BESS Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 26, 2021

P. Brissette
A. Sajassi
L. Burdet, Ed.
Cisco Systems
D. Voyer
Bell Canada

February 22, 2021

EVPN Multi-Homing Mechanism for Layer-2 Gateway Protocols
draft-brissette-bess-evpn-l2gw-proto-06

Abstract

The existing EVPN multi-homing load-balancing modes defined are Single-Active and All-Active. Neither of these multi-homing mechanisms adequately ethernet-segments facing access networks with Layer-2 Gateway protocols such as G.8032, (M)STP, REP, MPLS-TP, etc. These loop-preventing Layer-2 protocols require a new multi-homing mechanism defined in this draft.

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 https://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 26, 2021.

Copyright Notice

Table of Contents

1. Introduction 2
1.1. Requirements Language 3
1.2. Terms and Abbreviations 3
2. Requirements 4
3. Solution 5
3.1. Single-Flow-Active redundancy mode 6
3.2. Fast-Convergence 7
3.2.1. Handling of Topology Change Notification (TCN) . . . 7
3.2.2. Propagating L2GW Protocol Events 7
3.2.3. MAC Flush and Invalidation Procedure 8
3.3. Backwards compatibility 8
3.3.1. The two-ESI solution 8
3.3.2. RFC7432 Remote PE 9
4. ESI-label Extended Community Extension 10
5. EVPN Inter-subnet Forwarding 10
6. Conclusion 11
7. Security Considerations 11
8. Acknowledgements 11
9. IANA Considerations 11
10. References 11
10.1. Normative References 11
10.2. Informative References 12
Authors' Addresses 12

1. Introduction

Existing EVPN Single-Active and All-Active multi-homing mechanisms do not address the additional requirements of loop-preventing Layer-2 gateway protocols such as G.8032, (M)STP, REP, MPLS-TP, etc.

These Layer-2 Gateway protocols require that a given L2 flow of a VLAN be only active on one of the PEs in the multi-homing group, while another L2 flow may be active on the other PE. This is in contrast with Single-Active redundancy mode where all flows of a VLAN are active on a single multi-homing PEs and it is also in contrast with All-Active redundancy mode where all flows of a VLAN are active on all PEs in the redundancy group.

This draft defines a new multi-homing mechanism "Single-Flow-Active" specifying that a VLAN can be active on all PEs in the redundancy group but each unique L2 flow of that VLAN can be active on only one

of the PEs in the redundancy group at a time. In fact, the Designated Forwarder election algorithm for these L2 Gateway protocols, is not per VLAN but rather for a given L2 flow. A selected PE in the redundancy group must be the only Designated Forwarder for a specific L2 flow, but the decision is not taken by the PE. The loop-prevention blocking scheme occurs in the access network, by the Layer-2 protocol.

EVPN multi-homing procedures need to be enhanced to support Designated Forwarder election for all traffic (both known unicast and BUM) on a per L2 flow basis. The Single-Flow-Active multi-homing mechanism also requires new EVPN considerations for aliasing, mass-withdraw, fast-switchover and [I-D.ietf-bess-evpn-inter-subnet-forwarding] as described in the solution section.

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 [RFC2119].

1.2. Terms and Abbreviations

AC: Attachment Circuit

BUM: Broadcast, Unknown unicast, Multicast

DF: Designated Forwarder

GW: Gateway

L2 Flow: A given flow of a VLAN, represented by (MAC-SA, MAC-DA)

L2GW: Layer-2 Gateway

MAC-IP: EVPN Route-Type 2 with non-zero IP field

G.8032: Ethernet Ring Protection

(M)STP: Multi-Spanning Tree Protocol

REP: Resilient Ethernet Protocol

TCN: Topology Change Notification

2. Requirements

The EVPN L2GW framework for L2GW protocols in Access-Gateway mode, consists of the following rules:

- Peering PEs MUST share the same ESI.
- The Ethernet-Segment DF election MUST NOT be performed and forwarding state MUST be dictated by the L2GW protocol. In gateway mode, both PEs are usually in forwarding state. In fact, the access protocol is responsible for operationally setting the forwarding state for each VLAN.
- Split-horizon filtering is NOT needed because L2GW protocol ensures there will never be a loop in the access network. The forwarding between peering PEs MUST also be preserved. In Figure 1, CE1/CE4 device may need reachability with CE2 device. ESI-filtering capability MUST be disabled. The ESI label extended comunity advertised to other peering PEs in the redundancy group MUST NOT be applied it if received.
- ESI label BGP Extended Community MUST support a new multi-homing mode named "Single-Flow-Active" corresponding largely to the single-active behaviour of [RFC7432], applied per L2 flow rather than per VLAN.
- Upon receiving ESI label BGP Extended Community with the single-flow-active load-balancing mode, remote PE MUST:
  - Disable ESI label processing
  - Disable aliasing (at Layer-2 and Layer-3 [I-D.ietf-bess-evpn-inter-subnet-forwarding])
- The Ethernet-Segment procedures in the EVPN core such as Ethernet A-D per ES and per Ethernet A-D per EVI routes advertisement/ withdraw, as well as MAC and MAC+IP advertisement, remains as explained in [RFC7432] and [I-D.ietf-bess-evpn-inter-subnet-forwarding].
- For fast-convergence, remote PE3 MAY set up two distinct backup paths on a per-flow basis:
  - { PE1 active, PE2 backup }
  - { PE2 active, PE1 backup }

The backup paths so created, operate as in [RFC7432] section 8.4 where the backup PE of the redundancy group MAY immediately be selected for forwarding upon detection of a specific subset of failures: Ethernet A-D per ES route withdraw, Active PE loss of reachability (via IGP detection). An Ethernet A-D per EVI withdraw MUST NOT result in automatic switching to the backup PE as only a subset of the hosts may be changing reachability to the Backup PE, and the remote cannot determine which.

- o MAC mobility procedures SHALL have precedence over backup path procedure in Single-Flow-Active for tracking host reachability.

3. Solution

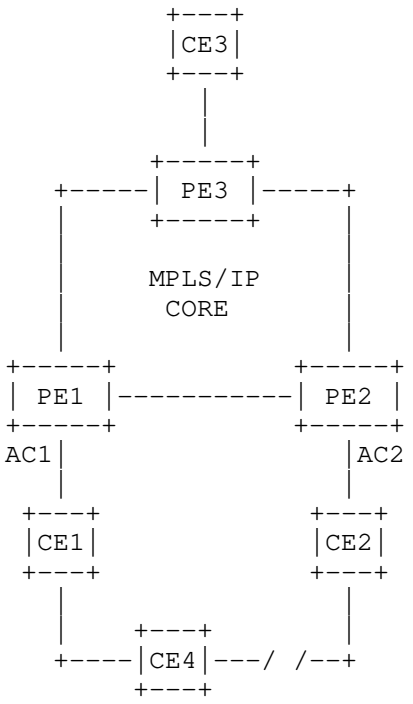


Figure 1: EVPN network with L2 access GW protocols

Figure 1 shows a typical EVPN network with an access network running a L2GW protocol, typically one of the following: G.8032, (M)STP, REP, MPLS- TP, etc. The L2GW protocol usually starts from AC1 (on PE1) up to AC2 (on PE2) in an open "ring" manner. AC1 and AC2 interfaces of PE1 and PE2 are participants in the access protocol.

The L2GW protocol is used for loop avoidance. In above example, the loop is broken on the right side of CE4.

3.1. Single-Flow-Active redundancy mode

PE1 and PE2 are peering PEs in a redundancy group, and sharing a same ESI. In the proposed Single-Flow-Active mode, load-balancing at PE1 and PE2 shares similarities with singular aspects of both Single-Active and All-Active. Designated Forwarder election must not compete with the L2GW protocol and must not result in blocked ports or portions of the access may become isolated. Additionally, the reachability between CE1/CE4 and CE2 is achieved with the forwarding path through the EVPN MPLS/IP core side. Thus, the ESI-Label filtering of [RFC7432] is disabled for Single-Flow-Active Ethernet segments.

Finally, PE3 behaves according to EVPN [RFC7432] rules for traffic to/from PE1/PE2. Peering PE, selected per L2 flow, is chosen by the L2GW protocol in the access, and is out of EVPN control.

From PE3 point of view, the L2 flows from PE3 destined to CE1/CE4 transit via edge node PE1 and the L2 flows destined to CE2 transit via edge node PE2. A specific unicast L2 flow never goes to both peering PEs. Therefore, aliasing of [RFC7432] Section 8.4 cannot be performed by PE3. That node operates in a single-active fashion for each of the unicast L2 flows.

The backup path of [RFC7432] Section 8.4 which is also setup for single-active rapid convergence on a per-VLAN basis, is not applicable here. For example, in Figure 1, if a failure happens between CE1 and CE4 the loop-prevention at the right of CE4 is released and:

- L2 flows coming from CE3 behind PE3 destined to CE1 still transit through edge device PE1, and shall not switch to PE2 as a backup path.

- L2 flows destined to CE4 on the other hand, may be backup switched to PE2 transit node.

On PE3, there is no way to know which L2 flow specifically is affected. During the transition time, PE3 may flood until unicast traffic recovers properly.

3.2. Fast-Convergence

3.2.1. Handling of Topology Change Notification (TCN)

In order to address rapid Layer-2 convergence requirement, topology change notification received from the L2GW protocols must be sent across the EVPN network to perform the equivalent of legacy L2VPN remote MAC flush.

The generation of TCN is done differently based on the access protocol. In the case of REP and G.8032, TCN gets generated in both directions and thus both of the dual-homing PEs receive it. However, with (M)STP, TCN gets generated only in one direction and thus only a single PE can receive it. That TCN is propagated to the other peering PE for local MAC flushing, and relaying back into the access.

In fact, PEs have no direct visibility on failures happening in the access network nor on the impact of those failures over the connectivity between CE devices. Hence, both peering PEs require to perform a local MAC flush on corresponding interfaces.

There are two options to relay the access protocol's TCN to the peering PE: in-band or out-of-band messaging. The first method is better for rapid convergence, and requires a dedicated channel between peering PEs. An EVPN-VPWS connection MAY be dedicated for that purpose, connecting the Untagged ACs of both PEs. The latter choice relies on the MAC Mobility BGP Extended Community applied to the Ethernet A-D per EVI route, detailed below. It is a slower method but has the advantage of avoiding a dedicated channel between peering PEs.

3.2.2. Propagating L2GW Protocol Events

Peering PE in Single Flow Active mode, upon receiving notification of a protocol convergence-event from access (such as TCN), MUST:

- As per legacy VPLS, perform a local MAC flush on the access-facing interfaces. An ARP probe is also sent for all hosts previously locally-attached.

- Advertise Ethernet A-D per EVI route along with the MAC Mobility BGP Extended Community, with incremented sequence number if previously advertised, in order to perform a remote MAC flush and steer L2 traffic to proper peering PE. The sequence number is incremented by one as a flushing indication to remote PEs.

- Ensure MAC and MAC+IP route re-advertisement, with incremented sequence number when host reachability is NOT moving to peering

PE. This is to ensure a re-advertisement of current MAC and MAC-IP which may have been flushed remotely upon MAC Mobility Extended Community reception. In theory, it should happen automatically since peering PE, receiving TCN from the access, performs local MAC flush on corresponding interface and will re-learn that local MAC or MAC+IP.

- Where an access protocol relies on TCN BPDU propagation to all participant nodes, a dedicated EVPN-VPWS connection MAY be used as an in-band channel to relay TCN between peering PEs. That connection may be auto-generated or can simply be configured by user.

3.2.3. MAC Flush and Invalidation Procedure

The MAC-Flush procedure described in [RFC7623] is borrowed, and the MAC mobility BGP Extended community is signaled along with the Ethernet A-D per EVI route from a PE in Single-Flow-Active mode.

When MAC Mobility BGP Extended Community is received on the Ethernet A-D per EVI route, it indicates to all remote PEs that all MAC addresses associated with that EVI/ESI are "flushed" i.e. must be unresolved.

Remote PEs, having previously received Ethernet A-D per ES with Single Flow Active indication from an originating PE, treat the MAC Mobility indication to simply invalidate the MAC entries for that originating PE on an EVI/ESI basis, similar to [RFC7432]’s mass-withdraw mechanism.

They remain unresolved until the remote PE receives a route update (or withdraw) for those MAC addresses. Note: the MAC may be re-advertised by the same PE, but also some are expected to have moved to a multi-homing peer, within the same ESI, due to the L2 protocol’s action.

The sequence number of the MAC Mobility extended community is of local significance from the originating PE, and is not used for comparison between peering PEs. Rather, it is used to signal via BGP successive MAC Flush requests from a given PE per EVI/ESI.

3.3. Backwards compatibility

3.3.1. The two-ESI solution

As a reference, an alternative solution which achieves some, but not all, of the requirements exists:

On the PE1 and PE2,

a. A single-homed (different) non-zero ESI, or zero-ESI, is used for each PE;

b. With no remote Ethernet-Segment routes received matching local ESI, each PE will be designated forwarder for all the local VLANs;

c. Each L2GW PE will send Ethernet A-D per ES and per EVI routes for its ESI if non-zero; and

d. When the L2GW PEs receive a MAC-Flush notification (STP TCN, G.8032 mac-flush, LDP MAC withdrawal etc.), they send an update of the Ethernet A-D per EVI route with the MAC Mobility extended community and a higher sequence number, using the procedure outlined in Section 3.2.3.

While this solution is feasible, it is considered to fall short of the requirements listed in Section 2, namely for all aspects meant to achieve fast-convergence.

3.3.2. RFC7432 Remote PE

A PE which receives an Ethernet A-D per ES route with the Single-Flow-Active bit set in the ESI-flags, and which does not support/ understand this bit, SHALL discard the bit and continue operating per [RFC7432] (All-Active). The operator should understand the usage of single-flow-active load-balancing mode else it is highly recommended to use the two-ESI approach as described in Section 3.3.1

The remote PE3 which does not support Single-Flow-Active redundancy mode as described, will ECMP traffic to peering PE1 and PE2 in the example topology above (Figure 1), per [RFC7432], Section 8.4 aliasing and load-balancing rules. PE1 and PE2, which support the Single-Flow-Active redundancy mode MUST setup redirections towards the PE at which the flow is currently active (sub-optimal Layer-2 forwarding and sub-optimal Layer-3 routing).

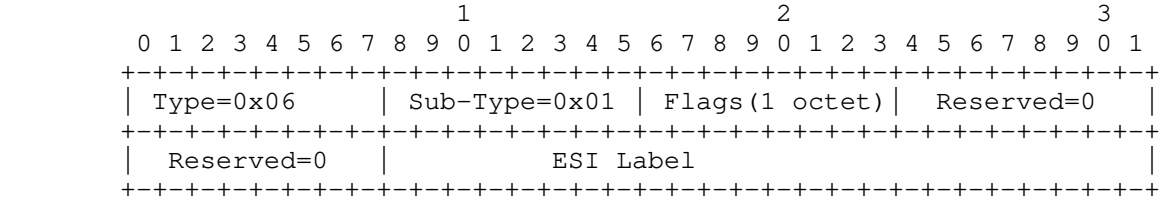
Thus, while PE3 will ECMP (on average) 50% of the traffic to the incorrect PE using [RFC7432] operation, PE1 and PE2 will handle this gracefully in Single-Flow-Active mode and redirect across peering pair of PEs appropriately.

No extra route or information is required for this. The [RFC7432] and [I-D.ietf-bess-evpn-inter-subnet-forwarding] route advertisements are sufficient.

4. ESI-label Extended Community Extension

In order to support the new EVPN load-balancing mode (single-flow-active), the ESI label Extended Community is updated.

The 1 octet flag field, part of the ESI Label Extended Community, is modified as follows:



Low-order bit: [7:0]
[2:0]- 000 = all-active,
001 = single-active,
010 = single-flow-active,
others = unassigned
[7:3]- Reserved

Figure 2: ESI Label BGP Extended Community

5. EVPN Inter-subnet Forwarding

EVPN Inter-subnet forwarding procedures in [I-D.ietf-bess-evpn-inter-subnet-forwarding] works with the current proposal and does not require any extension. Host routes continue to be installed at PE3 with a single remote nexthop, no aliasing.

