INTERNET-DRAFT
Updates: 6074 (if approved)
Intended Status: Standards Track
Expires: August 23, 2015

Avik Bhattacharya
Apratim Mukherjee
Ixia

February 19, 2015

Provisioning, Auto-Discovery, and Signaling in L2VPNs for IPv6 Remote PE draft-abhattacharya-bess-l2vpn-ipv6-remotepe-03

Abstract

L2VPN Signaling specification defines the semantic structure of the endpoint identifiers required by each model. It discusses the distribution of these identifiers by the discovery process, especially when such discovery is based on the Border Gateway Protocol (BGP). This document updates the end point encoding for BGP-Based Auto-Discovery and specifies a format for NLRI encoding for IPv6 PE Address. This document also specifies a new type of attachment identifier to carry IPv6 address as AII in LDP FEC 0x81. This document updates RFC6074.

Status of this Memo

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

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html

This Internet-Draft will expire on August 23, 2015.

Copyright and License Notice

Table of Contents

1 Introduction

[RFC6074] specifies a number of L2VPN provisioning models, and further specifies the semantic structure of the endpoint identifiers required by each model. It discusses the distribution of these identifiers by the discovery process, especially when discovery is based on the Border Gateway Protocol (BGP). It then specifies how the endpoint identifiers are carried in the two signaling protocols that are used to set up PWs, the Label Distribution Protocol (LDP), and the Layer 2 Tunneling Protocol version 3 (L2TPv3) [RFC6074]. This document updates Section 3.2.2.1 of RFC 6074 (BGP-Based Auto-Discovery) and specifies a format for NLRI encoding that allows to carry also an IPv6 PE Address.This document also specifies a new type of attachment identifier to carry IPv6 address as AII in LDP FEC 0x81. This gap in the specification of L2VPN in IPv6 only MPLS Network is also recognized in section 3.3.1 of [RFC7439].

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 RFC 2119 [RFC2119].

2 BGP NLRI Format for the IPv6 PE Address

Section 3.2.2.1 of [RFC6074] specifies the BGP advertisement for a particular VSI at a given PE will contain:

- an NLRI of AFI = L2VPN, SAFI = VPLS, encoded as RD:PE_addr
- a BGP next hop equal to the loopback address of the PE
- an Extended Community Attribute containing the VPLS-id
- an Extended Community Attribute containing one or more RTs.

The format for the NLRI encoding defined in Section 3.2.2.1 of [RFC6074] is:

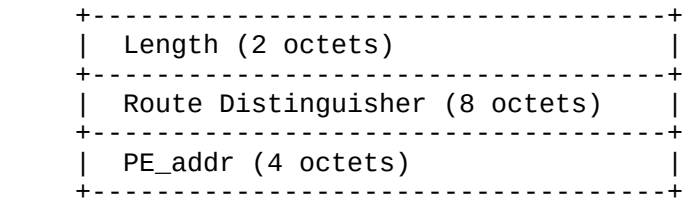


Figure 1: NLRI encoding in [RFC6074]

In this format the size of the PE_addr is defined as 4 octets which can carry only IPv4 addresses. In a situation where the route is originating from a BGP end point running on an IPv6 address, the PE_addr in the NLRI needs to carry that IPv6 address. The updated format for the NLRI encoding is depicted in Figure 2.

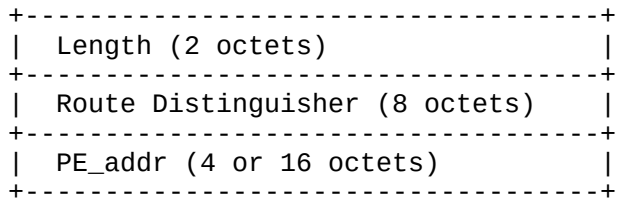


Figure 2: Updated NLRI encoding

The length field MUST contain the sum of the length of the Length field(2), the length of the Route Distinguisher (8) and the length of the 4 or 16 octet PE_addr field.

The type of the PE_addr can be derived by the receiving node by subtracting the fixed length of the Route Distinguisher and the Length field from the value of the received Length. An IPv4 PE_addr should be used to initiate adjacency of the underlying signaling protocol if it supports IPv4. An IPv6 PE_addr should be used to initiate adjacency of the underlying signaling protocol if it supports IPv6. (such as LDPv6)

3 Discussion on Route Distinguisher (RD) and Route Target (RT)

Note that RD and RT can be in format AS 2byte + 4 byte Assigned Number or IP 4 byte + 2 byte Assigned Number [RFC4364]. Just like RD or RT cannot carry 4 byte AS numbers , they also cannot utilize 16 byte IPv6 Address. Updates to RD and RT to operate in a pure IPv6 environment is outside the scope of this document.

4 Using IPv6 Remote PE address for signaling using LDP

Section 5.3.2 of [RFC4447] specifies the format of encoding for Generalized ID FEC Element (FEC 0x81)which is used for signaling in LDP. This document specifies a new type for AII carrying IPv6 address as TAII or SAII. (See Section 7)

An FEC 0x81 TLV MUST contain SAII and TAII of the same type i.e. either type 1 or type 2.

5 Interoperability in a mixed IPv4/IPv6 Network

If a VPLS instance is reachable though both IPv4 and IPv6 loopback in a PE node then the BGP instance(s) of that PE node MUST advertise the VPLS route using both NLRIs - one with IPv4 PE_addr and another with IPv6 PE_addr.

While signaling a TAII in type 2 format, the LDP implementation MUST use SAII also in type 2 format. The value of the SAII MAY be set from the IPv6 loopback address on which the BGP session is established.

While signaling a TAII type over an LDP session, on which it has already signaled with the other TAII type but with the same AGI, it SHOULD use the same label value in the Label Mapping for both TAII types.

On receiving an FEC 0x81 TLV in a Label Advertisement with a TAII type, the LDP implementation MAY lookup if on the same LDP session it has received a Label Mapping with the other TAII type but for the same AGI. If yes then it MUST store the Label Mapping but MAY choose not to install the label. If it chooses not to do the lookup stated above then it MUST install the received label.

If the LDP implementation chooses to do the lookup stated above during receipt of the Label Mapping, on receiving an FEC 0x81 TLV in a Label Withdraw with a TAII type, the LDP implementation MUST lookup if on the same LDP session it has received another Label Mapping with other TAII type but same AGI. If yes then it MUST install the stored Label Mapping and keep using that thereafter. (Along with taking necessary actions for processing the Label Withdraw as specified in [RFC5036])

6 Security Considerations

There is no additional security impact in addition to what is mentioned in [RFC6074].

7 IANA Considerations

This document requires a new AII type to be used in Generalized ID FEC (0x81). IANA already maintains a registry of name "Attachment Individual Identifier(AII) Type" specified by [RFC4446].

The following value is suggested for assignment:

AII Type	Length	Description
0x02	16	A 128 bit unsigned number local identifier.

8. Acknowledgments

Thanks to Mohamed Boucadair for his valuable suggestions.

9 References

9.1 Normative References

[RFC2119] S. Bradner,"Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997, <http://www.rfc-editor.org/info/rfc2119>.

[RFC6074] E. Rosen, B. Davie, "Provisioning, Auto-Discovery, and Signaling in Layer 2 Virtual Private Networks (L2VPNs)", BCP 74, RFC 6074, January 2011, <http://www.rfc-editor.org/info/rfc6074>.

[RFC4446] Martini, L., "IANA Allocations for Pseudowire Edge to Edge Emulation (PWE3)", BCP 116, RFC 4446, April 2006, <http://www.rfc-editor.org/info/rfc4446>.

[RFC4447] Martini, L., Rosen, E., El-Aawar, N., Smith, T., and G. Heron, "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", RFC 4447, April 2006, <http://www.rfc-editor.org/info/rfc4447>.

[RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007,<http://www.rfc-editor.org/info/rfc5036>.

[RFC7439] W. George, C. Pignataro,, "Gap Analysis for Operating IPv6-Only MPLS Networks", RFC 7439, January 2015,<http://www.rfc-editor.org/info/rfc7439>.

9.2 Informative References

[RFC4364] E. Rosen, "BGP/MPLS IP Virtual Private Networks (VPNs)", BCP 78, RFC 4364, February 2006, <http://www.rfc-editor.org/info/rfc4364>.

Authors' Addresses

Avik Bhattacharya
Ixia
Infinity Building, Tower 2, Floor -4
Sector 5, Saltlake,
Kolkata, West Bengal, India - 700091.

EMail: abhattacharya@ixiacom.com

Apratim Mukherjee
Ixia
Infinity Building, Tower 2, Floor -4
Sector 5, Saltlake,
Kolkata, West Bengal, India - 700091.

EMail: amukherjee@ixiacom.com

INTERNET-DRAFT
Intended Status: Standard Track

Sami Boutros
VMware

Patrice Brissette
Ali Sajassi
Cisco Systems

Daniel Voyer
Bell Canada

John Drake
Juniper Networks

Expires: December 31, 2017

June 29, 2017

EVPN-VPWS Service Edge Gateway
draft-boutros-bess-evpn-vpws-service-edge-gateway-04

Abstract

This document describes how a service node can dynamically terminate EVPN virtual private wire transport service (VPWS) from access nodes and offer Layer 2, Layer 3 and Ethernet VPN overlay services to Customer edge devices connected to the access nodes. Service nodes using EVPN will advertise to access nodes the L2, L3 and Ethernet VPN overlay services it can offer for the terminated EVPN VPWS transport service. On an access node an operator can specify the L2 or L3 or Ethernet VPN overlay service needed by the customer edge device connected to the access node that will be transported over the EVPN-VPWS service between access node and service node.

Status of this Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

Copyright and License Notice

Table of Contents

1 Introduction

This document describes how a service node can act as a gateway terminating dynamically EVPN virtual private wire service (VPWS) from access nodes and offering Layer 2, EVPN and Layer 3 VPN overlay services to Customer edge devices connected to the access nodes.

The service node would initially advertise using EVPN the different L2, L3 and Ethernet VPN overlay services that can be transported from access nodes over an EVPN-VPWS transport service.

The service node would advertise EVPN-VPWS per EVI Ethernet A-D routes with the Ethernet Segment Identifier field set to 0 and the Ethernet tag ID set to (0xFFFFFFFF wildcard), all those routes will be associated with the EVPN-VPWS service edge RT that will be imported by other service edge PEs, each route will have a unique RD and will be associated with another RT corresponding to the L2, L3 or Ethernet VPN overlay service that can be transported over the EVPN-VPWS transport service.

The access nodes will advertise EVPN-VPWS per EVI Ethernet A-D with the Ethernet Segment Identifier field set to 0 for single home customer edge CE device and set to the CE’s ESI and the Ethernet Tag field is set to the VPWS service instance identifier. The route will have a unique RD and will be associated with an RT corresponding to the L2, L3 or Ethernet VPN overlay service that will be transported over the EVPN-VPWS transport service.

If more than one service node advertise the ability to terminate the EVPN-VPWS transport service and offer the L2, L3 or Ethernet VPN service required by CE device connected to a given access node, then all service node(s) will perform a DF election based on HWR algorithm using {Ethernet tag-id, Service node IP addresses} to determine which service node will be the primary service node to to terminate the VPWS service and offer the L2, L3 or Ethernet overlay service for the customer edge, All active and single active redundancy can be offered.

The Service PE node that is a DF for a given VPWS service ID MUST respond to the Eth A-D route per EVI from the access node by sending its own Eth A-D per EVI route and by setting the same VPWS service instance ID and downstream assigned MPLS label to be used by Access node. When access node receives this Eth A-D route per EVI from the service node, it binds the two side of EVCs together.

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 RFC 2119 [RFC2119].

2. Requirements

This section describes the requirements specific to this draft. These requirements are in addition to the ones described in [EVPN-REQ], [EVPN], and [EVPN-VPWS].

2.1 Auto-Discovery

A service node needs to support the following functionality of auto-discovery:

(R1a) A service node PE MUST be agnostic of all access nodes PEs connected on the same access network.

(R1b) A service node PE MUST advertise its associated overlay VRF(L2 and/or L3) to all service nodes PEs connected on the same network.

(R1c) A service node PE MUST resolve received overlay VRF(L2 and/or L3) from other service nodes with local configuration. The information is used to select proper service node PE for a given EVPN-VPWS connection from an access PE.

(R1d) A service node PE MUST accept EVPN-VPWS connection from any access node PE which require one of the service node PE available L2 or L3 overlay service.

2.2 Scalability

(R2a) A single service node PE can be associated with many access node PEs. The following requirements give a quantitative measure.

(R2b) A service node PE MUST support thousand(s) head-end connections for a a given access node PE connecting to different overlay VRF services on that service node.

(R2c) A service node PE MUST support thousand(s) head-end connections to many access node PEs.

2.3 Head-end

(R3a) A service node PE MUST support L2 and/or L3 head-end functionality.

(R3b) A service node PE SHALL support auto-configuration of L2 and/or

L3 head-end functionality.

2.5 Multi-homing

TBD

2.5 Fast Convergence

TBD

3. Benefits

This section describes some of the major benefits of EVPN-VPWS service edge gateway solution. This list is not considered as exhaustive.

Majors benefits are:

- An easy and scalable mechanism for tunneling (head-end) customer traffic into a common IP/MPLS network infra structure
- Auto-provision features such as QOS access lists (ACL), tunnel preference, bandwidth, L3VPN on a per head-end interface basis
- reduces CAPEX in the access or aggregation network and service PE
- Auto configuration of head-end functionality:

  Configuring other Layer3 parameters, such as VRF and IP addresses, are optional for the head-end to be functional. However, they are required for Layer3 services to be operational (head-end L3 termination).

- Auto-discovery of access nodes by service nodes. Hence, there is no need to change any service node configuration when a new access node is being added to the access network.

4. Solution Overview

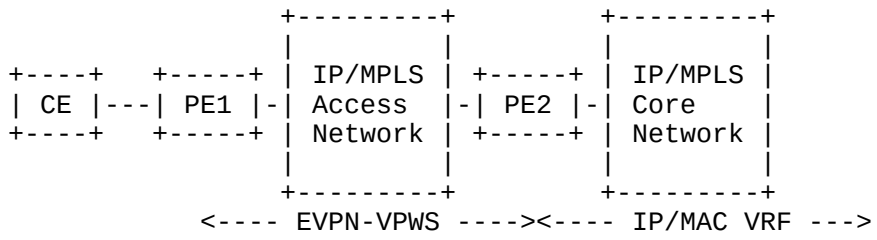


Figure 1: EVPN-VPWS Service Edge Gateway.
AN: Access node
SE: Service Edge node.

EVPN-VPWS Service Edge Gateway Operation

At the service edge node, the EVPN Per-EVI Ethernet A-D routes will be advertised with the ESI set to 0 and the Ethernet tag-id set to (wildcard 0xFFFFFFF). The Ethernet A-D routes will have a unique RD and will be associated with 2 BGP RT(s), one RT corresponding to the underlay EVI i.e. the EVPN VPWS transport service that’s configured only among the service edge nodes, and one corresponding to the L2, L3 or EVPN overlay service.

At the access nodes, the EVPN per-EVI Ethernet A-D routes will be advertised as described in [draft-ietf-bess-evpn-vpws] with the ESI field is set to 0 and for single homed CEs and to the CE’s ESI for multi-homed CE’s and the Ethernet Tag field will be set to the VPWS service instance identifier that identifies the EVPL or EPL service. The Ethernet-AD route will have a unique RD and will be associated with one BGP RT corresponding to the L2, L3 or EVPN overlay service that will be transported over this EVPN VPWS transport service.

Service edge nodes on the underlay EVI will determine the primary service node terminating the VPWS transport service and offering the L2, L3 or Ethernet VPN service by running the on HWR algorithm as described in [draft-mohanty-l2vpn-evpn-df-election] using weight [VPWS service identifier, Service Edge Node IP address]. This ensure that service node(s) will consistently pick the primary service node even after service node failure. Upon primary service node failure, all other remaining services nodes will choose another service node correctly and consistently.

Single-sided signaling mechanism is used. The Service PE node that is a DF for accepts to terminate the VPWS transport service from an access node, the primary service edge node shall:- Dynamically create an interface to terminate the service and shall attach this interface to the overlay VPN service required by the access node to service its

customer edge device.- Responds to the Eth A-D route per EVI from the access node by sending its own Eth A-D per EVI route by setting the same VPWS service instance ID and downstream assigned MPLS label to be used by the access node.

When access node receives this Eth A-D route per EVI from the service edge node, it binds the two side of EVCs together and it now knows what primary/backup service nodes to forward the traffic to.

The service edge node shall support per features such as QoS, ACL, etc. for the EVPN VPWS transport service it terminates.

4.1 Multi-homing

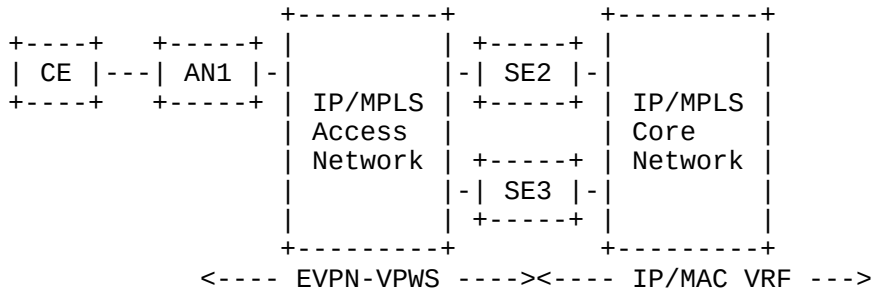


Figure 2: EVPN-VPWS SEG Multi-homing (same ASN)
AN: Access node
SE: Service Edge node.

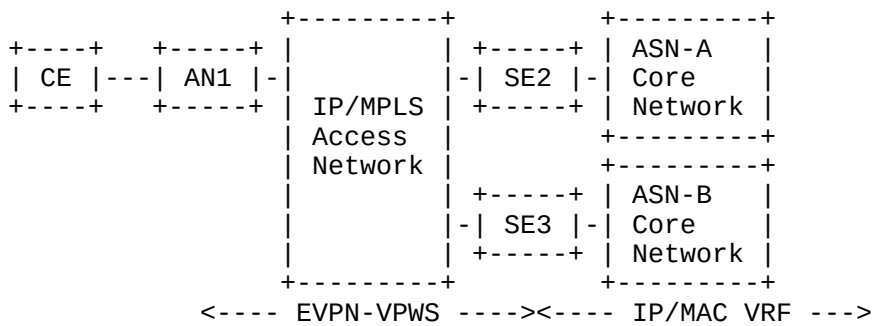


Figure 3: EVPN-VPWS SEG Multi-homing (different ASN)
AN: Access node
SE: Service Edge node.

Both All-active and single active redundancy can be supported.

A backup service node can be preprogrammed in data plane on an access node in order to switch traffic and based on how fast the data plane detect the failure of the primary service node traffic on an access node can switch to the backup node.

4.2 Applicability to IP-VPN TBD

5 Failure Scenarios TBD

6 Acknowledgements TBD.

7 Security Considerations

This document does not introduce any additional security constraints.

8 IANA Considerations

TBD.

9 References

9.1 Normative References

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

9.2 Informative References

[RFC7209] A. Sajassi, R. Aggarwal et. al., "Requirements for Ethernet VPN".

[EVPN] A. Sajassi, R. Aggarwal et. al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-11.txt.

[EVPN-VPWS] S. Boutros et. al., "EVPN-VPWS", draft-ietf-bess-evpn-vpws-00.txt.

Authors' Addresses

Sami Boutros
VMware, Inc.
Email: sboutros@vmware.com

Patrice Brissette
Cisco
Email: pbrisset@cisco.com

Ali Sajassi
Cisco
Email: sajassi@cisco.com

Daniel Voyer
Bell Canada
Email: daniel.voyer@bell.ca

John Drake
Juniper Networks
Email: jdrake@juniper.net

BESS WG
Internet-Draft
Updates: 6625 (if approved)
Intended status: Standards Track
Expires: August 19, 2016

A. Dolganow
J. Kotalwar
Alcatel-Lucent
E. Rosen, Ed.
Z. Zhang
Juniper Networks, Inc.

February 16, 2016

Explicit Tracking with Wild Card Routes in Multicast VPN
draft-dolganow-bess-mvpn-expl-track-02

Abstract

The MVPN specifications provide procedures to allow a multicast ingress node to invoke "explicit tracking" for a multicast flow or set of flows, thus learning the egress nodes for that flow or set of flows. However, the specifications are not completely clear about how the explicit tracking procedures work in certain scenarios. This document provides the necessary clarifications. It also specifies a new, optimized explicit tracking procedure. This new procedure allows an ingress node, by sending a single message, to request explicit tracking of each of a set of flows, where the set of flows is specified using a wildcard mechanism. This document updates RFC6625.

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 19, 2016.

Copyright Notice

Table of Contents

1. Introduction

[RFC6513] and [RFC6514] define the "Selective Provider Multicast Service Interface Auto-Discovery route" (S-PMSI A-D route). By originating one of these BGP routes, an ingress node advertises that it is transmitting a particular multicast flow. In the terminology of those RFCs, each flow is denoted by (C-S,C-G), where C-S is an IP source address and C-G is an IP multicast address, both in the address space of a VPN customer. The (C-S,C-G) of the multicast flow is encoded into the Network Layer Reachability Information (NLRI) of the S-PMSI A-D route.

Additionally, each S-PMSI A-D route contains a PMSI Tunnel attribute (PTA), which identifies a tunnel through the provider backbone network (a "P-tunnel"). If a P-tunnel is identified in the PTA of a given S-PMSI A-D route, the originator of that route is advertising that it will transmit the flow identified in the NLRI through the tunnel identified in the PTA.

[RFC6513] and [RFC6514] also define a procedure that allows an ingress node to determine the set of egress nodes that have requested to receive a particular flow from that ingress node. The ability of an ingress node to identify the egress nodes for a particular flow is known as "explicit tracking". An ingress node requests explicit tracking by setting a flag (the "Leaf Information Required" flag, or LIR) in the PTA. When an egress node receives an S-PMSI A-D route with LIR set, the egress node originates a Leaf A-D route whose NLRI contains the NLRI from the corresponding S-PMSI A-D route. In this way, the egress node advertises that it has requested to receive the particular flow identified in the NLRI of that S-PMSI A-D route.

[RFC6513] and [RFC6514] also allow an ingress node to originate an S-PMSI A-D route whose PTA has LIR set, but which does not identify any P-tunnel. This mechanism can be used when it is desired to do explicit tracking of a flow without at the same time binding that flow to a particular P-tunnel.

[RFC6625] (and other RFCs that update it) extends the specification of S-PMSI A-D routes, and allows an S-PMSI A-D route to encode a wildcard in its NLRI. Either the C-S or the C-G or both can be replaced by wildcards. These routes are known as (C-*,C-S) S-PMSI A-D routes, or as (C-S,C-*) S-PMSI A-D routes, or as (C-*,C-*) S-PMSI A-D routes, depending on whether the C-S or C-G or both have been replaced by wildcards. These routes are known jointly as "wildcard S-PMSI A-D routes".

One purpose of this document is to clarify the way that the explicit tracking procedures of [RFC6513] and [RFC6514] are applied when wildcard S-PMSI A-D routes are used.

In addition, this document addresses the following scenario, which is not addressed in [RFC6513], [RFC6514], or [RFC6625]. Suppose an ingress node originates an S-PMSI A-D route whose NLRI specifies, for example, (C-*,C-*) (i.e., both C-S and C-G are replaced by wildcards), and whose PTA identifies a particular P-tunnel. Now suppose that the ingress node wants explicit tracking for each individual flow that it transmits (following the procedures of [RFC6625] on that P-tunnel.

In this example, if the ingress node sets LIR in the PTA of the wildcard S-PMSI A-D route, each egress node that needs to receive a flow from the ingress node will respond with a Leaf A-D route whose NLRI specifies contains the (C-*,C-*) wildcard. This allows the ingress node to determine the set of egress nodes that are receiving flows from the ingress node. However, it does not allow the ingress node to determine which flows are being received by which egress nodes.

If the ingress node needs to determine which egress nodes are receiving which flows, it needs to originate an S-PMSI A-D route for each individual (C-S,C-G) flow that it is transmitting, and it needs to set LIR in the PTA of each such route. However, since all the flows are being sent through the tunnel identified in the (C-*,C-*) S-PMSI A-D route, there is no need to identify a tunnel in the PTA of each (C-S,C-G) S-PMSI A-D route. Per [RFC6514], the PTA of the (C-S,C-G) S-PMSI A-D routes can specify "no tunnel information". This procedure allows explicit tracking of individual flows, even though all those flows are assigned to tunnels in widlcard S-PMSI A-D routes.

Howver, this procedure requires several clarifications:

- The procedures of [RFC6625] do not clearly state how to handle an S-PMSI A-D route if its NLRI contains wild cards, but its PTA specifies "no tunnel info".

- If it is desired to send a set of flows through the same tunnel (where that tunnel is advertised in a wildcard S-PMSI A-D route), but it is also desired to explicitly track each individual flow transmitted over that tunnel, one has to send an S-PMSI A-D route (with LIR set in the PTA) for each individual flow. It would be more optimal if the ingress node could just send a single wildcard S-PMSI A-D route binding the set of flows to a particular tunnel, and have the egress nodes respond with Leaf A-D routes for each individual flow.

- [RFC6513] and [RFC6514] support the notion of "segmented P-tunnels", where "segmentation" occurs at ASBRs; [RFC7524] extends the notion segmented P-tunnels so that segmentation can occur at ABRs. One can think of a segmented P-tunnel as passing through a number of "segmentation domains". In each segmentation domain, a given P-tunnel has an ingress node and a set of egress nodes. The explicit tracking procedures allow an ingress node of a particular segmentation domain to determine, for a particular flow or set of flows, the egress nodes of that segmentation domain. This has given rise to two further problems:

  * The explicit tracking procedures do not allow an ingress node to "see" past the boundaries of the segmentation domain.

    This particular problem is not further addressed in this revision of this document.

  * The prior specifications do not make it very clear whether an egress node, upon receiving an S-PMSI A-D route whose PTA specifies "no tunnel information", is expected to forward the

S-PMSI A-D route, with the same PTA, to the next segmentation domain. This document provides the necessary clarifications.

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL", when and only when appearing in all capital letters, are to be interpreted as described in [RFC2119].

2. The Explicit Tracking Flags

Prior specifications define one flag in the PTA, the "Leaf Info Required" (LIR) flag, that is used for explicit tracking.

This document defines a new flag in the flags field of the PMSI Tunnel attribute. This new flag is known as the "Leaf Info Required per Flow" bit (LIR-pF). This flag MAY be set in the PTA of a (C-*,C-*), (C-*,C-G), or (C-S,C-*) S-PMSI A-D route. (Use of this flag in a PTA carried by other routes is outside the scope of this document.) Support for this flag is OPTIONAL.

The action taken by an egress node when the LIR-pF bit is set is detailed in Section 5.

If the LIR-pF flag is set in a given PTA, the LIR flag of that PTA SHOULD also be set. (By setting LIR as well as LIR-pF, one forces a a response to be sent an egress node that does not support LIR-pF, and it is possible to tell from that response that the egress node does not support LIR-pF.)

3. Match for Tracking vs. Match for Reception

RFC6625 (and other RFCs or RFCs-to-be that update RFC6625) specify a set of rules for finding the S-PMSI A-D route that is the "match for reception" for a given (C-S,C-G) or (C-*,C-G) state. These rules do not take into account the fact that some S-PMSI A-D routes may not be carrying PTAs at all, or may be carrying PTAs that do not identify any P-tunnel. (A PTA that does not identify any P-tunnel is one whose "tunnel type" field has been set to "no tunnel information", as specified in Section 5 of [RFC6514].)

The definition of "match for reception" in [RFC6625] is hereby modified as follows:

When finding the "match for reception" for a given (C-S,C-G) or (C-*,C-G), ignore any S-PMSI A-D route that has no PTA, or whose PTA specifying "no tunnel information".

