Internet Engineering Task Force (IETF)
Request for Comments: 6754
Category: Standards Track
ISSN: 2070-1721

Y. Cai
Microsoft
L. Wei
H. Ou
Cisco Systems, Inc.
V. Arya
S. Jethwani
DIRECTV Inc.

October 2012

Protocol Independent Multicast Equal-Cost Multipath (ECMP) Redirect

Abstract

A Protocol Independent Multicast (PIM) router uses the Reverse Path Forwarding (RPF) procedure to select an upstream interface and router in order to build forwarding state. When there are equal-cost multipaths (ECMPs), existing implementations often use hash algorithms to select a path. Such algorithms do not allow the spread of traffic among the ECMPs according to administrative metrics. This usually leads to inefficient or ineffective use of network resources. This document introduces the ECMP Redirect, a mechanism to improve the RPF procedure over ECMPs. It allows ECMP selection to be based on administratively selected metrics, such as data transmission delays, path preferences, and routing metrics.

Status of This Memo

This is an Internet Standards Track document.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Further information on Internet Standards is available in Section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc6754.

Table of Contents

1. Introduction

A PIM router uses the RPF procedure to select an upstream interface and a PIM neighbor on that interface to build forwarding state. When there are equal-cost multipaths (ECMPs) upstream, existing implementations often use hash algorithms to select a path. Such algorithms do not allow the spread of traffic among the ECMP according to administrative metrics. This usually leads to inefficient or ineffective use of network resources. This document introduces the ECMP Redirect, a mechanism to improve the RPF procedure over ECMP. It allows ECMP selection to be based on administratively selected metrics, such as data transmission delays, path preferences, and routing metrics, or a combination of metrics.

ECMPs are frequently used in networks to provide redundancy and to increase available bandwidth. A PIM router selects a path in the ECMP based on its own implementation-specific choice. The selection is a local decision. One way is to choose the PIM neighbor with the highest IP address; another is to pick the PIM neighbor with the best hash value over the destination and source addresses.

While implementations supporting ECMP have been deployed widely, the existing RPF selection methods have weaknesses. The lack of administratively effective ways to allocate traffic over alternative paths is a major issue. For example, there is no straightforward way to tell two downstream routers to select either the same or different RPF neighbor routers for the same traffic flows.

With the ECMP Redirect mechanism introduced here, the upstream routers use a PIM ECMP Redirect message to instruct the downstream routers on how to tiebreak among the upstream neighbors. The PIM ECMP Redirect message conveys the tiebreak information based on metrics selected administratively.

2. 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 [RFC2119].

This document uses terms defined in [RFC4601] to describe actions taken by PIM routers.

The following terms have special significance for ECMP Redirect:

- Equal-Cost Multipath (ECMP). In this document, the term "ECMP" refers to parallel, single-hop, equal-cost links between adjacent nodes.

- ECMP Bundle. An ECMP bundle is a set of PIM-enabled interfaces on a router, where all interfaces belonging to the same bundle share the same routing metric. The next hops for the ECMP are all one hop away.

  There can be one or more ECMP bundles on any router, while one individual interface can only belong to a single bundle. ECMP bundles are created on a router via configuration.

- RPF. RPF stands for Reverse Path Forwarding.

- Upstream. Towards the root of the multicast forwarding tree. An upstream router refers to a router that is forwarding, or potentially capable of forwarding, data packets onto interfaces in an ECMP bundle.

  When there are multiple routers forwarding packets onto interfaces in the ECMP bundle, all these routers are called upstream routers.

- Downstream. Away from the root of the multicast forwarding tree. A downstream router is a router that uses an interface in the ECMP bundle as an RPF interface for a multicast forwarding entry.

3. Overview

The existing PIM Assert mechanism allows the upstream router to detect the existence of multiple forwarders for the same multicast flow onto the same downstream interface. The upstream router sends a PIM Assert message containing a routing metric for the downstream routers to use for tiebreaking among the multiple upstream forwarders on the same RPF interface.

With ECMP interfaces between the downstream and upstream routers, the PIM ECMP Redirect mechanism works in a similar way, but extends the ability to resolve the selection of forwarders among different interfaces in the ECMP.

When a PIM router downstream of the ECMP interfaces creates a new (*,G) or (S,G) entry, it will populate the RPF interface and RPF neighbor information according to the rules specified by [RFC4601]. This router will send its initial PIM Joins to that RPF neighbor.

