

L3VPN Routing Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: August 17, 2014

H. Jeng  
AT&T  
L. Jalil  
Verizon  
R. Bonica  
Y. Rekhter  
Juniper Networks  
K. Patel  
Cisco Systems  
L. Yong  
Huawei Technologies  
February 13, 2014

Covering Prefixes Outbound Route Filter for BGP-4  
draft-bonica-l3vpn-orf-covering-prefixes-01

Abstract

This document defines a new ORF-type, called the "Covering Prefixes ORF (CP-ORF)". CP-ORF is applicable in Virtual Hub-and-Spoke VPNs. It also is applicable in BGP/MPLS Ethernet VPN (EVPN) Networks.

Requirements Language

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 RFC 2119 [RFC2119].

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 17, 2014.

## Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Problem Statement . . . . .	2
1.1. Terminology . . . . .	2
2. CP-ORF Encoding . . . . .	3
3. Processing Rules . . . . .	5
4. Applicability In Virtual Hub-and-Spoke VPNs . . . . .	8
5. Applicability In Network Virtualization Overlays . . . . .	10
6. Clean-up . . . . .	12
7. IANA Considerations . . . . .	12
8. Security Considerations . . . . .	12
9. Acknowledgements . . . . .	13
10. Normative References . . . . .	13
Authors' Addresses . . . . .	14

## 1. Problem Statement

A BGP [RFC4271] speaker can send Outbound Route Filters (ORF) [RFC5291] to a peer. The peer uses ORFs to filter routing updates that it sends to the BGP speaker. Using ORF, a speaker can realize a "route pull" paradigm in BGP, in which the speaker, on demand, pulls certain routes from the peer.

This document defines a new ORF-type, called the "Covering Prefixes ORF (CP-ORF)". CP-ORF is applicable in Virtual Hub-and-Spoke VPNs [RFC7024] [RFC4364]. It also is applicable BGP/MPLS Ethernet VPN (EVPN) [I-D.ietf-l2vpn-evpn] Networks.

## 1.1. Terminology

This document uses the following terms:

- o Address Family Indicator (AFI) - defined in [RFC4760]

- o Subsequent Address Family Indicator (SAFI) - defined in [RFC4760]
- o VPN IP Default Route - defined in [RFC7024].
- o V-Hub - defined in [RFC7024].
- o V-Spoke - defined in [RFC7024].
- o BGP/MPLS Ethernet VPN (EVPN) - defined in [I-D.ietf-l2vpn-evpn]
- o EVPN Instance (EVI) - defined in [I-D.ietf-l2vpn-evpn]
- o Default MAC Route (DMR) - An EVPN Route with MAC Address length equal to 0. See Section 10.2.1 of [I-D.ietf-l2vpn-evpn] for details.
- o Default Gateway (DMG) - An EVPN PE that advertises a DMR

## 2. CP-ORF Encoding

[RFC5291] augments the BGP ROUTE-REFRESH message so that it can carry ORF entries. When the ROUTE-REFRESH message carries ORF entries, it includes the following fields:

- o AFI [IANA.AFI]
- o SAFI [IANA.SAFI]
- o When-to-refresh (IMMEDIATE or DEFERRED)
- o ORF Type
- o Length (of ORF entries)

The ROUTE-REFRESH message also contains a list of ORF entries. Each ORF entry contains the following fields:

- o Action (ADD, REMOVE, or REMOVE-ALL)
- o Match (PERMIT or DENY)

The ORF entry may also contain Type-specific information. Type-specific information is present only when the Action is equal to ADD or REMOVE. It is not present when the Action is equal to REMOVE-ALL.

When the BGP ROUTE-REFRESH message carries CP-ORF entries, the following conditions MUST be true:

- o ORF Type MUST be equal to CP-ORF. (The value of CP-ORF is TBD. See Section 7 for details.)
- o The AFI MUST be equal to IPv4, IPv6 or L2VPN
- o If the AFI is equal to IPv4 or IPv6, SAFI MUST be equal to MPLS-labeled VPN address
- o If the AFI is equal to L2VPN, the SAFI MUST be equal to BGP EVPN
- o Match field MUST be equal to PERMIT

Figure 1 depicts the encoding of the CP-ORF type-specific information.

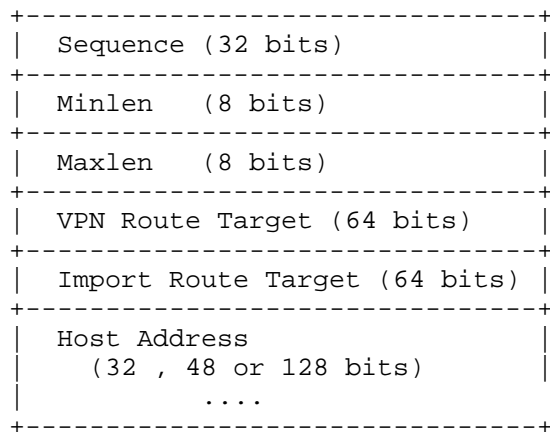


Figure 1: CP-ORF Type-specific Encoding

The Sequence field specifies the relative ordering of the entry among all CP-ORF entries.