We also introduce a new notion: the "match for tracking". This differs from the "match for reception" as follows:

> For a given C-flow ((C-S,C-G) or (C-*,C-G)) the "match for tracking" is chosen as follows. Ignore any S-PMSI A-D route that has no PTA. Also ignore any S-PMSI A-D route whose PTA specifies "no tunnel information", but does not have either LIR or LIR-pF set. (In particular, DO NOT ignore an S-PMSI A-D route that has a PTA specifying "no tunnel information", but whose LIR or LIR-pF bits are set). Then apply the rules (from [RFC6625] and other documents that that update it) for finding the "match for reception". The result (if any) is the match for tracking".

We will clarify this with a few examples. In these examples, we assume that there is only one segmentation domain. In this case, the ingress and egress nodes are Provider Edge (PE) routers.

Suppose a given PE router, PE1, has chosen PE2 as the "upstream PE" ([RFC6513]) for a given flow (C-S1,C-G1). And suppose PE1 has installed the following two routes that were originated by PE2:

- Route1: A (C-*,C-*) S-PMSI A-D route, whose PTA specifies a tunnel.

- Route2: A (C-S1,C-G1) S-PMSI A-D route, whose PTA specifies "no tunnel info" and has LIR set.

Route1 is (C-S1,C-G1)’s match for reception, and Route2 is (C-S1,C-G1)’s match for tracking.

Note that if there is no installed S-PMSI A-D route for (C-S2,C-G2), then Route1 would be (C-S2,C-G2)’s match for reception and also its match for tracking. Also note that if a match for tracking does not have the LIR flag or the LIR-pF flag set, no explicit tracking information will be generated. See Section 5.

As another example, suppose PE1 has installed the following two routes that were originated by PE2:

- Route1: A (C-*,C-*) S-PMSI A-D route (irrespective of whether the PTA specifies a tunnel)

- Route2: A (C-S1,C-G1) S-PMSI A-D route whose PTA specifies a tunnel.

Then Route2 is both the "match for reception" and the "match for tracking" for (C-S1,C-G1).

Note that for a particular C-flow, PE1's match for reception might be the same route as its match for tracking, or its match for reception might be a "less specific" route than its match for tracking. But its match for reception can never be a "more specific" route than its match for tracking.

4. Ingress Node Initiation of Tracking

An ingress node that needs to initiate explicit tracking for a particular flow or set of flows can do so by performing one of the following procedures:

1. An ingress node can initiate explicit tracking for (C-S1,C-G1) by originating an S-PMSI A-D route that identifies (C-S1,C-G1) in its NLRI, including a PTA in that route, and setting the LIR flag in that PTA. The PTA may specify a particular tunnel, or may specify "no tunnel info".

   However, the PTA of the (C-S1,C-G1) S-PMSI A-D route SHOULD NOT specify "no tunnel info" unless the ingress node also originates an A-D route carrying a PTA that specifies the tunnel to be used for carrying (C-S1,C-G1) traffic. Such a route could be an I-PMSI A-D route, a (C-*,C-G1) S-PMSI A-D route, a (C-S1,C-*) S-PMSI A-D route, or a (C-*,C-*) S-PMSI A-D route. (There is no point in requesting explicit tracking for a given flow if there is no tunnel on which the flow is being carried.)

   Further, if the ingress node originates a wildcard S-PMSI A-D route carrying a PTA specifying the tunnel to be used for carrying (C-S1,C-G1) traffic, and if that PTA has the LIR-pF bit set, then explicit tracking for (C-S1,C-G1) is requested by that S-PMSI A-D route. Thus the ingress node SHOULD NOT originate a (C-S1,C-G1) S-PMSI A-D route whose PTA specifies "no tunnel info"; such a route would not provide any additional functionality.

   To terminate explicit tracking that has been initiated by an S-PMSI A-D route whose PTA specifies "no tunnel info", the ingress node withdraws the route.

   To terminate explicit tracking that has been initiated by an S-PMSI A-D route whose PTA specifies a tunnel, the ingress node re-originates the route without the LIR flag set.

2. The following procedure can be used if (and only if) it is known that the egress nodes support the optional LIR-pF flag. If the ingress node originates a wildcard S-PMSI A-D route, it can initiate explicit tracking for the individual flows that match

the wildcard route by setting the LIR-pF flag in the PTA of the wildcard route. If an egress node needs to receive one or more flows for which that wildcard route is a match for tracking, the egress node will originate a Leaf A-D route for each such flow, as specified in Section 5.2).

When following this procedure, the PTA of the S-PMSI A-D route may specify a tunnel, or may specify "no tunnel info". The choice between these two options is determined by considerations that are outside the scope of this document.

To terminate explicit tracking that has been initiated by an S-PMSI A-D route whose PTA specifies "no tunnel info", the ingress node withdraws the route.

To terminate explicit tracking that has been initiated by an S-PMSI A-D route whose PTA specifies a tunnel, the ingress node re-originates the route without the LIR flag set

5. Egress Node Response to the Match for Tracking

5.1. General Egress Node Procedures

There are four cases to consider:

1. With regard to a particular (C-S,C-G) or (C-*,C-G) multicast state, the egress node’s match for tracking is same as its match for reception, and neither LIR nor LIR-pF flags are on.

   In this case, the egress node does not originate a Leaf A-D route in response to the match for reception/tracking, and there is no explicit tracking of the flow. This document specifies no new procedures for this case.

2. With regard to a particular (C-S,C-G) or (C-*,C-G) multicast state, the egress node’s match for tracking is the same as its match for reception, LIR is set, but LIR-pF is not set.

   In this case, a Leaf A-D route is originated by the egress node, corresponding to the S-PMSI A-D route that is the match for reception/tracking. Construction of the Leaf A-D route is as specified in [RFC6514]; this document specifies no new procedures for this case.

3. With regard to a particular (C-S,C-G) or (C-*,C-G) multicast state, the egress node’s match for tracking is the same as its match for reception, and LIR-pF is set. The egress PE MUST follow whatever procedures are required by other specifications,

based on the match for reception. If the egress PE supports the LIR-pF flag, it MUST also follow the procedures of Section 5.2.

4. With regard to a particular (C-S,C-G) or (C-*,C-G) multicast state, the egress node's match for tracking is not the same as its match for reception. This can only happen if the match for tracking has a PTA specifying "no tunnel info", with either LIR or LIR-pF set. In this case, the egress node must respond, separately, BOTH to the match for tracking and to the match for reception.

   When responding to the match for reception, the egress node MUST ignore the LIR-pF flag. However, the LIR flag is processed normally per the procedures for the match for reception.

   If the match for tracking has LIR set and if either (a) the egress node does not support LIR-pF, or (b) LIR-pF is not set, then the egress node must respond to the match for tracking, following procedures specified in other documents for the case where LIR is set.

   If the match for tracking has LIR-pF set, and the egress node supports the LIR-pF flag, the egress node must originate one or more Leaf A-D routes, as specified in Section 5.2.

   Note that if LIR is set in the PTA of the match for reception, the egress node may need to originate one or more Leaf A-D routes corresponding to the match for tracking, as well as originating a Leaf A-D route corresponding to the match for reception.

5.2. Responding to the LIR-pF Flag

To respond to a match for tracking that has LIR-pF set, an egress node originates one or more Leaf A-D routes.

Suppose the egress node has multicast state for a (C-S,C-G) or a (C-*,C-G) flow, and has determined a particular S-PMSI A-D route, which has the LIR-pF flag set, to be the match for tracking for that flow. Then if the egress node supports the LIR-pF flag, it MUST originate a Leaf A-D route whose NLRI identifies that particular flow. Note that if a single S-PMSI A-D route (with wild cards) is the match for tracking for multiple flows, the egress PE may need to originate multiple Leaf A-D routes, one for each such flow. We say that, from the perspective of a given egress node, a given S-PMSI A-D route tracks the set of flows for which it is the match for tracking. Each of the Leaf A-D routes originated in response to that S-PMSI A-D route tracks a single such flow.

The NLRI of each the Leaf A-D route that tracks a particular flow is constructed as follows. The "route key" field of the NLRI will have the following format:

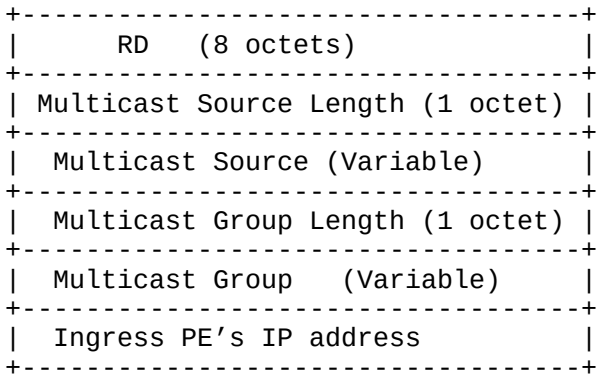


Figure 1: NLRI of S-PMSI A-D Route

- The "ingress PE" address is taken from the "originating router" field of the NLRI of the S-PMSI A-D route that is the match for tracking.

- The multicast source and group fields specify the S and G of one of the flow being tracked by this Leaf A-D route. If a (C-*,C-G) is being tracked by this Leaf A-D route, the source field is omitted, and its length is set to 0.

- The RD field is constructed as follows:

  * Take the RD value from the NLRI of the S-PMSI A-D route.

  * Add 16 to the second octet of the RD.

Note that, per RFC4364, every RD begins with a two-octet type field that is either 0, 1, or 2. By adding 16 to the second octet of the RD, we force the type field to be 16, 17, or 18. The presence of one of these values will indicate that the Leaf A-D route was constructed in response to a less specific S-PMSI A-D route that had the LIR-pF bit set. (That is, it distinguishes the routes from "ordinary" MVPN Leaf A-D routes.)

The encoding of these Leaf A-D routes is similar to the encoding of the Leaf A-D routes described in section 6.2.2 of [RFC7524], which were designed for the support of "global table multicast". However,

that document sets the RD to either 0 or -1; following the procedures of this document, the RD will never be 0 or -1. Therefore Leaf A-D routes constructed according to the procedures of this section can always be distinguished from the Leaf A-D routes constructed according to the procedures of section 6.2.2 of [RFC7524]. Also, Leaf A-D routes constructed according to the procedures of this section are VPN-specific routes, and will always carry an IP-address-specific Route Target, as specified in [RFC6514].

If a Leaf A-D route is originated as a response to a match for tracking whose PTA specifies "no tunnel info", a PTA SHOULD NOT be attached to the Leaf A-D route; if a PTA is attached, it MUST specify "no tunnel info".

In the case where the match for tracking and the match for reception are the same, the PTA of the match may have both the LIR and the LIR-pF flags set. This may cause the egress node to originate one Leaf A-D route in response to the LIR bit, and one or more Leaf A-D routes in response to the LIR-pF bit. A PTA SHOULD NOT be attached to the Leaf A-D routes that are originated in response to the LIR-pF bit.

When a Leaf A-D route constructed according to the procedures of this section is received, it MUST be processed by the node identified in its IP-address-specific Route Target, even though its "route key" field does not correspond to the NLRI of any S-PMSI A-D route.

Of course, an egress node that originates such Leaf A-D routes needs to remember which S-PMSI A-D route caused these Leaf A-D routes to be originated; if that S-PMSI A-D route is withdrawn, those Leaf A-D routes MUST be withdrawn.

Similarly, a Leaf A-D route needs to be withdrawn (either implicitly or explicitly) if the egress node changes its Upstream Multicast Hop (UMH) ([RFC6513]) for the flow that is identified in the Leaf A-D route’s NLRI, or if the egress node that originated the route no longer needs to receive the flow identified in the NLRI of the route.

Note that an egress node may acquire (C-S,C-G) state or (C-*,C-G) state after it has already received the S-PMSI A-D that is the match for tracking for that state. In this case, a Leaf A-D route needs to be originated at that time, and the egress node must remember that the new Leaf A-D route corresponds to that match for tracking.

Note that if a particular S-PMSI A-D route is a match for tracking but not a match for reception, the LIR bit in its PTA is ignored if the LIR-pF bit is set.

5.3. When the Egress Node is an ABR or ASBR

When segmented P-tunnels are used, the ingress and egress nodes may be ABRs or ASBRs. An egress ABR/ASBR that receives and installs an S-PMSI A-D route also forwards that route. If the PTA of an installed S-PMSI A-D route specifies a tunnel, the egress ABR/ASBR MAY change the PTA to specify a different tunnel type (as discussed in [RFC6514] and/or [RFC7524]).

However, if the PTA of the installed S-PMSI A-D route specifies "no tunnel info", the egress ABR/ASBR MUST pass the PTA along unchanged when it forwards the S-PMSI A-D route. (That is, a PTA specifying "no tunnel info" MUST NOT be changed into a PTA specifying a tunnel.) Furthermore, if the PTA specifies "no tunnel info", the LIR and LIR-pF flags in the PTA MUST be passed along unchanged.

In the case where the egress node is a PE, it will know whether it needs to receive a given flow by virtue of its having received a PIM or IGMP Join for that flow from a CE. In the case where the egress node is not a PE, but rather an ABR or ASBR, it will not know whether it needs to receive a given flow unless it receives a Leaf A-D route whose NLRI specifies that flow and whose IP-address-specific RT specifies an address of the egress node. Therefore an egress ABR/ASBR MUST NOT originate a Leaf A-D route for a given flow UNLESS it has an installed Leaf A-D route for that flow, received from further downstream.

This will ensure that an egress ABR/ASBR only sends a Leaf A-D route in response to a "match for tracking" if it is on the path to an egress PE for the flow(s) identified in the corresponding S-PMSI A-D route.

Then we can establish the following rule for egress ABRs/ASBRs. Suppose an egress ABR/ASBR receives an S-PMSI A-D route whose NLRI is X, and whose PTA (a) specifies "no tunnel info" and (b) has LIR set. The egress ABR/ASBR should not immediately originate a Leaf A-D route in response. Rather it should wait until it receives a Leaf A-D route whose NLRI contains X in the "route key" field. If it receives such a Leaf A-D route, it redistributes that route, but first it changes that route’s RT. The "global administrator" field of the modified RT will be set to the IP address taken either from the S-PMSI A-D route’s next hop field, or from its Segmented P2MP Next Hop Extended Community. (This is the same rule that is used for when the PTA does specify a tunnel type.)

6. Acknowledgments

The authors wish to thank Robert Kebler for his ideas and comments.

7. IANA Considerations

The LIR-pF flag needs to be added to the "P-Multicast Service Interface Tunnel (PMSI Tunnel) Attribute Flags" in the "Border Gateway Protocol (BGP) Parameters" registry. This registry is defined in [PTA_Flags]. The requested value is Bit Position 2. This document should be the reference.

IANA is requested to allocate three new types from the Route Distinguisher Type Field registry:

- Administrator field is two-byte Autonomous System Number. To be used only in certain MCAST-VPN Leaf A-D routes.
- Administrator field is four-byte IP Address. To be used only in certain MCAST-VPN Leaf A-D routes.
- Administrator field is four-byte Autonomous System Number. To be used only in certain MCAST-VPN Leaf A-D routes.

The requested values are 16, 17, and 18 respectively.

8. Security Considerations

The Security Considerations of [RFC6513] and [RFC6514] apply.

By setting the LIR-pF flag in a single wildcard S-PMSI A-D route, a large number of Leaf A-D routes can be elicited. If this flag is set when not desired (through either error or malfeasance), a significant increase in control plane overhead can result.

9. References

9.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <http://www.rfc-editor.org/info/rfc2119>.

[RFC6513] Rosen, E., Ed. and R. Aggarwal, Ed., "Multicast in MPLS/ BGP IP VPNs", RFC 6513, DOI 10.17487/RFC6513, February 2012, <http://www.rfc-editor.org/info/rfc6513>.

[RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, DOI 10.17487/RFC6514, February 2012, <http://www.rfc-editor.org/info/rfc6514>.

[RFC6625] Rosen, E., Ed., Rekhter, Y., Ed., Hendrickx, W., and R. Qiu, "Wildcards in Multicast VPN Auto-Discovery Routes", RFC 6625, DOI 10.17487/RFC6625, May 2012, <http://www.rfc-editor.org/info/rfc6625>.

9.2. Informative References

[PTA_Flags] Rosen, E. and T. Morin, "Registry and Extensions for P-Multicast Service Interface Tunnel Attribute Flags", internet-draft draft-ietf-bess-pta-flags-02, February 2016.

[RFC7524] Rekhter, Y., Rosen, E., Aggarwal, R., Morin, T., Grosclaude, I., Leymann, N., and S. Saad, "Inter-Area Point-to-Multipoint (P2MP) Segmented Label Switched Paths (LSPs)", RFC 7524, DOI 10.17487/RFC7524, May 2015, <http://www.rfc-editor.org/info/rfc7524>.

Authors' Addresses

Andrew Dolganow
Alcatel-Lucent
600 March Rd.
Ottawa, Ontario K2K 2E6
Canada

Email: andrew.dolganow@alcatel-lucent.com

Jayant Kotalwar
Alcatel-Lucent
701 East Middlefield Rd
Mountain View, California 94043
United States

Email: jayant.kotalwar@alcatel-lucent.com

Eric C. Rosen (editor)
Juniper Networks, Inc.
10 Technology Park Drive
Westford, Massachusetts 01886
United States

Email: erosen@juniper.net

Zhaohui Zhang
Juniper Networks, Inc.
10 Technology Park Drive
Westford, Massachusetts 01886
United States

Email: zzhang@juniper.net

BESS

Internet Draft
Intended status: Standard Track
Expires: November 2015

Weiguo Hao
Lucy Yong
Qiandeng Liang
Huawei

May 12, 2015

Handshaking mechanism for DF election
draft-hao-bess-evpn-df-handshaking-02.txt

Abstract

In [EVPN], in the DF re-election transient period, due to Ethernet Segment route transmission timer and timer clock discrepancy on each PEs, dual DF PEs will co-exist, traffic loop and disruption will be incurred. In MHN case, the consequences are particularly serious. Handshaking mechanism for DF election is introduced in this draft to resolve the problem.

Status of this Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/1id-abstracts.html

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

Table of Contents

1. Introduction

[EVPN] is a L2VPN solution using BGP for distributing customer/client MAC address reachability information over the core MPLS/IP network. EVPN provides flexible redundancy modes for both multi-homed device(MHD) and multi-homed network(MHN) scenarios. In MHD case, a CE node is normally accessed to a set of PE nodes leveraging multi-chassis Ethernet link aggregation groups(LAGs), both all-active and single-active redundancy mode can be achieved for the CE node. In MHN case, an Ethernet network, rather than a single device, is multi-homed to a group of PEs, only single-active redundancy mode can be achieved for the Ethernet network.

No matter it is all-active or single-active case, DF election mechanism is used to avoid packet duplication to local Ethernet Segment(ES) from a remote PE and loop among local PEs connecting to same ESI. The Designated Forwarder (DF) PE in (PBB-)EVPN networks is the PE that is responsible for sending multicast, broadcast and unknown unicast traffic to a multi-homed CE, on a given Ethernet Tag on a particular Ethernet Segment (ES). The DF PE is selected based on the list of PEs that advertise the Ethernet Segment Identifier (ESI) to the EVPN network.

The DF election procedure defined in [EVPN] is as follows:

1. When a PE discovers the ESI of the attached Ethernet Segment, it advertises an Ethernet Segment route with the associated ES-Import extended community attribute.

2. The PE then starts a timer (default value = 3 seconds) to allow the reception of Ethernet Segment routes from other PE nodes connected to the same Ethernet Segment.

3. The receiver PE(s) also starts a timer when the Ethernet Segment route is received. This timer value should be same across all PEs connected to the same Ethernet Segment.

4. When the timer expires, each PE starts DF election process independently using same algorithm. The PE elected as DF for a given EVPN instance will unblock the multi-destination traffic in the egress direction towards the Segment immediately, while the non-DF PEs will block the traffic immediately.

If one CE device or network is accessed to multiple PEs, one PE failure and recovery will trigger EVPN DF re-election. Because each PE relies on independent timer expiring to trigger local DF election process, due to Ethernet Segment route transmission timer and timer clock discrepancy on each PEs, in the DF re-election transient period, dual DF PEs will co-exist, traffic loop and disruption will be incurred. In MHN case, the consequences are particularly serious and will be described in detail in section 2.

2. Problems in MHN scenario

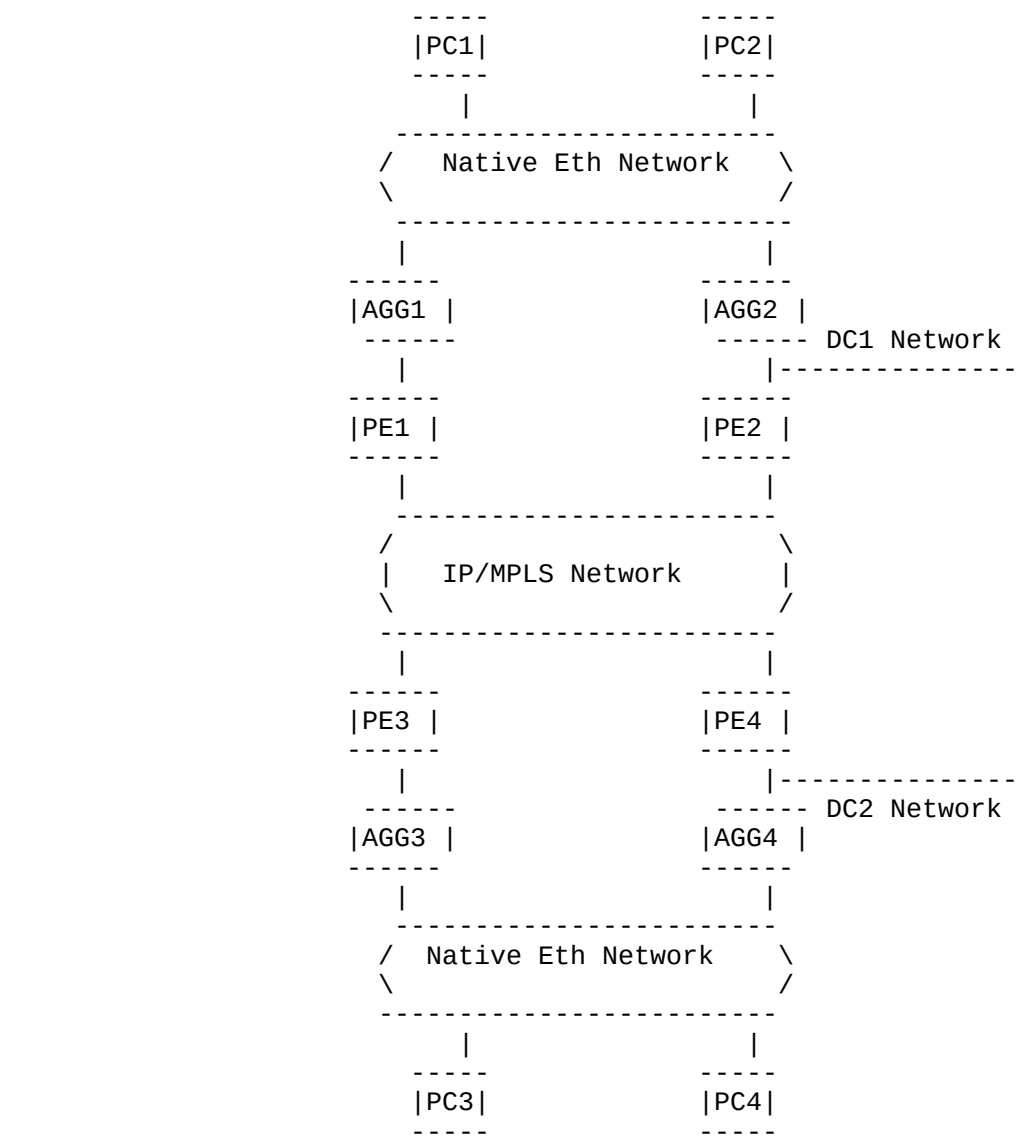

PC1
PC2
Native Eth Network
AGG1
AGG2
DC1 Network
PE1
PE2
IP/MPLS Network
PE3
PE4
DC2 Network
AGG3
AGG4
Native Eth Network
PC3
PC4

Figure 1 DC Network interconnecting using EVPN

In figure 1, both DC1 and DC2 networks are native Ethernet network and are multi-homed to two EVPN PEs in single-active mode respectively, i.e., DC 1 is connected to PE1 and PE2 while DC 2 is connected to PE3 and PE4. In regular time, PE1 and PE3 are DF PE, PE2 and PE4 are non-DF PE. PC1 to PC4 belong to same VLAN.

When PE3 fails, PE4 will take over the duties of DF PE. After some time, PE3 recovers and starts DF PE re-election process. In the re-election process, due to Ethernet Segment route transmission timer and timer clock discrepancy on PE3 and PE4, both PE3 and PE4 will act as DF PE in short transient time. During the transient time, the following data plane and control plane process will be performed.

Data plane:

1. PC1 in DC1 sends BUM traffic to PE1.

2. PE1 sends the traffic to PE3 (and PE4). Because both PE3 and PE4 are DF PE at this time, PE3 (and PE4) will forward the BUM traffic to DC2 local network.

3. PE4 (and PE3) will receive the BUM traffic from DC2 local access network and learn the PC1’s MAC through data plane.

4. PE4 (and PE3) will forward the BUM traffic to PE1.

5. PE1 will forward the BUM traffic to DC1 network. MAC flip-flop of PC1 will occur on the switches in DC1.

Control plane:

1. When PE4 (and PE3) learns PC1’s MAC from DC2 local network, PE4 (and PE3) will announce the MAC of PC1 to PE1 using BGP control plane.

2. When PE1 receives PC1’s MAC from PE4 (and PE3) through BGP, it will populate the MAC route of PC1 into local MAC-VRF, MAC flip-flop of PC1 on PE1 will occur.

When PC1's MAC flip-flop occurs on DC1 local switches and PE1, the regular traffic to PC1 in DC1 local network will be disrupted.

If there are multiple PCs sending BUM traffic proactively at the transient time, multiple MAC addresses fluctuation will occur. This will impose extra control plane burden on all related PEs and induce regular traffic disruption to these PCs. So in the MHN scenario, dual DF PEs in transient period will have serious consequences, it should be resolved.

To resolve the problem of dual DF PEs in transient period, a new handshaking mechanism for DF election is introduced in this draft.

3. Conventions used in this document

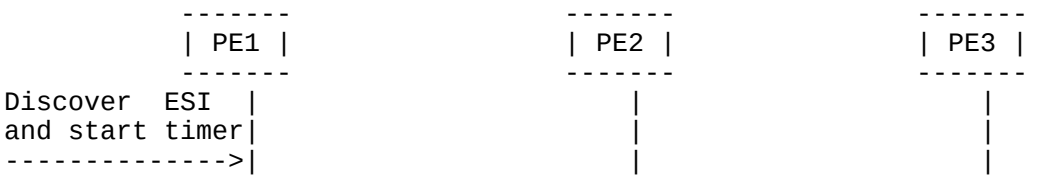
The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

Ethernet Segment (ES): If a multi-homed device or network is connected to two or more PEs via a set of Ethernet links, then that set of links is referred to as an 'Ethernet segment'.

Ethernet Segment Identifier (ESI): A unique non-zero identifier that identifies an Ethernet Segment is called an 'Ethernet Segment Identifier'.