However, leveraging the same-ESI on both L2GW PEs enables ARP/ND synchronization procedures which are defined for All-Active redundancy in [I-D.ietf-bess-evpn-inter-subnet-forwarding]. In steady-state, on PE2 where a host is not locally-reachable the routing table will reflect PE1 as the destination. However, with ARP/ND synchronization based on a common ESI, the ARP/ND cache may be pre-populated with the local AC as destination for the host, should an AC failure occur on PE1. This achieves fast-convergence.

When a host moves to PE2 from the PE1 L2GW peer, the MAC mobility sequence number is incremented to signal to remote peers that a 'move' has occurred and the routing tables must be updated to PE2. This is required when an Access Protocol is running where the loop is broken between two CEs in the access and the L2GWs, and the host is no longer reachable from the PE1-side but now from the PE2-side of the access network.

6. Conclusion

EVPN Multi-Homing Mechanism for Layer-2 Gateway Protocols solves a true problem due to the wide legacy deployment of these access L2GW protocols in Service Provider networks. The current draft has the main advantage to be fully compliant with [RFC7432] and [I-D.ietf-bess-evpn-inter-subnet-forwarding].

7. Security Considerations

The same Security Considerations described in [RFC7432] and [I-D.ietf-bess-evpn-inter-subnet-forwarding] remain valid for this document.

8. Acknowledgements

Authors would like to thank Thierry Couture for valuable review and inputs with respect to access protocol deployments related to procedures proposed in this document.

9. IANA Considerations

There are no IANA considerations.

10. References

10.1. Normative References

[I-D.ietf-bess-evpn-inter-subnet-forwarding]
Sajassi, A., Salam, S., Thoria, S., Drake, J., and J. Rabadan, "Integrated Routing and Bridging in EVPN", draft-ietf-bess-evpn-inter-subnet-forwarding-11 (work in progress), October 2020.

[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>.

[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>.

[RFC7623] Sajassi, A., Ed., Salam, S., Bitar, N., Isaac, A., and W. Henderickx, "Provider Backbone Bridging Combined with Ethernet VPN (PBB-EVPN)", RFC 7623, DOI 10.17487/RFC7623, September 2015, <https://www.rfc-editor.org/info/rfc7623>.

10.2. Informative References

[RFC6378] Weingarten, Y., Ed., Bryant, S., Osborne, E., Sprecher, N., and A. Fulignoli, Ed., "MPLS Transport Profile (MPLS-TP) Linear Protection", RFC 6378, DOI 10.17487/RFC6378, October 2011, <https://www.rfc-editor.org/info/rfc6378>.

Authors' Addresses

Patrice Brissette
Cisco Systems
Ottawa, ON
Canada

Email: pbrisset@cisco.com

Ali Sajassi
Cisco Systems
USA

Email: sajassi@cisco.com

Luc Andre Burdet (editor)
Cisco Systems
Ottawa, ON
Canada

Email: lburdet@cisco.com

Daniel Voyer
Bell Canada
Montreal, QC
Canada

Email: daniel.voyer@bell.ca

BESS WorkGroup
Internet-Draft
Intended status: Standards Track
Expires: August 26, 2021

D. Rao
S. Agrawal
C. Filsfils
K. Talaulikar
Cisco Systems

February 22, 2021

BGP Color-Aware Routing(CAR)
draft-dskc-bess-bgp-car-00

Abstract

This document describes a BGP based routing solution to establish end-to-end intent-aware paths across a multi-domain service provider transport network. This solution is called BGP Color-Aware Routing (BGP CAR).

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 https://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 26, 2021.

Copyright Notice

Table of Contents

1. Introduction 3
1.1. Objectives 3
1.2. Requirements Language 3
2. Concepts 3
2.1. Color 4
2.2. Colored vs Color-Aware 4
2.3. Color Domains 4
2.4. BGP Color-Aware Routing 5
3. BGP extensions for CAR 5
3.1. Why a new SAFI is required 5
3.2. Data Model of New SAFI 5
3.3. Extensible, future-proof encoding 6
3.4. BGP CAR Family 6
3.4.1. BGP CAR SAFI NLRI Format 7
3.4.2. CAR NLRI Type 8
3.4.3. Local-Color-Mapping (LCM) Extended Community 12
3.5. BGP transport CAR Route Origination 13
3.6. BGP CAR Next-Hop Processing 13
3.6.1. Validation 13
3.6.2. Resolution 14
3.7. AIGP Metric Computation 15
3.8. Multiple color domains 15
4. Steering a Colored Service Route onto an (E, C) BGP CAR route 17
4.1. E2E BGP transport CAR intent realized using IGP FA . . . 17
4.2. E2E BGP transport CAR intent realized using SR Policy . . 19
4.3. BGP transport CAR intent realized in a section of the network 21
4.4. Transit network domains that do not support CAR 23
5. Color Mapping Scenarios 24
5.1. Single color domain containing network domains with N:N color distribution 24
5.2. Single color domain containing network domains with N:M color distribution 25
5.3. Multiple color domains 25
6. Intent Use-cases 26
7. Scaling 26
7.1. Data plane does not have to scale to Colors * PEs 27
7.1.1. Inter-Domain Hop by hop BGP CAR for PE routes 27
7.1.2. Hierarchical Design with Next-hop self at ingress domain BR 29
7.1.3. Hierarchical Design with Next Hop Unchanged at ingress domain BR 31
7.2. Automated Emulated-Pull Model to learn BGP CAR (PE, C) . 33

7.2.1. Subscription based BGP CAR Signaling 34
7.3. Additional Design Options 36
7.3.1. Anycast SID for transit inter-domain nodes 36
7.3.2. Anycast SID for transport color endpoints i.e PEs . . 36
7.4. Convergence 36
8. Interworking Scenarios 37
9. Fault Handling 37
10. IANA Considerations 37
10.1. BGP CAR NLRI Types Registry 37
10.2. BGP CAR NLRI TLV Registry 38
10.3. Guidance for Designated Experts 38
10.4. BGP Extended Community Registry 38
11. Security Considerations 38
12. Acknowledgements 39
13. References 39
13.1. Normative References 39
13.2. Informative References 41
Authors' Addresses 42

1. Introduction

1.1. Objectives

- Address the Transport problem statement and requirements described in [dskc-bess-bgp-car-problem-statement]

- Define an inter-domain BGP-based Color-Aware Routing proposal to steer traffic for a C-colored service route V/v from a PE onto a BGP color-aware path to (PE, C)

  * Provide an alternative to the SR-PCE based design [I-D.ietf-spring-segment-routing-policy]

1.2. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

2. Concepts

A refresher on core concepts used in this document, some of which are described in [BGP-CAR-Problem-Statement]

2.1. Color

The solution must reuse the color concept defined in [I-D.ietf-spring-segment-routing-policy]. The color is a 32-bit numerical value that, today, associates an SR-policy with an intent (e.g., low latency).

2.2. Colored vs Color-Aware

- Colored: Egress PE PE2 colored its BGP VPN route V/v to indicate the intent that it requests for the traffic bound to V/v.
- Color-Aware: a new BGP solution which signals multiple "ways" to reach a given destination (e.g. PE2)
- Steering a colored VPN route to a color-aware route
  - If PE2 signals a VPN route V/v with color C
  - If PE1 installs that VPN route
  - If PE1 learns about a BGP Color-Aware Route R/r to PE2 for color C
  - Then PE1 steers packets destined to V/v via R/r

2.3. Color Domains

A domain (or network domain) generally refers to a unit of isolation or hierarchy in the network topology; for example, access, metro and or core domains. From a routing perspective, a domain may have a distinct IGP area or instance; or a distinct BGP ASN.

With the use of a 'Color' to represent intent, it is useful to describe the distinct concept of a color domain.

A color domain refers to a collection of one or more network domains with a single, consistent color-to-intent mapping.

When a route gets distributed into a domain with a different color-to-intent mapping scheme, the color associated with the route needs to be mapped to the locally assigned value in that domain.

Deployments under a single authority are expected to use the same color-to-intent mapping across all network domains.

A solution must distinguish the actual protocol boundaries (IGP, ASN) from the color domain boundaries.

2.4. BGP Color-Aware Routing

In the remainder of this document, the BGP Color-Aware Routing Solution is referred to as BGP CAR.

3. BGP extensions for CAR

This section analyzes the requirements for BGP CAR and proposes extensions, specifically for Transport Color-Aware-Routing

3.1. Why a new SAFI is required

- Existing BGP SAFI for BGP-LU (AFI 1 or 2 and SAFI 4) signals transport destination (likely PE loopback) with just an IP prefix in NLRI.
  - BGP CAR needs to signal multiple "ways" to reach a transport destination, each for a different intent or color; i.e., it needs a Color-Aware NLRI
- Hence, a new SAFI is needed for BGP Transport CAR which can encode IP prefix and Color

3.2. Data Model of New SAFI

The essential elements of the data model for the transport CAR SAFI are as follows:

- NLRI Key: E, C
  - E: IPv4/IPv6 prefix: Prefix is unique in inter-domain network.
  - Color: Distinguishes per-intent instances of a prefix. Additionally, it signals the intent provided by with the route in originator color domain. 32-bit value as per [I-D.ietf-spring-segment-routing-policy]
- NLRI non key data
  - To encode multiple encapsulations with efficient packing
    - MPLS label stack
    - Label Index (hint for label allocation from SRGB - same as BGP SR Prefix SID Attr Label Index TLV)
    - SRv6 SID(s)

+ Etc.

- Next-Hop
  - BGP Next-Hop
- AIGP Metric
  - To accumulate color/intent specific metric across domains
  - AIGP Attribute provides extensibility via TLVs, enabling definition of additional metric semantics for a color as needed for an intent
- Local-Color-Mapping Extended-Community (LCM-EC)
  - 32-bit Color value
  - Optional, used when a CAR route propagates across domains with different or inconsistent color-to-intent mapping schemes

The detailed protocol operations for these elements are described in later sections.

3.3. Extensible, future-proof encoding

Since a new SAFI is required, it is prudent to define an extensible encoding so that additional use-cases can be supported in future, without imposing limitations

Key design aspects for an extensible encoding:

Encode a NLRI (Route) Type field. This provides extensibility to add new NLRI formats for new route-types

Encode a key length field. This enables handling unsupported route-types opaquely, enabling transitivity via RRs

Define non-key NLRI data using TLVs. This enables flexible and efficient encoding of data such as multiple encapsulations

3.4. BGP CAR Family

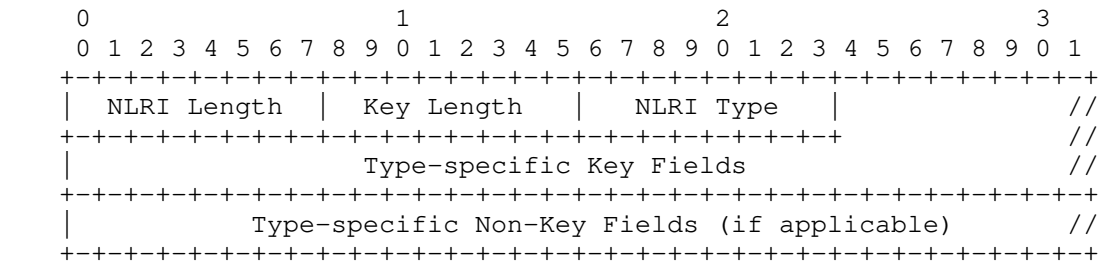
BGP CAR leverages the BGP multi-protocol extensions [RFC4760] and uses the MP_REACH_NLRI and MP_UNREACH_NLRI attributes for routes updates by using the SAFI value TBD1 along with AFI 1 for IPv4 prefixes and AFI 2 for IPv6 prefixes.

BGP speakers MUST use BGP Capabilities Advertisement to ensure support for processing of BGP CAR updates. This is done as specified in [RFC4760], by using capability code 1 (multi-protocol BGP), with AFI 1 and 2 (as required) and SAFI TBD1.

The sub-sections below specify the generic encoding of the BGP CAR NLRI followed by the encoding for specific NLRI types introduced in this document.

3.4.1. BGP CAR SAFI NLRI Format

The generic format for the BGP CAR Address-Family NLRI is shown below:



where:

- NLRI Length: 1 octet field that indicates the length in octets of the NLRI excluding the NLRI Length field itself.
- Key Length: 1 octet field that indicates the length in octets of the NLRI type-specific key fields. Key length MUST be at least 2 less than the NLRI length.
- NLRI Type: 1 octet field that indicates the type of the BGP Transport CAR NLRI.
- Type-Specific Key Fields: Depend on the NLRI type and of length indicated by the Key Length.
- Type-Specific Non-Key Fields: optional and variable depending on the NLRI type. The NLRI encoding allows for encoding of specific non-key information associated with the route (i.e. the key) as part of the NLRI for efficient packing of BGP updates.

The indication of the key length enables BGP Speakers to determine the key portion of the NLRI and use it along with the NLRI Type field in an opaque manner for handling of unknown or unsupported NLRI

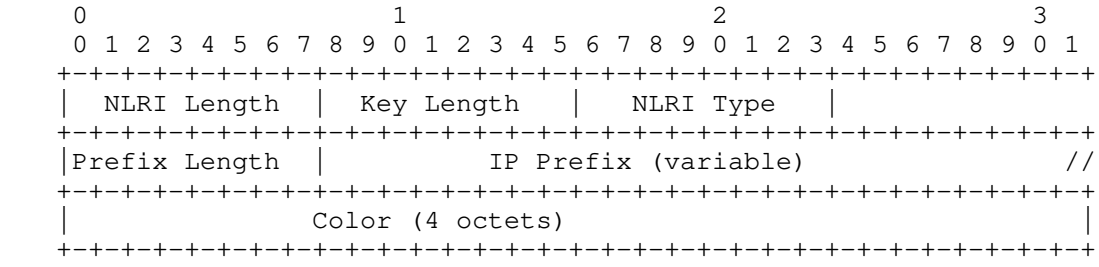
types. This can help Route Reflectors (RR) to propagate NLRI types introduced in the future in a transparent manner.

The NLRI encoding allows for encoding of specific non-key information associated with the route (i.e. the key) as part of the NLRI for efficient packing of BGP updates.

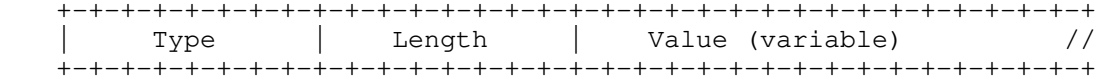
The non-key portion of the NLRI MUST be omitted while carrying it within the MP_UNREACH_NLRI when withdrawing the route advertisement.

3.4.2. CAR NLRI Type

The Color-Aware Routes NLRI Type is used for advertisement of color-aware routes and has the following format:



Followed by optional TLVs encoded as below:



where:

- NLRI Length: variable
- Key Length: variable
- NLRI Type: 1
- Type-Specific Key Fields: as below
  - Prefix Length: 1 octet field that carries the length of prefix in bits. Length MUST be less than or equal to 32 for IPv4 (AFI=1) and less than or equal to 128 for IPv6 (AFI=2).
  - IP Prefix: IPv4 or IPv6 prefix (based on the AFI). A variable size field that contains the most significant octets of the prefix, i.e., 1 octet for prefix length 1 to 8, 2 octets for

prefix length 9 to 16, 3 octets for prefix length 17 up to 24, 4 octets for prefix length 25 up to 32, and so on. The size of the field MUST be less than or equal to 4 for IPv4 (AFI=1) and less than or equal to 16 for IPv6 (AFI=2).