When the RPF neighbor router receives the Join message and finds that the receiving interface is one of the ECMP interfaces, it will check if the same flow is already being forwarded out of another ECMP interface. If so, this RPF neighbor router will send a PIM ECMP Redirect message onto the interface the Join was received on. The PIM ECMP Redirect message contains the address of the desired RPF

neighbor, an Interface ID [RFC6395], and the other parameters used as tiebreakers. In essence, a PIM ECMP Redirect message is sent by an upstream router to notify downstream routers to redirect PIM Joins to the new RPF neighbor via a different interface. When the downstream routers receive this message, they SHOULD trigger PIM Joins toward the new RPF neighbor specified in the packet.

This PIM ECMP Redirect message has similar functions as the existing PIM Assert message:

1. It is sent by an upstream router.

2. It is used to influence the RPF selection by downstream routers.

3. A tiebreaker metric is used.

However, the existing Assert message is used to select an upstream router within the same multi-access network (such as a LAN), while the Redirect message is used to select both a network and an upstream router.

One advantage of this design is that the control messages are only sent when there is a need to "rebalance" the traffic. This reduces the amount of control traffic.

4. Applicability

The use of ECMP Redirect applies to shared trees or source trees built with procedures described in [RFC4601]. The use of ECMP Redirect in PIM Dense Mode [RFC3973] or in Bidirectional PIM [RFC5015] is not considered in this document.

The enhancement described in this document can be applicable to a number of scenarios. For example, it allows a network operator to use ECMPs and have the ability to perform load splitting based on bandwidth. To do this, the downstream routers perform RPF selection with bandwidth, instead of IP addresses, as a tiebreaker. The ECMP Redirect mechanism assures that all downstream routers select the desired network link and upstream router whenever possible. Another example is for a network operator to impose a transmission delay limit on certain links. The ECMP Redirect mechanism provides a means for an upstream router to instruct a downstream router to choose a different RPF path.

This specification does not dictate the scope of applications of this mechanism.

5. Protocol Specification

5.1. Sending ECMP Redirect

ECMP Redirects are sent by an upstream router in a rate-limited fashion, under either of the following conditions:

- It detects a PIM Join on a non-desired outgoing interface.
- It detects multicast traffic on a non-desired outgoing interface.

In both cases, an ECMP Redirect is sent to the non-desired interface. An outgoing interface is considered "non-desired" when:

- The upstream router is already forwarding the same flow out of another interface belonging to the same ECMP bundle.
- The upstream router is not yet forwarding the flow out any interfaces of the ECMP bundle, but there is another interface with more desired attributes.

An upstream router MAY choose not to send ECMP Redirects if it becomes aware that some of the downstream routers are unreachable via some links in ECMP bundle.

An upstream router uses the Neighbor Address or the Interface ID field in the ECMP Redirect message to indicate the interface it wants traffic to be directed to. This Neighbor Address MUST be associated with an interface in the same ECMP bundle as the ECMP Redirect message's outgoing interface. If the Interface ID field is ignored, this Neighbor Address field uniquely identifies a LAN and an upstream router to which a downstream router SHOULD redirect its Join messages, and an ECMP Redirect message MUST be discarded if the Neighbor Address field in the message does not match the cached neighbor address.

The Interface ID field is used in IPv4 when one or more RPF neighbors in the ECMP bundle are unnumbered, or in IPv6 where link-local addresses are in use. For other IPv4 usage, this field is zeroed when sent, and ignored when received. If the Router ID part of the Interface ID is zero, the field MUST be ignored. See [RFC6395] for details of its assignment and usage in PIM Hellos. If the Interface ID is not ignored, the receiving router of this message MUST use the Interface ID, instead of Neighbor Address, to identify the new RPF neighbor. Additionally, an ECMP Redirect message MUST be discarded if the Interface ID field in the message does not match the cached Interface ID.

5.2. Receiving ECMP Redirect

When a downstream router receives an ECMP Redirect, and detects that the desired RPF path from its upstream router's point of view is different from its current one, it should choose to join the newly suggested path and prune from the current path. The exact order of such actions is implementation specific.

If a downstream router receives multiple ECMP Redirects sent by different upstream routers, it SHOULD use the Preference, Metric, or other fields as specified below as the tiebreakers to choose the most preferred RPF interface and neighbor. The tiebreak procedure is the same as that used in PIM Assert processing described by [RFC4601].