4. Handshaking mechanism for DF election in EVPN network

In figure 1, initially PE2 and PE3 boot up and have already finished DF election process, later PE1 boots up and is elected as DF, the timing diagram from PE1 boots up is as follows:



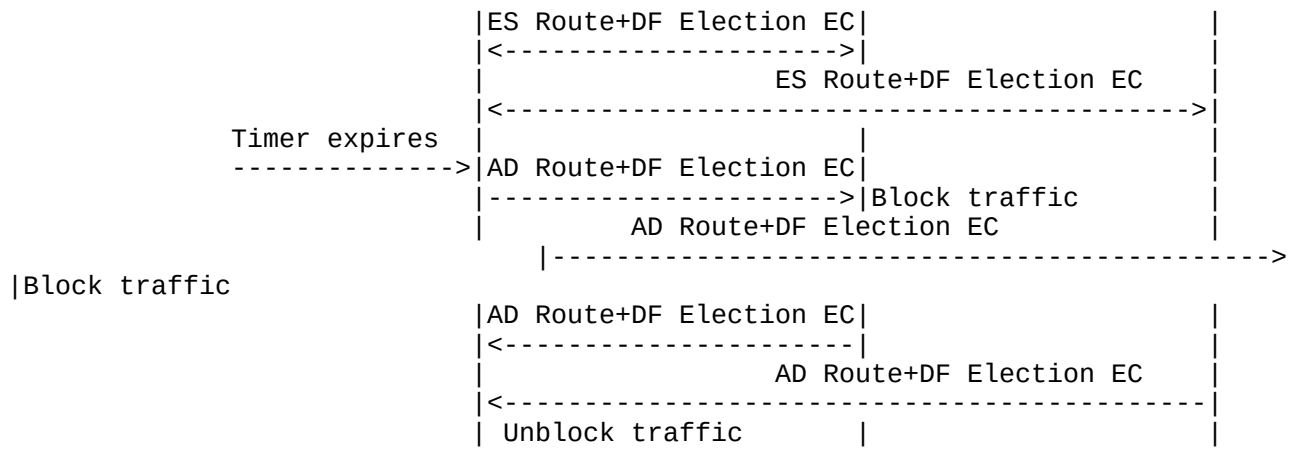


Figure 2 Handshaking DF Election Timing Diagram

1. When PE1 discovers the ESI of the attached Ethernet Segment, it advertises an Ethernet Segment route with the associated ES-Import extended community attribute. If the PE supports DF election handshaking mechanism, the ES route MUST associate with a new DF election BGP extended community with a DF handshaking capability Flag to show that the advertising PE itself supports DF election handshaking mechanism.

2. PE1 then starts a timer (default value = 3 seconds) to allow the reception of Ethernet Segment routes from other PE(PE2 and PE3) nodes connecting to the same Ethernet Segment. This timer value should be same across all PEs connected to the same Ethernet Segment.

3. When the timer expires, each PE knows all member PEs connecting to the same ESI and perform DF election algorithm independently as defined in [EVPN]. If all member PEs connecting to same ESI support DF handshaking mechanism, for the DF PE, at this time it doesn’t unblock multi-destination traffic in the egress direction towards the Segment directly. For the non-DF PEs, they also should keep forwarding state unchanged(Because PE2 is old DF PE, so it keep forwarding multi-destination traffic in the egress direction). The DF PE of PE1 should shake hands with other non-DF PEs to ensure unblocking action on DF PE and blocking action on non-DF PE enforced at the same time. The elected DF PE of PE1 notifies other non-DF PEs an Ethernet Auto-Discovery (A-D) route associated with a DF election

BGP extended community with DF Flag to show that the advertising PE1 itself is DF PE.

4. When each non-DF PE(PE2 and PE3) receives the A-D route from DF PE of PE1, the non-DF PE will block multi-destination traffic in the egress direction, then it advertises an Ethernet Auto-Discovery (A-D) route with the DF election extended community to DF PE. The DF election extended community carries a non-DF Flag to show that the advertising PE itself is non-DF PE and has blocked multi-destination traffic in the egress direction towards the Segment.

5. When the DF PE1 receives the A-D route with DF election extended community from all other non-DF PEs, the DF PE can start unblock multi-destination traffic in the egress direction towards the Segment. Because all non-DF PEs have blocked the traffic, so now no packet duplication and loop among local PEs will occur.

When a DF election happens, there may be other uncompleted DF election process existed among same PEs connecting to same Ethernet Segments at the same time. In order to ensure that all PEs negotiate for the newest DF election, it is necessary to introduce a sequence number into the DF election extended community attribute. The sequence number is generated on DF PE corresponding to each DF election process, non-DF PEs don’t generate the number and use same sequence number generated by DF PE. Each non-DF PE should stop negotiating old DF election when it receives a A-D route with larger sequence number.

Through the handshaking mechanism, when a DF PE is elected, it notifies all non-DF PEs block traffic immediately, only when all non-DF PEs block traffic successfully, DF PE can unblock traffic, so it can effectively reduce traffic disruption , avoid potential issues of packet duplication and loop incurred by all-active access.

5. Handshaking mechanism for DF election in PBB-EVPN network

In PBB-EVPN network, Ethernet Auto-Discovery Routes are not needed.

DF election extended community should be carried with ES routes. The ES routes are used for DF election handshaking among the PBB-EVPN PEs connecting to same ESI.

6. Network Migration Analysis

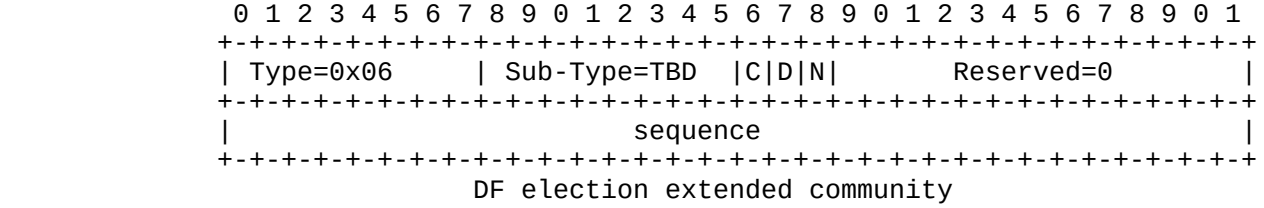
To support handshaking mechanism for DF election, software upgrade in all member PEs connecting to an ESI is needed.

In ESI member PE discovery phase, if a member PE doesn't support the new DF election handshaking mechanism, it advertises ES route without the new DF election extended capability, in this case the PEs connecting to the same ESI use old DF election method as defined in [EVPN]. Only if all member PEs connecting to same ESI support the new DF election handshaking capability, the new DF election method can be used for the ESI.

7. DF election extended community

This extended community is a new transitive extended community with the Type field of 0x06 and the Sub-Type of TBD. It may be advertised along with ES routes or A-D routes.

The DF election Extended Community is encoded as an 8-octet value as follows:



DF election extended community

o C. If C flag is one, it indicates that the advertising PE supports DF election handshaking mechanism. It's used in ESI member PE discovery phase for capability announcement.

o D. If D flag is one, it indicates that the advertising PE is the DF, the non-DF PE connecting to same ESI can block egress multi-destination traffic immediately when it receives ES or AD route with the DF election extended community from DF PE.

o N. If N flag is one, it indicates that the advertising PE is non-DF and has blocked multi-destination traffic in the egress direction towards the Segment.

o Sequence. The sequence number is used to ensure that PEs connecting to same ESI are negotiating for same DF election. In ES member PEs discovery phase, the sequence number is 0. In DF election handshaking phase, the sequence number should be non-zero and is generated by DF PE.

8. Security Considerations

NA.

9. IANA Considerations

IANA has allocated the following EVPN Extended Community sub-types in [RFC7153].

SUB-TYPE VALUE	NAME
0x00	MAC Mobility
0x01	ESI Label
0x02	ES-Import Route Target

IANA is requested to create and maintain a new registry for "EVPN DF Election Extended Community Sub-Types".

SUB-TYPE VALUE	NAME
0x03	DF Election

10. References

10.1. Normative References

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

10.2. Informative References

[2] [EVPN] Sajassi et al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-11.txt, work in progress, October, 2014

11. Acknowledgments

Authors like to thank Lili Wang, Ziqing Cao, Feng Qian for his valuable inputs.

Authors' Addresses

Weiguo Hao
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China
Phone: +86-25-56623144
Email: haoweiguo@huawei.com

Lucy Yong
Huawei Technologies
Phone: +1-918-808-1918
Email: lucy.yong@huawei.com

Qiandeng Liang
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China
Email: liangqiandeng@huawei.com

BESS

Internet Draft

Weiguo Hao
Lucy Yong
S. Hares
Huawei
R. Raszuk
Mirantis Inc.
L. Fang
Osama Zia
Microsoft
Shahram Davari
Broadcom
Andrew Qu
MediaTec

Intended status: Standard Track
Expires: November 2015

May 19, 2015

Inter-AS Option B between NVO3 and BGP/MPLS IP VPN network
draft-hao-bess-inter-nvo3-vpn-02.txt

Abstract

This draft describes the solution of inter-as option-B connection between NVO3 network and MPLS/IP VPN network. The ASBR located in NVO3 network is called ASBR-d, the control plane and data plane procedures at ASBR-d are specified in this document, there are some differences from traditional option-B ASBR defined in [RFC 4364].

Status of this Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

Copyright Notice

Table of Contents

1. Introduction

In cloud computing era, multi-tenancy has become a core requirement for data centers. Since NVO3 can satisfy multi-tenancy key requirements, this technology is being deployed in an increasing number of cloud data center network. NVO3 focuses on the construction of overlay networks that operate over an IP (L3) underlay transport network. It can provide layer 2 bridging and layer 3 IP service for each tenant. VXLAN and NVGRE are two typical NVO3 technologies. NVO3 overlay network can be controlled through centralized NVE-NVA architecture or through distributed BGP VPN protocol.

NVO3 has good scaling properties from relatively small networks to networks with several million tenant systems (TSs) and hundreds of thousands of virtual networks within a single administrative domain. In NVO3 network, 24-bit VNID is used to identify different virtual networks, theoretically 16M virtual networks can be supported in a data center. In a data center network, each tenant may include one or more layer 2 virtual network and in normal cases each tenant corresponds to one routing domain (RD). Normally each layer 2 virtual network corresponds to one or more subnets.

To provide cloud service to external data center client, data center networks should be connected with WAN networks. BGP MPLS/IP VPN has already been widely deployed at WAN networks. Normally internal data center and external MPLS/IP VPN network belongs to different autonomous system(AS). This requires the setting up of inter-as connections at Autonomous System Border Routers(ASBRs) between NVO3 network and external MPLS/IP network.

Currently, a typical connection mechanism between a data center network and an MPLS/IP VPN network is similar to Inter-AS Option-A of RFC4364, but it has scalability issue if there is huge number of tenants in data center networks. To overcome the issue, inter-as Option-B between NVO3 network and BGP MPLS/IP VPN network is proposed in this draft.

2. Conventions used in this document

Network Virtualization Edge (NVE) - An NVE is the network entity that sits at the edge of an underlay network and implements network virtualization functions.

Tenant System - A physical or virtual system that can play the role of a host, or a forwarding element such as a router, switch,

firewall, etc. It belongs to a single tenant and connects to one or more VNs of that tenant.

VN - A VN is a logical abstraction of a physical network that provides L2 network services to a set of Tenant Systems.

RD - Route Distinguisher. RDs are used to maintain uniqueness among identical routes in different VRFs, The route distinguisher is an 8-octet field prefixed to the customer’s IP address. The resulting 12-octet field is a unique "VPN-IPv4" address.

RT - Route targets. It is used to control the import and export of routes between different VRFs.

3. Reference model

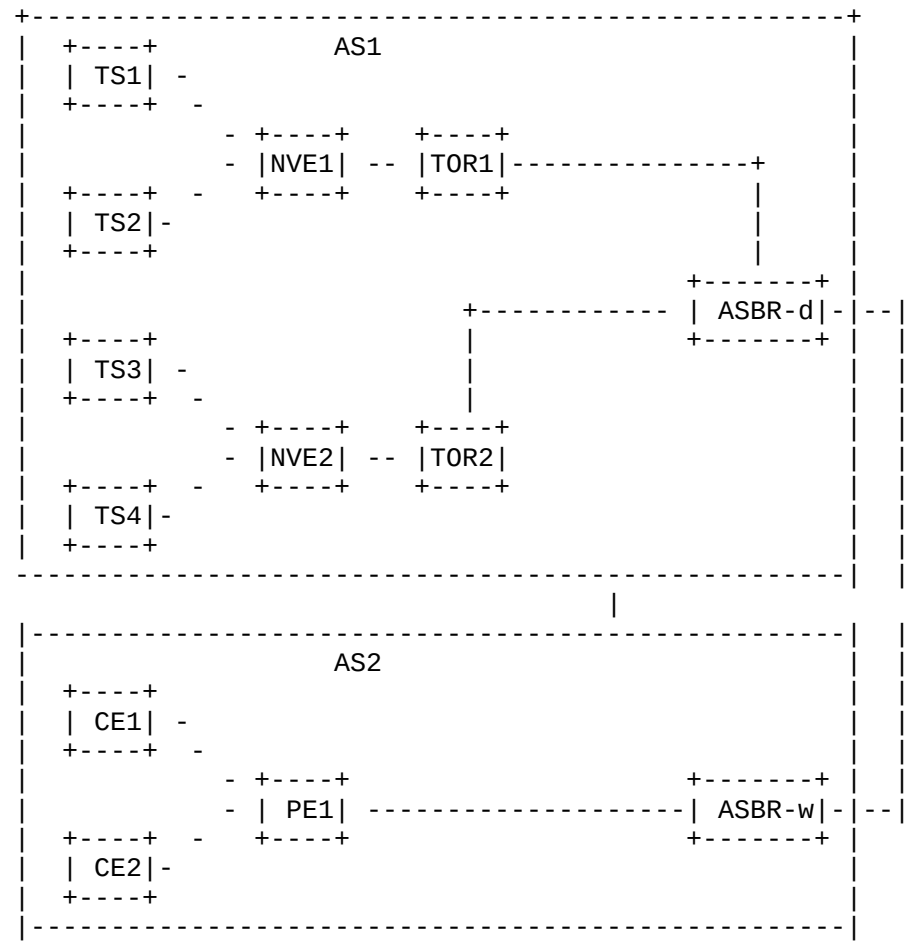


Figure 1 Reference model

Figure 1 shows an arbitrary Multi-AS VPN interconnectivity scenario between NVO3 network and BGP MPLS/IP VPN network. NVE1, NVE2, and ASBR-d forms NVO3 overlay network in internal DC. TS1 and TS2 connect to NVE1, TS3 and TS4 connect to NVE2. PE1 and ASBR-w forms MPLS IP/VPN network in external DC. CE1 and CE2 connect to PE1. The NVO3 network belongs to AS 1, the MPLS/IP VPN network belongs to AS 2.

There are two tenants in NVO3 network, TSs in tenant 1 can freely communicate with CEs in VPN-Red, TSs in tenant 2 can freely communicate with CEs in VPN-Green. TS1 and TS3 belong to tenant 1, TS2 and TS4 belong to tenant 2. CE1 belongs to VPN-Red, CE2 belongs to VPN-Green. VNID 10 and VNID 20 are used to identify tenant 1 and tenant 2 respectively. CE1 and CE2 have local IP prefix of 10.1.1.1/24 and 20.1.1.1/24 respectively.

4. Option-A inter-as solution overview

In Option-A inter-as solution, peering ASBRs are connected by multiple sub-interfaces, each ASBR acts as a PE, and thinks that the other ASBR is a CE. Virtual routing and forwarding (VRF)data bases (RIB/FIB) are configured at AS border routers (ASBR-d and ASBR-w) so that each ASBRs associate each such sub-interface with a VRF and use EBGP to distribute unlabeled IPv4 addresses to each other. In the data-plane, VLANs are used for tenant traffic separation. ASBR-d terminates NVO3 encapsulation for inter-subnet traffic from TS in internal DC to CE in external DC.

Option-A inter-as solution has following issues:

1. Up to 16 million (16M) gateway interfaces (virtual/physical) and 16M EBGP session need to exist between the ASBRs.

2. UP to 16M VRFs need to be supported on border routers.

3. Several million routing entries need to be supported on border routers.

Inter-as option-B between NVO3 network and MPLS IP/VPN network can be used to address these issues. As option-B proposed in this draft is for multi-as interconnection between heterogeneous networks, so there are some differences from traditional Inter-AS Option-B of RFC4364.

5. Vanilla Option-B inter-as solution overview

Similar to the solution described in section 10, part (b) of [RFC4364] (commonly referred to as Option-B) peering ASBRs are connected as private peers that are enabled to receive Labeled packets from trusted peers. An MP-BGP session is used to distribute the labeled VPN prefixes between the ASBRs. In data plane, the traffic that flows between the ASBRs is placed in MPLS tunnels. Traffic separation among different VPNs between the ASBRs relies on MPLS VPN Label. The advantage of this option is that it’s more

scalable, as there is no need to have separate interface and BGP session per VPN/Tenant.

As for the routing distribution process from DC to WAN side, MPLS VPN Label is allocated on ASBR-d per VN per NVE. As for the routing distribution process from WAN to DC side, VNID is allocated on ASBR-d per MPLS VPN Label receiving from per ASBR-w. As for the data plane process, NVO3 tunnel and MPLS VPN tunnel are stitched at ASBR-d. From DC to WAN side, NVO3 tunnel is terminated, VNID and MPLS VPN Label switching is performed by looking up outgoing forwarding table in section 6.1.2. From WAN to DC side, MPLS VPN tunnel is terminated, MPLS VPN Label and NVO3 tunnel switching is performed by looking up incoming forwarding table in section 6.1.1. ASBR-w has no difference with traditional RFC4364 based Option-B behavior, no VRF is created on the ASBR-d.

6. Vanilla Inter-As Option-B procedures

Each NVE operates as a layer 3 gateway for local connecting TS(s). Operators may configure single and unique VNID for each tenant network on all NVEs or configure NVEs to locally allocate VNID for each tenant on the NVEs, the VNID is called VNID-t.

Routing information for each tenant should be synchronized between NVO3 and MPLS VPN network. In internal DC NVO3 network, routing information synchronization between NVE and ASBR-d can be through either: a) BGP MPLS/IP VPN protocol running between the NVEs and the ASBR-d or b) NVE-NVA architecture.

In case a), it is a coupled solution, the NVE entity normally resides on hardware network device like TOR switch. VRFs can be created on each NVE to isolate IP routing information in control plane and IP forwarding process in data plane between different tenants, each VRF has its own IP routing table. The BGP routes are originated on NVE with either implied nexthop address of the BGP router or self-nexthop set.

In case b), it is a decoupled solution, the NVE entity normally resides on vSwitch. VRFs are created on NVA only for control plane information isolation between different tenants, while in data plane, unified flow tables are used for all tenants on each NVE.

6.1. Using BGP MPLS/IP VPN protocol

Each NVE is a BGP speaker. Operators configure VRF and RD/RT for each tenant network on each NVE. BGP MPLS/IP VPN protocol extension is running between NVEs and ASBR-d utilizing the [BGP Remote-Next-

Hop] attribute which specifies a set of remote tunnels (1 to N) that occur between two BGP speakers.

When an NVE advertises a prefix with RD/RT, tunnel encapsulation and VNID-t are carried in BGP update message [BGP Remote-Next-Hop]. The NVE BGP receiver imports the prefix according RD/RT and maintains the mapping of prefix and VNID plus tunnel encapsulation(For VXLAN and NVGRE, they are outer destination IP address and inner destination MAC) in VRF.

[Note: the [BGP Remote-Next-Hop] is a work-in-progress that is an individual draft. The IDR WG may modify this draft or adopt another that provides a similar mechanism to support remote next-hops. This draft will follow the IDR adoption of a remote next-hop solution.]

6.1.1. DC to WAN direction

1. NVE1 and NVE2 operate as a layer 3 gateway for local connecting TSs. They learn the local TS's IP Address via ARP or other mode and then advertise local TS's IP Address with local NVE's NVO3 tunnel end points information to ASBR-d using [BGP Remote-Next-Hop]. The routing information from NVE1 and NVE2 are as follows.

Node	IP Prefix	RD	RT	VNID-t
NVE1	TS1/32	RD-A	RT-A	10
NVE1	TS2/32	RD-B	RT-B	20
NVE2	TS3/32	RD-A	RT-A	10
NVE2	TS4/32	RD-B	RT-B	20

Table 1 Routing information from NVE

2. When ASBR-d receives routing information from each NVE, it allocates MPLS VPN Label per tenant (VNID-t) per NVE and the RD and RT remain the same (see table 2 below for examples). Then the ASBR-d advertises the VPN route with new allocated MPLS VPN Label to ASBR-w. The allocated MPLS VPN label and its corresponding <NVE, VNID-t> pair forms incoming forwarding table which is used to forward MPLS traffic from WAN to DC side. As an example the incoming forwarding table on ASBR-d could be as follows:

MPLS VPN Label	NVE + VNID
1000	NVE1 + 10
2000	NVE1 + 20
1001	NVE2 + 10
2001	NVE2 + 20

Table 2 Incoming forwarding table

6.1.2. WAN to DC direction

1. When ASBR-d receives routing information from ASBR-w, ASBR-d allocates VNID-d for each VPN Label, and then ASBR-w advertises the VPN route with new allocated VNID-d to each NVE (NVE1 and NVE2). The role of the VNID-d is similar to the role of Incoming VPN Label in traditional MPLS VPN Option-B based ASBR defined in [RFC 4364], it has local significance on ASBR-d, each VNID corresponds to a MPLS VPN Label received from peer ASBR-w. The allocated VNID-d and its corresponding out VPN Label forms an outgoing forwarding table which is used to forward NVO3 traffic from DC to WAN side. Assuming ASBR-d receives VPN Label 3000 and 4000 from ASBR-w allocated for VPN-Red and VPN-Green at PE1 respectively, the outgoing forwarding table on ASBR-d is as follows:

Node	IP Address	RD	RT	MPLS VPN Label
PE1	10.1.1.1/24	RD-A	RT-A	3000
PE1	20.1.1.1/24	RD-B	RT-B	4000

Table 3 Routing information from PE1

VNID	Out VPN Label
10000	3000
10001	4000

Table 4 Outgoing forwarding table

2. When each local NVE receives routing information from ASBR-d, it matches the Route Target Attribute in BGP MPLS/IP VPN protocol with local VRF's import RT configuration and populates local VRF with these matched VPN routes (see table 3 above).

6.2. Data plane procedures

This section describes the step by step procedures of data forward between TS1 and CE1 for either: a) DC to WAN direction IP data flows, or b) WAN to DC direction IP data flows.

6.2.1. DC to WAN direction

1. TS1 sends traffic to NVE1, the destination IP is CE1's IP address.

2. NVE1 looks up tenant 1's IP forwarding table, then it gets NVO3 tunnel encapsulation information. The destination outer address is ASBR-d's IP address, VNID is 10000 allocated by ASBR-d for VPN route of CE1 received from ASBR-w. NVE1 performs NVO3 encapsulation and sends the traffic to ASBR-d.

3. ASBR-d decapsulates NVO3 encapsulation and gets VNID 10000. Then it looks up outgoing forwarding table based on the VNID and gets MPLS VPN label 3000. Finally it pushes MPLS VPN label for the IP traffic and sends it to ASBR-w.

4. Then the traffic is forwarded to CE1 through regular MPLS VPN forwarding process.

6.2.2. WAN to DC direction

1. CE1 sends traffic to PE1, destination IP is TS1's IP address. The traffic is forwarded to ASBR-d through regular MPLS VPN forwarding process. The incoming MPLS VPN label at ASBR-d is 1000 allocated by ASBR-d for tenant 1 at NVE1.

2. ASBR-d looks up incoming forwarding table and gets NVO3 encapsulation, then performs NVO3 encapsulation and sends the traffic to NVE1. The destination outer IP is NVE1's IP, VNID is 10 corresponding to tenant 1.

3. NVE1 decapsulates NVO3 encapsulation, gets local IP forwarding table relying on VNID 10, and then sends the traffic to TS1.

6.2.3. Data plane NVE Operations summary

Each NVE maintains a lookup table per tenant, i.e. VNID-t and the received mappings from ASBR-d for each tenant. For the prefix that is from inside DC, the inner/outer mapping entry is the prefix <-> remote NVE IP address. For the prefix that is from outside DC, the inner/outer mapping entry is the prefix <-> VNID-d + ASBR1-d IP address.

When receiving a packet from a tenant system locally, NVE performs a lookup in the corresponding tenant table for the destination address on the packet. If the prefix results to single IP address, NVE will encapsulate the packet with VNID-t and IP address as outer IP address. If the prefix results to a VNID and IP address, NVE will encapsulate the packet with the VNID and IP address as outer IP address.

When receiving a packet from NVO3, NVE decapsulates the packet and find the attached tenant system based on the VNID and destination address on the packet, forward the decapsulated packet to the tenant system.

6.3. NVE-NVA architecture

In this architecture, the NVE control plane and forwarding functionality are decoupled. All NVEs in NVO3 network don't need support distributed BGP VPN protocol [BGP Remote-Next-Hop], these NVEs have only data plane functionality and are controlled by

centralized NVA using openflow, ovsdb, i2rs, etc. The NVA runs IBGP VPN protocol for all the NVEs with ASBR-d utilizing the [BGP Remote-Next-Hop] attribute to pass along the tunnel endpoints and encapsulations associated with each NVE. The ASBR-d runs EBGP VPN protocol with peer ASBR-w. ASBR-d allocates MPLS VPN Label per tenant per NVE.

NVA maintains all tenant information, and originates BGP routes with the appropriate RD and AD. The NVA tenant information includes VNID-t to identify each tenant and the corresponding RD and RT. This information can be statically configured by operators or dynamically notified by cloud management systems. This information also includes all TS's MAC/IP address and its attached NVE information.

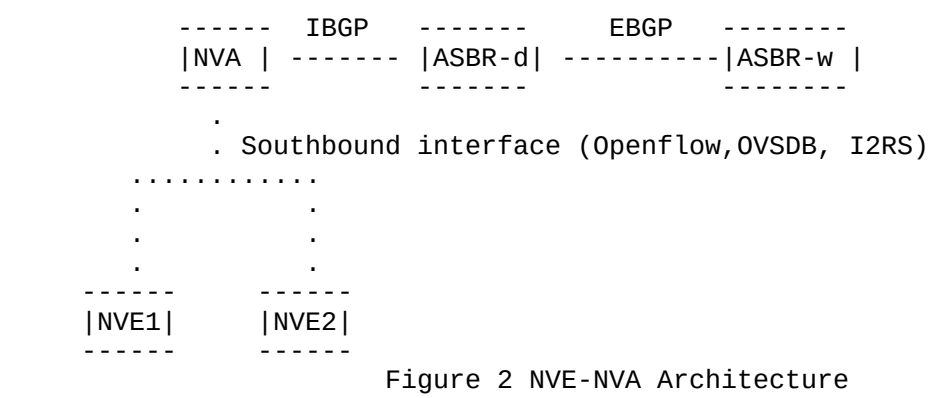


Figure 2 NVE-NVA Architecture

6.3.1. DC to WAN direction

1. NVA advertises all internal data center VPN routing information to ASBR-d, which includes RD, RT, VNID-t, IP prefix and the attached NVE IP address. The VNID-t and NVE IP address are used for traffic NVO3 encapsulation from ASBR-d to NVE.

2. ASBR-d allocates MPLS VPN Label per VNID per NVE and generates incoming forwarding table same as Table 2.

6.3.2. WAN to DC direction

1. ASBR-d receives VPN routing information from peer ASBR-w. ASBR-d allocates VNID-d, for each MPLS VPN Label receiving from ASBR-w and generates outgoing forwarding table same as Table 4. Then it advertises the VPN route to NVA, which includes RD, RT, VNID-l, IP prefix, and set itself as next hop. The VNID and ASBR-d IP address are used for traffic NVO3 encapsulation from NVE to ASBR-d.

2. NVA matches local Route Target configuration, imports VPN route to each tenant, and downloads flow table to corresponding NVE.

7. Partial Option-B solution

In vanilla option-B solution, each NVE need to maintain routing items corresponding to IP prefix located outside data center for north-south bound traffic forwarding. If there are some VPNs which have large number of IP prefix, it will cause much pressure on local NVEs. In this case, partial Option-B solution can be used.

In partial Option-B solution, default route is used for north-south bound traffic on each NVE. The traffic from each NVE will be forwarded to ASBR-d using NVO3 encapsulation, VNID is used to identify tenant VRF at ASBR-d. ASBR-d terminates the NVO3 encapsulation, looks up local VRF’s IP routing table, then performs MPLS encapsulation and sends to peer ASBR-w.