* Color: 4 octets that contains color value associated with the prefix. It distinguish different instances of a prefix. Additionally, it signals the intent associated with the route in originator color domain.

o Type-Specific Non-Key Fields: specified in the form of optional TLVs as below:

* Type: 1 octet field that contains the type of the non-key TLV

* Length: 1 octet field that contains the length of the value portion of the non-key TLV in terms of octets

* Value: variable length field as indicated by the length field and to be interpreted as per the type field.

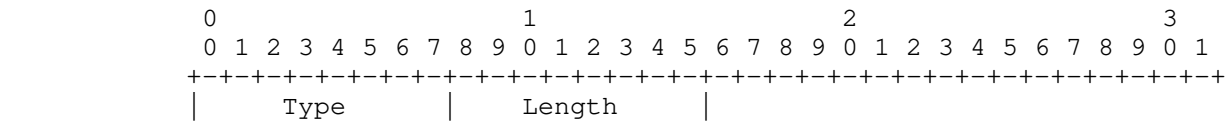
The prefix is routable across the administrative domain where BGP Transport CAR is deployed. It is possible that the same prefix is originated by multiple BGP Transport CAR speakers in the case of anycast addressing or multi-homing.

The Color is introduced to enable multiple route advertisements for the same prefix. The color is associated with an intent (e.g. low-latency) in originator color-domain.

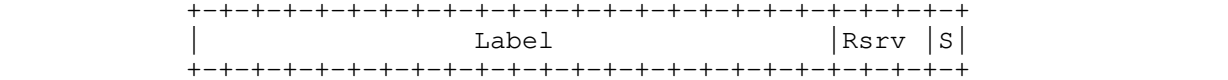
The following sub-sections specify the non-key TLVs associated with the Color-Aware Routes NLRI type.

3.4.2.1. Label TLV

The Label TLV is used for advertisement of color-aware routes along with their MPLS labels and has the following format:



Followed by one (or more) Labels encoded as below:



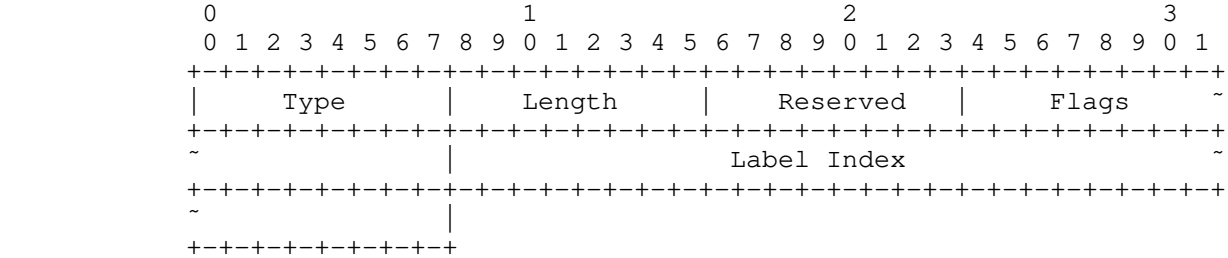
where:

- Type : 1
- Length: variable, MUST be a multiple of 3
- Label Information: multiples of 3 octet fields to convey the MPLS label(s) associated with the advertised color-aware route. It is used for encoding a single label or a stack of labels as per procedures specified in [RFC8277].

When a BGP Transport CAR speaker is propagating the route further after setting itself as the nexthop, it allocates a local label for the specific prefix and color combination which it updates in this TLV. It also MUST program a label cross-connect that would result in the label swap operation for the incoming label that it advertises with the label received from its best-path router(s).

3.4.2.2. Label Index TLV

The Label Index TLV is used for advertisement of Segment Routing MPLS (SR-MPLS) Segment Identifier (SID) [RFC8402] information associated with the labelled color-aware routes and has the following format:



where:

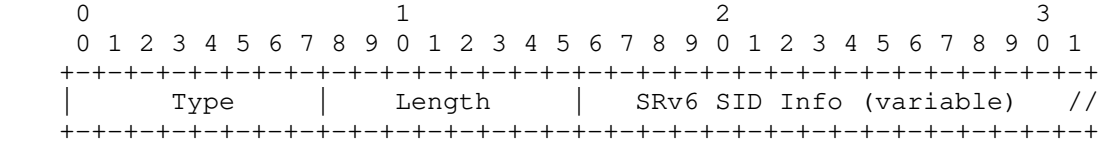
- Type : 2
- Length: 7
- Reserved: 1 octet field that MUST be set to 0 and ignored on receipt.
- Flags: 2 octet field that maps to the Flags field of the Label-Index TLV of the BGP Prefix SID Attribute [RFC8277].
- Label Index: 4 octet field that maps to the Label Index field of the Label-Index TLV of the BGP Prefix SID Attribute [RFC8277].

This TLV provides the equivalent functionality as Label-Index TLV of [RFC8669] for Transport CAR in SR-MPLS deployments. The BGP Prefix SID Attribute SHOULD be omitted from the labeled color-aware routes when the attribute is being used to only convey the Label Index TLV for better BGP packing efficiency.

When a BGP Transport CAR speaker is propagating the route further after setting itself as the nexthop, it allocates a local label for the specific prefix and color combination. When the received update has the Label Index TLV, it SHOULD use that hint to allocate the local label from the SR Global Block (SRGB) using procedures as specified in [RFC8669].

3.4.2.3. SRv6 SID TLV

BGP Transport CAR can be also used to setup end-to-end color-aware connectivity using Segment Routing over IPv6 (SRv6) [RFC8402]. [I-D.ietf-spring-srv6-network-programming] specifies the SRv6 Endpoint behaviors (e.g. End PSP) which MAY be leveraged for BGP CAR with SRv6.The SRv6 SID TLV is used for advertisement of color-aware routes along with their SRv6 SIDs and has the following format:



where:

- Type : 3
- Length: variable, MUST be either less than or equal to 16, or be a multiple of 16

- SRv6 SID Information: field of size as indicated by the length that either carries the SRv6 SID(s) for the advertised color-aware route as one of the following:

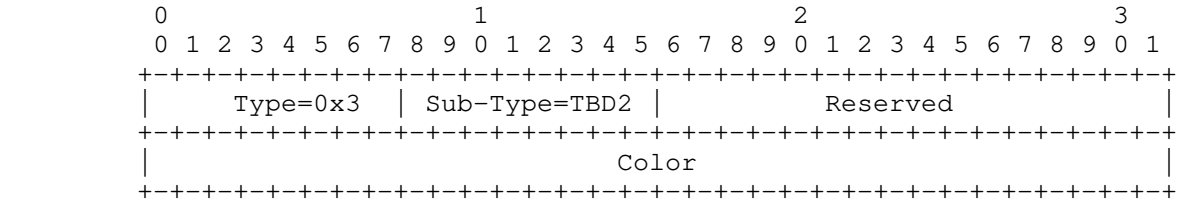
  * A single 128-bit SRv6 SID or a stack of 128-bit SRv6 SIDs

  * A transposed portion (refer [I-D.ietf-bess-srv6-services]) of the SRv6 SID that MUST be of size in multiples of one octet and less than 16.

The BGP color-aware route update for SRv6 MUST include the BGP Prefix-SID attribute along with the TLV carrying the SRv6 SID information as specified in [I-D.ietf-bess-srv6-services] when using the transposition scheme of encoding for packing efficiency of BGP updates.

3.4.3. Local-Color-Mapping (LCM) Extended Community

This document defines a new BGP Extended Community called "LCM". The LCM is a Transitive Opaque Extended Community with the following encoding:



where:

- Type: 0x3

- Sub-Type: TBD2.

- Reserved: 2 octet of reserved field that MUST be set to zero on transmission and ignored on reception.

- Color: 4-octet field that carries the 32-bit color value.

When CAR route crosses the original color domain boundary, LCM EC is added. LCM EC associate the local color mapping for the intent (e.g. low latency) in transit or remote color domain. Note: reminder "BGP CAR needs to signal multiple "ways" to reach a transport destination, each for a different intent or color". Original BGP CAR route (E, C) still signal multiple "ways" to reach E, but once LCM EC is added, intent is carried in it and not by C in NLRI.

The LCM EC MAY be used for filtering of BGP CAR routes and/or for applying routing policies for the intent.

3.5. BGP transport CAR Route Origination

- BGP CAR routes may be originated from a node via local injection (e.g., loopback)
  - Routes will be advertised with Implicit-NULL (or equivalent), and optionally may include Label-Index
- BGP Transport CAR routes may also be originated from a node, sourced from another mechanism
  - IGP Flexible Algorithm(FA) [I-D.ietf-lsr-flex-algo] redistribution
    - FA identifier mapping to BGP transport CAR color or vice versa by local policy. This will allow redistribution of prefixes, prefix SID between FA and BGP CAR
  - SR Policy [I-D.ietf-spring-segment-routing-policy]
    - An SR Policy is identified through the tuple (color, E) where color is a 32-bit numerical value that associates the SR Policy with an intent (e.g. low-latency). When color of SR policy maps directly into BGP CAR color because of same intent or through some local configuration, endpoint of policy can be advertised in BGP Transport CAR to rest of network for end to end color-aware transport connectivity.
  - BGP-LU [RFC8277]
    - Redistribution between BGP-LU and BGP CAR color table and vice versa. Most likely(but not limited) color represents best effort intent in BGP CAR domain. This provide connectivity between BGP-LU only domain and BGP CAR domain with best effort color-awareness.

3.6. BGP CAR Next-Hop Processing

3.6.1. Validation

- Validation of BGP Next-Hop: Reachability verified via underlying routing control plane. Local policy should be provided to verify it
  - Strictly within intent of BGP CAR route i.e "color"

    * Default routing table

    * Skip it when updates are propagated out of band

- Validation of Encapsulation: Validate data-plane availability of encapsulation before using and propagating further.

- Validation of the intent: Validate the intent provided by the underlying transport (e.g., via OAM), where applicable.

3.6.2. Resolution

BGP color-aware routes may be resolved over various intra-domain and inter-domain mechanisms that provide connectivity to the BGP next-Hop with the desired intent

- Leverage the notion of "color" in NLRI or LCM-EC to determine the matching intent-aware mechanism and instance.

- Leverage ODN/AS mechanisms where needed, for instance to use SR-PCE for an SR-policy to the BGP next-hop

- Flexible for all encapsulations

    * (SR-)MPLS

    * SRv6, IPv4/IPv6, etc.

- Flexible over various underlay mechanisms

    * SR Policy: Color from BGP CAR route and policy endpoint from BGP CAR Next hop

    * IGP Flexible Algorithm: Color from BGP CAR mapped to Flex Algo by configuration.

    * IGP/BGP best effort (SR, LDP, RSVP-TE, BGP-LU etc.)

    * BGP CAR in hierarchical CAR design

- Support selection preference among available mechanisms

- Fallback to a different color or best effort path

3.7. AIGP Metric Computation

- BGP CAR nodes update the Accumulated IGP (AIGP) Attribute as the BGP CAR route propagates across the network.
- The value set (or appropriately incremented) in the AIGP TLV corresponds to the metric associated with the underlying intent of the color. Example. when the color is associated with a low-latency path, the metric value is set based on the delay metric.
  - Information regarding the metric type used by the underlying intra-domain mechanism can also be set
- If BGP CAR routes traverse across a discontinuity in the transport path for a given intent, add penalty in accumulated IGP
- If BGP CAR routes traverse across a discontinuity in the transport path for a given intent, the AIGP TLV is used to indicate this e.g. with a discontinuity bit.
- AIGP metric computation is recursive.
- To avoid continuous IGP metric churn causing end to end BGP CAR churn, implementation should provide thresholds to trigger AIGP update.
- Additional AIGP extensions may be defined to signal state for specific use-cases.
  - MSD along the BGP CAR advertisement.
  - Minimum MTU along the BGP CAR advertisement.

3.8. Multiple color domains

- When BGP CAR routes get distributed to a domain with a different color-to-intent mapping, the color signaled must be re-mapped to the local color being used within the receiving domain
- A key requirement to consider is the separation and independence of the administrative authority in different color domains.
  - Each color domain needs to use its own local color. The route can traverse multiple such color domains where the color mappings change
- This requirement is addressed by the following steps :

    * The NLRI of the CAR route is never changed

        + E is globally unique. Hence even if C is local-domain significant, E-C in that order is globally unique

    * Each color domain needs to use its own local color. The route can traverse multiple such color domains where the color mappings change

        + To address this requirement, a border node in a color domain encodes its local color mapping in a Local-Color-Mapping Extended-Community when sending the route to a peer in a different color domain

        + The border routers within the receiving domain map the received LCM-EC Color value to a local color assigned for that intent and rewrite the LCM-EC

        + The nodes within the receiving domain use the local color encoded in the LCM-EC for next-hop resolution and BGP CAR route installation

- The LCM-EC is only used when a CAR route needs to be distributed across a color domain boundary. The likely case (color consistency) is supported with the simplest and most efficient scheme (E, C) key and no LCM-EC.

- Example: When going from a domain D1 to a domain D2 where D1 uses the color scheme is the NLRI but D2 uses another color scheme, then on the peering session from D1 to D2, D1 on egress or D2 on ingress inserts the LCM-EC which carries the mapped local color that will be used in D2. When the route travels from D2 to a domain D3 which uses the color scheme in the NLRI then either the LCM-EC is kept but its internal C is remapped to the color scheme of D3 or the LCM-EC is removed

- Color intent encoded in the service routes in the Color Ext-community should also be re-mapped consistently

- A color boundary is typically well-defined, at a BGP peering session on a border Router, and at a service/transport RR.

- A color domain may extend across one or more BGP ASNs

4. Steering a Colored Service Route onto an (E, C) BGP CAR route

BGP colored service routes (i.e., containing Color extended community [I-D.ietf-idr-tunnel-encaps]) resolve over BGP transport CAR routes i.e. (E, C), conceptually identical to the steering mechanism used with SR Policies.

All steering options are supported: Automated, on-demand steering, per-destination, per-flow, CO-only

Co-existence with SR-policy based steering is also supported

By default, when BGP CAR is enabled, a BGP CAR route will be preferred.

Similarly, if an IGP Flex-Algo route exists, typically for an intra-domain endpoint, it is preferred over a BGP CAR route to the same endpoint.

A node may support a local policy to set the preferences between different mechanisms.

