BESS
Internet-Draft
Updates: 7432, 6514, 7582 (if approved)
Intended status: Standards Track
Expires: October 29, 2018

Z. Zhang
E. Rosen
W. Lin
Juniper Networks
Z. Li
Huawei Technologies
I. Wijnands
Cisco Systems
April 27, 2018

MVPN/EVPN Tunnel Aggregation with Common Labels
draft-zzhang-bess-mvpn-evpn-aggregation-label-01

Abstract

The MVPN specifications allow a single Point-to-Multipoint (P2MP) tunnel to carry traffic of multiple VPNs. The EVPN specifications allow a single P2MP tunnel to carry traffic of multiple Broadcast Domains (BDs). These features require the ingress router of the P2MP tunnel to allocate an upstream-assigned MPLS label for each VPN or for each BD. A packet sent on a P2MP tunnel then carries the label that is mapped to its VPN or BD. (In some cases, a distinct upstream-assigned is needed for each flow.) Since each ingress router allocates labels independently, with no coordination among the ingress routers, the egress routers may need to keep track of a large number of labels. The number of labels may need to be as large (or larger) than the product of the number of ingress routers times the number of VPNs or BDs. However, the number of labels can be greatly reduced if the association between a label and a VPN or BD is made by provisioning, so that all ingress routers assign the same label to a particular VPN or BD. New procedures are needed in order to take advantage of such provisioned labels. These new procedures also apply to Multipoint-to-Multipoint (MP2MP) tunnels. This document updates RFCs 6514, 7432 and 7582 by specifying the necessary procedures.

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.

Copyright Notice

Table of Contents

1. Terminologies 3
2. Introduction 3
  2.1. Problem Description 4
  2.2. Proposed Solution 5
    2.2.1. MP2MP Tunnels 6
    2.2.2. Segmented Tunnels 6
    2.2.3. Summary of Label Allocation Methods 8
3. Specification 9
  3.1. Context Label Space ID Extended Community 9
  3.2. Procedures 10
4. IANA Considerations 11
5. Acknowledgements 11
6. Contributors 11
7. References 11
  7.1. Normative References 11
  7.2. Informative References 12
Authors' Addresses 13

1. Terminologies

Familiarity with MVPN/EVPN protocols and procedures is assumed. Some terminologies are listed below for convenience.

- BUM: Broadcast, Unknown Unicast, or Multicast (traffic).
- BD: Broadcast Domain.
- PMSI: Provider Multicast Service Interface - a pseudo interface for a PE to send overlay/customer multicast traffic via underlay/provider tunnels. Includes I/S-PMSI (often referred to as x-PMSI) for Inclusive/Selective-PMSI.
- IMET: Inclusive Multicast Ethernet Tag route. An EVPN specific name for I-PMSI A-D route.
- ESI: Ethernet Segment Identifier.

2. Introduction

MVPN can use P2MP tunnels (set up by RSVP-TE, mLDP, or PIM) to transport customer multicast traffic across a service provider's backbone network. Often, a given P2MP tunnel carries the traffic of only a single VPN. There are however procedures defined that allow a single P2MP tunnel to carry traffic of multiple VPNs. In this case, the P2MP tunnel is called an "aggregate tunnel". The PE router that is the ingress node of an aggregate P2MP tunnel allocates an "upstream-assigned MPLS label" [RFC5331] for each VPN, and each packet sent on the P2MP tunnel carries the upstream-assigned MPLS label that the ingress PE has bound to the packet's VPN.

Similarly, EVPN can use P2MP tunnels (set up by RSVP-TE, mLDP, or PIM) to transport BUM traffic (Broadcast traffic, Unicast traffic with an Unknown address, or Multicast traffic), across the provider network. Often a P2MP tunnel carries the traffic of only a single BD. However, there are procedures defined that allow a single P2MP tunnel to be an "aggregate tunnel" that carries traffic of multiple BDs. The procedures are analogous to the MVPN procedures -- the PE router that is the ingress node of an aggregate P2MP tunnel allocates an upstream-assigned MPLS label for each BD, and each packet sent on the P2MP tunnel carries the upstream-assigned MPLS label that the ingress PE has bound to the packet's BD.

MVPN and EVPN can also use BIER [RFC 8279] to transmit multicast traffic or BUM traffic [I-D.ietf-bier-mvpn] [I-D.ietf-bier-evpn]. Although BIER does not explicitly set up P2MP tunnels, from the perspective of MVPN/EVPN, the use of BIER transport is very similar

to the use of aggregate P2MP tunnels. When BIER is used, the PE transmitting a packet (the "BFIR" [RFC 8279]) must allocate an upstream-assigned MPLS label for each VPN or BD, and the packets transmitted using BIER transport always carry the label that identifies their VPN or BD. (See [BIER-MVPN] and [BIER-EVPN] for the details.) In the remainder of this document, we will use the term "aggregate tunnels" to include both P2MP tunnels and BIER transport.