For the traffic from WAN to DC, ASBR-d needs to maintain all TS’s IP addresses and their attached NVE device in corresponding VRF. When the ASBR-d receives MPLS traffic from peer ASBR-w, MPLS encapsulation is terminated, looks up local VRF’s IP routing table, then performs NVO3 encapsulation and sends to local destination NVE.

From control plane perspective, EBGP VPN connection is terminated at ASBR-d, which means the ASBR doesn’t allocate new VNID-d for each MPLS VPN Label and advertise it to peer NVE in local AS, VRF is created on the ASBR-d, the VPN route from WAN side populates to local VRF.

8. Inter-as option comparisons

The document describes several inter-as implementation options between ASBR-d and ASBR-w. The following table illustrates the comparison among the implementation options.

	Option-A	Partial Option-B	Vanilla Option-B
Sub-interface	Yes	No	No
VRF	Yes	Yes	No
Scalability	Worst	Middle	Best
Hardware Implementation at ASBR-d	No Upgrade	No Upgrade	Need Upgrade

Table 5 Inter-as option comparisons

Option-A design uses a regular VPN handoff between ASBR-d and ASBR-w. A sub-interface is required per a NVO instance in between. Both border routers perform the VRF lookup. Thus, the solution has a scalability concern. Existing hardware supports this solution.

Partial Option-B does not require sub-interfaces between ASBR-d and ASBR-w, only ASBR-d performs the VRF lookup, so it has better scalability than option A. Existing hardware can support this solution.

In the vanilla Option-B solution, there is no sub-interface between border routers and no VRF table on ASBR-d and ASBR-w. Tunnel stitching is performed on the ASBR-d. Thus this solution has the best scalability. From hardware perspective, the vanilla option-B needs ASBR-d hardware upgrade to support the tunnel stitching.

9. Security Considerations

Similar to the security considerations for inter-as Option-B in [RFC4364] the appropriate trust relationship must exist between NVO3 network and MPLS/IP VPN network. VPN-IPv4 routes in NVO3 network should neither be distributed to nor accepted from the public Internet, or from any BGP peers that are not trusted. For other general VPN Security Considerations, see [RFC4364].

10. IANA Considerations

This document requires no IANA actions. RFC Editor: Please remove this section before publication.

11. References

11.1. Normative References

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

[2] [RFC4364] E. Rosen, Y. Rekhter, " BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

[3] [RFC5512] P. Mohapatra, E. Rosen, " The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", RFC5512, April 2009

11.2. Informative References

[4] [NVA] D.Black, etc, "An Architecture for Overlay Networks (NVO3)", draft-ietf-nvo3-arch-01, February 14, 2014

[5] [BGP Remote-Next-Hop] G. Van de Velde,etc, ’’BGP Remote-Next-Hop’’, draft-vandevelde-idr-remote-next-hop-05, January, 2014

[6] [RFC7047] B. Pfaff, B. Davie,’’The Open vSwitch Database Management Protocol’’, RFC 7047, December 2013

[7] [OpenFlow1.3]OpenFlow Switch Specification Version 1.3.0 (Wire Protocol 0x04). June 25, 2012. (https://www.opennetworking.org/images/stories/downloads/sdn-resources/onf-specifications/openflow/openflow-spec-v1.3.0.pdf)

12. Acknowledgments

Authors like to thank Xiaohu Xu, Liang Xia, Shunwan Zhang, Yizhou Li, Lili Wang for his valuable inputs.

Authors’ Addresses

Weiguo Hao
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China
Email: haoweiguo@huawei.com

Lucy Yong
Huawei Technologies
Phone: +1-918-808-1918
Email: lucy.yong@huawei.com

Susan Hares
Huawei Technologies
Phone: +1-734-604-0323
Email: shares@ndzh.com.

Robert Raszuk
Mirantis Inc.
615 National Ave. #100
Mt View, CA 94043
USA
Email: robert@raszuk.net

Luyuan Fang
Microsoft
Email: lufang@microsoft.com

Osama Zia
Microsoft
Email: osamaz@microsoft.com

Shahram Davari
Broadcom
Email: Davari@Broadcom.com

Andrew Qu
MediaTec
Email: andrew.qu@mediatek.com

BESS Working Group
Internet-Draft
Updates: 6514 (if approved)
Intended status: Standards Track
Expires: November 4, 2016

E. Rosen
Juniper Networks, Inc.
T. Morin
Orange

May 3, 2016

Registry and Extensions for P-Multicast Service Interface Tunnel Attribute Flags
draft-ietf-bess-pta-flags-03.txt

Abstract

The BGP-based control procedures for Multicast Virtual Private Networks make use of a BGP attribute known as the "P-Multicast Service Interface (PMSI) Tunnel" attribute. The attribute contains a one-octet "Flags" field. The purpose of this document is to establish an IANA registry for the assignment of the bits in this field. Since the Flags field contains only eight bits, this document also defines a new BGP Extended Community, "Additional PMSI Tunnel Attribute Flags", that can be used to carry additional flags for the PMSI Tunnel attribute. This document updates RFC 6514.

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 November 4, 2016.

Copyright Notice

Table of Contents

1. Introduction

A BGP attribute known as the "P-Multicast Service Interface (PMSI) Tunnel" attribute is defined in [RFC6514]. This attribute contains a one-octet "Flags" field. Only one flag is defined in that RFC, but there is now a need to define additional flags. However, that RFC did not create an IANA registry for the assignment of bits in the "Flags" field. This document creates a registry for that purpose. In addition, there may be a need to define more than eight flags. Therefore this document defines a new BGP Extended Community, "Additional PMSI Tunnel Attribute Flags", that can be used to carry additional flags for the PMSI Tunnel attribute. A registry is also created for this Extended Community, allowing IANA to assign flag bits from the Extended Community's six-octet value field.

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. Extending the PMSI Tunnel Attribute Flags Field

In [RFC6514], only a single octet in the "PMSI Tunnel" attribute is defined to carry bit flags. This allows eight flags, which is unlikely to be sufficient for all future applications.

This document defines a new Transitive Opaque Extended Community ([RFC4360], [RFC7153]), "Additional PMSI Tunnel Attribute Flags". It also defines a new bit flag in the "PMSI Tunnel" attribute Flags field, called the "Extension" flag.

The "Additional PMSI Tunnel Attribute Flags" Extended Community MUST NOT be carried by a given BGP UPDATE message unless the following conditions both hold:

- o the given BGP UPDATE message is also carrying a "PMSI Tunnel" attribute, and
- o the "Extension" flag of that "PMSI Tunnel" attribute's "Flags" field is set.

The six-octet value field of the "Additional PMSI Tunnel Attribute Flags" Extended Community is considered to be a string of 48 bit flags. As shown in Figure 1, the leftmost bit (the most significant bit of the most significant octet) is bit 0, and the rightmost bit (the least significant bit of the least significant octet) is bit 47.

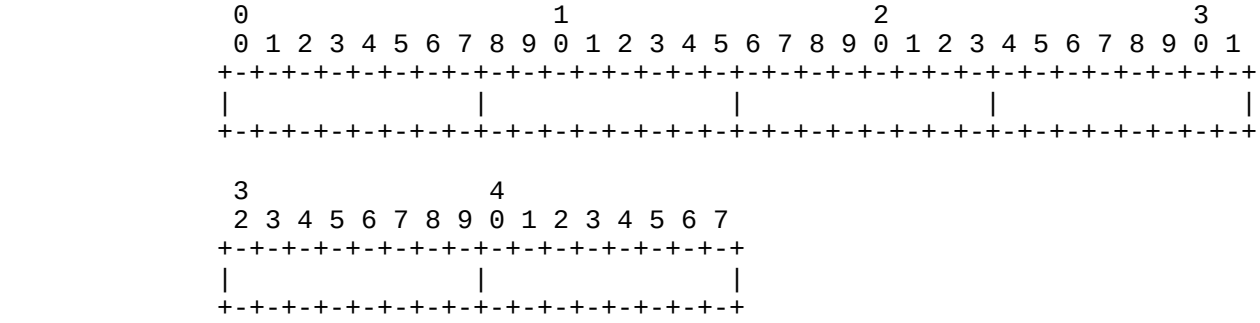


Figure 1: Value Field of the Additional PMSI Tunnel Attribute Flags Extended Community

A BGP speaker MUST NOT attach more than one "Additional PMSI Tunnel Attribute Flags" Extended Community to a given BGP UPDATE. If a given BGP UPDATE already contains an "Additional PMSI Tunnel Attribute Flags" Extended Community, a BGP speaker MUST NOT attach any additional such Extended Communities.

If a BGP speaker receives a BGP UPDATE with more than one "Additional PMSI Tunnel Attribute Flags" Extended Communities attached, only the flag settings in first occurrence of the Extended Community are significant. Flag settings in subsequent occurrences of the Extended Community MUST be ignored. When propagating the UPDATE, all instances of the Extended Community other than the first SHOULD be removed.

Suppose a BGP speaker receives an UPDATE message that contains a "PMSI Tunnel" attribute, but does not contain an "Additional PMSI

Tunnel Attribute Flags" Extended Community. If the "Extension" flag of the "PMSI Tunnel" attribute is set, the UPDATE is considered to be malformed, and the "treat-as-withdraw" procedure of [RFC7606] MUST be applied.

If a BGP speaker receives an UPDATE message that contains one or more "Additional PMSI Tunnel Attribute Flags" Extended Communities, but either (a) that UPDATE message does not contain a PMSI Tunnel attribute, or (b) the Extension flag of the PMSI Tunnel attribute is not set, then the Extended Community(ies) SHOULD be removed and SHOULD NOT be redistributed. The BGP UPDATE message MUST be processed (and if necessary, redistributed) as if the Extended Community(ies) had not been present.

A BGP speaker that supports the current document, but does not recognize a particular flag (either in the" PMSI Tunnel" attribute "Flags" field or in the "Additional PMSI Tunnel Attribute Flags" Extended Community) MUST simply ignore that flag. If the BGP speaker propagates either the PMSI Tunnel attribute or the "Additional PMSI Tunnel Attribute Flags" Extended Community or both along with the UPDATE message, it SHOULD leave the setting of the flag unchanged.

It is possible that a particular application will require all members of a particular set of BGP speakers to support a particular flag. How it is determined whether all such BGP speakers support that flag is outside the scope of this document.

In some situations, a BGP speaker may need to modify or replace the "PMSI Tunnel" attribute before propagating an UPDATE. If the "Extension" flag of the "PMSI Tunnel" attribute was set before the attribute is modified or replaced, but that flag is no longer set after the attribute is modified or replaced, any "Additional PMSI Tunnel Attribute Flags" Extended Communities MUST be removed before the UPDATE is propagated. If the PMSI Tunnel attribute is removed entirely before an UPDATE is propagated, the "Additional PMSI Tunnel Attribute Flags" Extended Communities (if any) MUST also be removed.

3. IANA Considerations

IANA is requested to create a new registry called "P-Multicast Service Interface (PMSI) Tunnel Attribute Flags" in the "Border Gateway Protocol (BGP) Parameters" registry.

Per [RFC6514] section 5, a "PMSI Tunnel" attribute contains a "Flags" octet. The Flags field is a single octet, with bits numbered, left-to-right, from 0 to 7. IANA is requested to initialize the registry as follows:

Bit Position (left to right)	Description	Reference
0	unassigned	
1	Extension	This document
2	unassigned	
3	unassigned	
4	unassigned	
5	unassigned	
6	unassigned	
7	Leaf Information Required (L)	RFC6514

PMSI Tunnel Attribute Flags

The registration procedure for this registry is Standards Action.

IANA is also requested to assign a codepoint, from the "First Come, First Served" range of the Transitive Opaque Extended Community Sub-Types registry, for "Additional PMSI Tunnel Attribute Flags". [TO BE REMOVED: This registration should take place at the following location: http://www.iana.org/assignments/bgp-extended-communities /bgp-extended-communities.xhtml#trans-opaque]

IANA is further requested to establish a registry for the bit flags carried in the "Additional PMSI Tunnel Attribute Flags" Extended Community. The bits shall be numbered 0-47, with 0 being the most significant bit and 47 being the least significant bit. The registration policy for this registry shall be "Standards Action". [TO BE REMOVED: The creation of the registry should take place at the following location: http://www.iana.org/assignments/bgp-extended-communities/bgp-extended-communities.xhtml] The initial registry should be as follows:

Bit Flag	Name	Reference
0-47	unassigned	

Additional PMSI Tunnel Attribute Flags

4. Acknowledgments

The authors wish to thank Martin Vigoureux for his review of this document. We also thank Christian Huitema and Alexey Melnikov for their review and comments.

5. Security Considerations

This document establishes an IANA registry, and defines a new Transitive Opaque Extended Community ([RFC4360], [RFC7153]).

Establishment of an IANA registry does not raise any security considerations.

While this document defines a new Extended Community for carrying bit flags, it does not define any of the bit flags in that Extended Community. Therefore no security considerations are raised.

This document defines a new flag, the "Extension" flag, in the "PMSI Tunnel" attribute. If a particular UPDATE contains "PMSI Tunnel" attribute with this flag set, but the UPDATE does not contain an "Additional PMSI Tunnel Attribute Flags" Extended Community, then the UPDATE is considered to be malformed, and the "treat-as-withdraw" procedure of [RFC7606] is invoked. Thus one can cause an UPDATE to be treated as a withdrawal by incorrectly setting this bit.

6. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <http://www.rfc-editor.org/info/rfc2119>.

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, DOI 10.17487/RFC4360, February 2006, <http://www.rfc-editor.org/info/rfc4360>.

[RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, DOI 10.17487/RFC6514, February 2012, <http://www.rfc-editor.org/info/rfc6514>.

[RFC7153] Rosen, E. and Y. Rekhter, "IANA Registries for BGP Extended Communities", RFC 7153, DOI 10.17487/RFC7153, March 2014, <http://www.rfc-editor.org/info/rfc7153>.

[RFC7606] Chen, E., Ed., Scudder, J., Ed., Mohapatra, P., and K. Patel, "Revised Error Handling for BGP UPDATE Messages", RFC 7606, DOI 10.17487/RFC7606, August 2015, <http://www.rfc-editor.org/info/rfc7606>.

Authors' Addresses

Eric C. Rosen
Juniper Networks, Inc.
10 Technology Park Drive
Westford, Massachusetts 01886
United States

Email: erosen@juniper.net

Thomas Morin
Orange
2, avenue Pierre-Marzin
22307 Lannion Cedex
France

Email: thomas.morin@orange.com

BESS
Internet-Draft
Intended status: Standards Track
Expires: September 10, 2015

R. Kebler, Ed.
J. Zhang
Juniper Networks
A. Dolganow
J. Kotalwar
Alcatel-Lucent
H. Sipra
Google

March 9, 2015

MVPN UMH Procedure Based on Source Active A-D Route
draft-kebler-bess-sa-pref-00

Abstract

This document define new procedures to use Source-Active A-D routes to influence the UMH selection procedures at a downstream PE in certain deployments. These procedures allow some greater flexibility to influence the UMH selection based on more than just the unicast route to the source.

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 September 10, 2015.

Copyright Notice

Table of Contents

1. Introduction

It may be desirable to influence the UMH selection result for a given customer multicast group, without influencing the UMH procedures for all the other customer groups with the same source. For example, if it is desirable for traffic to be chosen for S1,G1 from ingress PE, and for S1,G2 for a different ingress PE, it is not possible to accomplish with the exisitng UMH procedures that are based solely on the Source address.

Consider the case when an Anycast source address is being used to source the content from two headends. If the content were preferred from one headend for certain groups, and the other headend for other groups based on some policy on the ingress PEs depending on the particular groups, then this would not be possible with a source based UMH method.

This document define new procedures to use Source-Active A-D routes to influence the UMH selection procedures at an egress PE, taking both the Source and Group into account to allow greater flexibility in the UMH procedures.

As defined in RFC 6514, An ingress PE will advertise a (C-S,C-G) Source Active A-D route if it receives a PIM Register message or MSDP message saying that C-S is a source for C-G. When advertising the Source-Active A-D route, a policy can be applied at the ingress PEs (e.g., BGP communities) to help influence the BGP route selection of the egress PEs. The ingress PE can be configured to include some communities to the Source-Active A-D routes based on that policy. The egress PEs can then be configured to set the route preference based on the received communities. The exact details on procedures

to influence BGP route selection are outside the scope of this document. The selected Source Active A-D route will then be used to influence the UMH selection.

2. Applicability

These procedures are applicable only when procedures in Section 10 of RFC 6513 are being used to "Eliminate PE-PE Distribution of (C-*,C-G) State". Furthermore, the procedures in this document are restricted to the case when the ingress PEs are configured either MSDP or as RP. The typical use-case would be an IPTV deployment when a headend is located behind a set of PEs and those PEs can be configured as RPs or MSDP peers. These procedures are not applicable for groups in the SSM range.

3. Procedure Details

RFC 6513 describes procedures to build the "UMH Route Candidate Set" and then select the single route from the set to be the "Selected UMH Route". The procedures are modified to prefer, from the "UMH Route Candidate Set", the Upstream PE that has advertised the best (as determined by the BGP route selection procedures) Source-Active A-D route.

It may not be obvious on how to match the UMH candidate to the originator of the Source-Active A-D route since the NLRI of the Source Active A-D route does not specify the originator of the route. For MVPN procedures, refer to the extranet draft [I-D.ietf-bess-mvpn-extranet] (section 7.4). For Global Table Multicast (GTM) procedures, refer to the GTM draft [I-D.ietf-bess-mvpn-global-table-mcast] (section 2.8.1).

If the UMH is selected solely based on best Source Active A-D route without considering the UMH Route Candidate Set as defined in RFC 6514, then it would have the drawback that a UMH may be chosen which does not have reachability to the source through a vrf interface. Also, it may take some time for an RP to determine that the source has stopped sending traffic and the unicast reachability may converge before the Source Active A-D routes are withdrawn. As a result, using the UMH Route Candidate Set as the base can improve the convergence on the egress PEs.

4. IANA Considerations

None

5. Security Considerations

There are no security considerations for this design other than what is already in the base MVPN specifications.

6. Acknowledgements

The authors want to thank Eric Rosen for his review and useful feedback.

7. Normative References

[I-D.ietf-bess-mvpn-extranet]
Rekhter, Y., Rosen, E., Aggarwal, R., Cai, Y., Henderickx, W., Morin, T., Muley, P., Qiu, R., and I. Wijnands, "Extranet Multicast in BGP/IP MPLS VPNs", draft-ietf-bess-mvpn-extranet-00 (work in progress), November 2014.

[I-D.ietf-bess-mvpn-global-table-mcast]
Zhang, J., Giuliano, L., Rosen, E., Subramanian, K., Pacella, D., and J. Schiller, "Global Table Multicast with BGP-MVPN Procedures", draft-ietf-bess-mvpn-global-table-mcast-00 (work in progress), November 2014.

[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.

Authors' Addresses

Robert Kebler (editor)
Juniper Networks
10 Technology Park Drive
Westford, MA 01886
USA

Email: rkebler@juniper.net

Jeffrey Zhang
Juniper Networks
10 Technology Park Drive
Westford, MA 01886
USA

Email: zzhang@juniper.net

Andrew Dolganow
Alcatel-Lucent
600 March Rd.
Ottawa, Ontario K2K 2E6
Canada

Email: andrew.dolganow@alcatel-lucent.com

Jayant Kotalwar
Alcatel-Lucent
701 East Middlefield Rd
Mountain View, California 94043
United States

Email: jayant.kotalwar@alcatel-lucent.com

Hassan Sipra
Google

Email: hisipra@google.com

BESS Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 21, 2016

S. Mohanty
K. Patel
A. Sajassi
Cisco Systems, Inc.
J. Drake
Juniper Networks, Inc.
A. Przygienda
Ericsson

October 19, 2015

A new Designated Forwarder Election for the EVPN
draft-mohanty-bess-evpn-df-election-02

Abstract

This document describes an improved EVPN Designated Forwarder Election (DF) algorithm which can be used to enhance operational experience in terms of convergence speed and robustness over a WAN deploying EVPN

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 April 21, 2016.

Copyright Notice

Table of Contents

1. Introduction

Ethernet MPLS VPN (EVPN) [RFC7432] is an emerging technology that is gaining prominence in Internet Service Provider IP/MPLS networks. In EVPN, mac addresses are disseminated as routes across the geographical area via the Border Gateway Protocol, BGP [RFC4271] using the familiar L3VPN model [RFC4364]. An EVPN instance that spans across PEs is defined as an EVI. Constrained Route Distribution [RFC4684] can be used in conjunction to selectively advertise the routes to where they are needed. One of the major advantages of EVPN over VPLS [RFC4761],[RFC6624] is that it provides a solution for minimizing flooding of unknown traffic and also provides all Active mode of operation so that the traffic can truly be multi-homed. In technologies such as EVPN or VPLS, managing Broadcast, Unknown Unicast and multicast traffic (BUM) is a key requirement. In the case where the customer edge (CE) router is multi-homed to one or more Provider Edge (PE) Routers, it is necessary that one and only one of the PE routers should forward BUM traffic into the core or towards the CE as and when appropriate.

Specifically, quoting Section 8.5, [RFC7432], Consider a CE that is a host or a router that is multi-homed directly to more than one PE in

an EVPN instance on a given Ethernet segment. One or more Ethernet Tags may be configured on the Ethernet segment. In this scenario only one of the PEs, referred to as the Designated Forwarder (DF), is responsible for certain actions:

a. Sending multicast and broadcast traffic, on a given Ethernet Tag on a particular Ethernet segment, to the CE.

b. Flooding unknown unicast traffic (i.e. traffic for which an PE does not know the destination MAC address), on a given Ethernet Tag on a particular Ethernet segment to the CE, if the environment requires flooding of unknown unicast traffic.

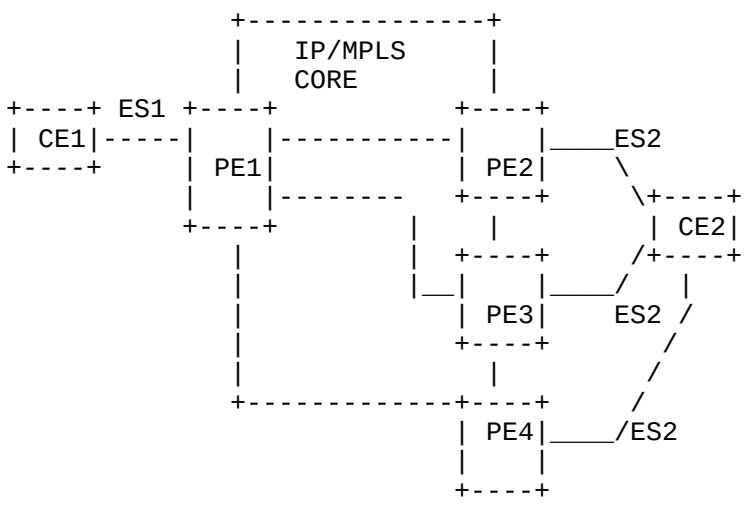


Figure 1 Multi-homing Network of E-VPN

Figure 1

Figure 1 illustrates a case where there are two Ethernet Segments, ES1 and ES2. PE1 is attached to CE1 via Ethernet Segment ES1 whereas PE2, PE3 and PE4 are attached to CE2 via ES2 i.e. PE2, PE3 and PE4 form a redundancy group. Since CE2 is multi-homed to different PEs on the same Ethernet Segment, it is necessary for PE2, PE3 and PE4 to agree on a DF to satisfy the above mentioned requirements.

Layer2 devices are particularly susceptible to forwarding loops because of the broadcast nature of the Ethernet traffic. Therefore

it is very important that in case of multi-homing, only one of the links be used to direct traffic to/from the core.

One of the pre-requisites for this support is that participating PEs must agree amongst themselves as to who would act as the Designated Forwarder. This needs to be achieved through a distributed algorithm in which each participating PE independently and unambiguously selects one of the participating PEs as the DF, and the result should be unanimously in agreement.

The DF election algorithm as described in [RFC7432] has some undesirable properties and in some cases can be somewhat disruptive and unfair. This document describes those issues and proposes a mechanism for dealing with those issues. These mechanisms do involve changes to the DF Election algorithm , but do not require any protocol changes to the EVPN Route exchange and have minimal changes to their content per se.

1.1. Finite State Machine

Since the specification in EVPN RFC [RFC7432] does leave several questions open as to the precise final state machine behavior of the DF election, the document also includes a section describing precisely the intended behavior. The finite state machine is presented in Section 7.1

1.2. 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].

2. The modulus based DF Election Algorithm

The default procedure for DF election at the granularity of (ESI,EVI) is referred to as "service carving". With service carving, it is possible to elect multiple DFs per Ethernet Segment (one per EVI) in order to perform load-balancing of multi-destination traffic destined to a given Segment. The objective is that the load-balancing procedures should carve up the EVI space among the redundant PE nodes evenly, in such a way that every PE is the DF for a disjoint set of EVIs.

The existing DF algorithm as described in the EVPN RFC(Section 8.5 [RFC7432]) is based on a modulus operation. The PEs to which the ES (for which DF election is to be carried out per vlan) is multi-homed from an ordered (ordinal) list in ascending order of the PE ip address values. Say, there are N PEs, P0, P1, ... PN-1 ranked as per

increasing IP addresses in the ordinal list; then for each vlan with ethernet tag v, configured on the ethernet segment ES1, PEx is the DF for vlan v on ES ES1 when x equals (v mod N). In the case of VLAN bundle only the lowest VLAN is used. In the case when the vlan density is high meaning there are significant number of vlans and the vlan-id or ethernet-tag is uniformly distributed, the thinking is that the DF election will be spread across the PEs hosting that ethernet segment and good service carving can be achieved.

3. Problems with the modulus based DF Election Algorithm

There are three fundamental problems with the current DF Election.

First, the algorithm will not perform well when the ethernet tag follows a non-uniform distribution, for instance when the ethernet tags are all even or all odd. In such a case let us assume that the ES is multi-homed to two PEs; all the vlans will only pick one of the PEs as the DF. This is very sub-optimal. It defeats the purpose of service carving as the DFs are not really evenly spread across. In this particular case, in fact one of the PEs does not get elected all as the DF, so it does not participate in the DF responsibilities at all. Consider another example where referring to Figure 1, lets assume that PE2, PE3, PE4 are in ascending order of the IP address; and each vlan configured on ES2 is associated with an Ethernet Tag of of the form (3x+1), where x is an integer. This will result in PE3 always be selected as the DF.

Even in the case when the ethernet tag distribution is uniform the instance of a PE being up or down results in re-computation ((v mod N-1) or (v mod N+1) as is the case); The resulting modulus value need not be uniformly distributed but subject to the primality of N-1 or N+1 as may be the case.

The third problem is one of disruption. Consider a case when the same Ethernet Segment is multi homed to a set of PEs. When the ES is down in one of the PEs, say PE1, or PE1 itself reboots, or the BGP process goes down or the connectivity between PE1 and an RR goes down, the effective number of PEs in the system now becomes N-1 and DFs are computed for all the vlans that are configured on that ethernet segment. In general, if the DF for a vlan v happens not to be PE1, but some other PE, say PE2, it is likely that some other PE will become the new DF. This is not desirable. Similarly when a new PE hosts the same Ethernet segment, the mapping again changes because of the mod operation. This results in needless churn. Again referring to Figure 1, say v1, v2 and v3 are vlans configured on ES2 with associated ethernet tags of value 999, 1000 and 10001 respectively. So PE1, PE2 and PE3 are also

the DFs for v1, v2 and v3 respectively. Now when PE3 goes down, PE2 will become the DF for v1 and PE1 will become the DF for v2.

One point to note is that the current DF election algorithm assumes that all the PEs who are multi-homed to the same Ethernet Segment and interested in the DF Election by exchanging EVPN routes have a V4 peering with each other or via a Route Reflector. This need not be the case as there can be a v6 peering and supporting the EVPN address-family.