The CP-ORF recipient uses the following fields to identify routes that match the CP-ORF:

- o Minlen
- o Maxlen
- o VPN Route Target
- o Host Address

See Section 3 for details.

The CP-ORF recipient marks routes that match CP-ORF with the Import Route Target before advertising those routes to the CP-ORF originator. See Section 3 for details.

If the ROUTE-REFRESH AFI is equal to IPv4, the length of the Host Address field is 32 bits. If the ROUTE-REFRESH AFI is equal to IPv6, the length of the Host Address field is 128 bits. If the ROUTE-REFRESH AFI is equal to L2VPN, the length of the Host Address field is 48 bits.

If the ROUTE-REFRESH AFI value is equal to IPv4 or IPv6, the following rules apply:

- o The value of Minlen MUST be less than or equal to the length of the Host Address field
- o The value of Maxlen MUST be less than or equal to the length of the Host Address field
- o The value of Minlen MUST be less than or equal to the value of Maxlen

If the ROUTE-REFRESH AFI is equal to L2VPN, the following rules apply:

- o The value of Minlen MUST be 48
- o The value of Maxlen MUST be 48

### 3. Processing Rules

According to [RFC4271], every BGP speaker maintains a single Loc-RIB. For each of its peers, the BGP speaker also maintains an Outbound Filter and an Adj-RIB-Out. The Outbound Filter defines policy that determines which Loc-RIB entries are processed into the corresponding Adj-RIB-Out. Mechanisms such as RT-Constraint [RFC4684] and ORF [RFC5291] enable a router's peer to influence the Outbound Filter. Therefore, the Outbound Filter for a given peer is constructed using a combination of the locally configured policy and the information received via RT-Constraint and ORF from the peer.

Using this model we can describe the operations of CP-ORF as follows:

When a BGP speaker receives a ROUTE-REFRESH message that contains a CP-ORF, and that ROUTE-REFRESH message that violates any of the encoding rules specified in Section 2, the BGP speaker MUST log the event and ignore the entire ROUTE-REFRESH message.

Otherwise, the BGP speaker processes each CP-ORF entry as indicated by the Action field. If the Action is equal to ADD, the BGP speaker adds the CP-ORF entry to the Outbound Filter associated with the peer in the position specified by the Sequence field. If the Action is equal to REMOVE, the BGP speaker removes the CP-ORF entry from the Outbound Filter. If the Action is equal to REMOVE-ALL, the BGP speaker removes all CP-ORF entries from the Outbound Filter.

Whenever the BGP speaker applies an Outbound Filter to a route contained by its Loc-RIB, it evaluates the route in terms of the CP-ORF entries first. It then evaluates the route in terms of the remaining, non-CP-ORF entries. The rules for the former are described below. The rules for the latter are outside the scope of this document.

The following route types can match a CP-ORF:

- o IPv4-VPN
- o IPv6-VPN
- o L2VPN (L2VPN MAC Advertisement only. See Section 8.2 of [I-D.ietf-l2vpn-evpn] for details.)

In order for an IPv4-VPN route or IPv6-VPN route to match a CP-ORF, all of the following conditions MUST be true:

- o the route carries an RT whose value is the same as the CP-ORF VPN Route Target
- o the route prefix length is greater than or equal to the CP-ORF Minlen plus 64 (i.e., the length of a VPN Route Distinguisher)
- o the route prefix length is less than or equal to the CP-ORF Maxlen plus 64 (i.e., the length of a VPN Route Distinguisher)
- o ignoring the Route Distinguisher, the leading bits of the route prefix are identical to the leading bits of the CP-ORF Host Address. CP-ORF Minlen defines the number of bits that must be identical.

The BGP speaker ignores Route Distinguishers when determining whether a prefix covers a host address. For example, assume that a CP-ORF carries the following information:

- o Minlen equal to 1
- o Maxlen equal to 32

- o Host Address equal to 192.0.2.1

Assume also that Loc-RIB contains routes for the following IPv4-VPN prefixes, and that all of these routes carry an RT whose value is the same as the CP-ORF VPN Route Target:

- o 1:0.0.0.0/64.
- o 2:192.0.2.0/88
- o 3:192.0.2.0/89

For the purposes of this evaluation, 2:192.0.2.0/88 and 3:192.0.2.0/89 cover 192.0.2.1. This is because the search algorithm ignores Route Distinguishers. However, 1:0.0.0.0/64 does not cover 192.0.2.1, because its length (64) is less than the CP-ORF Minlen (1) plus the length of an L3VPN Route Distinguisher (64).