The following sub-sections illustrate example scenarios of Colored Service Route Steering over E2E BGP CAR resolving over different intra-domain mechanisms

4.1. E2E BGP transport CAR intent realized using IGP FA

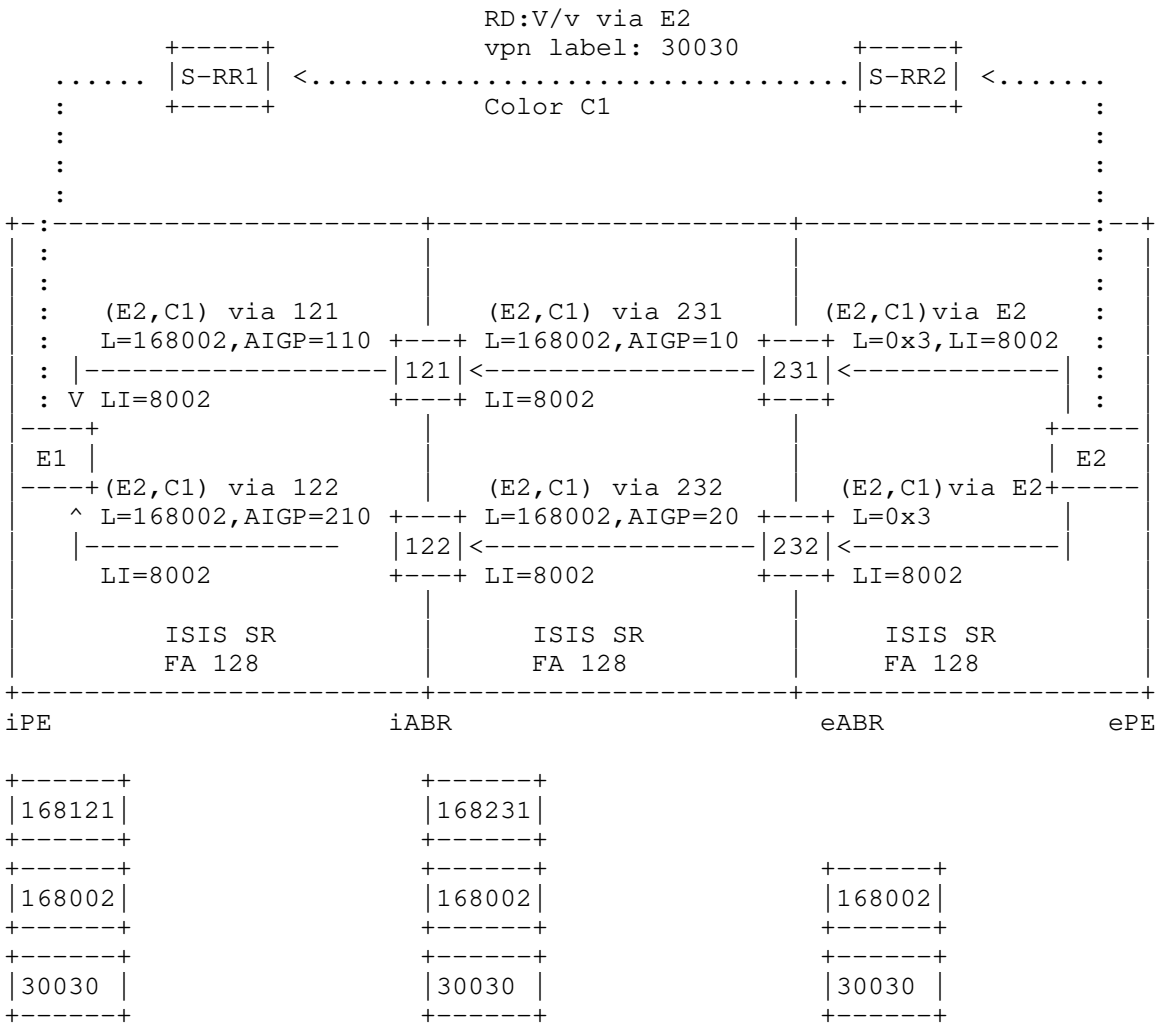


Figure 1: BGP FA Aware transport CAR path

Use case: Provide end to end intent for service flows.

- With reference to the topology above:
  - IGP FA 128 is running in each domain.
  - Egress PE E2 advertises a VPN route RD:V/v colored with (color extended community) C1 to steer traffic to BGP transport CAR (E2, C1). VPN route propagates via service RRs to ingress PE E1.

    * BGP CAR route (E2, C1) with next-hop, label-index and label as shown above are advertised through border routers in each domain.

    * Local policy on each hop maps intent C1 to resolve CAR route next-hop over IGP FA 128 of the domain. AIGP attribute influences BGP CAR route best path decision as per [RFC7311]. BGP CAR label swap entry is installed that goes over FA 128 LSP to next-hop providing intent in each IGP domain. Update AIGP metric to reflect FA 128 metric to next-hop.

    * Ingress PE E1 learns CAR route (E2, C1). It steers colored VPN route RD:V/v into (E2, C1)

  o Important:

    * IGP FA 128 top label provides intent in each domain.

    * BGP CAR label (e.g. 168002) carries end to end intent. Thus stitches intent over intra domain FA 128.

4.2. E2E BGP transport CAR intent realized using SR Policy

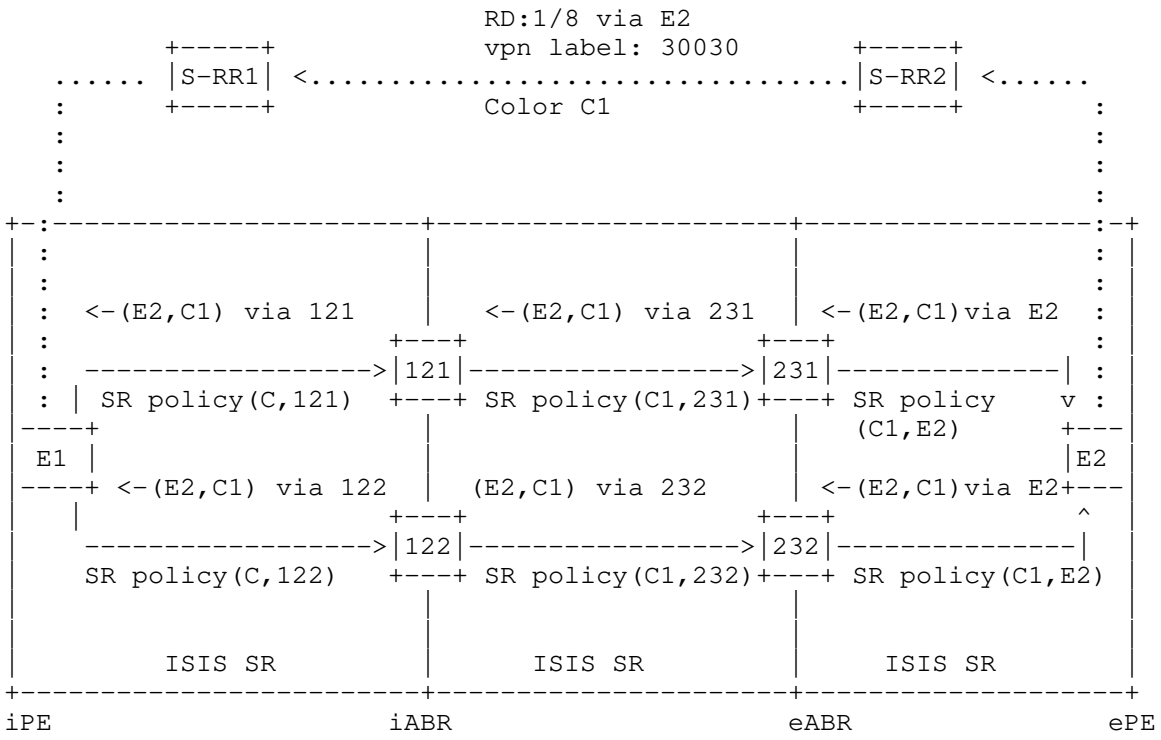


Figure 2: BGP SR policy Aware transport CAR path

Use case: Provide end to end intent for service flows

- With reference to the topology above:
  - SR Policy provide intra domain intent.
  - Egress PE E2 advertises a VPN route RD:V/v colored with (color extended community) C1 to steer traffic to BGP transport CAR (E2, C1). VPN route propagates via service RRs to ingress PE E1.
  - BGP CAR route (E2, C1) with next-hop, label-index and label as shown above are advertised through border routers in each domain.
  - Local policy on each hop maps intent C1 to resolve CAR route next-hop over an SR policy(C1, next-hop). BGP CAR label swap entry is installed that goes over SR policy segment list.

    * Ingress PE E1 learns CAR route (E2, C1). It steers colored VPN route RD:V/v into (E2, C1).

  o Important:

    * SR policy provides intent in each domain.

    * BGP CAR label (e.g. 168002) carries end to end intent. Thus stitches intent over intra domain SR policies.

4.3. BGP transport CAR intent realized in a section of the network

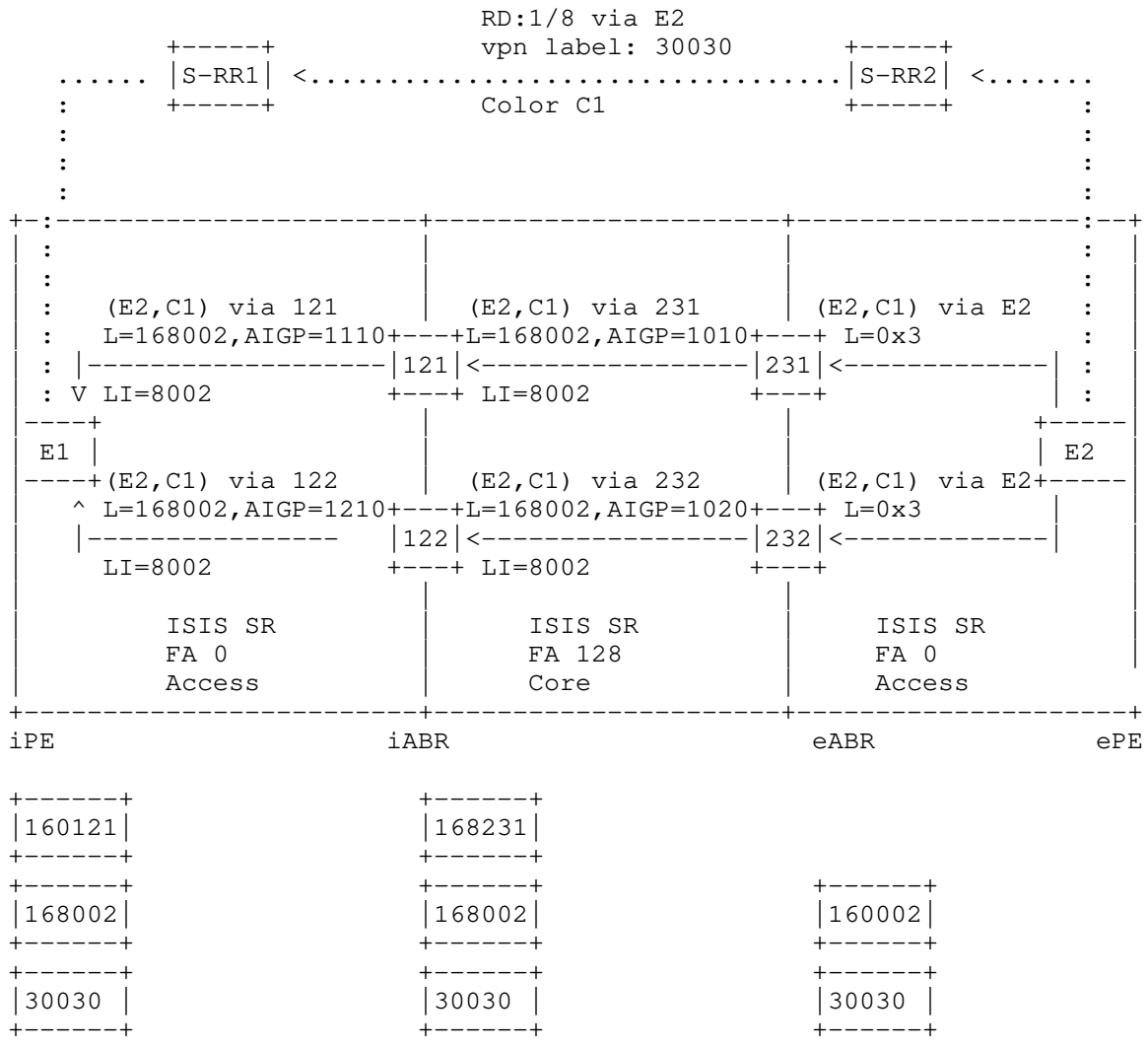


Figure 3: BGP Hybrid FA Aware transport CAR path

Use case: Provide intent for service flows only in Core domain.

- With reference to the topology above:
  - IGP FA 128 is only enabled in Core (e.g. WAN network). Access only has base algo 0.
  - Egress PE E2 advertises a VPN route RD:V/v colored with (color extended community) C1 to steer traffic to BGP transport CAR

(E2, C1). VPN route propagates via service RRs to ingress PE E1.

* BGP CAR route (E2, C1) with next-hop, label-index and label as shown above are advertised through border routers in each domain.

* Local policy on 231 and 232 maps intent C1 to resolve CAR route next-hop over IGP base algo 0 in right access domain. BGP CAR label swap entry is installed that goes over algo 0 LSP to next-hop. Update AIGP metric to reflect algo 0 metric to next-hop most likely with additional penalty.

* Local policy on 121 and 122 maps intent C1 to resolve CAR route next-hop learnt from Core domain over IGP FA 128. BGP CAR label swap entry is installed that goes over FA 128 LSP to next-hop providing intent in Core IGP domain.

* Ingress PE E1 learns CAR route (E2, C1). It maps intent C1 to resolve CAR route next-hop over IGP base algo 0. It steers colored VPN route RD:V/v into (E2, C1)

o Important:

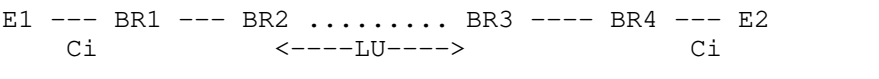
* IGP FA 128 top label provides intent in Core domain.

* BGP CAR label (e.g. 168002) carries intent from PEs which is realized in core domain

4.4. Transit network domains that do not support CAR

o In a brownfield deployment, color-aware paths between two PEs may need to go through a transit domain that does not support CAR. Example include an MPLS LDP network with IGP best-effort; or a BGP-LU based multi-domain network. MPLS LDP network with best effort IGP can adopt above scheme. Below is the example for BGP LU.

o Reference topology:



* Network between BR2 and BR3 comprises of multiple BGP-LU hops (over IGP-LDP domains).

* E1, BR1, BR4 and E2 are enabled for BGP CAR, with Ci colors

    * BR1 and BR2 are directly connected; BR3 and BR4 are directly connected

- BR1 and BR4 form an over-the-top peering (via RRs as needed) to exchange BGP CAR routes
- BR1 and BR4 also form direct BGP-LU sessions to BR2 and BR3 respectively, to establish labeled paths between each other through the BGP-LU network
- BR1 recursively resolves the BGP CAR next-hop for CAR routes learnt from BR4 via the BGP-LU path to BR4
- BR1 signals the transport discontinuity to E1 via the AIGP TLV, so that E1 can prefer other paths if available
- BR4 does the same in the reverse direction
- Thus, the color-awareness of the routes and hence the paths in the data plane are maintained between E1 and E2, even if the intent is not available within the BGP-LU island
- A similar design can be used for going over network islands of other types

5. Color Mapping Scenarios

There are a variety of deployment scenarios that arise w.r.t different color mappings in an inter-domain environment. This section attempts to enumerate them to provide clarity into the usage of the color related protocol constructs.

5.1. Single color domain containing network domains with N:N color distribution

All network domains (ingress, egress and all transit domains) are enabled for the same N colors

A color may of course be realized by different technologies in different domains as described above

The N intents are both signaled end-to-end via BGP CAR routes; as well as realized in the data plane

Section 4.1 is an example of this case

5.2. Single color domain containing network domains with N:M color distribution

Certain network domains may not be enabled for some of the colors, but may still be required to provide transit.

When a (E, C) route traverses a domain where color C is not available, the operator may decide to use a different intent of color c that is available in that domain to resolve the next-hop and establish a path through the domain

- The next-hop resolution may occur via paths of any intra-domain protocol or even via paths provided by BGP CAR
- The next-hop resolution color c may be defined as a local policy at ingress or transit nodes of the domain
- It may also be automatically signaled from egress border nodes by attaching a color extended community with value c to the BGP CAR routes

Hence, routes of N colors may be resolved via a smaller set of M colored paths in a transit domain, while preserving the original intent end-to-end.

Any ingress PE that installs a service (VPN) route with a color C, must have C enabled locally to install IP routes to (E, C) and resolve the service route next-hop

A degenerate case of these scenario is where a transit domain does not support any color. Section 4.3 describes an example of this case

5.3. Multiple color domains

When the routes are distributed between domains with different color-to-intent mapping schemes, both N:N and N:M ratios are possible, although an N:M mapping is more likely to occur.

Reference topology:

```
D1 ----- D2 ----- D3
 C1      C2       C3
```

- C1 in D1 maps to C2 in D2 and to C3 in D3
- BGP CAR is enabled in all three domains

The reference topology above is used to elaborate on the design described in Section-X

When the route originates in color domain D1 and gets advertised to a different color domain D2, following procedures apply:

The original intent in BGP CAR route is preserved; i.e. route is (E, C1)

A BR of D1 attaches LCM-EC with value C1 when advertising to a BR in D2

A BR in D2 receiving (E, C1) maps C1 in received LCM-EC to local color, say C2

Within D2, this LCM-EC value of C2 is used instead of the Color in CAR route NLRI (E, C1). This applies to all procedures described in the earlier section for a single color domain, such as next-hop resolution and route installation.