Mathematically, a conventional hash function maps a key k to a number i representing one of m hash buckets through a function h(k) i.e. i=h(k). In the EVPN case, h is simply a modulo-m hash function viz. h(v) = v mod N, where N is the number of PEs that are multi-homed to the Ethernet Segment in discussion. It is well-known that for good hash distribution using the modulus operation, the modulus N should be a prime-number not too close to a power of 2 [CLRS2009]. When the effective number of PEs changes from N to N-1 (or vice versa); all the objects (vlan v) will be remapped except those for which v mod N and v mod (N-1) refer to the same PE in the previous and subsequent ordinal rankings respectively.

From a forwarding perspective, this is a churn, as it results in programming the CE and PE side ports as blocking or non-blocking at potentially all PEs when the DF changes either because (i) a new PE is added or (ii) another one goes down or loses connectivity or else cannot take part in the DF election process for whatever reason. This draft addresses this problem and furnishes a solution to this undesirable behavior.

4. Highest Random Weight

Highest Random Weight (HRW) as defined in [HRW1999] is originally proposed in the context of Internet Caching and proxy Server load balancing. Given an object name and a set of servers, HRW maps a request to a server using the object-name (object-id) and server-name (server-id) rather than the state of the server states. HRW forms a hash out of the server-id and the object-id and forms an ordered list of the servers for the particular object-id. The server for which the hash value is highest, serves as the primary responsible for that particular object, and the server with the next highest value in that hash serves as the backup server. HRW always maps a given object object name to the same server within a given cluster; consequently it can be used at client sites to achieve global consensus on object-server mappings. When that server goes down, the backup server becomes the responsible designate.

Choosing an appropriate hash function that is statistically oblivious to the key distribution and imparts a good uniform distribution of the hash output is an important aspect of the algorithm,. Fortunately many such hash functions exist. [HRW1999] provides pseudorandom functions based on Unix utilities rand and srand and easily constructed XOR functions that perform considerably well. This imparts very good properties in the load balancing context. Also each server independently and unambiguously arrives at the primary server selection. HRW already finds use in multicast and ECMP [RFC2991],[RFC2992].

In the existing DF algorithm Section 2, whenever a new PE comes up or an existing PE goes down, there is a significant interval before the change is noticed by all peer PEs as it has to be conveyed by the BGP update message involving the type-4 route. There is a timer to batch all the messages before triggering the service carving procedures. When the timer expires, each PE will build the ordered list and follow the procedures for DF Election. In the proposed method which we will describe shortly this "jittered" behavior is retained.

5. HRW and Consistent Hashing

HRW is not the only algorithm that addresses the object to server mapping problem with goals of fair load distribution, redundancy and fast access. There is another family of algorithms that also addresses this problem; these fall under the umbrella of the Consistent Hashing Algorithms [CHASH]. These will not be considered here.

6. HRW Algorithm for EVPN DF Election

The applicability of HRW to DF Election can be described here. Let DF(v) denote the Designated Forwarder and BDF(v) the Backup Designated forwarder for the ethernet tag V, where v is the vlan, Si is the IP address of server i and weight is a pseudorandom function of v and Si. In case of a vlan bundle service, v denotes the lowest vlan similar to the ’lowest vlan in bundle’ logic of [RFC7432].

1. DF(v) = Si: Weight(v, Si) >= Weight(V, Sj) , for all j. In case of a tie, choose the PE whose IP address is numerically the least.

2. BDF(v) = Sk: Weight(v, Si) >= Weight(V, Sk) and Weight(v, Sk) >= Weight(v, Sj). in case of tie choose the PE whose IP address is numerically the least.

Since the Weight is a Pseudorandom function with domain as a concatenation of (v, S), it is an efficient deterministic algorithm

which is independent of the Ethernet Tag V sample space distribution. Choosing a good hash function for the pseudorandom function is an important consideration for this algorithm to perform provably better than the existing algorithm. As mentioned previously, such functions are described in the HRW paper. We take as candidate hash functions two of the ones that are preferred in [HRW1999].

1. Wrand(v, Si) = (1103515245((1103515245.Si+12345)XOR D(v))+12345)(mod 2^31) and

2. Wrand2(v, Si) = (1103515245((1103515245.D(v)+12345)XOR Si)+12345)(mod 2^31)

Here D(v) is the 31-bit digest of the ethernet-tag v and Si is address of the ith server. The server’s IP address length does not matter as only the low-order 31 bits are modulo significant. Eventually we plan to choose one of the two candidate hash functions as the preferred one.

A point to note is that the the domain of the Weight function is a concatenation of the ethernet-tag and the PE IP-address, and the actual length of the server IP address (whether V4 or V6) is not really relevant, so long as the actual hash algorithm takes into consideration the concatenated string. The existing algorithm in [RFC7432] as is cannot employ both V4 and V6 neighbor peering address.

HRW solves the disadvantage pointed out in Section 3 and ensures

- with very high probability that the task of DF election for respective vlans is more or less equally distributed among the PEs even for the 2 PE case

- If a PE, hosting some vlans on given ES, but is neither the DF nor the BDF for that vlan, goes down or its connection to the ES goes down, it does not result in a DF and BDF reassignment the other PEs. This saves computation, especially in the case when the connection flaps.

- More importantly it avoids the needless disruption case (c) that are inherent in the existing modulus based algorithm

- In addition to the DF, the algorithm also furnishes the BDF, which would be the DF if the current DF fails.

7. Protocol Considerations

Note that for the DF election procedures to be globally convergent and unanimous, it is necessary that all the participating PEs agree on the DF Election algorithm to be used. It is not possible that some PEs continue to use the existing modulus based DF election and some newer PEs use the HRW. For brownfield deployments and for interoperability with legacy boxes, its is important that all PEs need to have the capability to fall back on the modulus algorithm. A PE (one with a newer version of the software) can indicate its willingness to support HRW by signaling a new extended community along with the Ethernet-Segment Route (Type-4). This extended community is explained in the next paragraph. When a PE receives the Ethernet-Segment Routes from all the other PEs for the ethernet segment in question, it checks to see if all the advertisements have the extended community attached; in the case that they do, this particular PE, and by induction all the other PEs proceed to do DF Election as per the HRW Algorithm. Otherwise if even a single advertisement for the type-4 route is not received with the extended community or the received DF types (including locally configured type) do not ALL match a single value, the default modulus algorithm is used as before. Also, the HRW algorithm needs to be executed after the "batching" time.

A new BGP extended community attribute [RFC4360] needs to be defined to identify the DF election procedure to be used for the Ethernet Segment. We propose to name this extended community as the DF Election Extended Community. It is a new transitive extended community where the Type field is 0x06, and the Sub-Type is to be defined. It may be advertised along with Ethernet Segment routes.

Each DF Election Extended Community is encoded as a 8-octet value as follows:

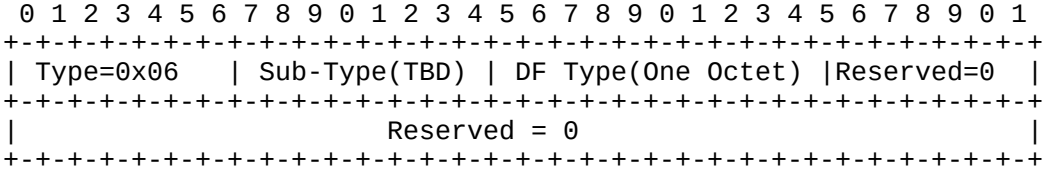


Figure 2

The DF Type state is encoded as one octet. A value of 0 means that the default (the mod based) DF election procedures are used and a value of 1 means that the HRW algorithm will be employed. A request

needs to registered with the IETF authority for the subtype [I-D.ietf-idr-extcomm-iana]

7.1. Finite State Machine

Per [RFC7432], the FSM described in Figure 3 is executed per ESI/VLAN in case of VLAN aware service or ESI/[VLANs in VLAN Bundle] in case of VLAN Bundle on each participating PE.

Observe that currently the VLANs are derived from local configuration and the FSM does not provide any protection against misconfiguration where same EVI,ESI combination has different set of VLANs on different participating PEs or one of the PEs elects to consider VLANs as VLAN bundle and another as separate VLANs for election purposes (service type mismatch).

The FSM is normative in the sense that any design or implementation MUST behave towards external peers and as observable external behavior (DF) in a manner equivalent to this FSM.

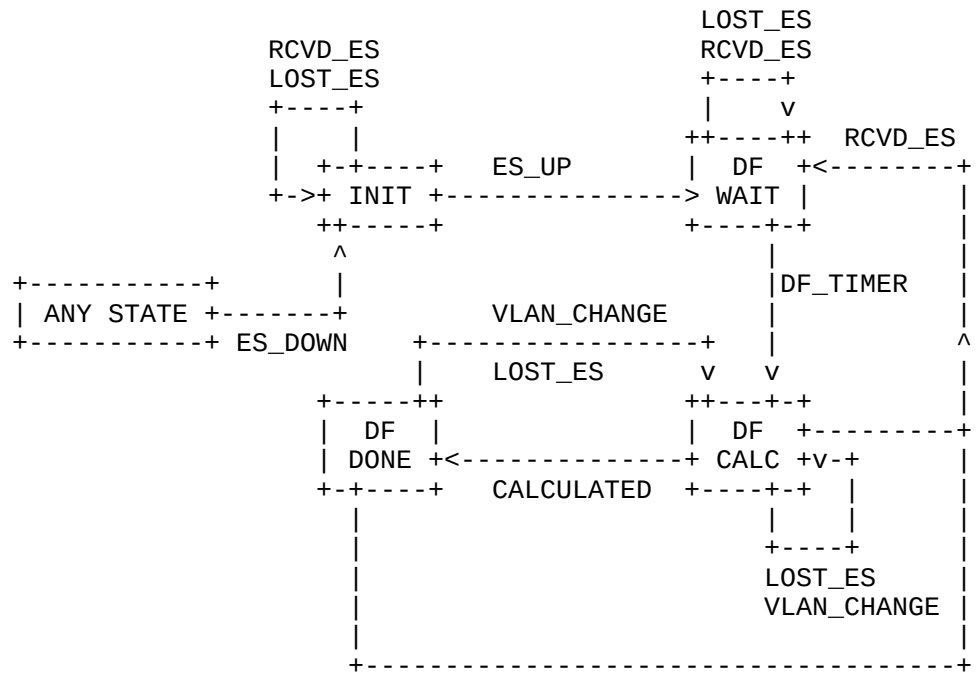


Figure 3

States:

1. INIT: Initial State

2. DF WAIT: State in which the participants waits for enough information to perform the DF election for the EVI/ESI/VLAN combination.

3. DF CALC: State in which the new DF is recomputed.

4. DF DONE: State in which the according DF for the EVI/ESI/VLAN combination has been elected.

Events:

1. ES_UP: The ESI has been locally configured as 'up'.

2. ES_DOWN: The ESI has been locally configured as 'down'.

3. VLAN_CHANGE: The VLANs configured in a bundle that uses the ESI changed. This event is necessary for VLAN bundles only.

4. DF_TIMER: DF Wait timer has expired.

5. RCVD_ES: A new or changed Ethernet Segment Route is received in a BGP REACH UPDATE. Receiving an unchanged UPDATE MUST NOT trigger this event.

6. LOST_ES: A BGP UNREACH UPDATE for a previously received Ethernet Segment route has been received. If an UNREACH is seen for a route that has not been advertised previously, the event MUST NOT be triggered.

7. CALCULATED: DF has been succesfully calculated.

According actions when transitions are performed or states entered/ exited:

1. ANY STATE on ES_DOWN: (i)stop DF timer (ii) assume non-DF for local PE

2. INIT on ES_UP: (i)do nothing

3. INIT on RCVD_ES, LOST_ES: (i)do nothing

4. DF_WAIT on entering the state: (i) start DF timer if not started already or expired (ii) assume non-DF for local PE

5. DF_WAIT on RCVD_ES, LOST_ES: do nothing

6. DF_WAIT on DF_TIMER: do nothing

7. DF_CALC on entering or re-entering the state: (i) rebuild according list and hashes and perform election (ii) FSM generates CALCULATED event against itself

8. DF_CALC on LOST_ES or VLAN_CHANGE: do nothing

9. DF_CALC on RCVD_ES: do nothing

10. DF_CALC on CALCULATED: (i) mark election result for VLAN or bundle

11. DF_DONE on exiting the state: (i)if RFC7432 election or new election and lost primary DF then assume non-DF for local PE for VLAN or VLAN bundle.

12. DF_DONE on VLAN_CHANGE or LOST_ES: do nothing

8. Auto-Derivation of ES-Import Route Target

Section 7.6 of RFC7432 describes how the value of the ES-Import Route Target for ESI types 1, 2, and 3 can be auto-derived by using the high-order six bytes of the nine byte ESI value. This document extends the same auto-derivation procedure to ESI types 0, 4, and 5.

9. Operational Considerations

TBD.

10. Security Considerations

This document raises no new security issues for EVPN.

11. Acknowledgements

The authors would like to thank Tamas Mondal, Sami Boutros, Jakob Heitz, Jorge Rabadan and Patrice Brissette for useful feedback and discussions.

12. References

12.1. Normative References

[HRW1999] Thaler, D. and C. Ravishankar, "Using Name-Based Mappings to Increase Hit Rates", IEEE/ACM Transactions in networking Volume 6 Issue 1, February 1998.

[I-D.ietf-idr-extcomm-iana] Rosen, E. and Y. Rekhter, "IANA Registries for BGP Extended Communities", draft-ietf-idr-extcomm-iana-02 (work in progress), December 2013.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <http://www.rfc-editor.org/info/rfc2119>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <http://www.rfc-editor.org/info/rfc4271>.

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, DOI 10.17487/RFC4360, February 2006, <http://www.rfc-editor.org/info/rfc4360>.

[RFC4761] Kompella, K., Ed. and Y. Rekhter, Ed., "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, DOI 10.17487/RFC4761, January 2007, <http://www.rfc-editor.org/info/rfc4761>.

[RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <http://www.rfc-editor.org/info/rfc7432>.

12.2. Informative References

[CHASH] Karger, D., Lehman, E., Leighton, T., Panigrahy, R., Levine, M., and D. Lewin, "Consistent Hashing and Random Trees: Distributed Caching Protocols for Relieving Hot Spots on the World Wide Web", ACM Symposium on Theory of Computing ACM Press New York, May 1997.

[CLRS2009] Cormen, T., Leiserson, C., Rivest, R., and C. Stein, "Introduction to Algorithms (3rd ed.)", MIT Press and McGraw-Hill ISBN 0-262-03384-4., February 2009.

[RFC2991] Thaler, D. and C. Hopps, "Multipath Issues in Unicast and Multicast Next-Hop Selection", RFC 2991, DOI 10.17487/RFC2991, November 2000, <http://www.rfc-editor.org/info/rfc2991>.

[RFC2992] Hopps, C., "Analysis of an Equal-Cost Multi-Path Algorithm", RFC 2992, DOI 10.17487/RFC2992, November 2000, <http://www.rfc-editor.org/info/rfc2992>.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <http://www.rfc-editor.org/info/rfc4364>.

[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, DOI 10.17487/RFC4684, November 2006, <http://www.rfc-editor.org/info/rfc4684>.

[RFC6624] Kompella, K., Kothari, B., and R. Cherukuri, "Layer 2 Virtual Private Networks Using BGP for Auto-Discovery and Signaling", RFC 6624, DOI 10.17487/RFC6624, May 2012, <http://www.rfc-editor.org/info/rfc6624>.

Authors' Addresses

Satya Ranjan Mohanty
Cisco Systems, Inc.
225 West Tasman Drive
San Jose, CA 95134
USA

Email: satyamoh@cisco.com

Keyur Patel
Cisco Systems, Inc.
225 West Tasman Drive
San Jose, CA 95134
USA

Email: keyupate@cisco.com

Ali Sajassi
Cisco Systems, Inc.
225 West Tasman Drive
San Jose, CA 95134
USA

Email: sajassi@cisco.com

John Drake
Juniper Networks, Inc.
1194 N. Mathilda Drive
Sunnyvale, CA 95134
USA

Email: jdrake@juniper.com

Antoni Przygienda
Ericsson
300 Holger Way
San Jose, CA 95134
USA

Email: antoni.przygienda@ericsson.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 3, 2016

T. Morin, Ed.
Orange
R. Kebler, Ed.
Juniper Networks

July 6, 2015

Multicast VPN fast upstream failover
draft-morin-bess-mvpn-fast-failover-02

Abstract

This document defines multicast VPN extensions and procedures that allow fast failover for upstream failures, by allowing downstream PEs to take into account the status of Provider-Tunnels (P-tunnels) when selecting the upstream PE for a VPN multicast flow, and extending BGP MVPN routing so that a C-multicast route can be advertized toward a standby upstream PE.

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 January 3, 2016.

Copyright Notice

Table of Contents

1. Introduction

In the context of multicast in BGP/MPLS VPNs, it is desirable to provide mechanisms allowing fast recovery of connectivity on different types of failures. This document addresses failures of

elements in the provider network that are upstream of PEs connected to VPN sites with receivers.

The sections 3 and 4 describe two independent mechanisms, allowing different levels of resiliency, and providing different failure coverage:

- o Section 3 describes local procedures allowing an egress PE (a PE connected to a receiver site) to take into account the status of P-Tunnels to determine the Upstream Multicast Hop (UMH) for a given (C-S, C-G). This method does not provide a "fast failover" solution when used alone, but can be used with the following sections for a "fast failover" solution.

- o Section 4 describes protocol extensions that can speed up failover by not requiring any multicast VPN routing message exchange at recovery time.

Moreover, section 5 describes a "hot leaf standby" mechanism, that uses a combination of these two mechanisms. This approach has similarities with the solution described in [I-D.mofrr] to improve failover times when PIM routing is used in a network given some topology and metric constraints.

2. Terminology

The terminology used in this document is the terminology defined in [RFC6513] and [RFC6514].

3. UMH Selection based on tunnel status

Current multicast VPN specifications [RFC6513], section 5.1, describe the procedures used by a multicast VPN downstream PE to determine what the upstream multicast hop (UMH) is for a said (C-S,C-G).

The procedure described here is an OPTIONAL procedure that consists of having a downstream PE take into account the status of P-tunnels rooted at each possible upstream PEs, for including or not including each said PE in the list of candidate UMHs for a said (C-S,C-G) state. The result is that, if a P-tunnel is "down" (see Section 3.1), the PE that is the root of the P-Tunnel will not be considered for UMH selection, which will result in the downstream PE to failover to the upstream PE which is next in the list of candidates.

A downstream PE monitors the status of the tunnels of UMHs that are ahead of the current one. Whenever the downstream PE determines that

one of these tunnels is no longer "known to down", the PE selects the UMH corresponding to that as the new UMH.

More precisely, UMH determination for a said (C-S,C-G) will consider the UMH candidates in the following order:

- first, the UMH candidates that either (a) advertise a PMSI bound to a tunnel, where the specified tunnel is not known to be down or (b) do not advertise any I- or S- PMSI applicable to the said (C-S,C-G) but have associated a VRF Route Import BGP attribute to the unicast VPN route for S (this is necessary to avoid considering some invalid UMH PEs that use a policy where no I-PMSI is advertized for a said VRF and where only S-PMSI are used, the S-PMSI advertisement being possibly done only after the upstream PE receives a C-multicast route for (C-S, C-G)/(C-*, C-G) to be carried over the advertized S-PMSI)

- second, the UMH candidates that advertise a PMSI bound to a tunnel that is "down" -- these will thus be used as a last resort to ensure a graceful fallback to the basic MVPN UMH selection procedures in the hypothetical case where a false negative would occur when determining the status of all tunnels

For a said downstream PE and a said VRF, the P-tunnel corresponding to a said upstream PE for a said (C-S,C-G) state is the S-PMSI tunnel advertized by that upstream PE for this (C-S,C-G) and imported into that VRF, or if there isn’t any such S-PMSI, the I-PMSI tunnel advertized by that PE and imported into that VRF.

Note that this documents assumes that if a site of a given MVPN that contains C-S is dual-homed to two PEs, then all the other sites of that MVPN would have two unicast VPN routes (VPN-IPv4 or VPN-IPv6) routes to C-S, each with its own RD.

3.1. Determining the status of a tunnel

Different factors can be considered to determine the "status" of a P-tunnel and are described in the following sub-sections. The procedure proposed here also allows that all downstream PEs don’t apply the same rules to define what the status of a P-tunnel is (please see Section 6), and some of them will produce a result that may be different for different downstream PEs. Thus what is called the "status" of a P-tunnel in this section, is not a characteristic of the tunnel in itself, but is the status of the tunnel, *as seen from a particular downstream PE*. Additionally, some of the following methods determine the ability of downstream PE to receive traffic on the P-tunnel and not specifically on the status of the P-tunnel itself. This could be referred to as "P-tunnel reception

status", but for simplicity, we will use the terminology of P-tunnel "status" for all of these methods.

Depending on the criteria used to determine the status of a P-tunnel, there may be an interaction with another resiliency mechanism used for the P-tunnel itself, and the UMH update may happen immediately or may need to be delayed. Each particular case is covered in each separate sub-section below.

3.1.1. mVPN tunnel root tracking

A condition to consider that the status of a P-tunnel is up is that the root of the tunnel, as determined in the PMSI tunnel attribute, is reachable through unicast routing tables. In this case the downstream PE can immediately update its UMH when the reachability condition changes.

This is similar to BGP next-hop tracking for VPN routes, except that the address considered is not the BGP next-hop address, but the root address in the PMSI tunnel attribute.

If BGP next-hop tracking is done for VPN routes, and the root address of a said tunnel happens to be the same as the next-hop address in the BGP autodiscovery route advertising the tunnel, then this mechanisms may be omitted for this tunnel, as it will not bring any specific benefit.

3.1.2. PE-P Upstream link status

A condition to consider a tunnel status as up can be that the last-hop link of the P-tunnel is up.

This method should not be used when there is a fast restoration mechanism (such as MPLS FRR [RFC4090]) in place for the link.

3.1.3. P2MP RSVP-TE tunnels

For P-Tunnels of type P2MP MPLS-TE, the status of the P-Tunnel is considered up if one or more of the P2MP RSVP-TE LSPs, identified by the P-Tunnel Attribute, are in up state. The determination of whether a P2MP RSVP-TE LSP is in up state requires Path and Resv state for the LSP and is based on procedures in [RFC4875]. In this case the downstream PE can immediately update its UMH when the reachability condition changes.

When signaling state for a P2MP TE LSP is removed (e.g. if the ingress of the P2MP TE LSP sends a PathTear message) or the P2MP TE LSP changes state from up to down as determined by procedures in

[RFC4875], the status of the corresponding P-Tunnel SHOULD be re-evaluated. If the P-Tunnel transitions from up to down state, the upstream PE, that is the ingress of the P-Tunnel, SHOULD not be considered a valid UMH.

3.1.4. Leaf-initiated P-tunnels

A PE can be removed from the UMH candidate list for a said (S,G) if the P-tunnel for this S,G (I or S , depending) is leaf triggered (PIM, mLDP), but for some reason internal to the protocol the upstream one-hop branch of the tunnel from P to PE cannot be built. In this case the downstream PE can immediately update its UMH when the reachability condition changes.

3.1.5. (S,G) counter information

In cases, where the downstream node can be configured so that the maximum inter-packet time is known for all the multicast flows mapped on a P-tunnel, the local per-(C-S,C-G) traffic counter information for traffic received on this P-tunnel can be used to determine the status of the P-tunnel.

When such a procedure is used, in context where fast restoration mechanisms are used for the P-tunnels, downstream PEs should be configured to wait before updating the UMH, to let the P-tunnel restoration mechanism happen. A configurable timer MUST be provided for this purpose, and it is recommended to provide a reasonable default value for this timer.

This method can be applicable for instance when a (S,G) flow is mapped on an S-PMSI.

In cases where this mechanism is used in conjunction with Hot leaf standby, then no prior knowledge of the rate of the multicast streams is required ; downstream PEs can compare reception on the two P-tunnels to determine when one of them is down.

3.1.6. BFD Discriminator

P-tunnel status can be derived from the status of a BFD session whose discriminator is advertized along with an x-PMSI A-D route.

3.1.6.1. Root PE Procedures

When it is desired to track the P-Tunnel status using BFD, the Root PE MUST include the BGP-BFD Attribute in the x-PMSI A-D Route.

If a P-Tunnel is already signaled, and then it is desired to track the P-Tunnel status using BFD, x-PMSI A-D Route must be re-sent with the same attributes as before, but the BGP-BFD Attribute MUST be included.

If P-Tunnel is already signaled, and P-Tunnel status tracked using BFD and it is desired to stop tracking P-Tunnel status using BFD, then x-PMSI A-D Route MUST be re-sent with the same attributes as before, but the BGP-BFD Attribute MUST be excluded.

3.1.6.2. Leaf PE Procedures

On receiving the BFD attribute in the x-PMSI A-D Route, the Leaf PE MUST associate the received discriminator with the P-Tunnel originating from the Root PE. Once the Leaf PE start getting the BFD probes from the Root PE with the said discriminator, the BFD session will be declared up and will then be used to track the health of the P-Tunnel.

If the Leaf PE does not receive BFD probes for a P-Tunnel from the Root PE for Detection Time, the BFD session would be brought down. And, it would declare the P-tunnel associated with the discriminator as down.

Leaf PE then can then initiate a switchover of the traffic from the Primary Tunnel, to the Standby Tunnel.

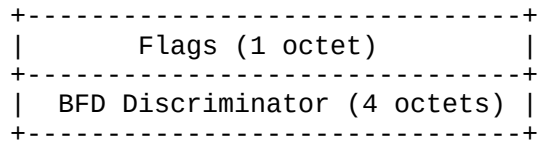
When Leaf PE’s P-Tunnel is already up, it receives new x-PMSI A-D Route with BGP-BFD attribute, it must accept the x-PMSI A-D Route and associate the discriminator with the P-tunnel. When the BFD probes are received with the said discriminator, the BFD session is declared up.

When Leaf PE’s P-Tunnel is already up, and is tracked with BFD, and it receives new x-PMSI A-D Route without BGP-BFD attribute, it must accept the x-PMSI A-D Route the BFD session should be declared admin down. Receiver node SHOULD not switch the traffic to the Standby P-tunnel.

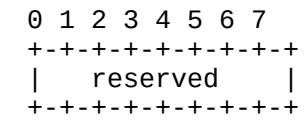
When such a procedure is used, in context where fast restoration mechanisms are used for the P-tunnels, leaf PEs should be configured to wait before updating the UMH, to let the P-tunnel restoration mechanism happen. A configurable timer MUST be provided for this purpose, and it is recommended to provide a reasonable default value for this timer.

3.1.6.3. BGP-BFD Attribute

This document defines and uses a new BGP attribute called the "BGP-BFD attribute". This is an optional transitive BGP attribute. The format of this attribute is defined as follows:



The Flags field has the following format:



3.1.7. Per PE-CE link BFD Discriminator

The following approach is proposed for fast failover on PE-CE link failures, in which UMH selection for a said (S,G) takes into account the state of a BFD session dedicated to the state of the upstream PE-CE link.

If this approach is enabled:

- each upstream PE: for each PE-CE link for which this protection is wanted, initiates a multipoint BFD session toward downstream PEs, with a trigger causing such a session to be torn down if the associated PE-CE link is detected as down.

- each upstream PE: for each prefix of a PE-CE link for which protection is wanted, advertizes a wildcard S-PMSI covering the sources inside this prefix, and signals along with this S-PMSI the multipoint BFD session discriminator associated with the PE-CE link. (note that all these S-PMSIs can perfectly use the same P-tunnel)

- each downstream PE: if an S-PMSI bound to a said (S,G) is signaled with a multipoint BFD session, then the upstream PE is considered during UMH selection for (S,G) if and only if the corresponding BFD session is up. Whenever the BFD session goes down the S-PMSI P-tunnel will be considered down and the downstream PE will switch to the backup P-tunnel. Note that the P-tunnel is considered down only for the (S,G) states that match to an S-PMSI signaling the BFD discriminator of a BFD session which is down

4. Standby C-multicast route

The procedures described below are limited to the case where the site that contains C-S is connected to exactly two PEs. The procedures require all the PEs of that MVPN to follow the single forwarder PE selection, as specified in [RFC6513]. The procedures assume that if a site of a given MVPN that contains C-S is dual-homed to two PEs, then all the other sites of that MVPN would have two unicast VPN routes (VPN-IPv4 or VPN-IPv6) routes to C-S, each with its own RD.

