Network Working Group
Request for Comments: 5496
Category: Standards Track

IJ. Wijnands
A. Boers
E. Rosen
Cisco Systems, Inc.

March 2009

The Reverse Path Forwarding (RPF) Vector TLV

Status of This Memo

This document specifies an Internet standards track protocol for the Internet community, and requests discussion and suggestions for improvements. Please refer to the current edition of the "Internet Official Protocol Standards" (STD 1) for the standardization state and status of this protocol. Distribution of this memo is unlimited.

Copyright Notice

Abstract

This document describes a use of the Protocol Independent Multicast (PIM) Join Attribute as defined in RFC 5384, which enables PIM to build multicast trees through an MPLS-enabled network, even if that network's IGP does not have a route to the source of the tree.

Table of Contents

1. Introduction

It is sometimes convenient to distinguish the routers of a particular network into two categories: "edge routers" and "core routers". The edge routers attach directly to users or to other networks, but the core routers attach only to other routers of the same network. If the network is MPLS-enabled, then any unicast packet that needs to travel outside the network can be "tunneled" via MPLS from one edge router to another. To handle a unicast packet that must travel outside the network, an edge router needs to know which of the other edge routers is the best exit point from the network for that packet's destination IP address. The core routers, however, do not need to have any knowledge of routes that lead outside the network; as they handle only tunneled packets, they only need to know how to reach the edge routers and the other core routers.

Consider, for example, the case where the network is an Autonomous System (AS), the edge routers are External Border Gateway Protocol (EBGP) speakers, the core routers may be said to constitute a "BGP-free core". The edge routers distribute BGP routes to each other, but not to the core routers.

However, when multicast packets are considered, the strategy of keeping the core routers free of "external" routes is more problematic. When using PIM Sparse-Mode (PIM-SM) [RFC4601], PIM Source-Specific Mode (PIM-SSM) [RFC4607], or Bidirectional PIM (BIDIR-PIM) [RFC5015] to create a multicast distribution tree for a particular multicast group, one wants the core routers to be full participants in the PIM protocol, so that multicasting can be done efficiently in the core. This means that the core routers must be able to correctly process PIM Join messages for the group, which in turn means that the core routers must be able to send the Join messages towards the root of the distribution tree. If the root of the tree lies outside the network's borders (e.g., is in a different AS), and the core routers do not maintain routes to external destinations, then the PIM Join messages cannot be processed, and the multicast distribution tree cannot be created.

In order to allow PIM to work properly in an environment where the core routers do not maintain external routes, a PIM extension is needed. When an edge router sends a PIM Join message into the core, it MUST include in that message a "Vector" that specifies the IP address of the next edge router along the path to the root of the multicast distribution tree. The core routers can then process the Join message by sending it towards the specified edge router (i.e., toward the Vector). In effect, the Vector serves as an attribute, within a particular network, for the root of the tree.

This document defines a new TLV in the PIM Join Attribute message [RFC5384]. It consists of a single Vector that identifies the exit point of the network.

2. Specification of Requirements

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 [RFC2119].

3. Use of the RPF Vector TLV

Before a router can start forwarding multicast packets, it is necessary to build a forwarding tree by sending PIM Joins hop-by-hop. Each router in the path creates a forwarding state and propagates the Join towards the root of the forwarding tree. The building of this tree is receiver driven. See Figure 1.

```
            ------------------ BGP -----------------
           |                                        |
[S]---( Edge 1)--(Core 1)---( Core )--(Core 2)---( Edge 2 )---[R]
               <--- (S,G) Join
```

Figure 1

In this example, the two edge routers are BGP speakers. The core routers are not BGP speakers and do not have any BGP distributed routes. The route to S is a BGP distributed route; hence, it is known to the edge but not to the core. The Edge 2 router determines the interface leading to S, and sends a PIM Join to the upstream router. In this example, though, the upstream router is a core router, with no route to S. Without the PIM extensions specified in this document, the core router cannot determine where the send the Join, so the tree cannot be constructed.

To allow the core router to participate in the construction of the tree, the Edge 2 router includes an "RPF (Reverse Path Forwarding) Vector" TLV in the PIM Join Attribute [RFC5384] of the PIM Join. In this example, the RPF Vector TLV will contain the IP address of Edge 1. Edge 2 forwards the PIM Join towards Edge 1. Each intermediate core router does its RPF check [RFC4601] on the address contained in the RPF Vector TLV (i.e., on the IP address of Edge 1), instead of doing the RPF check on the address S. This allows the tree to be constructed.

3.1. Attribute and Shared Tree Joins

In the example above, we build a source tree to illustrate the attribute behavior. Use of the attribute is, however, not restricted to the construction of source trees. It may also be used to construct a shared tree. In this case, the RPF Vector TLV contains the IP address of a Rendezvous Point (RP). Procedures defined in this document for (S,G) Joins are equally applicable to (*,G) and (*,*,RP) Joins unless otherwise noted.

3.2. Attribute and Bootstrap Messages

There is no way to carry an RPF Vector TLV in a Bootstrap Router (BSR) bootstrap message. The procedures in this document do not define a way for BSR messages to be forwarded across a core in which the BSP IP address is not routable.