If an upstream router receives an ECMP Redirect, it SHOULD NOT change its forwarding behavior even if the ECMP Redirect makes it a less preferred RPF neighbor on the receiving interface.

5.3. Transient State

During a transient network outage with a single link cut in an ECMP bundle, a downstream router may lose connection to its RPF neighbor and the normal ECMP Redirect operation may be interrupted temporarily. In such an event, the following actions are RECOMMENDED.

The downstream router SHOULD select a new RPF neighbor. Among all ECMP upstream routers, the preferred selection is the one on the LAN that the previous RPF neighbor resided on.

If there is no upstream router reachable on the LAN that the previous RPF neighbor resided on, the downstream router will select a new RPF neighbor on a different LAN. Among all ECMP upstream routers, the one that served as RPF neighbor before the link failure is preferred. Such a router can be identified by the Router ID, which is part of the Interface ID in the PIM ECMP Redirect Hello option.

During normal ECMP Redirect operations, when PIM Joins for the same (*,G) or (S,G) are received on a different LAN, an upstream router will send ECMP Redirect to prune the non-preferred LAN. Such ECMP Redirects during partial network outage can be suppressed if the upstream router decides that the non-preferred PIM Join is from a router that is not reachable via the preferred LAN. This check can be performed by retrieving the downstream router's Router ID, using the source address in the PIM Join, and searching neighbors on the preferred LAN for one with the same Router ID.

5.4. Interoperability

If a PIM router supports this specification, it MUST send the PIM ECMP Redirect Hello Option in its PIM Hello messages.

A PIM router sends ECMP Redirects on an interface only when it detects that all neighbors on that interface have sent this Hello option. If a PIM router detects that any of its neighbors on an ECMP bundle does not support this Hello option, it SHOULD NOT send ECMP Redirects to interfaces in that bundle; however, it SHOULD still process any ECMP Redirects received from interfaces in that same bundle.

If a PIM router does not support this specification, it will ignore the PIM ECMP Redirect Hello Options and ECMP Redirects in the PIM packets that it receives.

