Network Working Group
INTERNET-DRAFT
Expiration Date: August 2006

Dino Farinacci
Yiqun Cai
cisco Systems

February 7, 2006

Anycast-RP using PIM
<draft-ietf-pim-anycast-rp-07.txt>

Status of this Memo

Copyright Notice

Abstract

This specification allows Anycast-RP (Rendezvous Point) to be used inside a domain that runs Protocol Independent Multicast (PIM) only. There are no other multicast protocols required to support Anycast-RP, such as MSDP, which has been used traditionally to solve this problem.

1.0 Introduction

Anycast-RP as described in [I1] is a mechanism ISP-based backbones have used to get fast convergence when a PIM Rendezvous Point (RP) router fails. To allow receivers and sources to Rendezvous to the closest RP, the packets from a source need to get to all RPs to find joined receivers.

This notion of receivers finding sources is the fundamental problem of source discovery which MSDP was intended to solve. However, if one would like to retain the Anycast-RP benefits from [I1] with less protocol machinery, removing MSDP from the solution space is an option.

This memo extends the Register mechanism in PIM so Anycast-RP functionality can be retained without using MSDP.

1.1 Terminology

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 [N2].

2.0 Overview

o A unicast IP address is chosen to use as the RP address. This address is statically configured, or distributed using a dynamic protocol, to all PIM routers throughout the domain.

o A set of routers in the domain are chosen to act as RPs for this RP address. These routers are called the Anycast-RP set.

o Each router in the Anycast-RP set is configured with a loopback interface using the RP address.

o Each router in the Anycast-RP set also needs a separate IP address, to be used for communication between the RPs.

- The RP address, or a prefix that covers the RP address, is injected into the unicast routing system inside of the domain.

- Each router in the Anycast-RP set is configured with the addresses of all other routers in the Anycast-RP set. This must be consistently configured in all RPs in the set.

3.0 Mechanism

The following diagram illustrates a domain using 3 RPs where receivers are joining to the closest RP according to where unicast routing metrics take them and 2 sources sending packets to their respective RPs.

The rules described in this section do not override the rules in [N1]. They are intended to blend with the rules in [N1]. If there is any question on the interpretation, precedent is given to [N1].

```
      S1-----RP1              RP2                RP3------S3
               / \              |
              /   \             |
             R1   R1'            R2
```

Assume the above scenario is completely connected where R1, R1', and R2 are receivers for a group, and S1 and S3 send to that group. Assume RP1, RP2 and RP3 are all assigned the same IP address which is used as the Anycast-RP address (let's say the IP address is RPA).

Note, the address used for the RP address in the domain (the Anycast-RP address) needs to be different than the addresses used by the Anycast-RP routers to communicate with each other.

The following procedure is used when S1 starts sourcing traffic:

o S1 sends a multicast packet.

o The DR directly attached to S1 will form a PIM Register message to send to the Anycast-RP address (RPA). The unicast routing system will deliver the PIM Register message to the nearest RP, in this case RP1.

o RP1 will receive the PIM Register message, decapsulate it, send the packet down the shared-tree to get the packet to receivers R1 and R1'.

o RP1 is configured with RP2 and RP3's IP address. Since the Register message did not come from one of the RPs in the anycast-RP set, RP1 assumes the packet came from a DR. If the Register is not addressed to the Anycast-RP address, an error has occurred and it should be rate-limited logged.

o RP1 will then send a copy of the Register message from S1's DR to both RP2 and RP3. RP1 will use its own IP address as the source address for the PIM Register message.

- RP1 MAY join back to the source-tree by triggering a (S1,G) Join message toward S1. However, RP1 MUST create (S1,G) state.

- RP1 sends a Register-Stop back to the DR. If, for some reason, the Register messages to RP2 and RP3 are lost, then when the Register suppression timer expires in the DR, it will resend Registers to allow another chance for all RPs in the Anycast-RP set to obtain the (S,G) state.

- RP2 receives the Register message from RP1, decapsulates it, and also sends the packet down the shared-tree to get the packet to receiver R2.

- RP2 sends a Register-Stop back to the RP1. RP2 MAY wait to send the Register-Stop if it decides to join the source-tree. RP2 should wait until it has received data from the source on the source-tree before sending the Register-Stop. If RP2 decides to wait, the Register-Stop will be sent when the next Register is received. If RP2 decides not to wait, the Register-Stop is sent now.

- RP2 MAY join back to the source-tree by triggering a (S1,G) Join message toward S1. However, RP2 MUST create (S1,G) state.

- RP3 receives the Register message from RP1, decapsulates it, but since there are no receivers joined for the group, it can discard the packet.

- RP3 sends a Register-Stop back to the RP1.

- RP3 creates (S1,G) state so when a receiver joins after S1 starts sending, RP3 can join quickly to the source-tree for S1.