When an egress PE receives a packet from an aggregate tunnel, it must look at the upstream-assigned label carried by the packet, and must interpret that label in the context of the ingress PE. Essentially, each ingress PE has its own "context label space" [RFC5331] from which it allocates its upstream-assigned labels. When an egress PE looks up the upstream-assigned label carried by a given packet, it looks it up in the context label space owned by the packet's ingress PE. How an egress PE identifies the ingress PE of a given packet depends on the tunnel type.

2.1. Problem Description

Note that these procedures may require a very large number of labels. Suppose an MVPN or EVPN deployment has 1001 PEs, each hosting 1000 VPN/BDs. Each ingress PE has to assign 1000 labels, and each egress PE has to be prepared to interpret 1000 labels from each of the ingress PEs. Since each ingress PE allocates labels from its own context label space, and the ingress PEs do not coordinate their label assignments, each egress PE must be prepared to interpret 1,000,000 upstream-assigned labels. This is an evident scaling problem.

At the present time, few if any MVPN/EVPN deployments use aggregate tunnels, so this problem has not surfaced. However, the use of aggregate tunnels is likely to increase due to the following two factors:

- In EVPN, a single customer ("tenant") may have a large number of BDs, and the use of aggregate RSVP-TE or mLDP P2MP tunnels may become important, since each tunnel creates state at the intermediate nodes.

- The use of BIER as transport for MVPN/EVPN is becoming more and more attractive and feasible.

Note there are pros and cons with traditional P2MP tunnel aggregation (vs. BIER), which are already discussed in Section 2.1.1 of [RFC6513]. This document simply specifies a way to increase label scaling when tunnel aggregation is used.

A similar problem also exists with EVPN ESI labels used for multi-homing. A PE attached to a multi-homed Ethernet Segment (ES) advertises an ESI label in its Ethernet Segment route for the ES. The PE imposes the label when it sends frames received from the ES to other PEs via a P2MP/BIER tunnel. A receiving PE that is attached to the source ES will know from the ESI label that the packet originated on the source ES, and thus will not transmit the packet on its local attachment circuit to that ES. From the receiving PE's point of view, the ESI label is (upstream-)allocated from the source PE's label space, so the receiving PE needs to maintain context label tables, one for each source PE, just like the VRF/BD label case above. If there are 1,001 PEs, each attached to 1,000 ESes, this can require each PE to understand 1,000,000 ESI labels. Notice that the issue exists even when no P2MP tunnel aggregation (i.e. one tunnel used for multiple BDs) is used.

2.2. Proposed Solution

The number of labels could be greatly reduced if a central authority assigned a label to each VPN, BD, or ES, and if all PEs used that same label to represent a given VPN , BD, or ES. Then the number of total number of labels needed would just be the sum of the number of VPNs, BD, and/or ESes.

One method of achieving this is to reserve a portion of the label space for assignment by a central authority. We refer to this reserved portion as the "Domain-wide Common Block" (DCB) of labels. This is analogous to the "Segment Routing Global Block" (SRGB) that is described in [I-D.ietf-spring-segment-routing]. The DCB is taken from the same label space that is used for downstream-assigned labels, but each PE would know not to allocate local labels from that space. A PE that is attached (via L3VPN VRF interfaces or EVPN Access Circuits) would know by provisioning which label from the DCB corresponds to which of its locally attached VPNs, BDs, or ESes. The definition of "domain" is loose - it simply includes all the routers that share the same DCB. In this document, it includes all PEs of an MVPN/EVPN network. (Though if tunnel segmentation [RFC 6514] is used, each segmentation region could have its own DCB. This will be explained in more detail later.) If these PEs share other common label blocks (e.g. SRGB) with other routers, the DCB MUST not intersect with those common label blocks or those routers MUST be considered as part of the "domain". However, the labels advertised by PEs for the purposes defined in this document will only rise to the top of the label stack when traffic arrives the PEs.

In some deployments, it may be impractical to allocate a DCB that is large enough to contain labels for all the VPNs/BDs/ESes. In this case, it may be necessary to allocate those labels from a context

label space. However, it is not necessary for each ingress PE to have its own context label space. Instead, one (or some small number) of context label spaces can be dedicated to such labels. Each ingress PE would be provisioned to know both the context label space identifier and the label for each VPN/BD/ES.