A colored service route V/v originated in domain D1 with next-hop E and color C1 will also have its color extended-community value re-mapped to C2, typically at a service RR

On an ingress PE in D2, V/v will resolve via C2

When a BR in D2 advertises the route to a BR in D3, a similar process is followed

6. Intent Use-cases

This section will describe how BGP CAR addresses the various intent use-cases described in [ref:dskc-bess-bgp-car-problem-statement]. Details will be added in a later revision of the document.

7. Scaling

A key requirement of [ref:dskc-bess-bgp-car-problem-statement] is scale, specifically:

- No intermediate node dataplane should need to scale to (Colors * PEs)
- No node should learn and install a BGP CAR route to (E,C) if it does not install a Colored service route to E
  - An intermediate node may learn a BGP CAR route to (E, C) in control plane if it is an inline RR to an ingress PE

* An intermediate node may learn and install a BGP CAR route to (E, C) if it is set up to be the next-hop for an ingress PE that installs the BGP CAR route

7.1. Data plane does not have to scale to Colors * PEs

Depending on the scale of the network as well as the constraints associated with the nodes at different tiers, an appropriate design should be adopted. Three design variations are illustrated below.

7.1.1. Inter-Domain Hop by hop BGP CAR for PE routes

Reference topology is shown below, with the BGP signaling and the resulting BGP and example IGP label stack at different hops

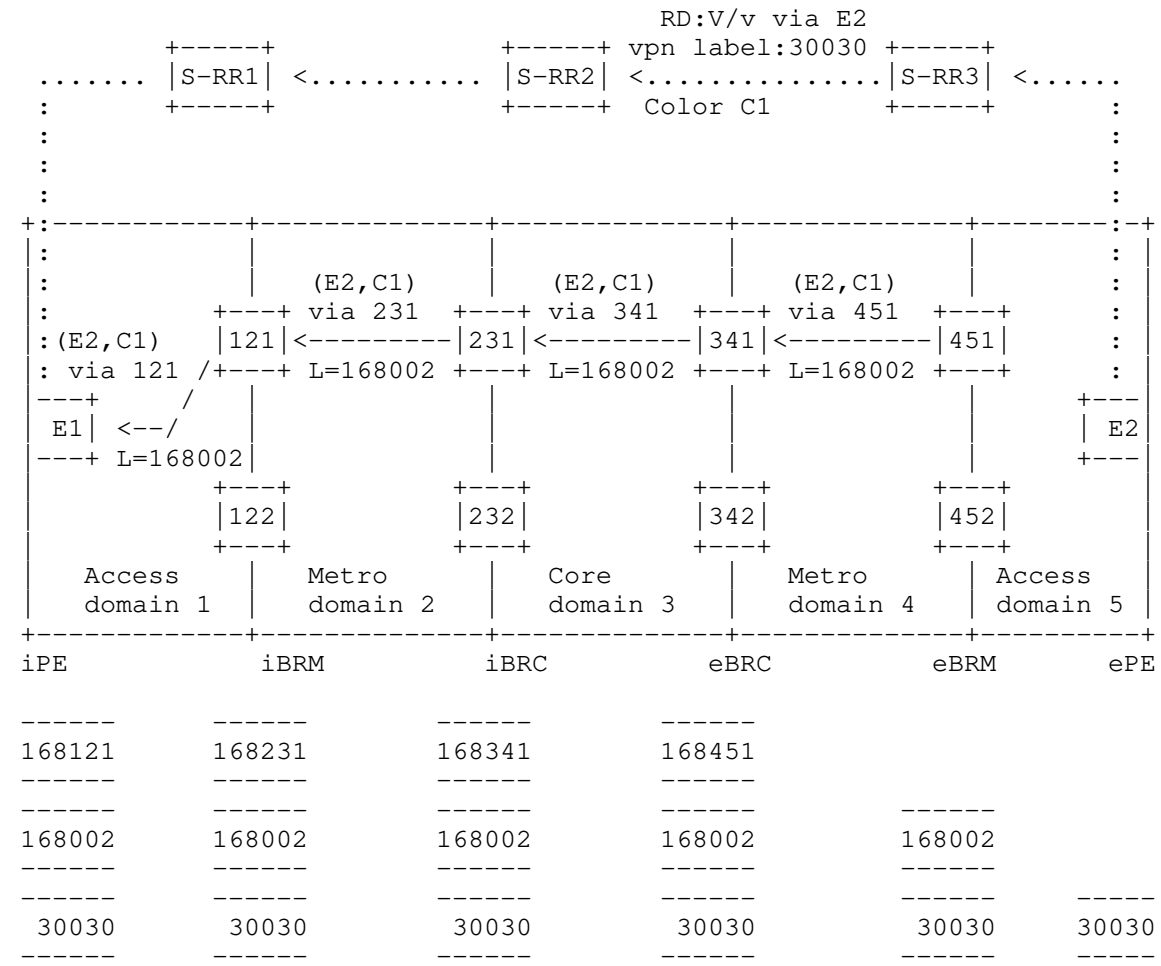


Figure 4: Single BGP transport CAR level

- With reference to the topology above:
  - Consider egress PE E2 advertises a VPN (service) route RD:V/v that propagates via service RRs to ingress PE E1.
  - A BGP CAR route (E2, C1) is advertised by egress BRM node 451. The route may be sourced locally, for instance by redistribution from an IGP-FA, and is distributed hop-by-hop through egress Metro, Core, ingress Metro to Access

    * Node 451, 341, 231 and 121 learns BGP CAR route (E2, C1). Each allocate local label and program swap entry in forwarding and set itself as next-hop.

    * E1 receives route. It recursively resolves (E2, C1) to build an outgoing label/SID stack to forward via nodes 121

  o This is the simplest design, with a single BGP transport CAR level

  o This results in the minimum label/SID stack at each inter-domain hop. However, it can significantly build up the scale overhead on the core BRs, and can easily exceed the FIB capacity as well as the MPLS label space on these nodes.

  o A subscription based Emulated-Pull solution is required with this flat design to enable all the intermediate nodes to be able to avoid learning and installing all the (PE, C) entries in the network.

7.1.2. Hierarchical Design with Next-hop self at ingress domain BR

Reference topology is shown below, with the BGP signaling and the resulting BGP and example IGP label stack at different hops

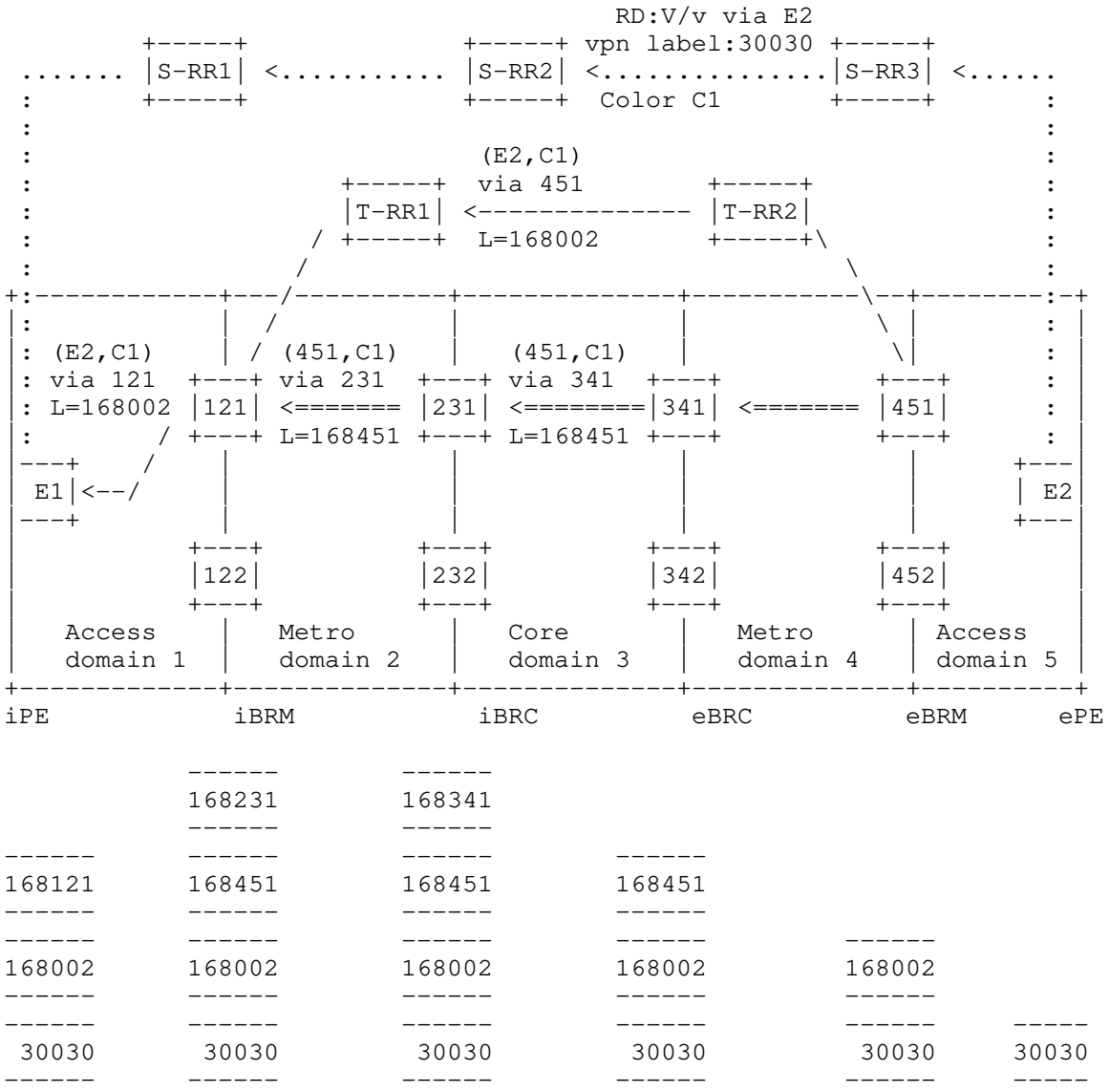


Figure 5: Heirarchical BGP transport CAR, NH at iBR

- With reference to the topology above:
  - Consider egress PE E2 advertises a VPN (service) route RD:V/v that propagates via service RRs to ingress PE E1.

* A BGP CAR route (E2, C1) is also advertised by egress BRM node 451. The route may be sourced locally, for instance by redistribution from an IGP-FA, and is distributed via a Transport RR plane.

* Ingress BRM node 121 learns about BGP CAR route (E2, C1) via node 451.

* Node 121 also learns about BGP CAR route (451, C1) via node 231.

* Node 121 advertise (E2, C1) received from T-RR to E1 with next-hop as it-self. It recursively resolves (E2, C1) to build an outgoing label/SID stack to forward traffic to (E1, C1) via (451, C1)

* (451, C1) is not advertised to node 121

* E1 receives route. It recursively resolves (E2, C1) to build an outgoing label/SID stack to forward via nodes 121

* Ingress BRM node 121 needs to install data plane entry for (451, C1), and for (E2, C1).

- This hierarchical design avoids the need for core BRs to learn and install entries for (PE, C)

- An ingress BR (e.g., node 121) advertises the received remote (PE, C) routes to it's local ingress PE, setting next-hop to itself

  * Hence, the ingress BR need to install (PE, C) entries for egress PEs that it's local ingress PEs have installed BGP CAR routes for, as well as support a swap and push operation.

- This design keeps simple label programming on the ingress PE i.e. like single BGP transport CAR level. It is not exposed to hierarchical BGP CAR design at ingress BRM

- A subscription based Emulated-Pull model should be used with this design if the ingress BR has limited FIB capacity, and should only learn and install the necessary subset of (PE, C) routes.

7.1.3. Hierarchical Design with Next Hop Unchanged at ingress domain BR

Reference topology is shown below, with the BGP signaling and the resulting BGP and example IGP label stack at different hops.

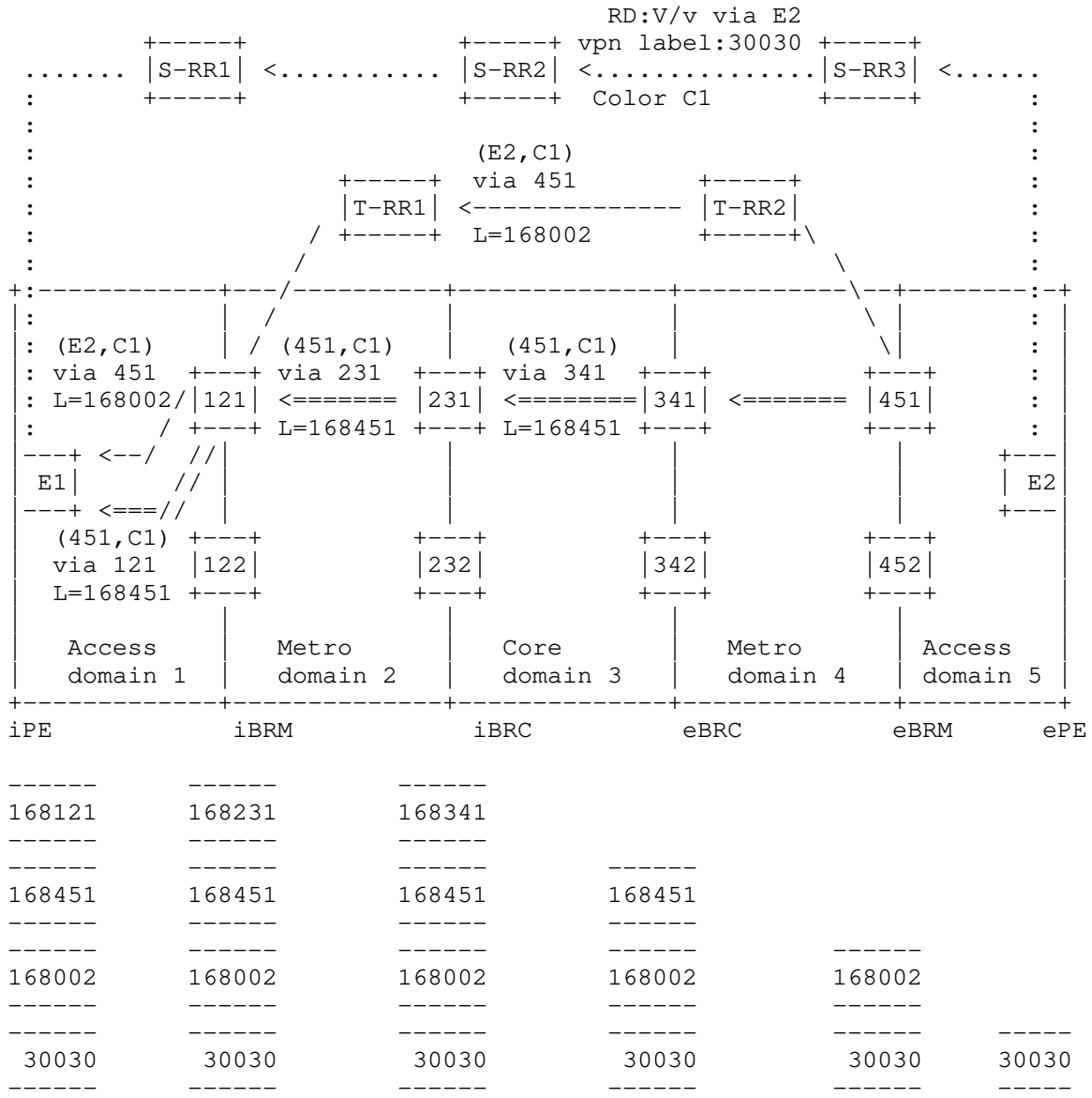


Figure 6: Heirarchical BGP transport CAR, NHU at iBR

- With reference to the topology above:
  - Consider egress PE E2 advertises a VPN (service) route RD:V/v that propagates via service RRs to ingress PE E1.