In order for an EVPN route match a CP-ORF, all of the following conditions MUST be true:

- o the route carries an RT whose value is the same as the CP-ORF VPN Route Target
- o the final 48 bits of the EVPN MAC Address are identical to the CP-ORF Host Address

If a route matches the selection criteria of a CP-ORF entry, and it does not violate any subsequent rule specified by the Outbound Filter (e.g., rules that reflect local policy, or rules that are due to RT-Constrains), the BGP speaker places the route into the Adj-RIB-Out. In Adj-RIB-Out, the BGP speaker adds the CP-ORF Import Route Target to the list of Route Targets that the route already carries. As a result of being placed in Adj-RIB-Out, the route is advertised to the peer associated with the Adj-RIB-Out.

Receiving CP-ORF entries with REMOVE or REMOVE-ALL Actions may cause a route that has previously been installed in a particular Adj-RIB-Out be excluded from that Adj-RIB-Out. In this case, as specified in [RFC4271], "the previously advertised route in that Adj-RIB-Out MUST be withdrawn from service by means of an UPDATE message".

[RFC5291] states that a BGP speaker should respond to a ROUTE REFRESH message as follows:

"If the When-to-refresh indicates IMMEDIATE, then after processing all the ORF entries carried in the message the speaker re-advertises to the peer routes from the Adj-RIB-Out associated with the peer that

have the same AFI/SAFI as what is carried in the message, and taking into account all the ORF entries for that AFI/SAFI received from the peer. The speaker **MUST** re-advertise all the routes that have been affected by the ORF entries carried in the message, but **MAY** also re-advertise the routes that have not been affected by the ORF entries carried in the message."

When the ROUTE-REFRESH message includes one or more CP-ORF entries, the BGP speaker **MUST** re-advertise routes that have been affected by ORF entries carried by the message. While the speaker **MAY** also re-advertise the routes that have not been affected by the ORF entries carried in the message, this memo **RECOMMENDS** not to re-advertise the routes that have not been affected.

#### 4. Applicability In Virtual Hub-and-Spoke VPNs

In a Virtual Hub-and-Spoke environment, VPN sites are attached to Provider Edge (PE) routers, where for a given VPN some of these PEs may act as V-hubs, while others as V-spokes. This memo assumes that PEs exchange VPN-IP routes using Route Reflectors (RRs).

This memo also assumes that RED-VPN sites are attached to PE routers, V-hub1 and V-spoke1. All of these devices advertise RED-VPN routes to a RR. They mark these routes with a route target, which we will call RT-RED.

V-hub1 serves the RED-VPN. Therefore, V-hub1 advertises a VPN IP default route for the RED-VPN to the RR, carrying the route target RT-RED-FROM-HUB1.

V-spoke1 establishes a BGP session with the RR, negotiating the CP-ORF capability, as well as the Multiprotocol Extensions Capability [RFC2858]. Upon establishment of the BGP session, the RR does not advertise any routes to V-spoke1. The RR will not advertise any routes until it receives either a ROUTE-REFRESH message or a BGP UPDATE message containing a Route Target Membership NLRI [RFC4684].

Immediately after the BGP session is established, V-spoke1 sends the RR a BGP UPDATE message containing a Route Target Membership NLRI. The Route Target Membership NLRI specifies RT-RED-FROM-HUB1 as its route target. In response to the BGP-UPDATE message, the RR advertises the VPN IP default route for the RED-VPN to V-spoke1. This route carries the route target RT-RED-FROM-HUB1. V-spoke1 subjects this route to its import policy and accepts it because it carries the route target RT-RED-FROM-HUB1.

Now, V-spoke1 begins normal operation, sending all of its RED-VPN traffic through V-hub1. At some point, V-spoke1 determines that it



might benefit from a more direct route to a destination. (Criteria by which V-spoke1 determines that it needs a more direct route are beyond the scope of this document.)