As long as C-S is reachable via both PEs, a said downstream PE will select one of the PEs connected to C-S as its Upstream PE with respect to C-S. We will refer to the other PE connected to C-S as the "Standby Upstream PE". Note that if the connectivity to C-S through the Primary Upstream PE becomes unavailable, then the PE will select the Standby Upstream PE as its Upstream PE with respect to C-S.

For readability, in the following sub-sections, the procedures are described for BGP C-multicast Source Tree Join routes, but they apply equally to BGP C-multicast Shared Tree Join routes failover for the case where the customer RP is dual-homed (substitute "C-RP" to "C-S").

4.1. Downstream PE behavior

When a (downstream) PE connected to some site of an MVPN needs to send a C-multicast route (C-S, C-G), then following the procedures specified in Section "Originating C-multicast routes by a PE" of [RFC6514] the PE sends the C-multicast route with RT that identifies the Upstream PE selected by the PE originating the route. As long as C-S is reachable via the Primary Upstream PE, the Upstream PE is the Primary Upstream PE. If C-S is reachable only via the Standby Upstream PE, then the Upstream PE is the Standby Upstream PE.

If C-S is reachable via both the Primary and the Standby Upstream PE, then in addition to sending the C-multicast route with an RT that identifies the Primary Upstream PE, the PE also originates and sends a C-multicast route with an RT that identifies the Standby Upstream

PE. This route, that has the semantic of being a 'standby' C-multicast route, is further called a "Standby BGP C-multicast route", and is constructed as follows:

- the NLRI is constructed as the original C-multicast route, except that the RD is the same as if the C-multicast route was built using the standby PE as the UMH (it will carry the RD associated to the unicast VPN route advertized by the standby PE for S)

- SHOULD carry the "Standby PE" BGP Community (this is a new BGP Community, see Section 7)

The normal and the standby C-multicast routes must have their Local Preference attribute adjusted so that, if two C-multicast routes with same NLRI are received by a BGP peer, one carrying the "Standby PE" attribute and the other one *not* carrying the "Standby PE" community, then preference is given to the one *not* carrying the "Standby PE" attribute. Such a situation can happen when, for instance due to transient unicast routing inconsistencies, two different downstream PEs consider different upstream PEs to be the primary one ; in that case, without any precaution taken, both upstream PEs would process a standby C-multicast route and possibly stop forwarding at the same time. For this purpose a Standby BGP C-multicast route MUST have the LOCAL_PREF attribute set to zero.

Note that, when a PE advertizes such a Standby C-multicast join for an (S,G) it must join the corresponding P-tunnel.

If at some later point the local PE determines that C-S is no longer reachable through the Primary Upstream PE, the Standby Upstream PE becomes the Upstream PE, and the local PE re-sends the C-multicast route with RT that identifies the Standby Upstream PE, except that now the route does not carry the Standby PE BGP Community (which results in replacing the old route with a new route, with the only difference between these routes being the presence/absence of the Standby PE BGP Community).

4.2. Upstream PE behavior

When a PE receives a C-multicast route for a particular (C-S, C-G), and the RT carried in the route results in importing the route into a particular VRF on the PE, if the route carries the Standby PE BGP Community, then the PE performs as follows:

when the PE determines that C-S is not reachable through some other PE, the PE SHOULD install VRF PIM state corresponding to this Standby BGP C-multicast route (the result will be that a PIM Join message will be sent to the CE towards C-S, and that the PE

will receive (C-S,C-G) traffic), and the PE SHOULD forward (C-S, C-G) traffic received by the PE to other PEs through a P-tunnel rooted at the PE.

Furthermore, irrespective of whether C-S carried in that route is reachable through some other PE:

a) based on local policy, as soon as the PE receives this Standby BGP C-multicast route, the PE MAY install VRF PIM state corresponding to this BGP Source Tree Join route (the result will be that Join messages will be sent to the CE toward C-S, and that the PE will receive (C-S,C-G) traffic)

b) based on local policy, as soon as the PE receives this Standby BGP C-multicast route, the PE MAY forward (C-S, C-G) traffic to other PEs through a P-tunnel independently of the reachability of C-S through some other PE. [note that this implies also doing (a)]

Doing neither (a), nor (b) for a said (C-S,C-G) is called "cold root standby".

Doing (a) but not (b) for a said (C-S,C-G) is called "warm root standby".

Doing (b) (which implies also doing (a)) for a said (C-S,C-G) is called "hot root standby".

Note that, if an upstream PE uses an S-PMSI only policy, it shall advertise an S-PMSI for an (S,G) as soon as it receives a C-multicast route for (S,G), normal or Standby ; i.e. it shall not wait for receiving a non-Standby C-multicast route before advertising the corresponding S-PMSI.

Section 9.3.2 of [RFC6514], describes the procedures of sending a Source-Active A-D result as a result of receiving the C-multicast route. These procedures should be followed for both the normal and Standby C-multicast routes.

4.3. Reachability determination

The standby PE can use the following information to determine that C-S can or cannot be reached through the primary PE:

- presence/absence of a unicast VPN route toward C-S

- supposing that the standby PE is an egress of the tunnel rooted at the Primary PE, the standby PE can determine the reachability of C-S through the Primary PE based on the status of this tunnel,

determined thanks to the same criteria as the ones described in Section 3.1 (without using the UMH selection procedures of Section 3)

- other mechanisms MAY be used

4.4. Inter-AS

If the non-segmented inter-AS approach is used, the procedures in section 4 can be applied.

When multicast VPNs are used in a inter-AS context with the segmented inter-AS approach described in section 8.2 of [RFC6514], the procedures in this section can be applied.

A pre-requisite for the procedures described below to be applied for a source of a said MVPN is:

- that any PE of this MVPN receives two Inter-AS I-PMSI auto-discovery routes advertized by the AS of the source (or more)

- that these Inter-AS I-PMSI autodiscovery routes have distinct Route Distinguishers (as described in item "(2)" of section 9.2 of [RFC6514]).

As an example, these conditions will be satisfied when the source is dual homed to an AS that connects to the receiver AS through two ASBR using auto-configured RDs.

4.4.1. Inter-AS procedures for downstream PEs, ASBR fast failover

The following procedure is applied by downstream PEs of an AS, for a source S in a remote AS.

Additionally to choosing an Inter-AS I-PMSI autodiscovery route advertized from the AS of the source to construct a C-multicast route, as described in section 11.1.3 [RFC6514] a downstream PE will choose a second Inter-AS I-PMSI autodiscovery route advertized from the AS of the source and use this route to construct and advertise a Standby C-multicast route (C-multicast route carrying the Standby extended community) as described in Section 4.1.

4.4.2. Inter-AS procedures for ASBRs

When an upstream ASBR receives a C-multicast route, and at least one of the RTs of the route matches one of the ASBR Import RT, the ASBR locates an Inter-AS I-PMSI A-D route whose RD and Source AS matches the RD and Source AS carried in the C-multicast route. If the match

is found, and C-multicast route carries the Standby PE BGP Community, then the ASBR performs as follows:

- if the route was received over iBGP ; the route is expected to have a LOCAL_PREF attribute set to zero and it should be re-advertized in eBGP with a MED attribute (MULTI_EXIT_DISC) set to the highest possible value (0xffff)

- if the route was received over eBGP ; the route is expected to have a MED attribute set of 0xffff and should be re-advertized in iBGP with a LOCAL_PREF attribute set to zero

Other ASBR procedures are applied without modification.

5. Hot leaf standby

The mechanisms defined in sections Section 4 and Section 3 can be used together as follows.

The principle is that, for a said VRF (or possibly only for a said C-S,C-G):

- downstream PEs advertise a Standby BGP C-multicast route (based on Section 4)

- upstream PEs use the "hot standby" optional behavior and thus will forward traffic for a said multicast state as soon as they have whether a (primary) BGP C-multicast route or a Standby BGP C-multicast route for that state (or both)

- downstream PEs accept traffic from the primary or standby tunnel, based on the status of the tunnel (based on Section 3)

Other combinations of the mechanisms proposed in Section 4) and Section 3 are for further study.

Note that the same level of protection would be achievable with a simple C-multicast Source Tree Join route advertized to both the primary and secondary upstream PEs (carrying as Route Target extended communities, the values of the VRF Route Import attribute of each VPN route from each upstream PEs). The advantage of using the Standby semantic for is that, supposing that downstream PEs always advertise a Standby C-multicast route to the secondary upstream PE, it allows to choose the protection level through a change of configuration on the secondary upstream PE, without requiring any reconfiguration of all the downstream PEs.

6. Duplicate packets

Multicast VPN specifications [RFC6513] impose that a PE only forwards to CEs the packets coming from the expected usptream PE (Section 9.1).

We highlight the reader’s attention to the fact that the respect of this part of multicast VPN specifications is especially important when two distinct upstream PEs are susceptible to forward the same traffic on P-tunnels at the same time in steady state. This will be the case when "hot root standby" mode is used (Section 4), and which can also be the case if procedures of Section 3 are used and (a) the rules determining the status of a tree are not the same on two distinct downstream PEs or (b) the rule determining the status of a tree depend on conditions local to a PE (e.g. the PE-P upstream link being up).

7. IANA Considerations

Allocation is expected from IANA for the BGP "Standby PE" community. (TBC)

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

8. Security Considerations

9. Acknowledgements

The authors want to thank Greg Reaume and Eric Rosen for their review and useful feedback.

10. Contributor Addresses

Below is a list of other contributing authors in alphabetical order:

Rahul Aggarwal
Arktan

Email: raggarwa_1@yahoo.com

Nehal Bhau
Alcatel-Lucent, Inc.
701 E Middlefield Rd
Mountain View, CA 94043
USA

Email: Nehal.Bhau@alcatel-lucent.com

Clayton Hassen
Bell Canada
2955 Virtual Way
Vancouver
CANADA

Email: Clayton.Hassen@bell.ca

Wim Henderickx
Alcatel-Lucent
Copernicuslaan 50
Antwerp 2018
Belgium

Email: wim.henderickx@alcatel-lucent.com

Pradeep Jain
Alcatel-Lucent, Inc.
701 E Middlefield Rd
Mountain View, CA 94043
USA

Email: pradeep.jain@alcatel-lucent.com

Jayant Kotalwar
Alcatel-Lucent, Inc.
701 E Middlefield Rd
Mountain View, CA 94043
USA

Email: Jayant.Kotalwar@alcatel-lucent.com

Praveen Muley
Alcatel-Lucent
701 East Middlefield Rd
Mountain View, CA 94043
U.S.A.

Email: praveen.muley@alcatel-lucent.com

Ray (Lei) Qiu
Juniper Networks
1194 North Mathilda Ave.
Sunnyvale, CA 94089
U.S.A.

Email: rqiu@juniper.net

Yakov Rekhter
Juniper Networks
1194 North Mathilda Ave.
Sunnyvale, CA 94089
U.S.A.

Email: yakov@juniper.net

Kanwar Singh
Alcatel-Lucent, Inc.
701 E Middlefield Rd
Mountain View, CA 94043
USA

Email: kanwar.singh@alcatel-lucent.com

11. References

11.1. Normative References

[I-D.ietf-bfd-multipoint]
Katz, D., Ward, D., and S. Pallagatti, "BFD for Multipoint Networks", draft-ietf-bfd-multipoint-06 (work in progress), January 2015.

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

[RFC4875] Aggarwal, R., Papadimitriou, D., and S. Yasukawa, "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, May 2007.

[RFC6513] Aggarwal, R., Bandi, S., Cai, Y., Morin, T., Rekhter, Y., Rosen, E., Wijnands, I., and S. Yasukawa, "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.

11.2. Informative References

[I-D.mofrr] Karan, A., Filsfils, C., Farinacci, D., Decraene, B., Leymann, N., and T. Telkamp, "Multicast only Fast Re-Route", draft-ietf-rtgwg-mofrr-08 (work in progress), February 2015.

[RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.

Authors' Addresses

Thomas Morin (editor)
Orange
2, avenue Pierre Marzin
Lannion 22307
France

Email: thomas.morin@orange-ftgroup.com

Robert Kebler (editor)
Juniper Networks
1194 North Mathilda Ave.
Sunnyvale, CA 94089
U.S.A.

Email: rkebler@juniper.net

BESS Workgroup
Internet Draft

J. Rabadan, Ed.
S. Sathappan
K. Nagaraj
Nokia

Intended status: Standards Track

Expires: January 4, 2018

July 3, 2017

Propagation of IPv6 Neighbor Advertisement Flags in EVPN
draft-snr-bess-evpn-na-flags-07

Abstract

The MAC/IP Advertisement route specified in [RFC7432] can optionally carry IPv4 and IPv6 addresses associated with a MAC address. Remote PEs can use this information to reply locally (act as proxy) to IPv4 ARP requests and IPv6 Neighbor Solicitation messages and reduce/suppress the flooding produced by the Address Resolution procedure. However, if the Neighbor information is learnt via EVPN, the PE would not know if a particular IPv6->MAC pair belongs to a host, a router or a host with an anycast address as this information is not carried in the MAC/IP route advertisements. This document proposes an OPTIONAL advertisement of the Flags defined in [RFC4861] along with the EVPN MAC/IP Advertisement routes, so that an EVPN PE implementing a proxy-ND function can reply to Neighbor Solicitations with the correct Flag information in Neighbor Advertisements.

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), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents

This Internet-Draft will expire on January 4, 2018.

Copyright Notice

Table of Contents

1. Introduction

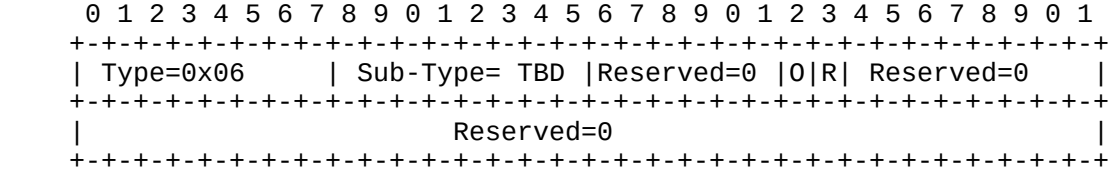
The MAC/IP Advertisement route specified in [RFC7432] can optionally carry IPv4 and IPv6 addresses associated with a MAC address. Remote PEs can use this information to reply locally (act as proxy) to IPv4 ARP requests and IPv6 Neighbor Solicitation messages and reduce/suppress the flooding produced by the Address Resolution procedure. However, if the Neighbor information is learned via EVPN, the PE would not know if a particular IPv6->MAC pair belongs to a host or a router as this information is not carried in the MAC/IP route advertisements.

This document proposes the OPTIONAL advertisement of the Flags defined in [RFC4861] along with the EVPN MAC/IP Advertisement routes, so that an EVPN PE implementing a proxy-ND function can issue Neighbor Advertisement messages conveying the correct Flag information.

The Flags are carried in the Neighbor Discovery (ND) EVPN Extended Community, as described in the following sections.

2. The EVPN Neighbor Discovery (ND) Extended Community

This document defines a new EVPN Extended Community with a Type field value of 0x06 and a Sub-Type TBD. It MAY be advertised along with EVPN MAC/IP Advertisement routes that carry an IPv6 address.



The following Flags are defined in the third octet of the Extended Community:

R - Router flag.

The low-order bit of the third octet is defined as the "Router flag". When set, the R-bit indicates that the IPv6->MAC pair advertised along with the MAC/IP Advertisement route belongs to a router. If the R-bit is zero, the IPv6-MAC pair belongs to a "host". The receiving PE implementing the proxy-ND function will use this information in Neighbor Advertisement messages for the associated IPv6 address.

O - Override flag

The second bit of the third octet is defined as the "Override flag". An egress PE will normally advertise IPv6->MAC pairs with the O-bit set, and only when IPv6 "anycast" is enabled in the EVI, the PE will send an IPv6->MAC pair with the O-bit = 0. The ingress PE will install the proxy-ND entry with the received O-bit and will use this information when replying to a Neighbor Solicitation for the IPv6 address.

3. Use of the EVPN ND Extended Community

An EVPN PE supporting a proxy-ND function and implementing the propagation of the Neighbor Advertisement Flags will follow this procedure:

a) Transmission of the EVPN ND Extended Community

A PE may learn the IPv6->MAC pair and its associated ND Flags in the management plane or snooping Neighbor Advertisement messages coming from the CE. Either way, the PE SHOULD send a MAC/IP Advertisement route including the learned IPv6->MAC pair and MAY send the ND Extended Community carrying its associated "R" and "O" Flags. This new Extended Community does not have any impact on the rest of the procedures described in [RFC7432], including the advertisement of the MAC Mobility Extended Community along with the MAC/IP Advertisement route.

b) Reception of the EVPN ND Extended Community

In addition to the procedures specified in [RFC7432] a PE receiving a MAC/IP Advertisement route containing an IPv6 address and the ND Extended Community SHOULD add the R and O Flags to the proxy-ND entry for the IPv6->MAC entry and use that information in Neighbor Advertisements when replying to a Solicitation for the IPv6 address.

A PE that implements the proxy-ND function SHOULD have an administrative option to define the default Flag to be used in case no EVPN ND Extended Community is received for a given IPv6->MAC entry.

4. Conventions used in this document

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].

In this document, these words will appear with that interpretation only when in ALL CAPS. Lower case uses of these words are not to be interpreted as carrying RFC-2119 significance.

In this document, the characters ">>" preceding an indented line(s) indicates a compliance requirement statement using the key words listed above. This convention aids reviewers in quickly identifying or finding the explicit compliance requirements of this RFC.

5. Security Considerations

The same security considerations described in [RFC7432] apply to this document.

6. IANA Considerations

This document requests the registration of a new EVPN Extended Community sub-type:

Sub-Type	Name	Reference
0x08	ND Extended Community	[this document]

7. References

7.1. Normative References

[RFC4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, DOI 10.17487/RFC4861, September 2007, <http://www.rfc-editor.org/info/rfc4861>.

[RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <http://www.rfc-editor.org/info/rfc7432>.

7.2. Informative References

8. Acknowledgments

Authors' Addresses

Jorge Rabadan (Editor)
Nokia
777 E. Middlefield Road
Mountain View, CA 94043 USA
Email: jorge.rabadan@nokia.com

Senthil Sathappan
Nokia
701 E. Middlefield Road
Mountain View, CA 94043 USA
Email: senthil.sathappan@nokia.com

Kiran Nagaraj
Nokia
701 E. Middlefield Road
Mountain View, CA 94043 USA
Email: kiran.nagaraj@nokia.com

BESS Workgroup
Internet Draft

Intended status: Informational

J. Rabadan, Ed.
S. Sathappan
K. Nagaraj
W. Henderickx
G. Hankins
Alcatel-Lucent

T. King
D. Melzer
DE-CIX

E. Nordmark
Arista Networks

Expires: April 8, 2016

October 6, 2015

Operational Aspects of Proxy-ARP/ND in EVPN Networks
draft-snr-bess-evpn-proxy-arp-nd-02

Abstract

The MAC/IP Advertisement route specified in [RFC7432] can optionally carry IPv4 and IPv6 addresses associated with a MAC address. Remote PEs can use this information to reply locally (act as proxy) to IPv4 ARP requests and IPv6 Neighbor Solicitation messages (or 'unicast-forward' them to the owner of the MAC) and reduce/suppress the flooding produced by the Address Resolution procedure. This EVPN capability is extremely useful in Internet Exchange Points (IXPs) and Data Centers (DCs) with large broadcast domains, where the amount of ARP/ND flooded traffic causes issues on routers and CEs, as explained in [RFC6820]. This document describes how the [RFC7432] EVPN proxy-ARP/ND function may be implemented to help IXPs and other operators deal with the issues derived from Address Resolution in large broadcast domains.

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), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Table of Contents

1. Terminology

BUM: Broadcast, Unknown unicast and Multicast layer-2 traffic.

ARP: Address Resolution Protocol.

GARP: Gratuitous ARP message.

ND: Neighbor Discovery Protocol.

NS: Neighbor Solicitation message.

NA: Neighbor Advertisement.

IXP: Internet eXchange Point.

IXP-LAN: it refers to the IXP's large Broadcast Domain to where Internet routers are connected.

DC: Data Center.

IP->MAC: it refers to an IP address associated to a MAC address. The entries may be of three different types: dynamic, static or EVPN-learned.

SN-multicast address: Refers to the Solicited-Node IPv6 multicast address used by NS messages.

NUD: Neighbor Unreachability Detection, as per [RFC4861].

DAD: Duplicate Address Detection, as per [RFC4861].

SLLA: Source Link Layer Address, as per [RFC4861].

TLLA: Target Link Layer Address, as per [RFC4861].

R-bit: Router Flag in NA messages, as per [RFC4861].

O-bit: Override Flag in NA messages, as per [RFC4861].

S-bit: Solicited Flag in NA messages, as per [RFC4861].

RT2: EVPN Route type 2 or MAC/IP Advertisement route, as per [RFC7432].

MAC or IP DA: MAC or IP Destination Address.

MAC or IP SA: MAC or IP Source Address.

AS-MAC: Anti-spoofing MAC.

2. Introduction

As specified in [RFC7432] the IP Address field in the MAC/IP Advertisement route may optionally carry one of the IP addresses associated with the MAC address. A PE may learn local IP->MAC pairs and advertise them in EVPN MAC/IP routes. The remote PEs may add those IP->MAC pairs to their Proxy-ARP/ND tables and reply to local ARP requests or Neighbor Solicitations (or 'unicast-forward' those packets to the owner MAC), reducing and even suppressing in some cases the flooding in the EVPN network.

EVPN and its associated Proxy-ARP/ND function are extremely useful in Data Centers (DCs) or Internet Exchange Points (IXPs) with large broadcast domains, where the amount of ARP/ND flooded traffic causes issues on routers and CEs. [RFC6820] describes the Address Resolution problems in Large Data Center networks.

This document describes how the [RFC7432] proxy-ARP/ND function may be implemented to help IXPs, DCs and other operators deal with the issues derived from Address Resolution in large broadcast domains.

2.1. The DC Use-Case

As described in [RFC6820] the IPv4 and IPv6 Address Resolution can create a lot of issues in large DCs. The amount of flooding that Address Resolution creates, as well as other associated issues can be mitigated with the use of EVPN and its proxy-ARP/ND function.

2.2. The IXP Use-Case

The implementation described in this document is especially useful in IXP networks.

A typical IXP provides access to a large layer-2 peering network, where (hundreds of) Internet routers are connected. Because of the requirement to connect all routers to a single layer-2 network the peering networks use IPv4 layer-3 addresses in length ranges from /21 to /24, which can create very large broadcast domains. This peering network is transparent to the Customer Edge (CE) devices and therefore floods any ARP request or NS messages to all the CEs in the network. Unsolicited GARP and NA messages are flooded to all the CEs too.

In these IXP networks, most of the CEs are typically peering routers and roughly all the BUM traffic is originated by the ARP and ND address resolution procedures. This ARP/ND BUM traffic causes significant data volumes that reach every single router in the peering network. Since the ARP/ND messages are processed in software processors and they take high priority in the routers, heavy loads of ARP/ND traffic can cause some routers to run out of resources. CEs disappearing from the network may cause Address Resolution explosions that can make a router with limited processing power fail to keep BGP sessions running.

The issue may be better in IPv6 routers, since ND uses SN-multicast address in NS messages, however ARP uses broadcast and has to be processed by all the routers in the network. Some routers may also be configured to broadcast periodic GARPs [RFC5227]. The amount of ARP/ND flooded traffic grows exponentially with the number of IXP participants, therefore the issue can only go worse as new CEs are added.

In order to deal with this issue, IXPs have developed certain solutions over the past years. One example is the ARP-Sponge daemon [ARP-Sponge]. While these solutions may mitigate the issues of Address Resolution in large broadcasts domains, EVPN provides new more efficient possibilities to IXPs. EVPN and its proxy-ARP/ND function may help solve the issue in a distributed and scalable way, fully integrated with the PE network.

3. Solution Requirements

The distributed EVPN proxy-ARP/ND function described in this document SHOULD meet the following requirements:

- The solution SHOULD support the learning of the CE IP->MAC entries on the EVPN PEs via the management, control or data planes. An implementation SHOULD allow to intentionally enable or disable

those possible learning mechanisms.

- The solution MAY suppress completely the flooding of the ARP/ND messages in the EVPN network, assuming that all the CE IP->MAC addresses local to the PEs are known or provisioned on the PEs from a management system. Note that in this case, the unknown unicast traffic can also be suppressed, since all the expected unicast traffic will be destined to known MAC addresses in the PE MAC-VRFs.

- The solution MAY reduce significantly the flooding of the ARP/ND messages in the EVPN network, assuming that some or all the CE IP->MAC addresses are learned on the data plane by snooping ARP/ND messages issued by the CEs.

- The solution MAY provide a way to refresh periodically the CE IP->MAC entries learned through the data plane, so that the IP->MAC entries are not withdrawn by EVPN when they age out unless the CE is not active anymore. This option helps reducing the EVPN control plane overhead in a network with active CEs that do not send packets frequently.

- The solution SHOULD provide a mechanism to detect duplicate IP addresses. In case of duplication, the detecting PE should not reply to requests for the duplicate IP. Instead, the PE should alert the operator and may optionally prevent any other CE from sending traffic to the duplicate IP.

- The solution MUST NOT change any existing behavior in the CEs connected to the EVPN PEs.

4. Solution Description

Figure 1 illustrates an example EVPN network where the Proxy-ARP/ND function is enabled.

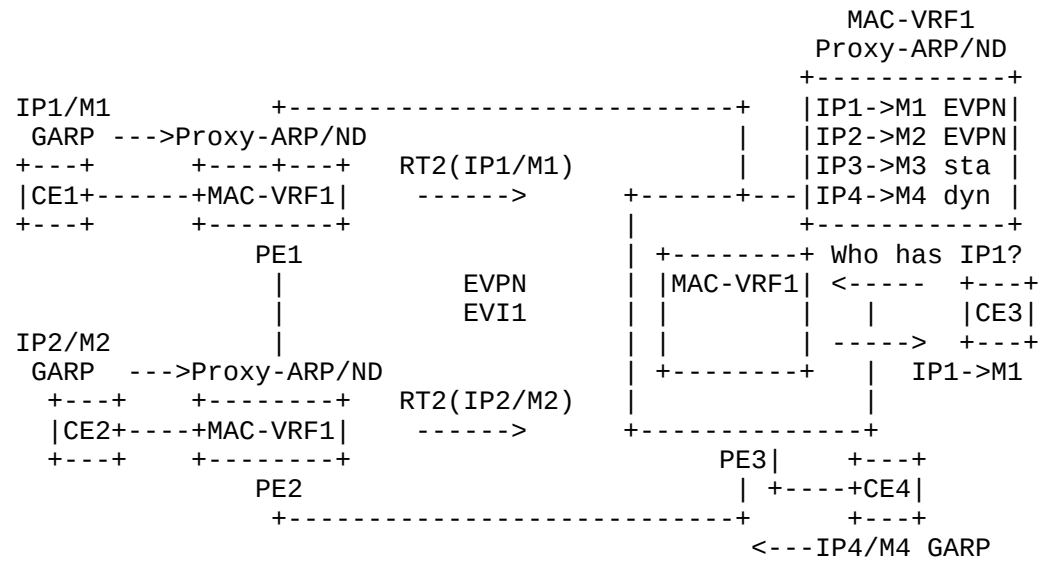


Figure 1 Proxy-ARP/ND network example

When the Proxy-ARP/ND function is enabled in the MAC-VRFs of the EVPN PEs, each PE creates a Proxy table specific to that MAC-VRF that can contain three types of Proxy-ARP/ND entries:

a) Dynamic entries: learned by snooping CE's ARP and ND messages. For instance, IP4->M4 in Figure 1.

b) Static entries: provisioned on the PE by the management system. For instance, IP3->M3 in Figure 1.

c) EVPN-learned entries: learned from the IP/MAC information encoded in the received RT2's coming from remote PEs. For instance, IP1->M1 and IP2->M2 in Figure 1.