3.3. The Vector Attribute

3.3.1. Inserting a Vector Attribute in a Join

In the example of Figure 1, when the Edge 2 router looks up the route to the source of the multicast distribution tree, it will find a BGP-distributed route whose "BGP next-hop" is Edge 1. Edge 2 then looks up the route to Edge 1 to find the next hop to the source, namely Core 2.

When Edge 2 sends a PIM Join to Core 2, it includes a Vector Attribute specifying the address of Edge 1. Core 2, and subsequent core routers, will forwarding the Join along the Vector (i.e., towards Edge 1) instead of trying to forward it towards S.

Whether an attribute is actually needed depends on whether the Core routers have a route to the source of the multicast tree. How the Edge router knows whether or not this is the case (and thus how the Edge router determines whether or not to insert an attribute field) is outside the scope of this document.

3.3.2. Processing a Received Vector Attribute

When processing a received PIM Join that contains a Vector Attribute, a router MUST first check to see if the Vector IP address is one of its own IP addresses. If so, the Vector Attribute is discarded, and not passed further upstream. Otherwise, the Vector Attribute is used to find the route to the source, and is passed along when a PIM Join is sent upstream. Note that a router that receives a Vector Attribute MUST use it, even if that router happens to have a route to the source. A router that discards a Vector Attribute MAY of course insert a new Vector Attribute. This would typically happen if a PIM Join needed to pass through a sequence of Edge routers, each pair of which is separated by a core that does not have external routes. In the absence of periodic refreshment, Vectors expire along with the corresponding (S,G) state.

3.3.3. Vector Attribute and Asserts

A PIM Assert message includes the routing protocol’s "metric" to the source of the tree. This information is used in the selection of the Assert winner. If a PIM Join is being sent towards a Vector, rather than towards the source, the Assert message MUST have the metric to the Vector instead of the metric to the source. The Assert message however does not have an attribute field and does not mention the Vector.

A router may change its upstream neighbor on a particular multicast tree as the result of receiving Assert messages. However, a Vector Attribute MUST NOT be sent in a PIM Join to an upstream neighbor that is chosen as the result of Assert processing, if that neighbor is different than the original upstream neighbor. Reachability of the Vector is only guaranteed by the router that advertises reachability to the Vector in its IGP. If the Assert winner upstream is not the real preferred next-hop, it is possible that the Assert winner does not know the path to the Vector. In the worst case the Assert winner has a route to the Vector that is on the same interface where the Assert was won. That will point the RPF interface to that interface and will result in the O-list being NULL. The Vector Attribute therefore MUST NOT be inserted if the RPF neighbor was chosen via an Assert process and the RPF neighbor is different from the RPF neighbor that would have been selected via the local routing table. In all other cases, the Vector MUST be included in the Join message.

3.3.4. Vector Attribute and Join Suppression

If a router receives a PIM Join on the upstream LAN interface for a particular multicast state, Join suppression may be applied if that PIM Join is targeted to the same upstream neighbor. Which router(s) will suppress their PIM Join is dependent on timing and is unpredictable. Downstream routers on a LAN MAY include different RPF Vectors in the PIM Joins. Therefore, an upstream router on that LAN may receive and use different RPF Vectors over time to reach the destination (depending on which downstream router(s) suppressed their Join). To make the upstream router behavior more predictable, the RPF Vector address MUST be used as additional condition to the Join suppression logic. Only if the RPF Vector in the PIM Join matches the RPF Vector in the multicast state, the suppression logic is applied. It is also possible to disable Join suppression on that LAN.

4. Vector Attribute TLV Format

```
   0                   1                   2                   3
   0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |F|S| Type      | Length        |        Value
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-.......
```

F bit
   Forward Unknown TLV. If this bit is set, the TLV is forwarded regardless of whether the router understands the Type. If the TLV is known, the F bit is ignored.

S bit
   Bottom of Stack. If this bit is set, then this is the last TLV in the stack.

Type
   The Vector Attribute type is 0.

Length
   Length depending on Address Family of Encoded-Unicast address.

Value
   Encoded-Unicast address.

5. IANA Considerations

IANA has assigned the value 0 to the RPF Vector in the "PIM Join Attribute Types" registry.

6. Security Considerations

Security of the RPF Vector Attribute is only guaranteed by the security of the PIM packet, so the security considerations for PIM Join packets as described in PIM-SM [RFC4601] apply here.

7. Acknowledgments

The authors would like to thank Yakov Rekhter and Dino Farinacci for their initial ideas on this topic and Su Haiyang for the comments on the document.

8. Normative References

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

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

[RFC4607] Holbrook, H. and B. Cain, "Source-Specific Multicast for IP", RFC 4607, August 2006.

[RFC5015] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", RFC 5015, October 2007.

[RFC5384] Boers, A., Wijnands, I., and E. Rosen, "The Protocol Independent Multicast (PIM) Join Attribute Format", RFC 5384, November 2008.

Authors' Addresses

IJsbrand Wijnands
Cisco Systems, Inc.
De kleetlaan 6a
Diegem 1831
Belgium

EMail: ice@cisco.com

Arjen Boers
Cisco Systems, Inc.
Avda. Diagonal, 682
Barcelona 08034
Spain

EMail: aboers@cisco.com

Eric Rosen
Cisco Systems, Inc.
1414 Massachusetts Avenue
Boxborough, Ma 01719

EMail: erosen@cisco.com