- RP1 processes the Register-Stop from each of RP2 and RP3. There is no specific action taken when processing Register-Stop messages.

The procedure for S3 sending follows the same as above but it is RP3 which sends a copy of the Register originated by S3's DR to RP1 and RP2. Therefore, this example shows how sources anywhere in the domain, associated with different RPs, can reach all receivers, also associated with different RPs, in the same domain.

4.0 Observations and Guidelines about this Proposal

- An RP will send a copy of a Register only if the Register is received from an IP address not in the Anycast-RP list (i.e. the Register came from a DR and not another RP). An implementation

MUST safeguard against inconsistently configured anycast-RP sets in each RP by copying the TTL from a Register message to the Register messages it copies and sends to other RPs.

- Each DR that PIM registers for a source will send the message to the anycast-RP address (which results in the packet getting to the closest physical RP). Therefore there are no changes to the DR logic.

- Packets flow to all receivers no matter what RP they have joined to.

- The source gets Registered to a single RP by the DR. It's the responsibility of the RP that receives the PIM Register messages from the DR (the closest RP to the DR based on routing metrics) to get the packet to all other RPs in the Anycast-RP set.

- Logic is changed only in the RPs. The logic change is for sending copies of Register messages. Register-Stop processing is unchanged. However, an implementation MAY suppress sending Register-Stop messages in response to a Register received from an RP.

- The rate-limiting of Register and Register-Stop messages are done end-to-end. That is from DR -> RP1 -> {RP2 and RP3}. There is no need for specific rate-limiting logic between the RPs.

- When topology changes occur, the existing source-tree adjusts as it does today according to [N1]. The existing shared-trees, as well, adjust as it does today according to [N1].

- Physical RP changes are as fast as unicast route convergence. Retaining the benefit of [I1].

- An RP that doesn't support this specification can be mixed with RPs that do support this specification. However, the non-supporter RPs should not have sources registering to it but may have receivers joined to it.

- If Null Registers are sent (Registers with an IP header and no IP payload), they MUST be replicated to all of the RPs in the Anycast-RP set so that source state remains alive for active sources.

- The number of RPs in the Anycast-RP set should remain small so the amount of non-native replication is kept to a minimum.

- Since the RP, who receives a Register from the DR, will send copies

of the Register to the other RPs at the same time it sends a Register-Stop to the DR, there could be packet loss and lost state in the other RPs until the time the DR sends Register messages again.

5.0 Interaction with MSDP running in an Anycast-PIM Router

The objective of this Anycast-PIM proposal is to remove the dependence on using MSDP. This can be achieved by removing MSDP peering between the Anycast RPs. However, to advertise internal sources to routers outside of a PIM routing domain and to learn external sources from other routing domains, MSDP may still be required.

5.1 Anycast-PIM Stub Domain Functionality

In this capacity, when there are internal sources that need to be advertised externally, an Anycast-RP which receives a Register message, either from a DR or an Anycast-RP, should process it as described in this specification as well as how to process a Register message as described in [N1]. That means an SA for the same internal source could be originated by multiple Anycast-RPs doing the MSDP peering. There is nothing inherently wrong with this other than the source is being advertised into the MSDP infrastructure from multiple places from the source domain. However, if this is not desirable, configuration of one or more (rather than all) Anycast-RP MSDP routers would allow only those routers to originate SAs for the internal source. And in some situations, there is a good possibility not all Anycast-RPs in the set will have MSDP peering sessions so this issue can be mitigated to a certain extent.

From an Anycast-RP perspective, a source should be considered internal to a domain, when it is discovered by an Anycast-RP through a received Register message. Regardless, if the Register message was sent by a DR, another Anycast-RP member, or the router itself.

For learning sources external to a domain, the MSDP SA messages could arrive at multiple MSDP-peering Anycast-RPs. The rules for processing an SA, according to [I1], should be followed. That is, if G is joined in the domain, an (S,G) join is sent towards the source. And if data accompanies the SA, each Anycast-PIM RP doing MSDP peering will

forward the data down each of their respective shared-trees.

The above assumes each Anycast-RP has external MSDP peering connections. If this is not the case, the Anycast-PIM routers with the MSDP peering connections would follow the same procedure as if a Data-Register or Null-Register was received from either a DR or another Anycast-RP. That is, they would send Registers to the other members of the Anycast-RP set.

If there is a mix of Anycast-RPs that do and do not have external MSDP peering connections, then the ones that do must be configured with the set that do not. So Register messages are sent only to the members of the Anycast-RP set that do not have external MSDP peering connections.