In order to discover a more direct route, V-spoke1 assigns a unique numeric identifier to the destination. V-spoke1 then sends a ROUTE-REFRESH message to the RR, containing the following information:

- o AFI is equal to IPv4 or IPv6, as appropriate
- o SAFI is equal to "MPLS-labeled VPN address"
- o When-to-refresh is equal IMMEDIATE
- o Action is equal to ADD
- o Match is equal to PERMIT
- o ORF Type is equal to CP-ORF
- o CP-ORF Sequence is equal to the identifier associated with the destination
- o CP-ORF Minlen is equal to 1
- o CP-ORF Maxlen is equal to 32 or 128, as appropriate
- o CP-ORF VPN Route Target is equal to RT-RED
- o CP-ORF Host Address is equal the destination address
- o CP-ORF Import Route Target is equal to RT-RED-FROM-HUB1

Upon receipt of the ROUTE-REFRESH message, the RR must ensure that it carries all routes belonging to the RED-VPN. In at least one special case, where all of the RR clients are V-spokes and none of the RR clients are V-hubs, the RR will lack some or all of the required RED-VPN routes. So, the RR sends a BGP UPDATE message containing a Route Target Membership NLRI for VPN-RED to all of its peers. This causes the peers to advertise VPN-RED routes to the RR, if they have not done so already.

Next, the RR adds the received CP-ORF to the Outbound Filter associated with V-spoke1. Using the procedures in Section 3, the RR determines whether any of the routes in its Loc-RIB satisfy the selection criteria of the newly updated Outbound Filter. If any routes satisfy the match criteria, they are added to the Adj-RIB-Out associated with V-spoke1. In Adj-RIB-Out, the RR adds RT-RED-FROM-

HUB1 to the list of Route Targets that the route already carries. Finally, RR advertises the newly added routes to V-spoke1. The advertised routes may specify either V-hub1 or any other node as the NEXT-HOP.

V-spoke1 subjects the advertised routes to its import policy and accepts them because they carry the route target RT-RED-FROM-HUB1.

V-spoke1 may repeat this process whenever it discovers another flow that might benefit from a more direct route to its destination.

## 5. Applicability In Network Virtualization Overlays

In an EVPN environment, Layer 2 networks are connected to Provider Edge (PE) devices. PE devices can be real or virtualized. Within a given EVPN, one or more EVPN Instances (EVI) can serve as a Default MAC Gateway (DMG). Each DMG advertises a Default MAC Route (DMR) to the rest of the EVIs in the EVPN. EVIs use the DMR to forward traffic destined to MAC addresses for which they do not have a corresponding MAC Advertisement Route.

For the purposes of example, assume the following:

- o Layer 2 Networks belonging to the RED-VPN are attached to PEs that support EVPN.
- o At any given point in time, an end-system that belongs to the RED-VPN communicates with only a small subset of other end-systems that belong to the RED-VPN. Therefore, at any given point in time, most of the PEs that serve the RED-VPN use only a small subset of the MAC Advertisement Routes in the RED-VPN.
- o One PE device serves as a DMG for the RED-VPN. We will call this device DMG 1. The RED-VPN EVI on DMG 1 is provisioned with RT-RED-FROM-HUB1 as its export RT, and RT-RED as its import RT.
- o Another PE device that hosts an EVI of the RED-VPN can not accommodate all RED-VPN MAC Advertisement routes. We will call this device Spoke 1. This EVI is provisioned with RT-RED as its export RT, and RT-RED-FROM-HUB1 as its import RT.
- o All PE devices that have EVIs of the RED-VPN advertise various EVPN routes, including MAC Advertisement Routes to one or more RRs.

DMG 1 serves the RED-VPN. Therefore, DMG 1 advertises a DMR for the RED-VPN to the RR, carrying the route target RT-RED-FROM-HUB1.

Spoke 1 establishes a BGP session with the RR, negotiating the CP-ORF capability, as well as the Multiprotocol Extensions Capability [RFC2858]. Upon establishment of the BGP session, the RR does not advertise any routes to Spoke 1. The RR will not advertise any routes until it receives either a ROUTE-REFRESH message or a BGP UPDATE message containing a Route Target Membership NLRI [RFC4684].

Immediately after the BGP session is established, Spoke 1 sends the RR a BGP UPDATE message containing a Route Target Membership NLRI. The Route Target Membership NLRI specifies RT-RED-FROM-HUB1 as its route target. In response to the BGP-UPDATE message, the RR advertises the DMR for the RED-VPN to Spoke 1. This route carries the route target RT-RED-FROM-HUB1. Spoke 1 subjects this route to its import policy and accepts it because it carries the route target RT-RED-FROM-HUB1.

Now, Spoke 1 begins normal operation, sending all of its RED-VPN traffic through DMG 1. At some point, Spoke 1 determines that it might benefit from a more direct route to a destination. (Criteria by which V-spoke1 determines that it needs a more direct route are beyond the scope of this document.)