As a high level example, the operation of the EVPN Proxy-ARP/ND function in the network of Figure 1 is described below. In this example we assume IP1, IP2 and IP3 are IPv4 addresses:

1. Proxy-ARP/ND is enabled in MAC-VRF1 of PE1, PE2 and PE3.

2. The PEs start adding dynamic, static and EVPN-learned entries to their Proxy tables:

   a. PE3 adds IP1->M1 and IP2->M2 based on the EVPN routes received from PE1 and PE2. Those entries were previously learned as dynamic entries in PE1 and PE2 respectively, and advertised in

BGP EVPN.
b. PE3 adds IP4->M4 as dynamic. This entry is learned by snooping the corresponding ARP messages sent by CE4.
c. An operator also provisions the static entry IP3->M3.

3. When CE3 sends an ARP Request asking for IP1, PE3 will:

a. Intercept the ARP Request and perform a Proxy-ARP lookup for IP1.
b. If the lookup is successful (as in Figure 1), PE3 will send an ARP Reply with IP1->M1. The ARP Request will not be flooded to the EVPN network or any other local CEs.
c. If the lookup is not successful, PE3 will flood the ARP Request in the EVPN network and the other local CEs.

As PE3 learns more and more host entries in the Proxy-ARP/ND table, the flooding of ARP Request messages is reduced and in some cases it can even be suppressed. In a network where most of the participant CEs are not moving between PEs and they advertise their presence with GARPs or unsolicited NA messages, the ARP/ND flooding as well as the unknown unicast flooding can practically be suppressed. In an EVPN-based IXP network, where all the entries are Static, the ARP/ND flooding is in fact totally suppressed.

The Proxy-ARP/ND function can be structured in six sub-functions or procedures:

1. Learning sub-function
2. Reply sub-function
3. Unicast-forward sub-function
4. Maintenance sub-function
5. Flooding reduction/suppression sub-function
6. Duplicate IP detection sub-function

A Proxy-ARP/ND implementation MAY support all those sub-functions or only a subset of them. The following sections describe each individual sub-function.

4.1. Learning Sub-Function

A Proxy-ARP/ND implementation SHOULD support static, dynamic and EVPN-learned entries.

Static entries are provisioned from the management plane. The provisioned static IP->MAC entry SHOULD be advertised in EVPN with a MAC Mobility extended community where the static flag is set to 1, as per [RFC7432]. A static entry MAY associate and IP to a list of

potential MACs, i.e. IP1->(MAC1,MAC2..MACN). When there is more than one MAC in the list of allowed MACs, the PE will not advertise any IP->MAC in EVPN until a local ARP/NA message or any other frame is received from the CE. Upon receiving traffic from the CE, the PE will check that the source MAC is included in the list of allowed MACs. Only in that case, the PE will activate the IP->MAC and advertise it in EVPN.

EVPN-learned entries MUST be learned from received valid EVPN MAC/IP Advertisement routes containing a MAC and IP address.

Dynamic entries are learned in different ways depending on whether the entry contains an IPv4 or IPv6 address:

a) Proxy-ARP dynamic entries:

They SHOULD be learned by snooping any ARP packet (Ethertype 0x0806) received from the CEs attached to the MAC-VRF. The Learning function will add the Sender MAC and Sender IP of the snooped ARP packet to the Proxy-ARP table. Note that MAC and IPs with value 0 SHOULD NOT be learned.

b) Proxy-ND dynamic entries:

They SHOULD be learned out of the Target Address and TLLA information in NA messages (Ethertype 0x86DD, ICMPv6 type 136) received from the CEs attached to the MAC-VRF. A Proxy-ND implementation SHOULD NOT learn IP->MAC entries from NS messages, since they don’t contain the R-bit Flag required by the Proxy-ND reply function. See section 4.1.1 for more information about the R-bit flag.

Note that if the O-bit is zero in the received NA message, the IP->MAC SHOULD only be learned in case IPv6 ’anycast’ is enabled in the EVI.

The following procedure associated to the Learning sub-function is recommended:

o When a new Proxy-ARP/ND EVPN or static active entry is learned (or provisioned), the PE SHOULD send an unsolicited GARP or NA message to the access CEs. The PE SHOULD send an unsolicited GARP/NA message for dynamic entries only if the ARP/NA message creating the entry was NOT flooded before. This unsolicited GARP/NA message makes sure the CE ARP/ND caches are updated even if the ARP/NS/NA messages from remote CEs are not flooded in the EVPN network.

Note that if a Static entry is provisioned with the same IP as an

existing EVPN-learned or Dynamic entry, the Static entry takes precedence.

4.1.1. Proxy-ND and the NA Flags

[RFC4861] describes the use of the R-bit flag in IPv6 Address Resolution:

- Nodes capable of routing IPv6 packets must reply to NS messages with NA messages where the R-bit flag is set (R-bit=1).
- Hosts that are not able to route IPv6 packets must indicate that inability by replying with NA messages that contain R-bit=0.

The use of the R-bit flag in NA messages has an impact on how hosts select their default gateways when sending packets off-link:

- Hosts build a Default Router List based on the received RAs and NAs with R-bit=1. Each cache entry has an IsRouter flag, which must be set based on the R-bit flag in the received NAs. A host can choose one or more Default Routers when sending packets off-link.
- In those cases where the IsRouter flag changes from TRUE to FALSE as a result of a NA update, the node MUST remove that router from the Default Router List and update the Destination Cache entries for all destinations using that neighbor as a router, as specified in [RFC4861] section 7.3.3. This is needed to detect when a node that is used as a router stops forwarding packets due to being configured as a host.

The R-bit and O-bit will be learned in the following ways:

- Static entries SHOULD have the R-bit information added by the management interface. The O-bit information MAY also be added by the management interface.
- Dynamic entries SHOULD learn the R-bit and MAY learn the O-bit from the snooped NA messages used to learn the IP->MAC itself.
- EVPN-learned entries SHOULD learn the R-bit and MAY learn the O-bit from the ND Extended Community received from EVPN along with the RT2 used to learn the IP->MAC itself. Please refer to [EVPN-NA-FLAGS]. If no ND extended community is received, the PE will add the default R-bit/O-bit to the entry. The default R-bit SHOULD be an administrative choice. The default O-bit SHOULD be 1.

Note that the O-bit SHOULD only be learned if ’anycast’ is enabled in the EVI. If so, Duplicate IP Detection must be disabled so that the

PE is able to learn the same IP mapped to different MACs in the same Proxy-ND table. If 'anycast' is disabled, NA messages with O-bit = 0 will not create a proxy-ND entry, hence no EVPN advertisement with ND extended community will be generated.

4.2. Reply Sub-Function

This sub-function will reply to Address Resolution requests/solicitations upon successful lookup in the Proxy-ARP/ND table for a given IP address. The following considerations should be taken into account:

a) When replying to ARP Request or NS messages, the PE SHOULD use the Proxy-ARP/ND entry MAC address as MAC SA. This is recommended so that the resolved MAC can be learned in the MAC FIB of potential Layer-2 switches seating between the PE and the CE requesting the Address Resolution.

b) A PE SHOULD NOT reply to a request/solicitation received on the same attachment circuit over which the IP->MAC is learned. In this case the requester and the requested IP are assumed to be connected to the same layer-2 switch/access network linked to the PE's attachment circuit, and therefore the requested IP owner will receive the request directly.

c) A PE SHOULD reply to broadcast/multicast Address Resolution messages, that is, ARP-Request, NS messages as well as DAD NS messages. A PE SHOULD NOT reply to unicast Address Resolution requests (for instance, NUD NS messages).

d) A PE SHOULD include the R-bit learned for the IP->MAC entry in the NA messages (see section 4.1.1). The S-bit will be set/unset as per [RFC4861]. The O-bit will be included if IPv6 'anycast' is enabled in the EVI and it is learned for the IP->MAC entry. If 'anycast' is enabled and there are more than one MAC for a given IP, the PE will reply to NS messages with as many NA responses as 'anycast' entries are in the proxy-ND table.

e) A PE SHOULD NOT reply to ARP probes received from the CEs. An ARP probe is an ARP request constructed with an all-zero sender IP address that may be used by hosts for IPv4 Address Conflict Detection [RFC5227].

f) A PE SHOULD only reply to ARP-Request and NS messages with the format specified in [RFC0826] and [RFC4861] respectively. Received ARP-Requests and NS messages with unknown options SHOULD be either forwarded (as unicast packets) to the owner of the requested IP (assuming the MAC is known in the proxy-ARP/ND table and MAC-VRF)

or discarded. An administrative option to control this behavior ('unicast-forward' or 'discard') SHOULD be supported. The 'unicast-forward' option is described in section 4.3.

4.3. Unicast-forward Sub-Function

As discussed in section 4.2. in some cases the operator may want to 'unicast-forward' certain ARP-Request and NS messages as opposed to reply to them. The operator SHOULD be able to activate this option with one of the following parameters:

a) unicast-forward always
b) unicast-forward unknown-options

If 'unicast-forward always' is enabled, the PE will perform a proxy-ARP/ND table lookup and in case of a hit, the PE will forward the packet to the owner of the MAC found in the proxy-ARP/ND table. This is irrespective of the options carried in the ARP/ND packet. This option provides total transparency in the EVI and yet reduces the amount of flooding significantly.

If 'unicast-forward unknown-options' is enabled, upon a successful proxy-ARP/ND lookup, the PE will perform a 'unicast-forward' action only if the ARP-Request or NS messages carry unknown options, as explained in section 4.2. As an example, this would allow to enable proxy-ND and Secure ND [RFC3971] in the same EVI. The 'unicast-forward unknown-options' configuration allows the support of new applications using ARP/ND in the EVI while still reducing the flooding at the same time.

4.4. Maintenance Sub-Function

The Proxy-ARP/ND tables SHOULD follow a number of maintenance procedures so that the dynamic IP->MAC entries are kept if the owner is active and flushed if the owner is no longer in the network. The following procedures are recommended:

a) Age-time

A dynamic Proxy-ARP/ND entry SHOULD be flushed out of the table if the IP->MAC has not been refreshed within a given age-time. The entry is refreshed if an ARP or NA message is received for the same IP->MAC entry. The age-time is an administrative option and its value should be carefully chosen depending on the specific use-case: in IXP networks (where the CE routers are fairly static)

the age-time may normally be longer than in DC networks (where mobility is required).

b) Send-refresh option

The PE MAY send periodic refresh messages (ARP/ND "probes") to the owners of the dynamic Proxy-ARP/ND entries, so that the entries can be refreshed before they age out. The owner of the IP->MAC entry would reply to the ARP/ND probe and the corresponding entry age-time reset. The periodic send-refresh timer is an administrative option and is recommended to be a third of the age-time or a half of the age-time in scaled networks.

An ARP refresh issued by the PE will be an ARP-Request message with the Sender’s IP = 0 sent from the PE’s MAC SA. If the PE has an IP address in the subnet, for instance on an IRB interface, then it MAY use it as a source for the ARP request (instead of Sender’s IP = 0). An ND refresh will be a NS message issued from the PE’s MAC SA and a Link Local Address associated to the PE’s MAC.

The refresh request messages should be sent only for dynamic entries and not for static or EVPN-learned entries. Even though the refresh request messages are broadcast or multicast, the PE SHOULD only send the message to the attachment circuit associated to the MAC in the IP->MAC entry.

The age-time and send-refresh options are used in EVPN networks to avoid unnecessary EVPN RT2 withdrawals: if refresh messages are sent before the corresponding MAC-VRF FIB and Proxy-ARP/ND age-time for a given entry expires, inactive but existing hosts will reply, refreshing the entry and therefore avoiding unnecessary MAC and MAC-IP withdrawals in EVPN. Both entries (MAC in the MAC-VRF and IP->MAC in Proxy-ARP/ND) are reset when the owner replies to the ARP/ND probe. If there is no response to the ARP/ND probe, the MAC and IP->MAC entries will be legitimately flushed and the RT2s withdrawn.

4.5. Flooding (to Remote PEs) Reduction/Suppression

The Proxy-ARP/ND function implicitly helps reducing the flooding of ARP Request and NS messages to remote PEs in an EVPN network. However, in certain use-cases, the flooding of ARP/NS/NA messages (and even the unknown unicast flooding) to remote PEs can be suppressed completely in an EVPN network.

For instance, in an IXP network, since all the participant CEs are well known and will not move to a different PE, the IP->MAC entries

may be all provisioned by a management system. Assuming the entries for the CEs are all provisioned on the local PE, a given Proxy-ARP/ND table will only contain static and EVPN-learned entries. In this case, the operator may choose to suppress the flooding of ARP/NS/NA to remote PEs completely.

The flooding may also be suppressed completely in IXP networks with dynamic Proxy-ARP/ND entries assuming that all the CEs are directly connected to the PEs and they all advertise their presence with a GARP/unsolicited-NA when they connect to the network.

In networks where fast mobility is expected (DC use-case), it is not recommended to suppress the flooding of unknown ARP-Requests/NS or GARPs/unsolicited-NAs. Unknown ARP-Requests/NS refer to those ARP-Request/NS messages for which the Proxy-ARP/ND lookups for the requested IPs do not succeed.

In order to give the operator the choice to suppress/allow the flooding to remote PEs, a PE MAY support administrative options to individually suppress/allow the flooding of:

o Unknown ARP-Request and NS messages.
o GARP and unsolicited-NA messages.

The operator will use these options based on the expected behavior in the CEs.

4.6. Duplicate IP Detection

The Proxy-ARP/ND function SHOULD support duplicate IP detection so that ARP/ND-spoofing attacks or duplicate IPs due to human errors can be detected.

ARP/ND spoofing is a technique whereby an attacker sends "fake" ARP/ND messages onto a broadcast domain. Generally the aim is to associate the attacker’s MAC address with the IP address of another host causing any traffic meant for that IP address to be sent to the attacker instead.

The distributed nature of EVPN and proxy-ARP/ND allows the easy detection of duplicated IPs in the network, in a similar way to the MAC duplication function supported by [RFC7432] for MAC addresses.

Duplicate IP detection monitors "IP-moves" in the Proxy-ARP/ND table in the following way:

o When an existing active IP1->MAC1 entry is modified, a PE starts an

M-second timer (default value of M=180), and if it detects N IP moves before the timer expires (default value of N=5), it concludes that a duplicate IP situation has occurred. An IP move is considered when, for instance, IP1->MAC1 is replaced by IP1->MAC2 in the Proxy-ARP/ND table.

- o In order to detect the duplicate IP faster, the PE MAY send a CONFIRM message to the former owner of the IP. A CONFIRM message is a unicast ARP-Request/NS message sent by the PE to the MAC addresses that previously owned the IP, when the MAC changes in the Proxy-ARP/ND table. The CONFIRM message uses a sender's IP 0.0.0.0 in case of ARP (if the PE has an IP address in the subnet then it MAY use it) and an IPv6 Link Local Address in case of NS. If the PE does not receive an answer within a given timer, the new entry will be confirmed and activated. In case of spoofing, for instance, if IP1->MAC1 moves to IP1->MAC2, the PE may send a unicast ARP-Request/NS message for IP1 with MAC DA= MAC1 and MAC SA= PE's MAC. This will force the legitimate owner respond if the move to MAC2 was spoofed, and make the PE issue another CONFIRM message, this time to MAC DA= MAC2. If both, legitimate owner and spoofer keep replying to the CONFIRM message, the PE will detect the duplicate IP within the M timer:

  - If the IP1->MAC1 pair was previously owned by the spoofer and the new IP1->MAC2 was from a valid CE, then the issued CONFIRM message would trigger a response from the spoofer.

  - If it were the other way around, that is, IP1->MAC1 was previously owned by a valid CE, the CONFIRM message would trigger a response from the CE.

    Either way, if this process continues, then duplicate detection will kick in.

- o Upon detecting a duplicate IP situation:

  a) The entry in duplicate detected state cannot be updated with new dynamic or EVPN-learned entries for the same IP. The operator MAY override the entry though with a static IP->MAC.

  b) The PE SHOULD alert the operator and stop responding ARP/NS for the duplicate IP until a corrective action is taken.

  c) Optionally the PE MAY associate an "anti-spoofing-mac" (AS-MAC) to the duplicate IP. The PE will send a GARP/unsolicited-NA message with IP1->AS-MAC to the local CEs as well as an RT2 (with IP1->AS-MAC) to the remote PEs. This will force all the CEs in the EVI to use the AS-MAC as MAC DA for IP1, and prevent

the spoofer from attracting any traffic for IP1. Since the AS-MAC is a managed MAC address known by all the PEs in the EVI, all the PEs MAY apply filters to drop and/or log any frame with MAC DA= AS-MAC. The advertisement of the AS-MAC as a "black-hole MAC" that can be used directly in the MAC-VRF to drop frames is for further study.

o The duplicate IP situation will be cleared when a corrective action is taken by the operator, or alternatively after a HOLD-DOWN timer (default value of 540 seconds).

The values of M, N and HOLD-DOWN timer SHOULD be a configurable administrative option to allow for the required flexibility in different scenarios.

For Proxy-ND, Duplicate IP Detection SHOULD only monitor IP moves for IP->MACs learned from NA messages with O-bit=1. NA messages with O-bit=0 would not override the ND cache entries for an existing IP. Duplicate IP Detection for IPv6 SHOULD be disabled when IPv6 'anycast' is activated in a given EVI.

5. Solution Benefits

The solution described in this document provides the following benefits:

a) The solution may suppress completely the flooding of the ARP/ND and unknown-unicast messages in the EVPN network, in cases where all the CE IP->MAC addresses local to the PEs are known and provisioned on the PEs from a management system.

b) The solution reduces significantly the flooding of the ARP/ND messages in the EVPN network, in cases where some or all the CE IP->MAC addresses are learned on the data plane by snooping ARP/ND messages issued by the CEs.

c) The solution reduces the control plane overhead and unnecessary BGP MAC/IP Advertisements and Withdrawals in a network with active CEs that do not send packets frequently.

d) The solution provides a mechanism to detect duplicate IP addresses and avoid ARP/ND-spoof attacks or the effects of duplicate addresses due to human errors.

6. Deployment Scenarios

Four deployment scenarios with different levels of ARP/ND control are

available to operators using this solution, depending on their requirements to manage ARP/ND: all dynamic learning, all dynamic learning with proxy-ARP/ND, hybrid dynamic learning and static provisioning with proxy-ARP/ND, and all static provisioning with proxy-ARP/ND.

6.1. All Dynamic Learning

In this scenario for minimum security and mitigation, EVPN is deployed in the peering network with the proxy-ARP/ND function shutdown. PEs do not intercept ARP/ND requests and flood all requests, as in a conventional layer-2 network. While no ARP/ND mitigation is used in this scenario, the IXP can still take advantage of EVPN features such as control plane learning and all-active multihoming in the peering network. Existing mitigation solutions, such as the ARP-Sponge daemon [ARP-Sponge] MAY also be used in this scenario.

Although this option does not require any of the procedures described in this document, it is added as baseline/default option for completeness.

6.2. Dynamic Learning with Proxy-ARP/ND

This scenario minimizes flooding while enabling dynamic learning of IP->MAC entries. The Proxy-ARP/ND function is enabled in the MAC-VRFs of the EVPN PEs, so that the PEs intercept and respond to CE requests.

The solution MAY further reduce the flooding of the ARP/ND messages in the EVPN network by snooping ARP/ND messages issued by the CEs.

PEs will flood requests if the entry is not in their Proxy table. Any unknown source MAC->IP entries will be learnt and advertised in EVPN, and traffic to unknown entries is discarded at the ingress PE.

6.3. Hybrid Dynamic Learning and Static Provisioning with Proxy-ARP/ND

Some IXPs want to protect particular hosts on the peering network while allowing dynamic learning of peering router addresses. For example, an IXP may want to configure static MAC->IP entries for management and infrastructure hosts that provide critical services. In this scenario, static entries are provisioned from the management plane for protected MAC->IP addresses, and dynamic learning with Proxy-ARP/ND is enabled as described in section 6.2 on the peering network.

6.4 All Static Provisioning with Proxy-ARP/ND

For a solution that maximizes security and eliminates flooding and unknown unicast in the peering network, all MAC-IP entries are provisioned from the management plane. The Proxy-ARP/ND function is enabled in the MAC-VRFs of the EVPN PEs, so that the PEs intercept and respond to CE requests. Dynamic learning and ARP/ND snooping is disabled so that traffic to unknown entries is discarded at the ingress PE. This scenario provides and IXP the most control over MAC->IP entries and allows an IXP to manage all entries from a management system.

6.5 Deployment Scenarios in IXPs

Nowadays, almost all IXPs installed some security rules in order to protect the IXP-LAN. These rules are often called port security. Port security summarizes different operational steps that limit the access to the IXP-LAN, to the customer router and controls the kind of traffic that the routers are allowed to be exchange (e.g., Ethernet, IPv4, IPv6). Due to this, the deployment scenario as described in 6.4 "All Static Provisioning with Proxy-ARP/ND" is the predominant scenario for IXPs.

In addition to the "All Static Provisioning" behavior, in IXP networks it is recommended to configure the Reply Sub-Function to ’discard’ ARP-Requests/NS messages with unrecognized options.

At IXPs, customers usually follow a certain operational life-cycle. For each step of the operational life-cycle specific operational procedures are executed.

The following describes the operational procedures that are needed to guarantee port security throughout the life-cycle of a customer with focus on EVPN features:

1. A new customer is connected the first time to the IXP:

   Before the connection between the customer router and the IXP-LAN is activated, the MAC of the router is white-listed on the IXP’s switch port. All other MAC addresses are blocked. Pre-defined IPv4 and IPv6 addresses of the IXP’s peering network space are configured at the customer router. The IP->MAC static entries (IPv4 and IPv6) are configured in the management system of the IXP for the customer’s port in order to support Proxy-ARP/ND.

   In case a customer uses multiple ports aggregated to a single logical port (LAG) some vendors randomly select the MAC address of the LAG from the different MAC addresses assigned to the ports. In this case the static entry will be used associated to a list of allowed MACs.

2. Replacement of customer router:

   If a customer router is about to be replaced, the new MAC address(es) must be installed in the management system besides the MAC address(es) of the currently connected router. This allows the customer to replace the router without any active involvement of the IXP operator. For this, static entries are also used. After the replacement takes place, the MAC address(es) of the replaced router can be removed.

3. Decommissioning a customer router

   If a customer router is decommissioned, the router is disconnected from the IXP PE. Right after that, the MAC address(es) of the router and IP->MAC bindings can be removed from the management system.

6.6 Deployment Scenarios in DCs

DCs normally have different requirements than IXPs in terms of Proxy-ARP/ND. Some differences are listed below:

a) The required mobility in virtualized DCs makes the "Dynamic Learning" or "Hybrid Dynamic and Static Provisioning" models more appropriate than the "All Static Provisioning" model.

b) IPv6 ’anycast’ may be required in DCs, while it is not a requirement in IXP networks. Therefore if the DC needs IPv6 ’anycast’ it will be explicitly enabled in the proxy-ND function, hence the proxy-ND sub-functions modified accordingly. For instance, if IPv6 ’anycast’ is enabled in the proxy-ND function, Duplicate IP Detection must be disabled.

c) DCs may require special options on ARP/ND as opposed to the Address Resolution function, which is the only one typically required in IXPs. Based on that, the Reply Sub-function may be modified to forward or discard unknown options.

7. Conventions Used in this Document

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].

In this document, these words will appear with that interpretation

only when in ALL CAPS. Lower case uses of these words are not to be interpreted as carrying RFC-2119 significance.

In this document, the characters ">>" preceding an indented line(s) indicates a compliance requirement statement using the key words listed above. This convention aids reviewers in quickly identifying or finding the explicit compliance requirements of this RFC.

8. Security Considerations

When EVPN and its associated Proxy-ARP/ND function are used in IXP networks, they only provide ARP/ND security and mitigation. IXPs MUST still employ security mechanisms that protect the peering network and SHOULD follow established BCPs such as the ones described in [Euro-IX BCP].

For example, IXPs should disable all unneeded control protocols, and block unwanted protocols from CEs so that only IPv4, ARP and IPv6 Ethertypes are permitted on the peering network. In addition, port security features and ACLs can provide an additional level of security.

9. IANA Considerations

No IANA considerations.

10. References

10.1. Normative References

[RFC7432]Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <http://www.rfc-editor.org/info/rfc7432>.

[RFC4861]Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, DOI 10.17487/RFC4861, September 2007, <http://www.rfc-editor.org/info/rfc4861>.

[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, DOI 10.17487/RFC0826, November 1982, <http://www.rfc-editor.org/info/rfc826>.

[RFC6820]Narten, T., Karir, M., and I. Foo, "Address Resolution Problems in Large Data Center Networks", RFC 6820, DOI 10.17487/RFC6820, January 2013, <http://www.rfc-editor.org/info/rfc6820>.

[RFC7342]Dunbar, L., Kumari, W., and I. Gashinsky, "Practices for Scaling ARP and Neighbor Discovery (ND) in Large Data Centers", RFC 7342, DOI 10.17487/RFC7342, August 2014, <http://www.rfc-editor.org/info/rfc7342>.

[RFC3971]Arkko, J., Ed., Kempf, J., Zill, B., and P. Nikander, "SEcure Neighbor Discovery (SEND)", RFC 3971, DOI 10.17487/RFC3971, March 2005, <http://www.rfc-editor.org/info/rfc3971>.

[RFC7432]Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <http://www.rfc-editor.org/info/rfc7432>.

[RFC5227]Cheshire, S., "IPv4 Address Conflict Detection", RFC 5227, DOI 10.17487/RFC5227, July 2008, <http://www.rfc-editor.org/info/rfc5227>.

10.2. Informative References

[ARP-Sponge] Wessel M. and Sijm N., Universiteit van Amsterdam, "Effects of IPv4 and IPv6 address resolution on AMS-IX and the ARP Sponge", July 2009.

[EVPN-ND-FLAGS] Sathappan S., Nagaraj K. and Rabadan J., "Propagation of IPv6 Neighbor Advertisement Flags in EVPN", draft-snr-bess-evpn-na-flags-02, Work in Progress, July 2015.

[Euro-IX BCP] https://www.euro-ix.net/pages/28/1/bcp_ixp.html

11. Acknowledgments

The authors want to thank Ranganathan Boovaraghavan, Sriram Venkateswaran, Manish Krishnan, Seshagiri Venugopal, Tony Przygienda and Robert Raszuk for their review and contributions. Thank you to Oliver Knapp as well, for his detailed review.

Authors' Addresses

Jorge Rabadan (Editor)
Alcatel-Lucent
777 E. Middlefield Road
Mountain View, CA 94043 USA
Email: jorge.rabadan@alcatel-lucent.com

Senthil Sathappan
Alcatel-Lucent
Email: senthil.sathappan@alcatel-lucent.com

Kiran Nagaraj
Alcatel-Lucent
Email: kiran.nagaraj@alcatel-lucent.com

Wim Henderickx
Alcatel-Lucent
Email: wim.henderickx@alcatel-lucent.com

Greg Hankins
Alcatel-Lucent
Email: greg.hankins@alcatel-lucent.com

Thomas King
DE-CIX Management GmbH
Lichtstrasse 43i, Cologne 50825, Germany
Email: thomas.king@de-cix.net

Daniel Melzer
DE-CIX Management GmbH
Lichtstrasse 43i, Cologne 50825, Germany
Email: daniel.melzer@de-cix.net

Erik Nordmark
Arista Networks
Email: nordmark@arista.com

BESS Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 7, 2015

Kishore Tiruveedhula, Ed.
Tapraj Singh
Juniper Networks
Ali Sajassi
Deepak Kumar
Cisco Systems
Luay Jalil
Verizon

March 6, 2015

YANG Data Model for PBB EVPN protocol
draft-tsingh-bess-pbb-evpn-yang-cfg-00

Abstract

This document defines a YANG data model that can be used to configure and manage PBB EVPN.

