MBONED Working Group — Dorian Kim
Internet Draft — Verio
David Meyer
Cisco Systems
Henry Kilmer
Dino Farinacci

Category: Informational
October, 1999

Anycast RP mechanism using PIM and MSDP

<draft-ietf-mboned-anycast-rp-00.txt>

1. Status of this Memo

This document is an Internet-Draft and is in full conformance with all provisions of Section 10 of RFC 2026.

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.

2. Abstract

This document describes a mechanism to allow for an arbitrary number of RPs per group in a single share-tree PIM-SM domain.

This memo is a product of the MBONE Deployment Working Group (MBONED) in the Operations and Management Area of the Internet Engineering Task Force. Submit comments to <mboned@ns.uoregon.edu> or the authors.

3. Copyright Notice

Copyright (C) The Internet Society (1999). All Rights Reserved.

4. Introduction

PIM-SM as currently defined allows for only a single active RP per group, and as such the decision of optimal RP placement can become problematic for a multi-regional network deploying PIM-SM.

The single active RP, or flat RP space design of PIM-SM has several implications, including traffic concentration, lack of scalable load balancing and redundancy between RPs, sub-optimal forwarding of multicast packets, and distant RP dependencies. These properties of PIM-SM have been demonstrated in recent native continental or inter-continental scale multicast deployments. As a result, it became clear that ISP backbones require a mechanism that allows definition of multiple active RPs per group in single PIM-SM domain. Further, any such mechanism should also addresses the issues addressed above.

The mechanism described here is intended to address the need for redundancy and load sharing among RPs in a domain. It is primarily intended for application within those networks which are using MBGP, MSDP and PIM-SM protocols for native multicast deployment, although it not limited to those protocols. In particular, Anycast RP is applicable in any PIM-SM network that also supports MSDP (MSDP is required so that the various RPs in the domain maintain a consistent view of the sources that are active). Note however, a domain deploying Anycast RP is not required to run MBGP.

5. Problem Definition

The anycast RP solution provides a solution for both redundancy and load balancing among any number of active RPs in a domain.

5.1. Traffic Concentration and Load Balancing Between RPs

While PIM-SM allows for multiple RPs to be defined for a given group, only one group to RP mapping can active at a given time. A traditional deployment mechanism for load balancing between multiple RPs covering the multicast group space is to split up the 224.0.0.0/4 space between multiple defined RPs. This is an acceptable solution as long as multicast traffic remains low, but has problems as multicast traffic increases, especially because the network operator defining group space split between RPs does not alway have a priori knowledge of traffic distribution between groups. This can be overcome via periodic reconfigurations, but operational considerations cause this type of solution to scale poorly. The other alternative to periodic reconfiguration is to split 224.0.0.0/4 space more finely between more RPs, but this solution can have the disadvantage of creating more complex RP configurations, along with the attendant operational problems when RPs are configured [CLUSTERS].

5.2. Sub-optimal Forwarding of Multicast Packets

When a single RP serves a given multicast group, all joins to that group will be sent to that RP regardless of the topological distance between the RP and the sources and receivers. Initial data will be sent towards the RP also until configured shortest path tree switch threshold is is reached, or the data will always be sent towards the RP if the network is configured to always use RP rooted shared tree. This holds true even if all the sources and the receivers are in any given single region, and RP is topologically distant from the sources and the receivers. This is an artifact of the dynamic nature of multicast group members, and of the fact that operators may not always have a priori knowledge of the topological placement of the group members.

Taken together, these effects can mean that (for example) although all the sources and receivers of a given group are in Europe, they are joining towards the RP in USA and the data will be traversing relatively expensive pipe(s) twice, once to get to RP, and back down the RP rooted tree again, creating inefficient use of expensive resources.

5.3. Distant RP Dependencies

As outlined above, single active RP per group may cause local sources and receivers to become dependent on a topologically distant RP. In case of a scenario where there are backup RPs configured, distant RP dependence can be created due to the failure of the primary RP, which is topologically closer, and may become exacerbated by switching to the backup RP, which may be even more distant topologically, which may lead to inferior performance, if not outright loss of connectivity to an RP serving the group, depending on the network condition at the given moment.

6. Solution

Given the problem set outlined above, a good solution would allow an operator to define multiple RPs per group, and distribute those RPs in a topologically significant manner to the sources and receivers.

6.1. Mechanisms

All the RPs serving a given group or set of groups are configured with identical unicast address, using a numbered interface on the RPs (frequently a logical interface such as a loopback is used). RPs then advertise group to RP mappings using this interface address. This will cause group members (senders) to join (register) towards the topologically closest RP. RPs MSDP peer with each other using the unique shared addresses. Note that if the router implementation chooses the shared address for the BGP router ID, then BGP peerings will not be established. As a result, care should be taken to avoid the ambiguity of the BGP router ID with the RP address (for example, if the logical address chosen is the highest IP address configured on the router, and the router implementation that automatically chooses a router ID based upon highest IP address assigned to interfaces). Finally, the solution described here can be implemented without any modification to existing protocols or their implementations.

6.2. Interaction with MSDP Peer-RPF check

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 or MBGP routing tables are 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". There are a few simple rules that govern how MSDP

Peer-RPF checks. These rules should be kept in mind when configuring Anycast RP:

6.2.1. Singly Homed MSDP Speaker

A singly homed MSDP speaker always accepts SA messages from its peer.

6.2.2. RP in SA is a MSDP Peer

A MSDP speaker always accepts SAs for which the RP in the SA message is a peer.

6.2.3. Router is itself RP in SA message

A MSDP speaker always rejects an SA from any peer if it the RP in the SA message.