  * A BGP CAR route (E2, C1) is also advertised by egress BRM node 451. The route may be sourced locally, for instance by redistribution from an IGP-FA, and is distributed via a Transport RR plane.

  * Ingress BRM node 121 learns about BGP CAR route (E2, C1) via node 451.

  * Node 121 also learns about BGP CAR route (451, C1) via node 231.

  * Node 121 advertises both routes to E1.

  * (E2, C1) is advertised with NH via node 451; i.e., next-hop unchanged

  * (451, C1) is advertised with next-hop 121 i.e., next-hop self and local label 16451

  * Hence, E1 receives both routes. It recursively resolves (E2, C1) to build an outgoing label/SID stack to forward traffic to E1, via nodes 121 and 451.

  * Ingress BRM node 121 only needs to install data plane entry for (451, C1), and not for (E2, C1).

- In summary, with this design:

  * Only E1 needs to learn and install (E2, C1) because it has to install a service route RD:V/v with next-hop E2, and associated with a Color C1

  * However, E1 incurs additional complexity to perform the additional recursion to build and program the label stack. The complexity increases when there are multiple paths to be load-balanced across.

7.2. Automated Emulated-Pull Model to learn BGP CAR (PE, C)

From [BGP-CAR-Problem-Statement], we remind:

- The SR-PCE solution natively supports a PULL model: when E1 installs a VPN route V/v via (E2, C1), E1 requests its serving SR-PCE to compute the SR Policy to (E2, C1). I.e. E1 does not learn unneeded SR policies.

- BGP Signaling is natively a PUSH model.

- Emulated-PULL refers to the ability for a BGP CAR node E1 to "subscribe" to (E2, C1) route such that only the related paths are signaled to E1.

7.2.1. Subscription based BGP CAR Signaling

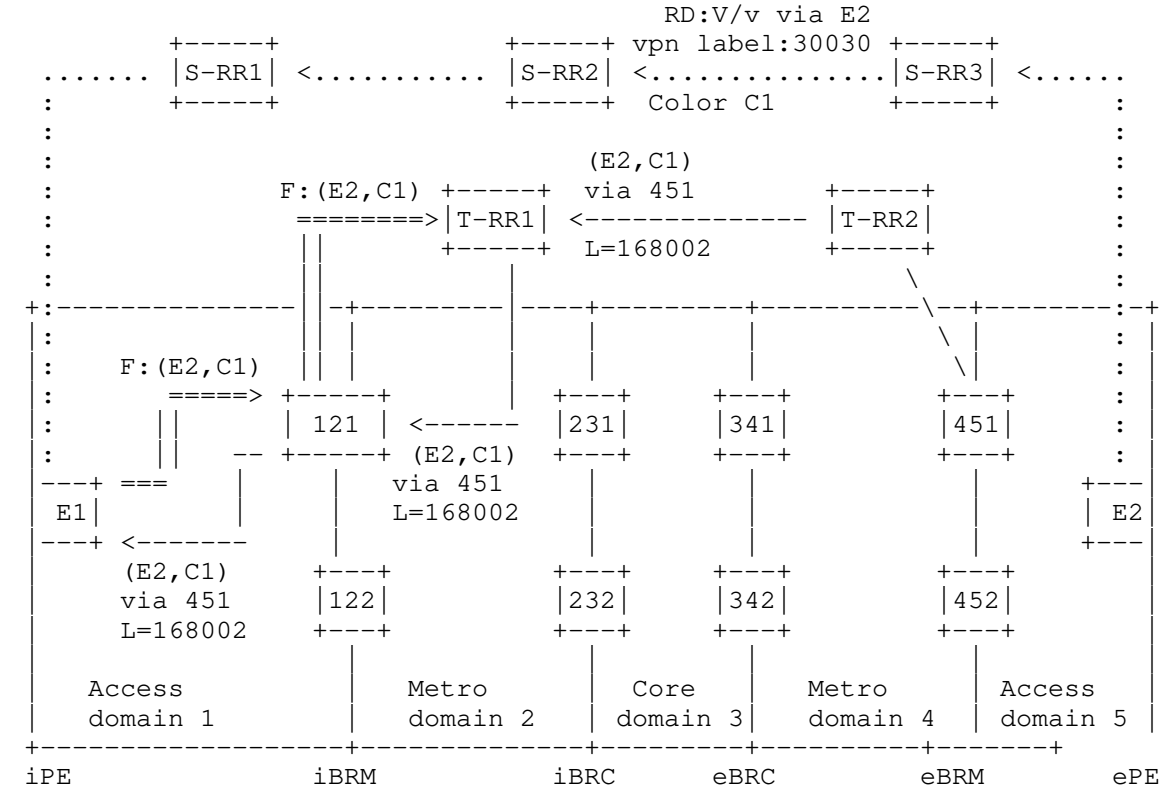


Figure 7: BGP transport CAR route Subscription

- Using the reference figure above that illustrates the use-case in section Figure 6

  * Ingress PE E1 subscribes to (E2, C1) using a BGP CAR filter route F (E2, C1), sent via Ingress BRM node 121

    + node 121 may act as an RR to E1

  * Node 121 propagates F(E2, C1) to Transport-RR T-RR1.

  * Assume Transport-RR has learnt routes for all PEs in network.

  * Based on received F(E2, C1), T-RR1 selectively sends (E2, C1) route to node 121, with Next-Hop of node 451 (i.e., egress BRM).

  * node 121 propagates the received (E2, C1) route to E1 that subscribed for it, with Next-Hop of node 451 (i.e., with BGP Next-Hop unchanged), and received label 168002.

  * Hence E1 learns (E2, C1) that it needs for resolving the received VPN route next-hop for colored route RD:V/v.

  * Note, redundant control flows that exist, for instance via node 122, are not shown above for simplicity.

- In addition, the subscription can be recursive triggered (not shown in the reference diagram above):

  * Upon receiving (E2, C1), E1 further subscribes to (451, C1) using a BGP CAR filter route F (451, C1) sent via node 121

  * Node 121 may not have learnt (451, C1), and hence propagates F (451, C1) to node 231

  * Assuming node 231 has learnt (451, C1), it will selectively send (451, C1) to node 121

  * Node 121 propagates received (451, C1) route to E1, with next-hop set to self and local label 168451

  * Node 121 also installs a data plane entry in this case for label 168451 and BGP recursive next-hop 231

  * Hence, E1 also learns (451, C1) that it needs for resolving the next-hop for (E2, C1)

  * This recursive subscription procedure can be used to minimize state further on ingress BRM nodes, if necessary

- The subscription based selective route signaling technique minimizes the state learnt and installed on both the ingress PEs as well as transit nodes.

  * The solution applies to all the design variants described in section Section 7.1

- This subscription-based selective route signaling has another benefit

* It minimizes routing state that nodes such as BRs or T-RRs need to push to each of their subscription clients

* When a remote node such as an egress BR or egress PE fails, the withdrawal of these routes can also be faster as a result, leading to faster convergence

- Details regarding the subscription based signaling will be described in a later version.

7.3. Additional Design Options

Other related well-known techniques that may be used to complement the solution design or provide an alternative as needed

7.3.1. Anycast SID for transit inter-domain nodes

Redundant BRs (e.g. egress BRMs) advertise their local domain's PE routes with same SID (based on label-index)

Anycast SID assigned to the egress BRMs abstracts state and hence avoids necessity to propagate failure of an egress BRM to ingress BRMs and PEs.

It also avoids traffic convergence issues for traffic from a remote ingress PE

7.3.2. Anycast SID for transport color endpoints i.e PEs

Anycast SID may be assigned to a redundant pair of PEs that have a common, dedicated set of service (VPN) attachments

Used with Anycast SID/static labels for services (e.g., per-VRF VPN label/SID)

This technique, similarly, abstracts state for the egress PEs and hence failure events from remote ingress PEs.

7.4. Convergence

Both existing and additional techniques are used to provide fast convergence for various network failure and change events

BGP Add-Path should be enabled for BGP CAR to signal multiple next hops through RR for fast convergence.

8. Interworking Scenarios

Details regarding various interworking scenarios will be added in a later version.

9. Fault Handling

This the fault management actions as described in [RFC7606] are applicable for handling of BGP update messages for BGP-CAR.

When the error determined allows for the router to skip the malformed NLRI(s) and continue processing of the rest of the update message, then it MUST handle such malformed NLRIs as 'Treat-as-withdraw'. In other cases, where the error in the NLRI encoding results in the inability to process the BGP update message (e.g. length related encoding errors), then the router SHOULD handle such malformed NLRIs as 'AFI/SAFI disable' when other AFI/SAFI besides BGP-CAR are being advertised over the same session. Alternately, the router MUST perform 'session reset' when the session is only being used for BGP-CAR.

10. IANA Considerations

IANA is requested to assign SAFI value TBD1 (BGP CAR) from the "SAFI Values" sub-registry under the "Subsequent Address Family Identifiers (SAFI) Parameters" registry with this document as a reference.

10.1. BGP CAR NLRI Types Registry

IANA is requested to create a "BGP CAR NLRI Types" sub-registry under the "Border Gateway Protocol (BGP) Parameters" registry with this document as a reference. The registry is for assignment of the one octet sized code-points for BGP CAR NLRI types and populated with the values shown below:

Type	NLRI Type	Reference
0	Reserved (not to be used)	[This document]
1	Color-Aware Routes NLRI	[This document]
2-255	Unassigned	

Allocations within the registry are to be made under the "Specification Required" policy as specified in [RFC8126]).

10.2. BGP CAR NLRI TLV Registry

IANA is requested to create a "BGP CAR NLRI TLV Types" sub-registry under the "Border Gateway Protocol (BGP) Parameters" registry with this document as a reference. The registry is for assignment of the one octet sized code-points for BGP-CAR NLRI non-key TLV types and populated with the values shown below:

Type	NLRI Type	Reference
0	Reserved (not to be used)	[This document]
1	Label TLV	[This document]
2	Label Index TLV	[This document]
3	SRv6 SID TLV	[This document]
4-255	Unassigned	

Allocations within the registry are to be made under the "Specification Required" policy as specified in [RFC8126]).

10.3. Guidance for Designated Experts

In all cases of review by the Designated Expert (DE) described here, the DE is expected to ascertain the existence of suitable documentation (a specification) as described in [RFC8126]. The DE is also expected to check the clarity of purpose and use of the requested code points. Additionally, the DE must verify that any request for one of these code points has been made available for review and comment within the IETF: the DE will post the request to the IDR Working Group mailing list (or a successor mailing list designated by the IESG). If the request comes from within the IETF, it should be documented in an Internet-Draft. Lastly, the DE must ensure that any other request for a code point does not conflict with work that is active or already published within the IETF.

10.4. BGP Extended Community Registry

IANA is requested to allocate the sub-type TBD2 for "Local Color Mapping (LCM)" under the "BGP Transitive Opaque Extended Community" registry under the "BGP Extended Community" parameter registry.

11. Security Considerations

TBD

12. Acknowledgements

The authors would like to acknowledge the review and inputs from many people.TBD

13. References

13.1. Normative References

[I-D.ietf-bess-srv6-services]
Dawra, G., Filsfils, C., Talaulikar, K., Raszuk, R., Decraene, B., Zhuang, S., and J. Rabadan, "SRv6 BGP based Overlay services", draft-ietf-bess-srv6-services-05 (work in progress), November 2020.

[I-D.ietf-idr-bgp-ipv6-rt-constrain]
Patel, K., Raszuk, R., Djernaes, M., Dong, J., and M. Chen, "IPv6 Extensions for Route Target Distribution", draft-ietf-idr-bgp-ipv6-rt-constrain-12 (work in progress), April 2018.

[I-D.ietf-idr-tunnel-encaps]
Patel, K., Velde, G., Sangli, S., and J. Scudder, "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-21 (work in progress), January 2021.

[I-D.ietf-lsr-flex-algo]
Psenak, P., Hegde, S., Filsfils, C., Talaulikar, K., and A. Gulko, "IGP Flexible Algorithm", draft-ietf-lsr-flex-algo-13 (work in progress), October 2020.

[I-D.ietf-spring-segment-routing-policy]
Filsfils, C., Talaulikar, K., Voyer, D., Bogdanov, A., and P. Mattes, "Segment Routing Policy Architecture", draft-ietf-spring-segment-routing-policy-09 (work in progress), November 2020.

[I-D.ietf-spring-srv6-network-programming]
Filsfils, C., Camarillo, P., Leddy, J., Voyer, D., Matsushima, S., and Z. Li, "SRv6 Network Programming", draft-ietf-spring-srv6-network-programming-28 (work in progress), December 2020.

[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>.

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

[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, <https://www.rfc-editor.org/info/rfc4684>.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <https://www.rfc-editor.org/info/rfc4760>.

[RFC5512] Mohapatra, P. and E. Rosen, "The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", RFC 5512, DOI 10.17487/RFC5512, April 2009, <https://www.rfc-editor.org/info/rfc5512>.

[RFC5701] Rekhter, Y., "IPv6 Address Specific BGP Extended Community Attribute", RFC 5701, DOI 10.17487/RFC5701, November 2009, <https://www.rfc-editor.org/info/rfc5701>.

[RFC7311] Mohapatra, P., Fernando, R., Rosen, E., and J. Uttaro, "The Accumulated IGP Metric Attribute for BGP", RFC 7311, DOI 10.17487/RFC7311, August 2014, <https://www.rfc-editor.org/info/rfc7311>.

[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, <https://www.rfc-editor.org/info/rfc7606>.

[RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 8126, DOI 10.17487/RFC8126, June 2017, <https://www.rfc-editor.org/info/rfc8126>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

[RFC8277] Rosen, E., "Using BGP to Bind MPLS Labels to Address Prefixes", RFC 8277, DOI 10.17487/RFC8277, October 2017, <https://www.rfc-editor.org/info/rfc8277>.

[RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <https://www.rfc-editor.org/info/rfc8402>.

[RFC8669] Previdi, S., Filsfils, C., Lindem, A., Ed., Sreekantiah, A., and H. Gredler, "Segment Routing Prefix Segment Identifier Extensions for BGP", RFC 8669, DOI 10.17487/RFC8669, December 2019, <https://www.rfc-editor.org/info/rfc8669>.

13.2. Informative References

[I-D.ietf-mpls-seamless-mpls] Leymann, N., Decraene, B., Filsfils, C., Konstantynowicz, M., and D. Steinberg, "Seamless MPLS Architecture", draft-ietf-mpls-seamless-mpls-07 (work in progress), June 2014.

[RFC3906] Shen, N. and H. Smit, "Calculating Interior Gateway Protocol (IGP) Routes Over Traffic Engineering Tunnels", RFC 3906, DOI 10.17487/RFC3906, October 2004, <https://www.rfc-editor.org/info/rfc3906>.

[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, <https://www.rfc-editor.org/info/rfc4271>.

[RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis", RFC 4272, DOI 10.17487/RFC4272, January 2006, <https://www.rfc-editor.org/info/rfc4272>.

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

[RFC6952] Jethanandani, M., Patel, K., and L. Zheng, "Analysis of BGP, LDP, PCEP, and MSDP Issues According to the Keying and Authentication for Routing Protocols (KARP) Design Guide", RFC 6952, DOI 10.17487/RFC6952, May 2013, <https://www.rfc-editor.org/info/rfc6952>.

[RFC7911] Walton, D., Retana, A., Chen, E., and J. Scudder, "Advertisement of Multiple Paths in BGP", RFC 7911, DOI 10.17487/RFC7911, July 2016, <https://www.rfc-editor.org/info/rfc7911>.