The MVPN/EVPN signaling defined in [RFC6514] and [RFC7432] assumes that certain MPLS labels are allocated from a context label space owned by a particular ingress PE. In this document, we augment the signaling procedures so that it is possible to signal that a particular label is from the DCB, rather than from an ingress PE's context label space. We also augment the signaling so that it is possible to indicate that a particular label is from an identified context label space that is different than the ingress PE's own context label space.

Notice that, the VPN/BD/ES-identifying labels from the DCB or from those few context label spaces are very similar to VNIs in VXLAN. Allocating a label from the DCB or from those a few context label spaces and communicating them to all PEs should not be different from allocating VNIs, and should be feasible in today's networks since controllers are used more and more widely.

2.2.1. MP2MP Tunnels

MP2MP tunnels present the same problem that can be solved the same way.

Per RFC 7582 ("MVPN: Using Bidirectional P-tunnels"), when MP2MP tunnels are used for MVPN, the root of the MP2MP tunnel may need to allocate and advertise "PE Distinguisher Labels". RFC 7582 states that these labels are upstream-assigned, from the label space used by the root node for its upstream-assigned labels.

It is REQUIRED by this document that the PE Distinguisher labels allocated by a particular node come from the same source that the node uses to allocate its VPN-identifying labels.

2.2.2. Segmented Tunnels

There are some additional issues to be considered when MVPN or EVPN is using "tunnel segmentation" (see [RFC6514], [RFC7524], and [EVPN-BUM] Sections 5 and 6).

2.2.2.1. Selective Tunnels

For "selective tunnels" (see [RFC6513] Sections 2.1.1 and 3.2.1, and [EVPN-BUM] Section 4), the procedures outlined above work only if tunnel segmentation is not used.

A selective tunnel carries one or more particular sets of flows to a particular subset of the PEs that attach to a given VPN or BD. Each set of flows is identified by a Selective PMSI A-D route [RFC6514]. The PTA of the S-PMSI route identifies the tunnel used to carry the corresponding set of flows. Multiple S-PMSI routes can identify the same tunnel.

When tunnel segmentation is applied to a S-PMSI, certain nodes are "segmentation points". A segmentation point is a node at the boundary between two "segmentation regions". Let's call these "region A" and "region B". A segmentation point is an egress node for one or more selective tunnels in region A, and an ingress node for one or more selective tunnels in region B. A given segmentation point must be able to receive traffic on a selective tunnel from region A, and label switch the traffic to the proper selective tunnel in region B.

Suppose one selective tunnel (call it T1) in region A is carrying two flows, Flow-1 and Flow-2, identified by S-PMSI route Route-1 and Route-2 respectively. However, it is possible that, in region B, Flow-1 is not carried by the same selective tunnel that carries Flow-2. Let's suppose that in region B, Flow-1 is carried by tunnel T2 and Flow-2 by tunnel T3. Then when the segmentation point receives traffic from T1, it must be able to label switch Flow-1 from T1 to T2, while also label switching Flow-2 from T1 to T3. This implies that Route-1 and Route-2 must signal different labels in the PTA.

In this case, it is not practical to have a central authority assign domain-wide unique labels to individual S-PMSI routes. To address this problem, all PEs can be assigned disjoint label blocks in those few context label spaces, and each will allocate labels for segmented S-PMSI independently from its assigned label block that is different from any other PE's. For example, PE1 allocates from label block [101~200], PE2 allocates from label block [201~300], and so on.

Allocating from disjoint label blocks can be used for VPN/BD/ES labels as well, though it does not address the original scaling issue, because there would be one million labels allocated from those a few context label spaces in the original example, instead of just one thousand common labels.

2.2.2.2. Per-PE/Region Tunnels

Similarly, for segmented per-PE (MVPN (C-*,C-*) S-PMSI or EVPN IMET) or per-AS/region (MVPN Inter-AS I-PMSI or EVPN per-Region I-PMSI) tunnels, labels need to be allocated per PMSI route. In case of per-PE PMSI route, the labels should be allocated from the label block allocated to the advertising PE. In case of per-AS/region PMSI route, different ASBR/RBRs attached to the same source AS/region will advertise the same PMSI route. The same label could be used when the same route is advertised by different ASBRs/RBRs, though a simpler way is for each ASBR/RBR to allocate its own label from the label block allocated to itself.