In order to discover a more direct route, Spoke 1 assigns a unique numeric identifier to the destination. Spoke 1 then sends a ROUTE-REFRESH message to the RR, containing the following information:

- o AFI is equal to L2VPN
- o SAFI is equal to BGP EVPN
- o When-to-refresh is equal IMMEDIATE
- o Action is equal to ADD
- o Match is equal to PERMIT
- o ORF Type is equal to CP-ORF
- o CP-ORF Sequence is equal to the identifier associated with the destination
- o CP-ORF Minlen is equal to 48
- o CP-ORF Maxlen is equal to 48
- o CP-ORF VPN Route Target is equal to RT-RED
- o CP-ORF Host Address is equal the destination address

- o CP-ORF Import Route Target is equal to RT-RED-FROM-HUB1

Next, the RR adds the received CP-ORF to the Outbound Filter associated with Spoke 1. Using the procedures in Section 3, the RR determines whether any of the MAC Advertisement routes in its Loc-RIB satisfy the selection criteria of the newly updated Outbound Filter. If any of these routes satisfy the match criteria, they are added to the Adj-RIB-Out associated with Spoke 1. In Adj-RIB-Out, the RR adds RT-RED-FROM-HUB1 to the list of Route Targets that the route already carries. Finally, RR advertises the newly added routes to Spoke 1. The advertised routes carry as their NEXT-HOP the address of the PE device from which the routes have been originated.

Spoke 1 subjects the the MAC Advertisement Routes received from RR to its import policy and accepts them because they carry the route target RT-RED-FROM-HUB1.

Spoke 1 may repeat this process whenever it discovers another flow that might benefit from a more direct route to its destination.

## 6. Clean-up

Each CP-ORF consumes memory and compute resources on the device that supports it. Therefore, in order to obtain optimal performance, BGP speakers periodically evaluate all CP-ORFs that they have originated and remove unneeded CP-ORFs. The criteria by which a BGP speaker identifies unneeded CP-ORF entries is a matter of local policy, and is beyond the scope of this document.

## 7. IANA Considerations

IANA is requested to assign a All Covering Prefixes ORF Type from the BGP Outbound Route Filtering (ORF) Types Registry.

## 8. Security Considerations

Each CP-ORF consumes memory and compute resources on the device that supports it. Therefore, a device supporting CP-ORF take the following steps to protect itself from oversubscription:

- o When negotiating the ORF capability, advertise willingness to receive the CP-ORF only to known, trusted iBGP peers
- o Enforce a per-peer limit on the number of CP-ORFs that can be installed at any given time. Ignore all requests to add CP-ORFs beyond that limit

## 9. Acknowledgements

The authors wish to acknowledge Han Nguyen and James Uttaro for their comments and contributions.

## 10. Normative References

- [I-D.ietf-l2vpn-evpn]  
Sajassi, A., Aggarwal, R., Henderickx, W., Balus, F., Isaac, A., and J. Uttaro, "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-04 (work in progress), July 2013.
- [IANA.AFI]  
IANA, "abbrev="Address Family Numbers"", <<http://www.iana.org/assignments/address-family-numbers/address-family-numbers.xhtml>>.
- [IANA.SAFI]  
IANA, "abbrev="Subsequent Address Family Identifiers (SAFI) Parameters"", <<http://www.iana.org/assignments/safi-namespace/safi-namespace.xhtml#safi-namespace-2>>.
- [RFC0826] Plummer, D., "Ethernet Address Resolution Protocol: Or converting network protocol addresses to 48.bit Ethernet address for transmission on Ethernet hardware", STD 37, RFC 826, November 1982.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2858] Bates, T., Rekhter, Y., Chandra, R., and D. Katz, "Multiprotocol Extensions for BGP-4", RFC 2858, June 2000.
- [RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
- [RFC4684] Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk, R., Patel, K., and J. Guichard, "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, November 2006.
- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.

- [RFC5291] Chen, E. and Y. Rekhter, "Outbound Route Filtering Capability for BGP-4", RFC 5291, August 2008.
- [RFC5292] Chen, E. and S. Sangli, "Address-Prefix-Based Outbound Route Filter for BGP-4", RFC 5292, August 2008.
- [RFC7024] Jeng, H., Uttaro, J., Jalil, L., Decraene, B., Rekhter, Y., and R. Aggarwal, "Virtual Hub-and-Spoke in BGP/MPLS VPNs", RFC 7024, October 2013.