5.5. Packet Format

5.5.1. PIM ECMP Redirect Hello Option

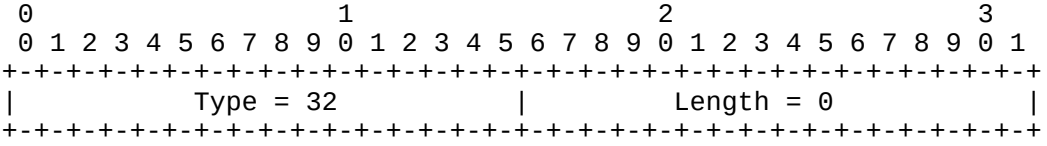

Figure 1: ECMP Redirect Hello Option

Type: 32

Length: 0

5.5.2. PIM ECMP Redirect Format

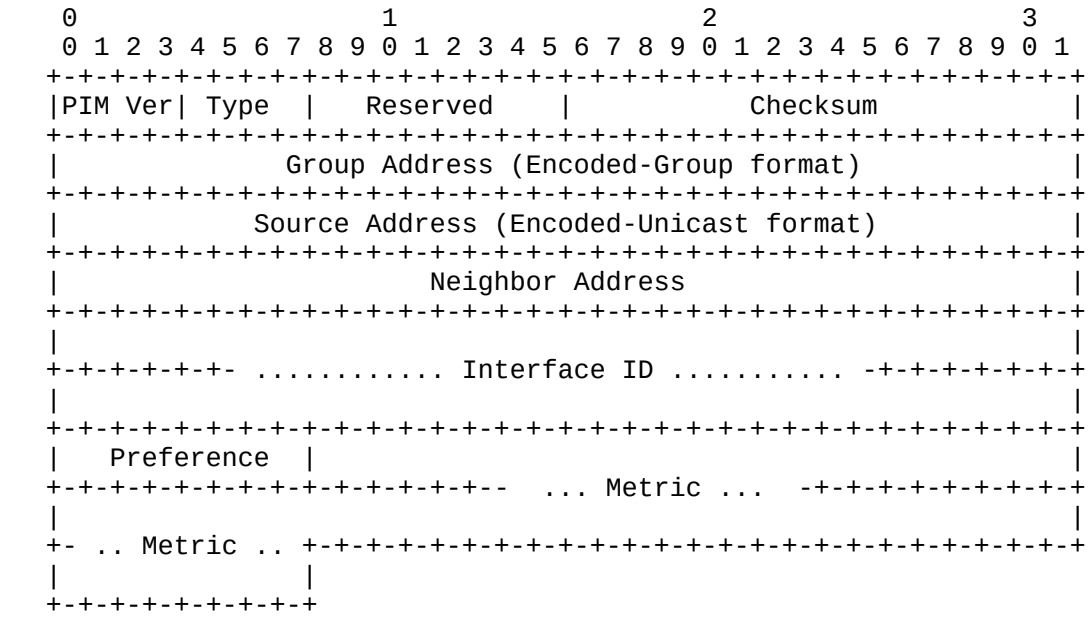

Figure 2: ECMP Redirect Message Format

PIM Ver: See Section 4.9 in [RFC4601].

Type: 11

Reserved: See Section 4.9 in [RFC4601].

Checksum: See Section 4.9 in [RFC4601].

Group Address (64 or 160 bits): Encoded-Group address as specified in Section 4.9.1 of [RFC4601].

Source Address (48 or 144 bits): Encoded-Unicast address as specified in Section 4.9.1 of [RFC4601].

Neighbor Address (32 or 128 bits): Address of desired upstream neighbor where the downstream receiver redirects PIM Joins.

Interface ID (64 bits): See [RFC6395] for details.

Preference (8 bits): The first tiebreaker when ECMP Redirects from multiple upstream routers are compared against each other. A numerically smaller value is preferred. A reserved value (15) is used to indicate the metric value following the Preference field is a Network Time Protocol (NTP) timestamp, encoded in the format specified in [RFC5905], taken at the moment the sending router started to forward out of this interface.

Metric (64 bits): The second tiebreaker if the Preference values are the same. A numerically smaller value is preferred. This Metric can contain path parameters defined by users. When the Preference and Metric values are the same, the Neighbor Address or Interface ID field is used as the third tiebreaker, depending on which field is used to identify the RPF neighbor; the bigger value wins.

6. IANA Considerations

A PIM-Hello Option Type (32) has been assigned to the PIM ECMP Redirect Hello Option.

In the PIM Message Types registry created by [RFC6166], a PIM Message Type (11) has been assigned to the ECMP Redirect message.

7. Security Considerations

Security of the ECMP Redirect is only guaranteed by the security of the PIM packet; the security considerations for PIM Assert packets as described in [RFC4601] apply here. Spoofed ECMP Redirect packets may cause the downstream routers to send PIM Joins to an undesired upstream router and trigger more ECMP Redirect messages. Security considerations for PIM packets described in [RFC4601] also apply to the new Hello option defined here.

8. Acknowledgements

The authors would like to thank Apoorva Karan for helping with the original idea, and Eric Rosen, Isidor Kouvelas, Toerless Eckert, Stig Venaas, Jeffrey Zhang, Bill Atwood, and Adrian Farrel for their review comments.

9. References

9.1. 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.

9.2. Informative References

[RFC3973] Adams, A., Nicholas, J., and W. Siadak, "Protocol Independent Multicast - Dense Mode (PIM-DM): Protocol Specification (Revised)", RFC 3973, January 2005.

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

[RFC5905] Mills, D., Martin, J., Ed., Burbank, J., and W. Kasch, "Network Time Protocol Version 4: Protocol and Algorithms Specification", RFC 5905, June 2010.

[RFC6166] Venaas, S., "A Registry for PIM Message Types", RFC 6166, April 2011.

[RFC6395] Gulrajani, S. and S. Venaas, "An Interface Identifier (ID) Hello Option for PIM", RFC 6395, October 2011.

Authors' Addresses

Yiqun Cai
Microsoft
1065 La Avenida
Mountain View, CA 94043
USA

EMail: yiqunc@microsoft.com

Liming Wei
Cisco Systems, Inc.
Tasman Drive
San Jose, CA 95134
USA

EMail: lwei@cisco.com

Heidi Ou
Cisco Systems, Inc.
Tasman Drive
San Jose, CA 95134
USA

EMail: hou@cisco.com

Vishal Arya
DIRECTV Inc.
2230 E Imperial Hwy
El Segundo, CA 90245
USA

EMail: varya@directv.com

Sunil Jethwani
DIRECTV Inc.
2230 E Imperial Hwy
El Segundo, CA 90245
USA

EMail: sjethwani@directv.com