Authors' Addresses

Dhananjaya Rao
Cisco Systems
USA

Email: dhrao@cisco.com

Swadesh Agrawal
Cisco Systems
USA

Email: swaagraw@cisco.com

Clarence Filsfils
Cisco Systems
Belgium

Email: cfilsfil@cisco.com

Ketan Talaulikar
Cisco Systems
India

Email: ketant@cisco.com

RIFT
Internet-Draft
Intended status: Standards Track
Expires: 26 August 2021

J. Head, Ed.
T. Przygienda
W. Lin
Juniper Networks
22 February 2021

RIFT Auto-EVPN
draft-head-rift-auto-evpn-00

Abstract

This document specifies procedures that allow an EVPN overlay to be fully and automatically provisioned when using RIFT as underlay and leveraging its no touch ZTP architecture.

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 https://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 26 August 2021.

Table of Contents

1. Introduction 3
1.1. Requirements Language 3
2. Design Considerations 4
3. System ID 4
4. Fabric ID 4
5. Auto-EVPN Device Roles 5
5.1. All Participating Nodes 5
5.2. ToF Nodes as Route Reflectors 5
5.3. Leaf Nodes 6
6. Auto-EVPN Variable Derivation 7
6.1. Auto-EVPN Version 8
6.2. MAC-VRF ID 8
6.3. Loopback Address 8
6.3.1. Leaf Nodes as Gateways 8
6.3.2. ToF Nodes as Route Reflectors 9
6.3.2.1. Route Reflector Election Procedures 9
6.4. Autonomous System Number 9
6.5. Cluster ID 10
6.6. Router ID 10
6.7. Route Target 10
6.8. Route Distinguisher 10
6.9. EVPN MAC-VRF Services 10
6.9.1. Untagged Traffic in Multiple Fabrics 11
6.9.1.1. VLAN 11
6.9.1.2. VNI 11
6.9.1.3. MAC Address 11
6.9.1.4. IPv6 IRB Gateway Address 11
6.9.1.5. IPv4 IRB Gateway Address 11
6.9.2. Tagged Traffic in Multiple Fabrics 11
6.9.2.1. VLAN 12
6.9.2.2. VNI 12
6.9.2.3. MAC Address 12
6.9.2.4. IPv6 IRB Gateway Address 12
6.9.2.5. IPv4 IRB Gateway Address 12
6.9.3. Tagged Traffic in a Single Fabric 12
6.9.3.1. VLAN 12
6.9.3.2. VNI 13
6.9.3.3. MAC Address 13
6.9.3.4. IPv6 IRB Gateway Address 13
6.9.3.5. IPv4 IRB Gateway Address 13
6.9.4. Traffic Routed to External Destinations 13
6.9.4.1. Route Distinguisher 13
6.9.4.2. Route Target 13
7. Acknowledgements 14
8. Security Considerations 14
9. References 14

9.1. Normative References 14
Appendix A. Appendix 14
A.1. RIFT LIE Schema 14
A.1.1. Auto-EVPN Version 14
A.1.2. Fabric ID 14
A.2. RIFT Node-TIE Schema 15
A.2.1. Auto-EVPN Version 15
A.2.2. Fabric ID 15
A.3. Variable Derivation 15
A.3.1. Random Seed Values 15
A.3.2. Fabric ID 15
A.3.3. Loopback Address 15
A.3.4. Autonomous System Number 15
A.3.5. Cluster ID 15
A.3.6. Router ID 15
A.3.7. Route Target 15
A.3.8. Route Distinguisher 16
A.3.9. VLAN 16
A.3.10. VNI 16
A.3.11. Gateway (MAC) 16
A.3.12. Gateway (IPv6) 16
A.3.13. Gateway (IPv4) 16
Authors' Addresses 16

1. Introduction

RIFT is a protocol that focuses heavily on operational simplicity. [RIFT] natively supports Zero Touch Provisioning (ZTP) functionality that allows each node in an underlay network to automatically derive its place in the topology and configure itself accordingly when properly cabled. RIFT can also disseminate Key-Value information contained in Key-Value Topology Information Elements (KV-TIEs). These KV-TIEs can contain any information and therefore be used for any purpose. Leveraging RIFT to provision EVPN overlays without any need for configuration and leveraging KV capabilities to easily validate correct operation of such overlay without a single point of failure would provide significant benefit to operators in terms of simplicity and robustness of such a solution.

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. Design Considerations

EVPN supports various service models, this document defines a method for the VLAN-Aware service model defined in [RFC7432]. Other service models may be considered in future revisions of this document.

Each model has its own set of requirements for deployment. For example, a functional BGP overlay is necessary to exchange EVPN NLRI regardless of the service model. Furthermore, the requirements are made up of individual variables, such as each node's loopback address and AS number for the BGP session. Some of these variables may be coordinated across each node in a network, but are ultimately locally significant (e.g. route distinguishers). Similarly, calculation of some variables will be local only to each device. RIFT contains currently enough topology information in each node to calculate all those necessary variables automatically.

Once the EVPN overlay is configured and becomes operational KV TIEs can be used to distribute state information to allow for validation of basic operational correctness without need for further tooling.

3. System ID

The 64-bit RIFT System ID that uniquely identifies a node as defined in [RIFT].

4. Fabric ID

RIFT operates on variants of Clos substrate which are commonly called an IP Fabric. Since EVPN VLANs can be either contained within one fabric or span them, Auto-EVPN introduces the concept of a Fabric ID into RIFT.

This section describes an optional extension to LIE packet schema in the form of a 16-bit Fabric ID that identifies a nodes membership within a particular fabric. Auto-EVPN capable nodes MUST support this extension but MAY not advertise it when not participating in Auto-EVPN. A non-present Fabric ID and value of 0 is reserved as ANY_FABRIC and MUST NOT be used for any other purpose.

Fabric ID MUST be considered in existing adjacency FSM rules so nodes that support Auto-EVPN can interoperate with nodes that do not. The LIE validation is extended with following clause and if it is not met, miscabling should be declared:

```
(if fabric_id is not advertised by either node OR
if fabric_id is identical on both nodes)
AND
(if auto_evpn_version is not advertised by either node OR
if auto_evpn_version is identical on both nodes)
```

The appendix details LIE (Appendix A.1.2) and Node-TIE (Appendix A.2.2) schema changes.

5. Auto-EVPN Device Roles

Auto-EVPN requires that each node understand its given role within the scope of the EVPN implementation so each node derives the necessary variables and provides the necessary overlay configuration. For example, a leaf node performing VXLAN gateway functions does not need to derive its own Cluster ID or learn one from the route reflector that it peers with.

5.1. All Participating Nodes

Not all nodes have to participate in Auto-EVPN but when they do they do assume EVPN roles and MUST derive according variables:

IPv6 Loopback Address
: Unique IPv6 loopback address used in BGP sessions.

Router ID
: The BGP Router ID.

Autonomous System Number
: The ASN for IBGP sessions.

Cluster ID
: The Cluster ID for Top-of-Fabric IBGP route reflection.

5.2. ToF Nodes as Route Reflectors

This section defines an Auto-EVPN role whereby some Top-of-Fabric nodes act as EVPN route reflectors. It is expected that route reflectors would establish IBGP sessions with leaf nodes in the same fabric. The typical route reflector requirements do not change, however determining which specific values to use requires further consideration. ToF nodes performing route reflector functionality MUST derive the following variables:

IPv6 RR Loopback Address
: The source address for IBGP sessions with leaf nodes in case ToF won election for one of the route reflectors in the fabric.

IPv6 RR Acceptable Prefix Range
Range of addresses acceptable by the route reflector to form a IBGP session. This range covers ALL possible IPv6 Loopback Addresses derived by other Auto EVPN nodes in the current fabric and other Auto-EVPN RRs addresses.

5.3. Leaf Nodes

Leaf nodes derive their role from realizing they are at the bottom of the fabric, i.e. not having any southbound adjacencies. Alternately, a node can assume a leaf node if it has only southbound adjacencies to nodes with explicit LEAF_LEVEL to allow for scenarios where RIFT leaves do NOT participate in Auto-EVPN.

Leaf nodes MUST derive the following variables:

IPv6 RR Loopback Adresses
Addresses of the RRs present in the fabric. Those addresses are used to build BGP sessions to the RR.

EVIs
Leaf node derives all the necessary variables to instantiate EVIs with layer-2 and optionally layer-3 functionality.

If a leaf node is required to perform layer-2 VXLAN gateway functions, it MUST be capable of deriving the following types of variables:

Route Distinguisher
The route distinguisher corresponding to a MAC-VRF that uniquely identifies each node.

Route Target
The route target that corresponds to a MAC-VRF.

MAC VRF name
This is an optional variable to provide a common MAC VRF name across all leaves.

Set of VLANs
Those are VLANs provisioned either within the fabric or allowing to stretch across fabrics.

For each VLAN derived in an EVI the following variables MUST be derived:

VLAN
The VLAN ID.

name
This is an optional variable to provide a common VLAN name across all leaves.

VNI
The VNI that corresponds to the VLAN ID. This will contribute to the EVPN Type-2 route.

IRB
Optional variables of the IRB for the VLAN if the leaf performs layer-3 gateway function.

If a leaf node is required to perform layer-3 VXLAN gateway functions, it MUST additionally be capable of deriving the following types of variables:

IP Gateway MAC Address
The MAC address associated with IP gateway.

IP Gateway Subnetted Address
The IPv4 and/or IPv6 gateway address including its subnet length.

Type-5 EVPN IP Prefix with ToFs performing gateway functionality can also be derived and will be described in a future version of this document.

6. Auto-EVPN Variable Derivation

As previously mentioned, not all nodes are required to derive all variables in a given network (e.g. a transit spine node may not need to derive any or participate in Auto-EVPN). Additionally, all derived variables are derived from RIFT's FSM or ZTP mechanism so no additional flooding beside RIFT flooding is necessary for the functionality.

It is also important to mention that all variable derivation is in some way based on combinations of System ID, MAC-VRF ID, Fabric ID, EVI and VLAN and MUST comply precisely with calculation methods specified in the Appendix section to allow interoperability between different implementations.

6.1. Auto-EVPN Version

This section describes extensions to both the RIFT LIE packet and Node-TIE schemas in the form of a 16-bit value that identifies the Auto-EVPN Version. Auto-EVPN capable nodes MUST support this extension, but MAY choose not to advertise it in LIEs and Node-TIEs when Auto-EVPN is not being utilized. The appendix describes LIE (Appendix A.1.1) and Node-TIE (Appendix A.2.1) schema changes in detail.

6.2. MAC-VRF ID

This section describes a variable MAC-VRF ID that uniquely identifies an instance of EVPN instance (EVI) and is used in variable derivation procedures. Each EVPN EVI MUST be associated with a unique MAC-VRF ID, this document does not specify a method for making that association or ensuring that they are coordinated properly across fabric(s).

6.3. Loopback Address

First and foremost, RIFT does not advertise anything more specific than the fabric default route in the southbound direction by default. However, Auto-EVPN nodes MUST advertise specific loopback addresses southbound to all other Auto-EVPN nodes so to establish MP-BGP reachability correctly in all scenarios.

Auto-EVPN nodes MUST derive a ULA-scoped IPv6 loopback address to be used as both the IBGP source address, as well as the VTEP source when VXLAN gateways are required. Calculation is done using the 6-bytes of reserved ULA space, the 2-byte Fabric ID, and the node's 8-byte System ID. Derivation of the System ID varies slightly depending upon the node's location/role in the fabric and will be described in subsequent sections.

IPv4 addresses MAY be supported, but it should be noted that they have a higher likelihood of collision.

The required algorithm can be found in the appendix (Appendix A.3.3).

6.3.1. Leaf Nodes as Gateways

Calculation is done using the 6-bytes of reserved ULA space, the 2-byte Fabric ID, and the node's 8-byte System ID.

6.3.2. ToF Nodes as Route Reflectors

ToF nodes acting as route reflectors MUST derive their loopback address according to the specific section describing the algorithm. Calculation is done using the 6-bytes of reserved ULA space, the 2-byte Fabric ID, and the 8-byte System ID of each elected route reflector.

6.3.2.1. Route Reflector Election Procedures

Four Top-of-Fabric nodes MUST be elected as an IBGP route reflector. Each ToF performs the election independently based on system IDs of other ToFs in the fabric obtained via southbound reflection. The route reflector election procedures are defined as follows:

1. ToF node with the highest System ID.

2. ToF node with the lowest System ID.

3. ToF node with the 2nd highest System ID.

4. ToF node with the 2nd lowest System ID.

This ordering is necessary to prevent a single node with either the highest or lowest System ID from triggering changes to route reflector loopback addresses as it would result in all BGP sessions dropping.

For example, if two nodes, ToF01 and ToF02 with System IDs 002c6af5a281c000 and 002c6bf5788fc000 respectively, ToF02 would be elected due to it having the highest System ID of the ToFs (002c6bf5788fc000). If a ToF determines that it is elected as route reflector, it uses the knowledge of its position in the list to derive route reflector v6 loopback address.

Considerations for multiplane route reflector elections will be included in future revisions.

6.4. Autonomous System Number

Nodes in each fabric MUST derive a private autonomous system number based on its Fabric ID so that it is unique across the fabric.

The required algorithm for 2-byte ASNs can be found in the appendix (Appendix A.3.4).

6.5. Cluster ID

Route reflector nodes in each fabric MUST derive a cluster ID that is based on its Fabric ID so that it is unique across the fabric. Implementations MAY choose to simply use the AS number as the cluster ID.

The required algorithm can be found in the appendix (Appendix A.3.5).

6.6. Router ID

Nodes MUST drive a Router ID that is based on both its System ID and Fabric ID so that it is unique to both.

The required algorithm can be found in the appendix (Appendix A.3.6).

6.7. Route Target

Nodes hosting EVPN EVIs MUST derive a route target extended community based on the MAC-VRF ID for each EVI so that it is unique across the network. Route targets MUST be of type 0 as per RFC4360.

For example, if given a MAC-VRF ID of 1, the derived route target would be "target:1"

The required algorithm can be found in the appendix (Appendix A.3.7).

6.8. Route Distinguisher

Nodes hosting EVPN EVIs MUST derive a type-0 route distinguisher based on its System ID and Fabric ID so that it is unique per MAC-VRF and per node.

The required algorithm can be found in the appendix (Appendix A.3.8).

6.9. EVPN MAC-VRF Services

It’s obvious that applications utilizing Auto-EVPN overlay services may require a variety of layer-2 and/or layer-3 traffic considerations. Variables supporting these services are also derived based on some combination of MAC-VRF ID, Fabric ID, and other constant values. Integrated Routing and Bridging (IRB) gateway address derivation also leverages a set of constant "random seed" values to provide additional entropy.

The required derivation procedures can be found in the appendix (Appendix A.3).

6.9.1. Untagged Traffic in Multiple Fabrics

This section defines a methods to derive unique VLAN, VNI, MAC, and gateway address values for deployments where untagged traffic is stretched across multiple fabrics.

6.9.1.1. VLAN

Untagged traffic stretched across multiple fabrics MUST derive VLAN tags based on MAC-VRF ID in conjunction with a constant value of 1 (i.e. MAC-VRF ID + 1).

6.9.1.2. VNI

Untagged traffic stretched across multiple fabrics MUST derive VNIs based on MAC-VRF ID and Fabric ID in conjunction with a constant value. These VNIs MUST correspond to EVPN Type-2 routes.

6.9.1.3. MAC Address

The MAC address MUST be a unicast address and also MUST be identical for any IRB gateways that belong to an individual bridge-domain across fabrics. The last 5-bytes MUST be a hash of the MAC-VRF ID and a constant value of 1 that is calculated using the previously mentioned random seed values.

6.9.1.4. IPv6 IRB Gateway Address

The derived IPv6 gateway address MUST be from a ULA-scoped range that will account for the first 6-bytes. The next 5-bytes MUST be the last bytes of the derived MAC address. Finally, the remaining 7-bytes MUST be ::0001.

6.9.1.5. IPv4 IRB Gateway Address

The derived IPv4 gateway address MUST be from a RFC1918 range, which accounts for the first octet. The next octet MUST a hash of the MAC-VRF ID and a constant value of 1 that is calculated using the previously mentioned random seed values. Finally, the remaining 2 octets MUST be 0 and 1 respectively.