The amount of Register traffic generated by this MSDP-peering RP would be equal to the number of active sources external to the domain. The Source-Active state would have to be conveyed to all other RPs in the Anycast-RP set since the MSDP-peering RP would not know about the group membership associated with the other RPs. To avoid this periodic control traffic, it is recommended that all Anycast-RPs be configured with external MSDP peering sessions so no RP in the Anycast-RP set will have to originate Register messages on behalf of external sources.

5.2 Anycast-PIM Transit Domain Functionality

Within a routing domain, it is recommended that an Anycast-RP set defined in this specification should not be mixed with MSDP peering among the members. In some cases, the source discovery will work but it may not be obvious to the implementations what sources are local to the domain and which are not. This may affect external MSDP advertisement of internal sources.

Having said that, this draft makes no attempt to connect MSDP peering domains together by using Anycast-PIM inside a transit domain.

6.0 IANA Considerations

This document makes no request to IANA.

Note to RFC Editor: this section may be removed on publication as an RFC.

7.0 Security Consideration

This section describes the security consideration for Register and Register-Stop messages between Anycast-RPs. For PIM messages between DR and RP, please see [N1].

7.1 Attack Based On Forged Messages

An attacker may forge a Register message using one of the addresses in the Anycast-RP list in order to achieve one or more of the following effects:

1. Overwhelm the target RP in a denial-of-service attack
2. Inject unauthorized data to receivers served by the RP
3. Inject unauthorized data and create bogus SA entries in other PIM domains if the target RP has external MSDP peerings

An attacker may also forge a Register-Stop message using one of the addresses in the Anycast-RP list. However, besides denial-of-service, the effect of such an attack is limited because an RP usually ignores Register-Stop messages.

7.2 Protect Register and Register-Stop Messages

The DOS attack using forged Register or Register-Stop messages can not be prevented. But the RP can still be protected. For example, the RP can rate-limit incoming messages. It can also choose to refuse to process any Register-Stop messages. The actual protection mechansim is implementation specific.

The distribution of unauthorized data and bogus Register messages can be prevented using the method described in section 6.3.2 of [N1]. When RP1 sends a copy of a register to RP2, RP1 acts as [N1] describes the DR and RP2 acts as [N1] describes the RP.

As described in [N1], an RP can be configured using a unique SA and SPI for traffic (Registers or Register-Stops) to each member of Anycast-RPs in the list, but this results in a key management problem; therefore, it may be preferable in PIM domains where all Rendezvous Points are under a single administrative control, to use the same authentication algorithm parameters (including the key) for all Registered packets in a domain.

8.0 Acknowledgments

The authors would like to thank Yiqun Cai and Dino Farinacci for prototyping this draft in the cisco IOS and Procket implementations, respectively.

The authors would like to thank John Zwiebel for doing interoperability testing of the two prototype implementations.

The authors would like to thank Thomas Morin from France Telecom for having an extensive discussion on Multicast the Registers to an SSM-based full mesh among the anycast-RP set. This idea may come in a subsequent Internet Draft.

And finally, the authors would like to thank the following for their comments on earlier drafts:

Greg Shepherd (Procket Networks (now cisco Systems))
Lenny Giuliano (Juniper Networks)
Prashant Jhingran (Huawei Technologies)
Pekka Savola (CSC/FUNET)
Bill Fenner (AT&T)
James Lingard (Data Connection)
Amit Shukla (Juniper Networks)
Tom Pusateri (Juniper Networks)

9.0 Author Information

Dino Farinacci
cisco Systems
dino@cisco.com

Yiqun Cai
cisco Systems
ycai@cisco.com

10.0 References

10.1 Normative References

[N1] Fenner, Handley, Holbrook, Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM):Protocol Specification (Revised)", Internet Draft draft-ietf-pim-sm-v2-new-11.txt, October 2004.

[N2] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

10.2 Informative References

[I1] Kim, Meyer, Kilmer, Farinacci, "Anycast RP mechanism using PIM and MSDP", RFC 3446, January 2003.

Appendix A - Possible Configuration Language

A possible set of commands to be used could be:

```
ip pim anycast-rp <anycast-rp-addr> <rp-addr>
```

Where:

<anycast-rp-addr> describes the Anycast-RP set for the RP which is assigned to the group range. This IP address is the address that first-hop and last-hop PIM routers use to register and join to.

<rp-addr> describes the IP address where Register messages copies are sent to. This IP address is any address assigned to the RP router not including the <anycast-rp-addr>.

Example:

From the illustration above, the configuration commands would be:

```
ip pim anycast-rp RPA RP1
ip pim anycast-rp RPA RP2
ip pim anycast-rp RPA RP3
```

Comment:

It may be useful to include the local router IP address in the command set so the above lines can be cut-and-pasted or scripted into all the RPs in the Anycast-RP set.

But the implementation would have to be aware of its own address and not inadvertently send a Register to itself.