In the rest of the document, we call the label allocated for a particular PMSI a (per-)PMSI label, just like we have (per-)VPN/BD/ES labels. Notice that using per-PMSI label in case of per-PE PMSI still has the original scaling issue associated with the upstream allocated label, so per-region PMSIs should be preferred. Within each AS/region, per-PE PMSIs are still used though they do not go across border and per-VPN/BD labels can still be used.

Note that, when a segmentation point re-advertise a PMSI route to the next segment, it does not need to re-advertise a new label unless the upstream or downstream segment uses Ingress Replication. [note - future revision may extend the applicability of this document to Ingress Replication as well]

2.2.2.3. Alternative to the per-PMSI Label Allocation

The per-PMSI label allocation in case of segmentation, whether for S-PMSI or for per-PE/Region I-PMSI, is for the segmentation points to be able to label switch traffic w/o having to do IP or MAC lookup in VRFs (the segmentation points typically do not have those VRFs at all). If the label scaling becomes a concern, altertatively the segmenation points could use (C-S,C-G) lookup in VRFs for flows identifyed by the S-PMSIs. This allows the S-PMSIs for the same VPN/BD to share the a VPN/BD-identifying label that leads to lookup in the VRFs. That label should be different from the label used in the per-PE/region I-PMSIs though, so that the segmentation points can label switch other traffic (not identified by those S-PMSIs). However, this moves the scaling problem from the number of labels to the number of (C-S/*,C-G) routes in VRFs on the segmentation points.

2.2.3. Summary of Label Allocation Methods

In summary, labels can be allocated and advertised the following ways:

1. A central authority allocates per-VPN/BD/ES labels from the DCB. PEs advertise the labels with an indication that they are from the DCB.

2. A central authority allocates per-VPN/BD/ES labels from a few common context label spaces, and allocate labels from the DCB to identify those context label spaces. PEs advertise the VPN/BD labels along with the context-identifying labels.

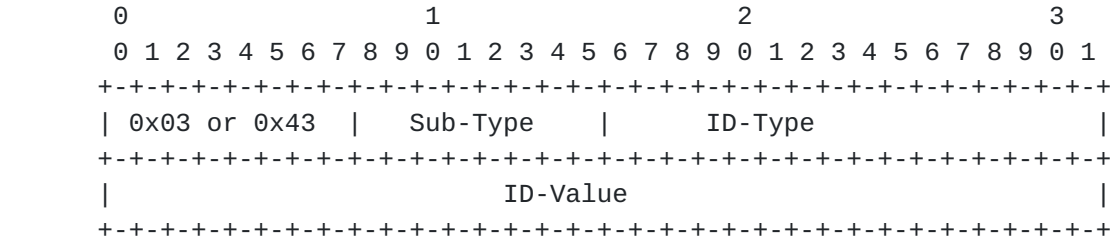
3. A central authority assigns disjoint label blocks from those a few context label spaces to each PE, and allocate labels from the DCB to identify the context label spaces. Each PE allocates labels from its assigned label block independently for its segmented S-PMSI, along with the context-identifying labels.

Option 1 is simplest, but it requires that all the PEs set aside a common label block for the DCB that is large enough for all the VPNs/BDs/ESes combined. Option 3 is needed only for segmented selective tunnels that are set up dynamically. Multiple options could be used in any combination depending on the deployment situation.

3. Specification

3.1. Context Label Space ID Extended Community

Context Label Space ID Extended Community is a new Transitive Opaque EC with the following structure:



- ID-Type: A 2-octet field that specifies the type of Label Space ID. In this document, the ID-Type is 0, indicating that the ID-Value field is a label.

- ID-Value: A 4-octet field that specifies the value of Label Space ID. When it is a label (with ID-Value 0), the most significant 20-bit is set to the label value.

3.2. Procedures

The protocol and procedures specified in this section need not be applied unless when BIER, or P2MP/MP2MP tunnel aggregation is used for MVPN/EVPN, or BIER/P2MP/MP2MP tunnels are used with EVPN multi-homing.

By means outside the scope of this document, each VPN/BD/ES is assigned a label from the DCB or one of those few context label spaces, and every PE that is part of the VPN/BD/ES is aware of the assignment. The ES label and the BD label MUST be assigned from the same source. If PE Distinguisher labels are used [RFC7582], they must be allocated from the same source as well.

In case of tunnel segmentation, each PE is also assigned a disjoint label block from one of those few context label spaces and it allocates labels for its segmented PMSI routes from its assigned label block.

When a PE originates an x-PMSI/IMET route, if the label is assigned from the DCB, a C-bit in the PTA's Flags field is set to indicate the label is from the DCB.