Authors' Addresses

Huajin Jeng  
AT&T

Email: [hj2387@att.com](mailto:hj2387@att.com)

Luay Jalil  
Verizon

Email: [luay.jalil@verizon.com](mailto:luay.jalil@verizon.com)

Ron Bonica  
Juniper Networks  
2251 Corporate Park Drive  
Herndon, Virginia 20170  
USA

Email: [rbonica@juniper.net](mailto:rbonica@juniper.net)

Yakov Rekhter  
Juniper Networks  
1194 North Mathilda Ave.  
Sunnyvale, California 94089  
USA

Email: [yakov@juniper.net](mailto:yakov@juniper.net)

Keyur Patel  
Cisco Systems  
170 W. Tasman Drive  
San Jose, California 95134  
USA

Email: [keyupate@cisco.com](mailto:keyupate@cisco.com)

Lucy Yong  
Huawei Technologies

Email: [lucy.yong@huawei.com](mailto:lucy.yong@huawei.com)

Network Working Group  
Internet-Draft  
Updates: 6514 (if approved)  
Intended status: Standards Track  
Expires: August 9, 2014

Zhang  
Rekhter  
Juniper Networks  
Dolganow  
Alcatel-Lucent  
February 5, 2014

Simulating "Partial Mesh of MP2MP P-Tunnels" with Ingress Replication  
draft-ietf-l3vpn-mvpn-bidir-ingress-replication-00.txt

## Abstract

RFC 6513 described a method to support bidirectional C-flow using "Partial Mesh of MP2MP P-Tunnels". This document describes how partial mesh of MP2MP P-Tunnels can be simulated with Ingress Replication, instead of a real MP2MP tunnel. This enables a Service Provider to use Ingress Replication to offer transparent BIDIR-PIM service to its VPN customers.

## 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 9, 2014.

## Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must



include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	3
1.1. Terminology . . . . .	3
2. Requirements Language . . . . .	4
3. Operation . . . . .	5
3.1. Control State . . . . .	5
3.2. Forwarding State . . . . .	7
4. Security Considerations . . . . .	9
5. IANA Considerations . . . . .	10
6. Acknowledgements . . . . .	11
7. Normative References . . . . .	12
Authors' Addresses . . . . .	13

## 1. Introduction

Section 11.2 of RFC 6513, "Partitioned Sets of PEs", describes two methods of carrying bidirectional C-flow traffic over a provider core without using the core as RPL or requiring Designated Forwarder election.

With these two methods, all PEs of a particular VPN are separated into partitions, with each partition being all the PEs that elect the same PE as the Upstream PE wrt the C-RPA. A PE must discard bidirectional C-flow traffic from PEs that are not in the same partition as the PE itself.

In particular, Section 11.2.3 of RFC 6513, "Partial Mesh of MP2MP P-Tunnels", guarantees the above discard behavior without using an extra PE Distinguisher label by having all PEs in the same partition join a single MP2MP tunnel dedicated to that partition and use it to transmit traffic. All traffic arriving on the tunnel will be from PEs in the same partition, so it will be always accepted.

RFC 6514 specifies BGP encodings and procedures used to implement MVPN as specified in RFC 6513, while the details related to MP2MP tunnels are specified in [draft-ietf-l3vpn-mvpn-bidir-05].

[draft-ietf-l3vpn-mvpn-bidir-05] assumes that an MP2MP P-tunnel is realized either via PIM-Bidir, or via MP2MP mLDP. Each of them would require signaling and state not just on PEs, but on the P routers as well. This document describes how the MP2MP tunnel can be simulated with a mesh of P2MP tunnels, each of which is instantiated by Ingress Replication. This does not require each PE on the MP2MP tunnel to send an S-PMSI A-D route for the P2MP tunnel that the PE is the root for, nor does it require each PE to send a Leaf A-D route to the root of each P2MP tunnel in the mesh.

With the use of Ingress Replication, this scheme has both the advantages and the disadvantages of Ingress Replication in general.

### 1.1. Terminology

This document uses terminology from [RFC6513], [RFC6514], and [draft-ietf-l3vpn-mvpn-bidir-05]. In particular, the following new term is defined:

- o C-G-BIDIR: A C-G where G is a Bidir-PIM group.

## 2. Requirements Language

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. Operation

In following sections, the originator of an S-PMSI A-D route or Leaf A-D route is determined from the "originating router's IP address" field of the corresponding route.

#### 3.1. Control State

If a PE, say PEx, is connected to a site of a given VPN, and PEx's next hop interface to some C-RPA is a VRF interface, then PEx MUST advertise a (C-\*,C-BIDIR) S-PMSI A-D route, regardless of whether it has any local Bidir-PIM join states corresponding to the C-RPA learned from its CEs. It MAY also advertise one or more (C-\*,C-G-BIDIR) S-PMSI A-D route, just like how any other S-PMSI A-D routes are triggered. Here the C-G-BIDIR refers to a C-G where G is a Bidir-PIM group, and the corresponding C-RPA is in the site that the PEx connects to. For example, the (C-\*,C-G-BIDIR) S-PMSI A-D routes could be triggered when the (C-\*, C-G-BIDIR) traffic rate goes above a threshold, and fan-out could also be taken into account. Note that this requires measuring the traffic in both directions, due to the nature of Bidir-PIM.