6.2.4. Router has default peers

If a MSDP speaker has one or more default peers configured, then it will accept an SA message if comes from the default peer for the RP in the SA message.

6.2.5. Complex MSDP Scenario

Consider routers R1, R2, and R3 form an Anycast RP mesh for an AS. C1 and C2 are customer routers (and the BGP session not multi-hop); RP is C1's RP. P1 is a peer router. The picture is as follows:

```
              MSDP/MBGP        Anycast RPs
         C1 ============= R1----------------R2
          |  SA(S,G,RP)                    /
          |                               /
          |  MBGP/MSDP        MSDP  MSDP /
          |                             /
          |                            /
         C2 SA(S',G',RP')             /
                                     /
                                   R3 ============== P1
                                        SA(S,G,RP)
                                         ------->
```

6.2.6. Internal MSDP Peering

When R1 sees SA(S,G,RP) from C1, it sees the next-hop toward prefix covering RP is through C1, so R1 accepts the SA. R1 will forward SA(S,G,RP) to R2 and R3, which will only accept SA(S,G,RP) if R1 is announcing (and is) the next-hop towards the originating RP each SA message. Note that operationally, this means next-hop needs to be the same as the MSDP connect-source.

This implies that if you want to pass on an SA internally, you have to be announcing the next-hop towards the AS that originates the prefix covering the originating RP. Note that the MSDP connect-source has to be the interface that is configured with the address of the next-hop.

Now, if C2 tries MSDP peer with R1 directly (cutting out transit provider C1), then C2's SA RPF fails at R1, because R1 expects SA message to come from a MSDP peer in the next AS in the AS-PATH towards the originating RP, which in this case would C1.

6.2.6.1. RULE

An internal MSDP peer will accept an SA message from another internal peer iff that peer is the advertiser of towards the prefix covering the RP which originated the SA.

6.2.7. External MSDP Peering

External peer P1 will accept an SA from R3 iff R3 comes from the next AS in the path. This breaks, for example, if P1 peers with C1.

6.2.7.1. RULE

An external MSDP peer will accept an SA message from another peer iff the peer is in the next AS in the path towards the AS originating the prefix covering the RP in the SA message.

6.3. Further Applications of Anycast RP mechanism

The solution described above can also be applied to external MSDP peers that are used to join two PIM-SM domains together. This can provide redundancy to the MSDP peering session, ease operational complexity as well as simplify configuration management. A side effect to be aware of with this design is that which of the configured MSDP sessions comes up will be determined via the unicast topology between two providers, and can be some what unpredictable. If any of the backup peering sessions resets, the active session will also reset.

7. Multicast State Scaling

```
    Let  k = m + r, where

    r = registering to an RP
    m = number internal sources learned through MSDP
    p = number of anycast (internal) MSDP peers

    For p = 1, m = 0

     0 receivers               ==> 1 (*,G) + 0 SAs
     Greater than 1 receiver   ==> k (S,G) + 0 SAs

    For p > 1, m != 0

     0 receivers               ==> 1 (*,G) + m SAs
     Greater than 1 receiver   ==> k (S,G) + m SAs
```

Importantly, the multicast state growth is O(k), where k is not a function of p, the number of anycast RP peers.

8. Security considerations

Since the solution described here makes heavy use of anycast addressing, care must be taken to avoid spoofing. In particular unicast routing and PIM RPs must be protected.

8.1. Unicast Routing

Both internal and external unicast routing can be weakly protected with keyed MD5 [RFC1828], as implemented in an internal protocol such as OSPF [RFC2382] or in BGP [RFC2385]. More generally, IPSEC [RFC1825] could be used to provide protocol integrity for the unicast routing system.

8.2. Multicast Protocol Integrity

The mechanisms described in [PIMAUTH] should be used to provide protocol message integrity protection and group-wise message origin authentication.

8.3. MSDP Peer Integrity

As is the the case for BGP, MSDP peers can be protected using keyed MD5 [RFC1828].

9. Acknowledgments

John Meylor, Dave Thaler and Tom Pusateri provided insightful comments on earlier versions for this idea.

10. References

[CLUSTERS] D. Farinacci, et. al., "Use of Anycast Clusters for Inter-Domain Multicast Routing", draft-ietf-farinacci-anycast-clusters-01.txt, March, 1998. ftp://ftpeng.cisco.com/ipmulticast/internet-drafts

[MSDP] D. Farinacci, et. al., "Multicast Source Discovery Protocol (MSDP)", draft-farinacci-msdp-00.txt, June, 1998.

[PIMAUTH] L. Wei, et al., "Authenticating PIM version 2 messages", draft-ietf-pim-v2-auth-00.txt, November, 1998.

[RFC1825] Atkinson, R., "IP Security Architecture", August 1995.

[RFC1828] P. Metzger and W. Simpson, "IP Authentication using Keyed MD5", RFC 1828, August, 1995.

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

[RFC2382] Moy, J., "OSPF Version 2", RFC 2382, April 1998.

[RFC2385] Herrernan, A., "Protection of BGP Sessions via the TCP MD5 Signature Option", RFC 2385, August, 1998.

[RFC2403] C. Madson and R. Glenn, "The Use of HMAC-MD5-96 within ESP and AH", RFC 2403, November, 1998.

11. Author's Address

Dorian Kim
Verio, Inc.
2361 Lancashire Dr. #2A
Ann Arbor, MI 48015
Email: dorian@blackrose.org

Hank Kilmer
Email: hank@rem.com

Dino Farinacci
Email: dino@dinof.net

David Meyer
Cisco Systems, Inc.
170 Tasman Drive
San Jose, CA, 95134
Email: dmm@cisco.com