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].

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 September 7, 2015.

Copyright Notice

Table of Contents

1. Introduction

This document defines a YANG data model for PBB EVPN protocol.

This yang data model includes both configuration of an PBB EVPN protocol instance as well as operational states.

2. Terminology

ISID : Instance Service Identifier

ESI : Ethernet Segment Identifier

BMAC : Backbone MAC

C-BD : Customer Bridge domain

B-BD : Backbone Bridge domain

B-Comp: B-Component instance

I-Comp: I-Component instance

CE : Customer Edge device

CIP : Customer Instance Port

PIP : Provider Instance Port

CBP : Customer Backbone Port

PBP : Provider Backbone Port

3. Design of the Data Model

The PBB EVPN YANG module is divided into two containers. One is I-Component container and other one is B-Component container.

The I-component is responsible for mapping of traffic from CE to the customer bridge domain. The customer bridge domain is mapped to an ISID. Within the I-component, there are two different ports, one is customer instance port and other one is provider instance port. The I-component container contains customer bridge domain information, ISID, the customer instance port (CIP) and provider instance port (PIP).

The B-component is responsible to learn and forward packets based on the Backbone MAC addresses (BMACs). Within the B-component, there are two different ports, one is customer backbone port (CBP) and provider backbone port (PBP). The B-component container which contains PBB EVPN specific information and backbone bridge domain information which maps the I-component traffic to B-Component towards the MPLS core.

The figure below describe the overall structure of the PBB EVPN YANG module:

```
module: pbb-evpn
   +--rw interfaces
   |  +--rw interface* [name]
   |     +--rw name     leafref
   |     +--rw esi?     string
   |     +--rw redudancy-mode?   boolean
   |     +--rw auto-source-bmac?   boolean
   |     +--rw source-bmac?   yang:mac-address
   +--rw logicalinterfaces
   |  +--rw interface* [name]
   |     +--rw name     leafref
   |     +--rw encap?   uint16
   +--rw b-comp-instance
   |  +--rw instance-name    string
   |  +--rw route-distinguisher?  string
   |  +--rw auto-route-target?   boolean
   |  +--rw route-target?         string
   |  +--rw protocol?   enumeration
   |  +--rw control-word?      boolean
   |  +--rw cbp interface [name]
   |  +--rw b-bridge-domain*
   |  |  +--rw nvlanid?   uint16
   |  |  +--rw isid*
   |  |     +--rw isid?   uint32
   |  |     +--rw extended?   boolean
   |   ...
   +--rw i-comp-instance
      +--rw pip interface [name]
      +--rw cbd*
      |  +--rw member-interface
      |  |  +--rw memberifs*
      |  +--rw nvlanid?    uint16
      |  +--rw isid?       uint32
      +--rw peer-b-component?   string
```

4. B-Component Configuration

The B-component configuration contains EVPN instance name, route distinguisher, route target and B-component bridge domain configuration.

```
           +--rw b-comp-instance*
              +--rw instance-name    string
              +--rw route-distinguisher   string
              +--rw auto-route-target?   boolean
              +--rw route-target
              +--rw protocol
              +--rw control-word
              +--rw cbp-interface
                 +--rw b-bridge-domain*
                    +.......
```

4.1. Backbone Bridge domain Configuration

The Backbone bridge domain contains the ISIDs and whether those ISIDs are to be extended to PBB EVPN core. The bridge domains which are not extended to PBB EVPN core can be used for local switching purpose.

```
     +--rw b-bridge-domain*
     |  +--rw nvlanid?   uint16
     |  +--rw isid*
     |     +--rw isid?      uint32
     |     +--rw extended?  boolean
```

5. I-Component Configuration

The I-component configuration contains customer bridge domain configuration and B-component instance name to map the I-component to B-component.

```
           +--rw i-comp-instance*
              +--rw pip-interface    [name]
              +--rw mapping-b-comp-instance-name [name]
                 +--rw cbd*
                    +.......
```

5.1. Customer Bridge domain Configuration

The customer bridge domain contains the mapping of interface to ISID.

```
        +--rw cbd*
        |  +--rw isid       uint32
        |  +--rw interface-name?  [name]
```

5.2. BMAC Configuration

For single home case, the multiple ISIDs in the customer bridge domains can share the same source BMAC. For the multi-homing cases, the source BMAC is associated to interface. The source BMAC can also be auto-derived based on LACP info.

```
        +--rw service-groups*
        |  +--rw service-group-name [uint32]
        |  +--rw isid*
        |  +--rw source-bmac
```

5.3. PBB EVPN Interface Configuration

PBB EVPN interface configuration includes the name of the interface, Ethernet Segment Identifier(ESI) and mode of interface, which tells single-active or active-active and source BMAC.

```
         +--rw interfaces
         |  +--rw interface*         [name]
         |     +--rw if_name         string
         |     +--rw esi_value       string
         |     +--rw redundancy-mode string
         |     +--rw source-bmac
```

6. YANG Module

<CODE BEGINS> file "ietf-pbb-evpn@2015-03-6.yang"

```
module pbbevpn {
  namespace "urn:juniper:params:xml:ns:yang:pbbevpn";
  // replace with IANA namespace when assigned

  prefix pevpn;

  import ietf-interfaces {
    prefix if;
    //rfc7223-YANG Interface Management
  }
/*
```

```
     import ietf-inet-types {
       prefix inet;
       //rfc6991
     }
   */

     import ietf-yang-types {
       prefix yang;
     }

     description
       "This YANG module defines the generic configuration data for
        PBB EVPN Service.

        Terms and Acronyms
        EVN: Ethernet Virtual Network
        EVPN: Ethernet VPN
        I-SID: Service Instance Identifier
        B-VID: Backbone VLAN ID
        C-MAC: Customer/Client MAC
        B-MAC: Backbone MAC
        BEB: Backbone Edge Bridge
        ES: Ethernet Segment
        ESI: Ethernet Segment Identifier
        LSP: Label Switched Path
        MP2MP: Multipoint to Multipoint
        MP2P: Multipoint to Point
        P2MP: Point to Multipoint
        P2P: Point to Point
        PE: Provider Edge
        EVPN: Ethernet VPN
        EVI: EVPN Instance
       ";

     revision 2015-03-06 {
       description
         "Initial revision.";
     }

     /*
      * Configuring Ethernet Segment
      */
     container interfaces {
       list interface {
         key "name";
         leaf name {
           type leafref {
             path "/if:interfaces/if:interface/if:name";
```

```
          }
        }
        leaf esi {
          description
            "Specify the ethernet segment ID.";

          config "true";
          type string {
            length "24";
            pattern "(^00([0-9a-fA-F]){2}\.(([0-9a-fA-F]){4}\.){3}
            (([0-9a-fA-F]){4})$)";
          }
        }

        leaf redudancy-mode {
          description
            "Specify Redundancy mode, value are all-active (false),
             single-active (true)";
          config "true";
          type boolean;
        }

        leaf auto-source-bmac {
          description
            "Specify auto derived mode (true) ,
             manual bmac config (false)";
          config "true";
          type boolean;
        }

        leaf source-bmac {
           type yang:mac-address;
        }
      } /* End of Interface */
    } /* End of Container */

    /*
     * Configuring Service Classification
     */
    container logicalinterfaces {
      list interface {
        key "name";
        leaf name {
          type leafref {
            path "/if:interfaces/if:interface/if:name";
          }
        }
        leaf encap {
```

```
          description
            "Vlan ID";
          config "true";
           type uint16 {
             range "1..4094";
           }
        } /* End encap */
      }
    }

    /*
     * Configuring I-component
     */
    container i-component {
      list bridge-domain {
        description
          "Customer Bridge Domain.";
        config "true";
        type uint16;

        container member-interface {
          description
            "member interface.";
          config true;
          list memberifs {
            description
             "member interfaces.";
             config true;
             type if:interface-ref;
          }
        } /* End of member if*/
        leaf nvlanid {
          description
            "Normalized Vlan ID";
          config "true";
           type uint16 {
             range "1..4094";
           }
        }
        leaf isid {
          description
            "I-SID";
          config "true";
           type uint32 {
             range "1..16777215";
           }
        }
      } /*End of List */
```

```
        leaf peer-b-component {
          description
            "Peer Backbone Component.";
          config true;
          type string {
            length "24";
            pattern "(^00([0-9a-fA-F]){2}\.(([0-9a-fA-F]){4}\.){3}
            (([0-9a-fA-F]){4})$)";
          }
        } /*end of peer-b-component */
      }

      /* Configuring Bcomponent */
      container b-component {
        list b-bridge-domain {
          description
            "Backbone Bridge Domain.";
          config "true";
          type uint16;
          leaf nvlanid {
            description
              "Normalized Vlan ID";
            config "true";
             type uint16 {
               range "1..4094";
             }
          }
          list isid {
            description
              "I-SID";
            config "true";
            leaf isid {
              description
                "I-SID";
              config "true";
               type uint32 {
                 range "1..16777215";
               }
            }
          }
        } /*End of List */

        leaf route-distinguisher {
          description
            "Route Distinguisher.";
          config true;
          type string;
        } /*end of route-distinguiher */
```

```
       leaf auto-route-target {
           description
             "Specify auto derived route target (true) ,
              manual route target (false)";
           config "true";
           type boolean;
       }

       leaf route-target {
         description
           "Route Target.";
           config true;
           type string;
       } /* end of route target. */

       leaf protocol {
         description
           "Protocol running on Backbone B-Comp.";
         config true;
         type enumeration {
           enum "evpn" {
             value 0;
           }
           enum "pbb-evpn" {
             value 1;
           }
         }
       }

       leaf control-word {
         description
           "Control Word.";
         config true;
         type boolean;
       }
     }
   }
```

<CODE ENDS>

7. Security Considerations

Configuration and state data defined in this document are designed to be accessed via the NETCONF protocol [RFC6241].

Authors recommends to implement NETCONF access control model ([RFC6536]) to restrict access to all or part of the configuration to

specific users. Access control to RPCs is also critical as RPC permits to clear protocol datastructures that would definitively impact the network service. This kind of RPC needs only to be used in specific cases by well-known experienced users.

Authors consider that all the configuration is considered as sensitive/vulnerable as well as RPCs. But security teams can decide to open some part of the configuration to less experienced users depending on the internal organization, for example:

- User FullWrite: would access to the whole data model. This kind of profile may be restricted to few experienced people.
- User PartialWrite: would only access to configuration part within /interfaces/interface. So this kind of profile is restricted to creation/modification/deletion of interfaces. This profile does not have access to RPC.
- User Read: would only access to state part.

Unauthorized access to configuration or RPC may cause high damages to the network service.

When configuring ISIS using the NETCONF protocol, authors recommends the usage of secure transport of NETCONF using SSH ([RFC6242]).

8. Contributors

Authors would like to thank Wen Lin for their major contributions to the draft.

9. Acknowledgements

TBD.

10. IANA Considerations

TBD.

11. References

11.1. Informative References

[I-D.ietf-l2vpn-evpn]
Sajassi, A., Aggarwal, R., Bitar, N., Isaac, A., and J. Uttaro, "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-11 (work in progress), October 2014.

[I-D.ietf-l2vpn-pbb-evpn]
Sajassi, A., Salam, S., Bitar, N., Isaac, A., Henderickx, W., and L. Jin, "PBB-EVPN", draft-ietf-l2vpn-pbb-evpn-09 (work in progress), October 2014.

11.2. Normative References

[I-D.ietf-netmod-routing-cfg]
Lhotka, L., "A YANG Data Model for Routing Management", draft-ietf-netmod-routing-cfg-15 (work in progress), May 2014.

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

[RFC6020] Bjorklund, M., "YANG - A Data Modeling Language for the Network Configuration Protocol (NETCONF)", RFC 6020, October 2010.

[RFC6241] Enns, R., Bjorklund, M., Schoenwaelder, J., and A. Bierman, "Network Configuration Protocol (NETCONF)", RFC 6241, June 2011.

[RFC6242] Wasserman, M., "Using the NETCONF Protocol over Secure Shell (SSH)", RFC 6242, June 2011.

[RFC6536] Bierman, A. and M. Bjorklund, "Network Configuration Protocol (NETCONF) Access Control Model", RFC 6536, March 2012.

Appendix A. Example: NETCONF <get> Reply

This section gives an example of a reply to the NETCONF <get> request for a device that implements the data model defined in this document. The example is written in XML.

Authors' Addresses

Kishore Tiruveedhula (editor)
Juniper Networks
10 Technology Park Drive
Westford MA 01886
USA

Email: kishoret@juniper.net

Tapraj Singh
Juniper Networks
1194 N Mathilda Ave
Sunnyvale CA 94089
USA

Email: tsingh@juniper.net

Ali Sajassi
Cisco Systems
170 West Tasman Drive
San Jose, CA 95134
USA

Email: sajassi@cisco.com

Deepak Kumar
Cisco Systems
510 McCarthy Blvd
Milpitas CA 95035
USA

Email: dekumar@cisco.com

Luay Jalil
Verizon
400 International Parkway
Richardson, TX 75081
USA

Email: luay.jalil@verizon.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: October 29, 2015

X. Xu
Huawei
C. Jacquenet
Orange
L. Fang
Microsoft

April 27, 2015

L3VPN Address Prefix Based Outbound Route Filter for BGP-4
draft-xu-bess-l3vpn-prefix-orf-02

Abstract

This document defines a new Outbound Router Filter (ORF) type for BGP, refered to as "L3VPN Address Prefix Outbound Route Filter", that can be used to perform L3VPN address-prefix-based route filtering. This ORF-type supports prefix-length- or range-based matching, wild-card-based address prefix matching, as well as the exact address prefix matching for L3VPN address families. The L3VPN Address Prefix ORF is applicable in the Virtual Subnet context.

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 October 29, 2015.

Copyright Notice

Table of Contents

1. Introduction

The Outbound Route Filtering (ORF) Capability defined in [RFC5291] provides a mechanism for a BGP speaker to send to its BGP peer a set of ORFs that can be used by its peer to filter its outbound routing updates to the speaker. The Address Prefix ORF defined in [RFC5292] is used to perform address-prefix-based route filtering. However, the Address Prefix ORF is not much suitable for L3VPN [RFC4364] route filtering since there is no Route-Target (RT) field contained in the Address Prefix ORF entry.

This document builds on [RFC5292] and defines a new ORF-type for BGP, referred to as "L3VPN Address Prefix Outbound Route Filter (L3VPN Address Prefix ORF)", that can be used to perform L3VPN address prefix-based route filtering. The L3VPN Address Prefix ORF supports prefix-length- or range-based matching, wild-card-based address prefix matching, as well as the exact address prefix matching for L3VPN address families. The L3VPN Address Prefix ORF is applicable to reduce the RIB size of PE routers in the Virtual Subnet [I-D.ietf-l3vpn-virtual-subnet] context.

1.1. 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].

2. Terminology

This memo makes use of the terms defined in [RFC5292] and [RFC4364].

3. L3VPN Address Prefix ORF Encoding

The ORF-Type for the L3VPN Address Prefix ORF-Type is TBD.

A L3VPN Address Prefix ORF entry includes a Route Target field in addition to those fields which have been contained in the Address Prefix ORF [RFC5292]. That's to say, a L3VPN Address Prefix ORF entry consists of the following fields <Sequence, Action, Match, Reserved, Route-Target, Minlen, Maxlen, Length, Prefix>. Note that the Prefix field here doesn't include the Route Distinguisher (RD) part of a L3VPN address prefix. For example, in the case of a VPNv4 address prefix, only the IPv4 address prefix part of that VPNv4 address prefix is contained in that Prefix field.

A L3VPN Address Prefix ORF entry is encoded as follows: the "Action", "Match" and "Reserved" fields of the entry are encoded in the common part [RFC5291], while the remaining fields of the entry are encoded in the "type specific part" [RFC5291], as shown in Figure 1. When the Action component of an ORF entry specifies REMOVE-ALL, the entry consists of only the common part.

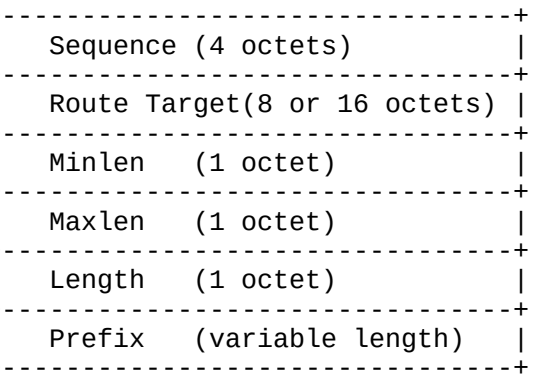


Figure 1: Type Specific Part of L3VPN Address Prefix ORF Entry Encoding

4. L3VPN Address Prefix ORF Matching

When performing route matching search on those L3VPN routes which are associated with the Route Target as specified in the received L3VPN Address Prefix ORF entries, the Address-Prefix-ORF-specific matching

rules as defined in [RFC5292] are almost preserved except that the RD SHOULD be ignored.

5. Acknowledgements

The authors would like to thank Mach Chen and Shunwan Zhuang for their comments on this document.

6. IANA Considerations

The ORF-type for the L3VPN Address Prefix ORF needs to be assigned by the IANA.

7. Security Considerations

This document does not introduce any new security considerations.

8. References

8.1. Normative References

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

[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.

8.2. Informative References

[I-D.ietf-l3vpn-virtual-subnet] Xu, X., Raszuk, R., Hares, S., Yongbing, F., Jacquenet, C., Boyes, T., and B. Fee, "Virtual Subnet: A L3VPN-based Subnet Extension Solution", draft-ietf-l3vpn-virtual-subnet-03 (work in progress), December 2014.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

Authors' Addresses

Xiaohu Xu
Huawei

Email: xuxiaohu@huawei.com

Christian Jacquenet
Orange

Email: christian.jacquenet@orange.com

Luyuan Fang
Microsoft

Email: lufang@microsoft.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: February 8, 2016

X. Xu
Huawei
S. Hares
Individual
Y. Fan
China Telecom
C. Jacquenet
Orange
T. Boyes
Bloomberg LP
B. Fee
Extreme Networks
August 7, 2015

RIB Reduction in Virtual Subnet
draft-xu-bess-virtual-subnet-rib-reduction-01

Abstract

Virtual Subnet is a BGP/MPLS IP VPN-based subnet extension solution which is intended for building Layer3 network virtualization overlays within and/or across data centers. This document describes a mechanism for reducing the RIB size of PE routers in the Virtual Subnet context.

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 February 8, 2016.

Table of Contents

1. Introduction

Virtual Subnet [I-D.ietf-bess-virtual-subnet] is a BGP/MPLS IP VPN [RFC4364] -based subnet extension solution which is intended for building Layer3 network virtualization overlays within and/or across data centers. In the Virtual Subnet context, since CE host routes of a given VPN instance need to be exchanged among PE routers participating in that VPN instance, the resulting routing table size of PE routers may become a big concern, especially in large-scale data center environment where they may need to install a huge amount of host routes into their routing tables.

[I-D.ietf-bess-virtual-subnet-fib-reduction] describes a method to reduce the FIB size of PE routers without any change to the RIB and the routing table. This FIB reduction approach is applicable in the case where the control plane of PE routers still needs to maintain all host routes of the attached VPN instances for some reason (e.g., to support multicast VPN service). In the case where the control plane of PE routers doesn't need to maintain all host routes of the attached VPN instances, the RIB size of PE routers can be reduced as well which would be beneficial for CPU and memory resource saving purpose. This document proposes a very simple RIB reduction mechanism. The basic idea of this mechanism is: remote host routes

are learnt by PE routers on demand by using the L3VPN Address Prefix ORF as described in [I-D.xu-bess-l3vpn-prefix-orf].

1.1. 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].

2. Terminology

This memo makes use of the terms defined in [RFC4364].

3. Solution Description

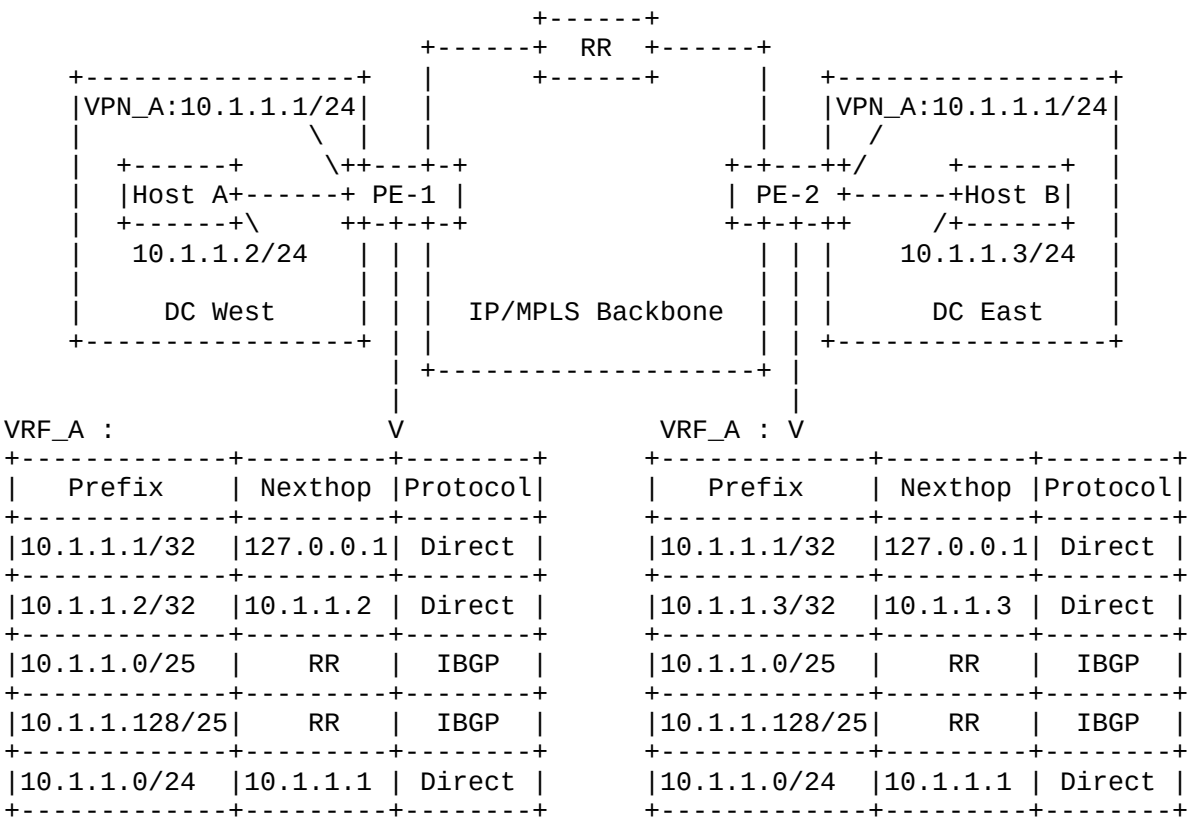


Figure 1: RIB Reduction Example

To reduce the RIB size of PE routers in the Virtual Subnet context, the L3VPN Address Prefix ORF mechanism is used to realize on-demand

route announcement. Take the VPN instance as shown in Figure 1 as an example, the RIB reduction procedures are described as follows:

1. PE routers as RR clients advertise host routes for their local CE hosts to the RR by using Rout Target (RT) ORF [RFC4364] (i.e., the RR is configured to advertise route refresh messages containing a RT-ORF entry corresponding to that VPN instance) or Route Target (RT) Constrain [RFC4684] (i.e., the RR is configured to advertise update messages containing RT membership information corresponding to that VPN instance). Those PE routers belonging to that VPN instance which don’t want to receive remote CE host routes of that VPN instance would notify the RR not to advertise any host route to them by using the L3VPN Address Prefix ORF mechanism (i.e., only requesting L3VPN routes with prefix length less than 32 (in the VPNv4 case) or 128 (in the VPNv6 case)).

2. Meanwhile, the RR is configured with static routes for more specific subnets (e.g., 10.1.1.0/25 and 10.1.1.128/25) corresponding to the extended subnet (e.g., 10.1.1.0/24) with next-hop being pointed to Null0 and then redistributes these routes to BGP. In the case where the RR is not available for transferring L3VPN traffic between PE routers for some reason (e.g., the RR is running on a server), a particular PE router other than the RR could be selected to advertise the above more specific subnet routes as long as that PE router has learnt all remote host routes belonging to that VPN instance.

3. Upon receiving a packet destined for a remote CE host from a local CE host, if there is no host route for that remote CE host in the FIB, the ingress PE router will forward the packet to the RR according to the longest-matching subnet routes learnt from the RR, which in turn forwards the packet to the relevant egress PE router according to the host route learnt from that egress PE router. As such, the RIB size of PE routers can be greatly reduced at the cost of path stretch.

4. In order to forward packets destined for that remote CE host directly to the corresponding egress PE router without any potential path stretch penalty, ingress PE routers could perform on-demand route learning of remote host routes by using one of the following options:

   A. Upon receiving an ARP request or Neighbor Solicitation (NS) message from a local CE host, if there is no CE host route for that target host in its RIB yetthe ingress PE router would request the corresponding CE host route for the target host from its RR by using the L3VPN Address Prefix ORF mechanism.

B. Upon receiving a packet whose longest-matching FIB entry is a particular more specific subnet routes (e.g., 10.1.1.0/25 and 10.1.1.128/25) learnt from the RR, a copy of this packet would be sent to the control plane while this original packet is forwarded as normal. The above copy sent to the control plane would trigger a route pull for that destination CE host. To provide robust protection against DoS attacks on the control plane, rate-limiting of the above packets sent to the control plane MUST be enabled.

5. RIB entries of remote CE host routes would expire if they have not been used for forwarding for a certain period of time. Once the expiration time for a given RIB entry is approaching, the PE router would notify its RR to remove the corresponding L3VPN Address Prefix ORF entry for that CE host route by using the L3VPN Address Prefix ORF mechanism.

4. Acknowledgements

TBD.

5. IANA Considerations

There is no requirement for any IANA action.

6. Security Considerations

This document doesn't introduce additional security risk to BGP/MPLS IP VPN, nor does it provide any additional security feature for BGP/MPLS IP VPN.

7. References

7.1. Normative References

[I-D.ietf-bess-virtual-subnet]
Xu, X., Raszuk, R., Jacquenet, C., Boyes, T., and B. Fee, "Virtual Subnet: A BGP/MPLS IP VPN-based Subnet Extension Solution", draft-ietf-bess-virtual-subnet-00 (work in progress), June 2015.

[I-D.xu-bess-l3vpn-prefix-orf]
Xu, X., Jacquenet, C., and L. Fang, "L3VPN Address Prefix Based Outbound Route Filter for BGP-4", draft-xu-bess-l3vpn-prefix-orf-02 (work in progress), April 2015.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <http://www.rfc-editor.org/info/rfc2119>.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <http://www.rfc-editor.org/info/rfc4364>.

[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, DOI 10.17487/RFC4684, November 2006, <http://www.rfc-editor.org/info/rfc4684>.

7.2. Informative References

[I-D.ietf-bess-virtual-subnet-fib-reduction] Xu, X., Jacquenet, C., Boyes, T., Fee, B., and W. Henderickx, "FIB Reduction in Virtual Subnet", draft-ietf-bess-virtual-subnet-fib-reduction-01 (work in progress), July 2015.

Authors' Addresses

Xiaohu Xu
Huawei

Email: xuxiaohu@huawei.com

Susan Hares
Individual

Email: shares@ndzh.com

Yongbing Fan
China Telecom

Email: fanyb@gsta.com

Christian Jacquenet
Orange

Email: christian.jacquenet@orange.com

Truman Boyes
Bloomberg LP

Email: tboyes@bloomberg.net

Brendan Fee
Extreme Networks

Email: bfee@enterasys.com