The S-PMSI A-D routes include a PMSI Tunnel Attribute (PTA) with tunnel type set to Ingress Replication, with Leaf Information Required flag set, with a downstream allocated MPLS label that other PEs in the same partition MUST use when sending relevant C-bidir flows to this PE, and with the Tunnel Identifier field in the PTA set to a routable address of the originator. The label may be shared with other P-tunnels, subject to the anti-ambiguity rules for extranet. For example, the (C-\*,C-BIDIR) and (C-\*,C-G-BIDIR) S-PMSI A-D routes originated by a given PE can optionally share a label.

If some other PE, PEy, receives and imports into one of its VRFs any (C-\*, C-BIDIR) S-PMSI A-D route whose PTA specifies an IR P-tunnel, and the VRF has any local Bidir-PIM join state that PEy has received from its CEs, and if PEy chooses PEx as its Upstream PE wrt the C-RPA for those states, PEy MUST advertise a Leaf A-D route in response. Or, if PEy has received and imported into one of its VRFs a (C-\*,C-BIDIR) S-PMSI A-D route from PEx before, then upon receiving in the VRF any local Bidir-PIM join state from its CEs with PEx being the Upstream PE for those states' C-RPA, PEy MUST advertise a Leaf A-D route.

The encoding of the Leaf A-D route is as specified in RFC 6514, except that the Route Targets are set to the same value as in the corresponding S-PMSI A-D route so that the Leaf A-D route will be imported by all VRFs that import the corresponding S-PMSI A-D route. This is irrespective of whether from a receiving PE, PEz's

perspective PEx (originator of the S-PMSI A-D route) is the Upstream PE or not. The label in the PTA of the Leaf A-D route originated by PEy MUST be allocated specifically for PEx, so that when traffic arrives with that label, the traffic can be associated with the partition (represented by the PEx). The label may be shared with other P-tunnels, subject to the anti-ambiguity rules for extranet. For example, the (C-\*,C-BIDIR) and (C-\*,C-G-BIDIR) S-PMSI A-D routes originated by a given PE can optionally share a label.

Note that RFC 6514 requires a PE/ASBR take no action with regard to a Leaf A-D route unless that Leaf A-D route carries an IP Address Specific RT identifying the PE/ASBR. This document removes that requirement when the route key of a Leaf A-D route identifies a (C-\*,C-BIDIR) or a (C-\*,C-G-BIDIR) S-PMSI.

To speed up convergence (so that PEy starts receiving traffic from its new Upstream PE immediately instead of waiting until the new Leaf A-D route corresponding to the new Upstream PE is received by sending PEs), PEy MAY advertise a Leaf A-D route even if it does not choose PEx as its Upstream PE wrt the C-RPA. With that, it will receive traffic from all PEs, but some will arrive with the label corresponding to its choice of Upstream PE while some will arrive with a different label, and the traffic in the latter case will be discarded.

Similar to the (C-\*,C-BIDIR) case, if PEy receives and imports into one of its VRFs any (C-\*,C-G-BIDIR) S-PMSI A-D route whose PTA specifies an IR P-tunnel, and PEy chooses PEx as its Upstream PE wrt the C-RPA, and it has corresponding local (C-\*,C-G-BIDIR) join state that it has received from its CEs in the VRF, PEy MUST advertise a Leaf A-D route in response. Or, if PEy has received and imported into one of its VRFs a (C-\*,C-G-BIDIR) S-PMSI A-D route before, then upon receiving its local (C-\*,C-G-BIDIR) join state from its CEs in the VRF, it MUST advertise a Leaf A-D route.

The encoding of the Leaf A-D route is as specified in RFC 6514, except that the Route Targets are set to the same as in the corresponding S-PMSI A-D route so that the Leaf A-D route will be imported by all VRFs that import the corresponding S-PMSI A-D route. This is irrespective of whether from the receiving PE, PEz's perspective PEx (originator of the S-PMSI A-D route) is the Upstream PE or not. The label in the PTA of the Leaf A-D route originated by PEy MUST be allocated specifically for PEx, so that when traffic arrives with that label, the traffic can be associated with the partition (represented by the PEx). The label may be shared with other P-tunnels, subject to the anti-ambiguity rules for extranet. For example, the (C-\*,C-BIDIR) and (C-\*,C-G-BIDIR) S-PMSI A-D routes originated by a given PE can optionally share a label.

Whenever the (C-\*,C-BIDIR) or (C-\*,C-G-BIDIR) S-PMSI A-D route is withdrawn, or if PEy no longer chooses the originator PEx as its Upstream PE wrt C-RPA and PEy only advertises Leaf A-D routes in response to its Upstream PE's S-PMSI A-D route, or if relevant local join state is pruned, PEy MUST withdraw the corresponding Leaf A-D route.

