Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 6, 2012

Yiqun Cai
Liming Wei
Heidi Ou
Cisco Systems, Inc.
Vishal Arya
Sunil Jethwani
DIRECTV Inc.

February 3, 2012

Protocol Independent Multicast ECMP Redirect
draft-ietf-pim-ecmp-02.txt

Abstract

A PIM router uses the RPF procedure to select an upstream interface and router to build forwarding state. When there are equal cost multiple paths (ECMP), 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 path selection to be based on administratively selected metrics, such as data transmission delays, path preferences and routing metrics.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

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."

This Internet-Draft will expire on August 6, 2012.

Copyright Notice

Table of Contents

1. Requirements Notation

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].

2. Introduction

A PIM [RFC4601] 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 multiple paths (ECMP) 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 path 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 tie-break among the upstream neighbors. The PIM ECMP Redirect message conveys the tie-break information based on metrics selected administratively.

2.1. 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 tie-breaking 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], along with other parameters used as tie breakers. 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; And
3. A tie breaker 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 need to "re-balance" the traffic. This reduces the amount of control traffic.

2.2. 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 "Protocol Independent Multicast - Dense Mode" [RFC3973] or in "Bidirectional Protocol Independent Multicast" [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 ECMP paths 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 tie breaker. 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.

3. Protocol Specification

3.1. 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 ECMP paths reside between the upstream and downstream routers over the ECMP bundle.

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.

3.2. Sending ECMP Redirect

ECMP Redirects are sent by a preferred upstream router in a rate limited fashion, under the following conditions,

- It detects a PIM Join on a non-desired outgoing interface; or
- 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 forwarding the flow yet 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 do not support the message, or are unreachable via some links in ECMP bundle.

3.3. 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 prune from the current path and join the new 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 tie breakers to choose the most preferred RPF interface and neighbor.

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

3.4. 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 down stream router may re-select a new RPF neighbor. Among all ECMP upstream routers, the one on the same LAN as the previous RPF neighbor is preferred.

If there is no upstream router reachable on the same LAN, the down stream router will select an 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'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.

3.5. Interoperability

If a PIM router supports this specification, it MUST send the Hello option ECMP-Redirect-Supported TLV in its PIM Hello messages. A PIM router sends ECMP Redirects on an interface only when it detects that all neighbors have sent this Hello option. If a PIM router detects that any of its neighbor does not support this Hello option, it MUST not send ECMP Redirects, however, it SHOULD still process any ECMP Redirects received.

3.6. Packet Format

3.6.1. PIM ECMP Redirect Hello Option

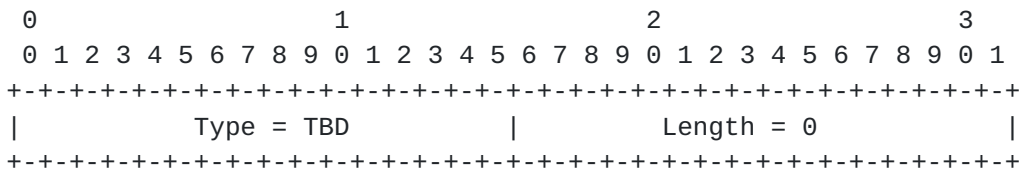

Figure 1: ECMP Redirect Hello Option

Type: TBD.
Length: 0

3.6.2. PIM ECMP Redirect Format

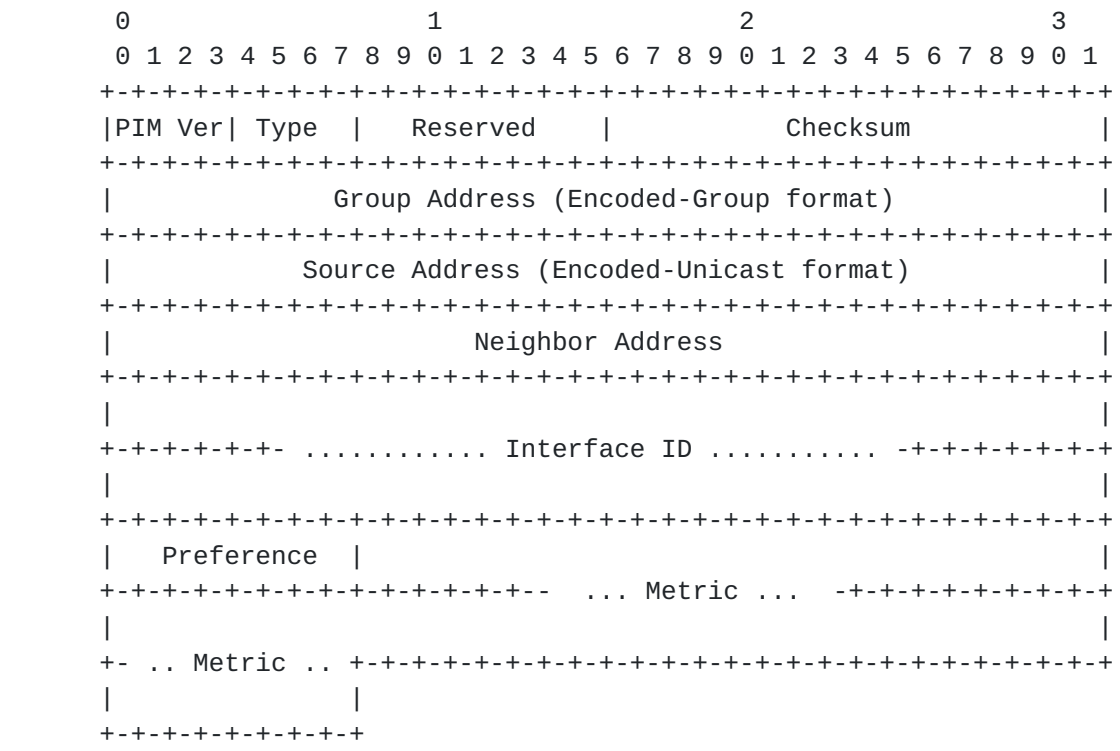

Figure 2: ECMP Redirect Message Format

Type: TBD

Neighbor Address (32/128 bits): Address of desired upstream neighbor where the downstream receiver should redirect PIM Joins. This 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 (see below) 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 cached neighbor address.

Interface ID (64 bits): This 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 zero'ed 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, and an ECMP Redirect message MUST be discarded if the "Interface ID" field in the message does not match the cached interface ID.

Preference (8 bits): The first tie breaker when ECMP Redirects from multiple upstream routers are compared against each other. Numerically smaller value is preferred. A reserved value (15) is used to indicate the metric value following the "Preference" field is a timestamp, taken at the moment the sending router started to forward out of this interface.

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

4. IANA Considerations

A PIM Hello Option Type is requested to be assigned to the PIM ECMP Redirect Hello Option. According to [HELLO-OPT], this document recommends 32 (0x20) as the "PIM ECMP Redirect Hello Option Type".

A PIM Message Type is requested to be assigned to the ECMP Redirect message. According to [RFC6166], the next available Type value is 11 (0xB).

5. 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.

6. Acknowledgement

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

7. References

7.1. Normative Reference

[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.

7.2. 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.

[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.

[HELLO-OPT] IANA, "PIM Hello Options", PIM-HELLO-OPTIONS per RFC4601 http://www.iana.org/assignments/pim-hello-options, October 2011.

Authors' Addresses

Yiqun Cai
Cisco Systems, Inc.
Tasman Drive
San Jose, CA 95134
USA

Email: ycai@cisco.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