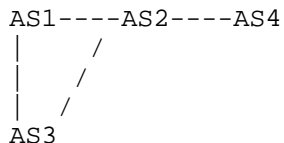
   In this case, AS4 receives a Source Active (SA) message, originated
   by AS1, from AS2. AS2 also has an MBGP peering with AS4. The MBGP
   first hop AS from AS4, in the best path to the originating RP, is
   AS2. The AS of the sending MSDP peer is also AS2. In this case, the
   peer-Reverse Path Forwarding 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 and 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 document,
   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
   document compliant code, AS1 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

   For 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 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 provider's AS. Providers, which desire more flexibility
in MSDP peering placement, commonly choose a few dedicated routers
within their core networks for the inter-domain MSDP peerings to
their customers. These core MSDP routers will also typically be in
the provider's intra-domain MSDP mesh group and be configured for
Anycast RP. All multicast routers in the provider's 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 [RFC4601] and/or the Bootstrap Router
(BSR) [BSR] dynamic RP mapping mechanisms could also be 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 the originating RP for the RPF
check. The MSDP peer address should be in the same AS as the AS of
the border router's MBGP peer. The MSDP peer address should be
advertised via MBGP.

For example, in the diagram below, if customer R1 router is MBGP
peering with the R2 router and if R1 is MSDP peering with the R3
router, then R2 and R3 must be in the same AS (or must appear, to
AS1, to be from the same AS in the event that private AS numbers are
deployed). 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, R2's MBGP address becomes the
announcer of the next hop for R3, towards the originating RP, and R3
must peer with that R2 address. Moreover, 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 its 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 a peer-RPF check on the BGP path information is suspended.
   These cases are:

   o There is only a single MSDP peer connection.

   o A default peer (default MSDP route) is configured.

   o The originating RP is directly connected.

   o A mesh group is used.

   o An implementation is used that allows for an MSDP peer-RPF check
      using an IGP.

   An enterprise will typically configure a unicast default route from
   its 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 that points 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 peer-RPF
   check.

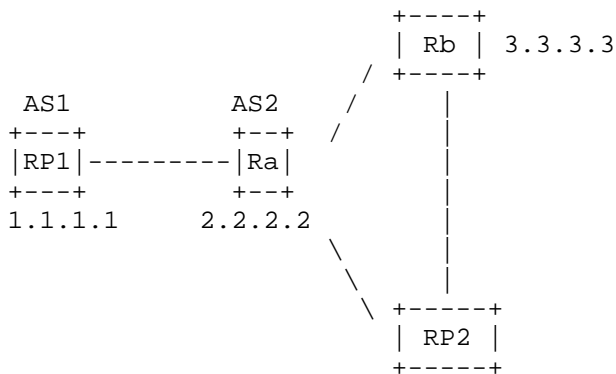2.4. MSDP Peering at a Multicast Exchange

   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 to 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.

   This deployment could also fail on an update from Ra to RP2 if RP2
   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 that 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:
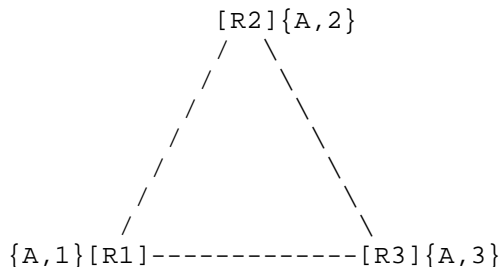
   o By configuring the MSDP peer as a mesh group peer.

   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 an MSDP mesh group is deployed, there is no RPF check
   on arriving SA messages when they are 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
   that is fully compliant with the latest MSDP document, 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

   Hierarchical 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 that 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, and 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 that 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 that R3 had two routers connecting its 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

BGP requires all iBGP speakers that are not route-reflector clients
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 router reflector (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
```
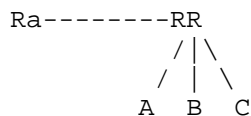
Ra is forwarding MSDP SAs to the route reflector RR. Routers A, B,
and C also MSDP peer with RR. When RR forwards the SA to A, B, and
C, these RR clients will accept the SA because RR is the announcer of
the next hop to the originating RP address.

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 document, 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, and 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 that deploy several RPs within their domains. 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 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. Security Considerations

An MSDP service should be secured by explicitly controlling the state that 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.

4.1. Filtering SA Messages

The process of originating SA messages should be filtered to ensure that only intended local sources are resulting in SA message origination. In addition, MSDP speakers should filter 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 include 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, an external SA filter list is recommended to help
   prevent unnecessary creation, forwarding, and caching of well-known
   domain local sources.

4.2. 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, an SA-cache state limit SHOULD be
   configured as a final safeguard to state spikes. When an MSDP
   peering has reached a stable state (i.e., when the peering has been
   established and the initial SA state has been transferred), it may
   also be desirable to configure a rate limiter for the creation of new
   SA state entries.

5. Acknowledgements

   The authors would like to thank Pekka Savola, John Zwiebel, Swapna
   Yelamanchi, Greg Shepherd, and Jay Ford for their feedback on earlier
   versions of this document.

6. References

6.1. Normative References

   [RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas,
             "Protocol Independent Multicast - Sparse Mode (PIM-SM):
             Protocol Specification (Revised)", RFC 4601, August 2006.

   [RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway
             Protocol 4 (BGP-4)", RFC 4271, January 2006.

   [RFC1918] Rekhter, Y., Moskowitz, B., Karrenberg, D., de Groot, G.,
             and E. Lear, "Address Allocation for Private Internets",
             BCP 5, RFC 1918, February 1996.

   [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate
             Requirement Levels", BCP 14, RFC 2119, March 1997.

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

   [RFC3446] Kim, D., Meyer, D., Kilmer, H., and D. Farinacci, "Anycast
             Rendevous Point (RP) mechanism using Protocol Independent
             Multicast (PIM) and Multicast Source Discovery Protocol
             (MSDP)", RFC 3446, January 2003.

   [RFC3618] Fenner, B. and D. Meyer, "Multicast Source Discovery
             Protocol (MSDP)", RFC 3618, October 2003.

6.2. Informative References

   [BSR]   Fenner, W., et. al., "Bootstrap Router (BSR) Mechanism for
             PIM Sparse Mode", Work in Progress, February 2003.

   [RFCED]   http://www.rfc-editor.org/policy.html

Authors' Addresses

   Mike McBride
   Cisco Systems

   EMail: mcbride@cisco.com

   John Meylor
   Cisco Systems

   EMail: jmeylor@cisco.com

   David Meyer

   EMail: dmm@1-4-5.net