### 3.2. Forwarding State

The following specification regarding forwarding state matches the "When an S-PMSI is a 'Match for Transmission'" and "When an S-PMSI is a 'Match for Reception'" rules for "Flat Partitioning" method in [draft-ietf-l3vpn-mvpn-bidir-05], except that the rules about (C-\*,C-\*) are not applicable, because this document requires that (C-\*,C-BIDIR) S-PMSI A-D routes are always originated for a VPN that supports C-Bidir flows.

For the (C-\*,C-G-BIDIR) S-PMSI A-D route that a PEy receives and imports into one of its VRFs from its Upstream PE wrt the C-RPA, or if PEy itself advertises the S-PMSI A-D route in the VRF, PEy maintains a (C-\*,C-G-BIDIR) forwarding state in the VRF, with the Ingress Replication provider tunnel leaves being the originators of the S-PMSI A-D route and all relevant Leaf-A-D routes. The relevant Leaf A-D routes are the routes whose Route Key field contains the same information as the MCAST-VPN NLRI of the (C-\*, C-G-BIDIR) S-PMSI A-D route advertised by the Upstream PE.

For the (C-\*,C-BIDIR) S-PMSI A-D route that a PEy receives and imports into one of its VRFs from its Upstream PE wrt a C-RPA, or if PEy itself advertises the S-PMSI A-D route in the VRF, it maintains appropriate forwarding states in the VRF for the ranges of bidirectional groups for which the C-RPA is responsible. The provider tunnel leaves are the originators of the S-PMSI A-D route and all relevant Leaf-A-D routes. The relevant Leaf A-D routes are the routes whose Route Key field contains the same information as the MCAST-VPN NLRI of the (C-\*, C-BIDIR) S-PMSI A-D route advertised by the Upstream PE. This is for the so-called "Sender Only Branches" where a router only has data to send upstream towards C-RPA but no explicit join state for a particular bidirectional group. Note that the traffic must be sent to all PEs (not just the Upstream PE) in the partition, because they may have specific (C-\*,C-G-BIDIR) join states that this PEy is not aware of, while there is no corresponding (C-\*,C-G-BIDIR) S-PMSI A-D and Leaf A-D routes.

For a (C-\*,C-G-BIDIR) join state that a PEy has received from its CEs in a VRF, if there is no corresponding (C-\*,C-G-BIDIR) S-PMSI A-D route from its Upstream PE in the VRF, PEy maintains a corresponding forwarding state in the VRF, with the provider tunnel leaves being

the originators of the (C-\*,C-BIDIR) S-PMSI A-D route and all relevant Leaf-A-D routes (same as the above Sender Only Branch case). The relevant Leaf A-D routes are the routes whose Route Key field contains the same information as the MCAST-VPN NLRI of the (C-\*, C-BIDIR) S-PMSI A-D route originated by the Upstream PE. If there is no (C-\*,C-BIDIR) S-PMSI A-D route from its Upstream PE either, then the provider tunnel has an empty set of leaves and PEy does not forward relevant traffic across the provider network.

#### 4. Security Considerations

This document raises no new security issues. Security considerations for the base protocol are covered in [RFC6514].



## 5. IANA Considerations

This document has no IANA considerations.

This section should be removed by the RFC Editor prior to final publication.

## 6. Acknowledgements

We would like to thank Eric Rosen for his comments, and suggestions of some texts used in the document.

## 7. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC6513] Rosen, E. and R. Aggarwal, "Multicast in MPLS/BGP IP VPNs", RFC 6513, February 2012.
- [RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, February 2012.
- [RFC5015] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", RFC 5015, October 2007.
- [I-D.ietf-l3vpn-mvpn-bidir]  
Rosen, E., Wijnands, I., Cai, Y., and A. Boers, "MVPN: Using Bidirectional P-Tunnels",  
draft-ietf-l3vpn-mvpn-bidir-06 (work in progress),  
October 2013.

Authors' Addresses

Jeffrey Zhang  
Juniper Networks  
10 Technology Park Dr.  
Westford, MA 01886  
US

Email: [zzhang@juniper.net](mailto:zzhang@juniper.net)

Yakov Rekhter  
Juniper Networks  
1194 North Mathilda Ave.  
Sunnyvale, CA 94089  
US

Email: [yakov@juniper.net](mailto:yakov@juniper.net)

Andrew Dolganow  
Alcatel-Lucent  
600 March Rd.  
Ottawa, ON K2K 2E6  
CANADA

Email: [andrew.dolganow@alcatel-lucent.com](mailto:andrew.dolganow@alcatel-lucent.com)