If the VPN/BD/PMSI label is assigned from one of those few context label spaces, a Context Label Space ID Extended Community is attached to the route. The ID-Type in the EC is set to 0 and the ID-Value is set to a label allocated from the DCB and identifies the context label space. When an ingress PE sends traffic, it imposes the DCB label that identifies the context label space after it imposes the label (that is advertised in the PTA's Label field of the x-PMSI/IMET route) for the VPN/BD and/or the label (that is advertised in the ESI Label EC) for the ESI, and then imposes the encapsulation for the transport tunnel.

When a PE receives an x-PMSI/IMET route with the Context Label Space ID EC, it programs its default MPLS forwarding table to map the label in the EC that identifies the context label space to a corresponding context label table in which the next label lookup is done for traffic that this PE receives.

The receiving PE then programs the label in the PTA or ESI Label EC into either the default mpls forwarding table (if the C-bit is set) or the context label table (if the Context Label Space ID EC is present) according to the x-PMSI/IMET route.

A PE MUST NOT both set the C-bit in the PTA of an x-PMSI/IMET route and attach the Context Label Space ID EC in the route. A PE MUST ignore a received route with both the C-bit set and the Context Label

Space ID EC attached. If neither C-bit is set nor the Context Label Space ID EC is attached, the label in the PTA or ESI Label EC is treated as the upstream allocated from the source PE's label space, and procedures in [RFC6514][RFC7432] must be followed.

In case of MPLS P2MP tunnels, if two x-PMSI/IMET routes specify the same tunnel, one of the following conditions MUST be met, so that a receiving PE can correctly intrerpret the label that follows the tunnel label in the right context.

- They MUST all have the C-bit set, or,
- They MUST all carry the Context Label Space ID EC, or,
- None of them has the C-bit set, or,
- None of them carry the Context Label Space ID EC.

4. IANA Considerations

This document introduces a C-bit in the Flags field of PTA. An IANA request will be submitted for bit 0x02 as the C-bit in the P-Multicast Service Interface (PMSI) Tunnel Attribute Flags registry. This is subject to approval/change.

This document introduces a new Transitive Opaque Extended Community "Context Label Space ID Extended Community". An IANA request will be submitted for sub-type value 0x15 (subject to approval/change) in the BGP Transitive Opaqaue Extended Community Sub-Types registry.

5. Acknowledgements

6. Contributors

The following also contributed to this document.

Selvakumar Sivaraj
Juniper Networks

Email: ssivaraj@juniper.net

7. References

7.1. 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, <https://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, <https://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, <https://www.rfc-editor.org/info/rfc6514>.

[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, <https://www.rfc-editor.org/info/rfc7432>.

[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, <https://www.rfc-editor.org/info/rfc7524>.

[RFC7582] Rosen, E., Wijnands, IJ., Cai, Y., and A. Boers, "Multicast Virtual Private Network (MVPN): Using Bidirectional P-Tunnels", RFC 7582, DOI 10.17487/RFC7582, July 2015, <https://www.rfc-editor.org/info/rfc7582>.

7.2. Informative References

[I-D.ietf-bess-evpn-bum-procedure-updates] Zhang, Z., Lin, W., Rabadan, J., Patel, K., and A. Sajassi, "Updates on EVPN BUM Procedures", draft-ietf-bess-evpn-bum-procedure-updates-03 (work in progress), April 2018.

[I-D.ietf-bier-evpn] Zhang, Z., Przygienda, T., Sajassi, A., and J. Rabadan, "EVPN BUM Using BIER", draft-ietf-bier-evpn-00 (work in progress), August 2017.

[I-D.ietf-bier-mvpn] Rosen, E., Sivakumar, M., Aldrin, S., Dolganow, A., and T. Przygienda, "Multicast VPN Using BIER", draft-ietf-bier-mvpn-11 (work in progress), March 2018.

[I-D.ietf-spring-segment-routing]
Filsfils, C., Previdi, S., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", draft-ietf-spring-segment-routing-15 (work in progress), January 2018.

[RFC5331] Aggarwal, R., Rekhter, Y., and E. Rosen, "MPLS Upstream Label Assignment and Context-Specific Label Space", RFC 5331, DOI 10.17487/RFC5331, August 2008, <https://www.rfc-editor.org/info/rfc5331>.

Authors' Addresses

Zhaohui Zhang
Juniper Networks

EMail: zzhang@juniper.net

Eric Rosen
Juniper Networks

EMail: erosen@juniper.net

Wen Lin
Juniper Networks

EMail: wlin@juniper.net

Zhenbin Li
Huawei Technologies

EMail: lizhenbin@huawei.com

IJsbrand Wijnands
Cisco Systems

EMail: ice@cisco.com