6.9.2. Tagged Traffic in Multiple Fabrics

This section defines a methods to derive unique VLAN, VNI, MAC, and gateway address values for deployments where tagged traffic is stretched across multiple fabrics.

6.9.2.1. VLAN

Tagged traffic stretched across multiple fabrics MUST derive VLAN tags based on MAC-VRF ID in conjunction with a constant value of 16 (i.e. MAC-VRF ID + 16).

6.9.2.2. VNI

Tagged traffic stretched across multiple fabrics MUST derive VNIs based on MAC-VRF ID and Fabric ID in conjunction with a constant value. These VNIs MUST correspond to EVPN Type-2 routes.

6.9.2.3. MAC Address

The MAC address MUST be a unicast address and also MUST be identical for any IRB gateways that belong to an individual bridge-domain across fabrics. The last 5-bytes MUST be a hash of the MAC-VRF ID and a constant value of 1 that is calculated using the previously mentioned random seed values.

6.9.2.4. IPv6 IRB Gateway Address

The derived IPv6 gateway address MUST be from a ULA-scoped range that will account for the first 6-bytes. The next 5-bytes MUST be the last bytes of the derived MAC address. Finally, the remaining 7-bytes MUST be ::0001.

6.9.2.5. IPv4 IRB Gateway Address

The derived IPv4 gateway address MUST be from a RFC1918 range, which accounts for the first octet. The next octet MUST a hash of the MAC-VRF ID and a constant value of 16 that is calculated using the previously mentioned random seed values. Finally, the remaining 2 octets MUST be 0 and 1 respectively.

6.9.3. Tagged Traffic in a Single Fabric

This section defines a methods to derive unique VLAN, VNI, MAC, and gateway address values for deployments where untagged traffic is contained within a single fabric.

6.9.3.1. VLAN

Tagged traffic contained to a single fabric MUST derive VLAN tags based on MAC-VRF ID and Fabric ID in conjunction with a constant value of 17 (i.e. MAC-VRF ID + Fabric ID + 17).

6.9.3.2. VNI

Tagged traffic contained to a single fabric MUST derive VNIs based on MAC-VRF ID and Fabric ID in conjunction with a constant value. These VNIs MUST correspond to EVPN Type-2 routes.

6.9.3.3. MAC Address

The MAC address MUST be a unicast address and also MUST be identical for any IRB gateways that belong to an individual bridge-domain across fabrics. The last 5-bytes MUST be a hash of the MAC-VRF ID and a constant value of 1 that is calculated using the previously mentioned random seed values.

6.9.3.4. IPv6 IRB Gateway Address

The derived IPv6 gateway address MUST be from a ULA-scoped range, which accounts for the first 6-bytes. The next 5-bytes MUST be the last bytes of the derived MAC address. Finally, the remaining 7-bytes MUST be ::0001.

6.9.3.5. IPv4 IRB Gateway Address

The derived IPv4 gateway address MUST be from a RFC1918 range, which accounts for the first octet. The next octet MUST a hash of the MAC-VRF ID and a constant value of 17 that is calculated using the previously mentioned random seed values. Finally, the remaining 2 octets MUST be 0 and 1 respectively.

6.9.4. Traffic Routed to External Destinations

6.9.4.1. Route Distinguisher

Nodes hosting IP Prefix routes MUST derive a type-0 route distinguisher based on its System ID and Fabric ID so that it is unique per IP-VRF and per node.

The required algorithm can be found in the appendix (Appendix A.3.8).

6.9.4.2. Route Target

Nodes hosting IP prefix routes MUST derive a route target extended community based on the MAC-VRF ID for each IP-VRF so that it is unique across the network. Route targets MUST be of type 0.

The required algorithm can be found in the appendix (Appendix A.3.7).

7. Acknowledgements

TBD

8. Security Considerations

This document introduces no new security concerns to RIFT or other specifications referenced in this document.

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, <https://www.rfc-editor.org/info/rfc2119>.

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

[RIFT] Przygienda, T., Sharma, A., Thubert, P., Rijsman, B., and D. Afanasiev, "RIFT: Routing in Fat Trees", Work in Progress, draft-ietf-rift-rift-12, May 2020.

Appendix A. Appendix

A.1. RIFT LIE Schema

A.1.1. Auto-EVPN Version

```
struct LIEPacket {
 ...
   /** It provides the optional ID of the configured fabric  */
   25: optional common.FabricIDType       fabric_id;
...
```

A.1.2. Fabric ID

```
...
struct LIEPacket {
 ...
   /** It provides optional version of EVPN ZTP as 256 * MAJOR + MINOR */
   26: optional i16                       auto_evpn_version;
...
```

A.2. RIFT Node-TIE Schema

A.2.1. Auto-EVPN Version

```
struct NodeTIEElement {
...
   /** It provides optional version of EVPN ZTP as 256 * MAJOR + MINOR */
   13: optional i16                          auto_evpn_version;
```

A.2.2. Fabric ID

```
 struct NodeTIEElement {
 ...
   /** It provides the optional ID of the Fabric configured */
   12: optional common.FabricIDType         fabric_id;
```

A.3. Variable Derivation

A.3.1. Random Seed Values

To be provided in future version of this document.

A.3.2. Fabric ID

To be provided in future version of this document.

A.3.3. Loopback Address

To be provided in future version of this document.

A.3.4. Autonomous System Number

To be provided in future version of this document.

A.3.5. Cluster ID

To be provided in future version of this document.

A.3.6. Router ID

To be provided in future version of this document.

A.3.7. Route Target

To be provided in future version of this document.

A.3.8. Route Distinguisher

To be provided in future version of this document.

A.3.9. VLAN

To be provided in future version of this document.

A.3.10. VNI

To be provided in future version of this document.

A.3.11. Gateway (MAC)

To be provided in future version of this document.

A.3.12. Gateway (IPv6)

To be provided in future version of this document.

A.3.13. Gateway (IPv4)

To be provided in future version of this document.

Authors' Addresses

Jordan Head (editor)
Juniper Networks
1137 Innovation Way
Sunnyvale, CA
United States of America

Email: jhead@juniper.net

Tony Przygienda
Juniper Networks
1137 Innovation Way
Sunnyvale, CA
United States of America

Email: prz@juniper.net

Wen Lin
Juniper Networks
10 Technology Park Drive
Westford, MA
United States of America

Email: wlin@juniper.net

SPRING
Internet-Draft
Intended status: Informational
Expires: August 26, 2021

S. Hegde
C. Bowers
Juniper Networks Inc.
X. Xu
Alibaba Inc.
A. Gulko
EdwardJones
A. Bogdanov
Google Inc.
J. Uttaro
ATT
L. Jalil
Verizon
M. Khaddam
Cox communications
A. Alston
Liquid Telecom
LM. Contreras
Telefonica

February 22, 2021

Seamless SR Problem Statement
draft-hegde-spring-mpls-seamless-sr-05

Abstract

This draft documents a set of use cases and requirements for end-to-end intent-based paths spanning multi-domain packet networks. The document explicitly focuses on use cases that require high scale and availability, which will likely benefit from distributed solutions. It is intended that the requirements in this document serve as a basis for future IETF work to develop distributed solutions for inter-domain intent-based transport paths.

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 https://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 26, 2021.

Table of Contents

1. Introduction 3
2. Large scale networks 4
2.1. Service provider networks 4
2.2. Cloud provider WAN networks 5
2.3. Data Center WAN Networks 6
3. Use Cases for Inter-domain Intent-based Transport 6
3.1. Inter-domain Data Sovereignty 6
3.2. Inter-domain Low-Latency Services 7
3.3. Network Mergers 7
3.4. Inter-domain Service Function Chaining 8
3.5. AS Confederation 8
3.6. Inter-domain Multicast Use cases 8
4. Requirements 9
4.1. AS and IGP Domain Models 9
4.1.1. Multiple ASes connected with E-BGP 9
4.1.2. Single AS multiple IGP domains 10
4.1.3. Single AS, Multiple IGP domains with no common border node 11
4.2. Transport tunneling Requirements 11
4.2.1. Unicast tunneling Requirements 11
4.2.2. Multicast tunneling Requirements 12

4.3. Inter-domain SLA Requirements 13
4.3.1. Latency, Delay Variation, and Link Loss Constraints . 13
4.3.2. Bandwidth Constraints 13
4.3.3. Link Inclusion/Exclusion Constraints 13
4.3.4. Node Inclusion/Exclusion Constraints 14
4.3.5. Domain Inclusion/Exclusion Constraints 14
4.3.6. Diverse Paths 14
4.3.7. Constraint applicability to a subset of domains . . . 15
4.3.8. Service function chaining 15
4.4. Multicast specific requirements 15
4.5. Interoperate with BGP-LU 15
4.6. Merger and Migration Requirements 16
4.6.1. Option A and Option B Usecases 16
4.6.2. Inter-Domain Intent Translation 16
4.6.3. Native Support for Best Effort Paths 16
4.6.4. Interoperate with Other tunneling Mechanisms 16
4.7. Scalability Requirements 16
4.8. Availability Requirements 17
4.9. Operations and Automation Requirements 17
4.10. Service Mapping Requirements 18
4.10.1. Traffic service mapping 18
4.10.2. 1 to N service mapping 19
4.11. Interaction with Other Approaches 19
5. Backward Compatibility 20
6. Security Considerations 20
7. IANA Considerations 20
8. Acknowledgements 20
9. Contributors 20
10. References 20
10.1. Normative References 20
10.2. Informative References 21
Authors' Addresses 25

1. Introduction

Evolving trends in wireless access technology, cloud applications, virtualization, and network consolidation all contribute to the increasing demands being placed on a common packet network. In order to meet these demands, a given network will need to scale horizontally in terms of its bandwidth, absolute number of nodes, and geographical extent. The same network will need to scale vertically in terms of the different services that it needs to simultaneously support.

In order to operate networks with large numbers of devices, network operators organize networks into multiple smaller network domains. Each network domain typically runs an IGP which has complete visibility within its own domain, but limited visibility outside of

its domain. Network operators will continue to use multiple domains to scale horizontally. These multi-domain networks will also need to scale vertically, to allow a common multi-domain network to carry all of an organization's services.

Evolving network requirements (e.g., 5G, native cloud) motivate network operators to deploy tunnels that span multiple AS's and maintain specific transport characteristics (e.g., bandwidth, latency). There is a need to provide flexible, scalable, and reliable end-to-end connectivity for multiple services across independent network domains.

2. Large scale networks

2.1. Service provider networks

Service Provider networks can contain many nodes distributed over a large geographic area. 5G networks can include as many as one million nodes, with the majority of those being radio access nodes. Radio and access nodes may be constrained by their memory and processing capabilities.

Service provider transport networks use multiple domains to support scalability. For this analysis, we consider a representative network design with four level of hierarchy: access domains, pre-aggregation domains, aggregation domains and a core. (See Figure 1). The separation of domains internal to the service provider can be performed by using either IGP or BGP.

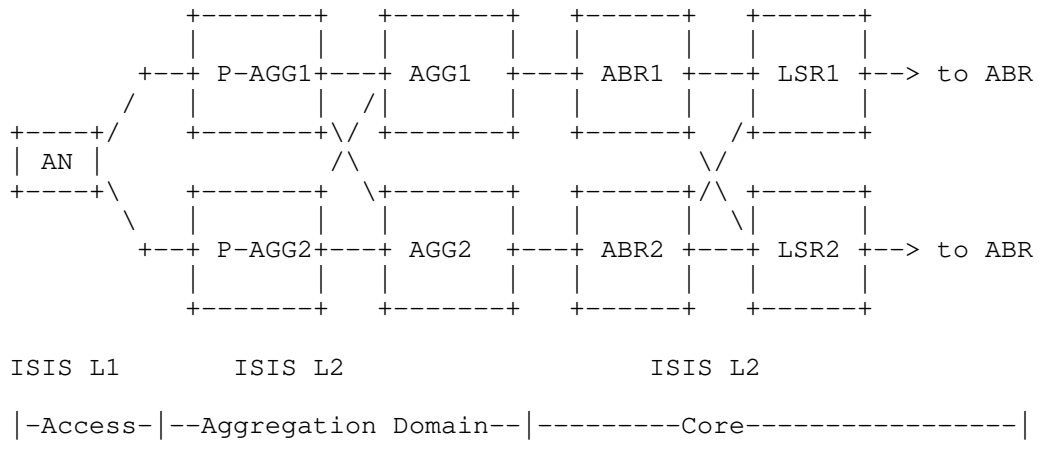


Figure 1: 5G network

5G networks support a variety of service use cases that require end-to-end slicing. In certain cases the end-to-end connectivity requires the ability to forward over intent-based paths. The inter-domain solution should support end-to-end Service Level Objectives(SLO) to allow the creation of network slices.

2.2. Cloud provider WAN networks

As WAN networks grow beyond several thousand nodes, it is often useful to divide the network into multiple IGP domains, as illustrated in Figure 2. Separate IGP domains increase service availability by establishing a constrained failure domain. Smaller IGP domains may also improve network performance and health by reducing the device scale profile (including protocol and FIB scale).

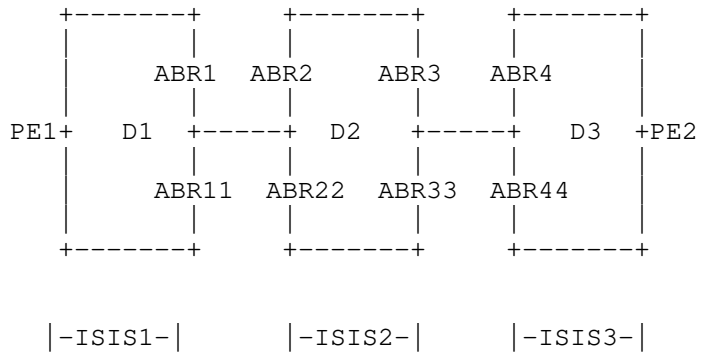


Figure 2: WAN Network

These large WAN networks often cross national boundaries. In order to meet data sovereignty requirements, operators need to maintain strict control over end-to-end traffic-engineered (TE) paths. A goal of a distributed inter-domain solution is to be able to create highly constrained inter-domain TE paths in a scalable manner.

Some deployments may use a centralized controller to acquire the topologies of multiple domains and build end-to-end constrained paths. This centralized approach can be scaled with hierarchical controllers. However, there is still significant risk of a loss of network connectivity to one or more controllers, which can result in a failure to satisfy the strict requirements of data sovereignty. The network should have pre-established TE paths end-to-end that don't rely on controllers in order to address these failure scenarios.

2.3. Data Center WAN Networks

Data centers are playing an increasingly important role in providing access to information and applications. Geographically diverse data centers usually connect via a high speed, reliable and secure DC WAN core network.

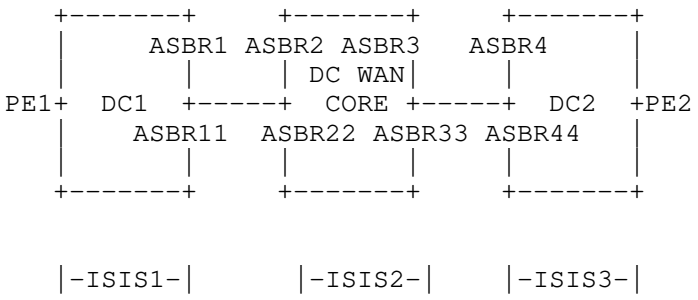


Figure 3: DCI Network

In many DC WAN deployments, applications require end-to-end path diversity and end-to-end low latency paths. The DC WAN networks may consist of large number of devices owing to global presence. In some DC WAN deployments the tunneling mechanisms used within the data centers are the same as those used in the DC WAN core. For example, a network may use MPLS in both data center and DC WAN core. Or it may use SRv6 in both data center and DC WAN core. This can simplify network deployments.

However, in some DC WAN deployments the traffic within data centers and the traffic over the DC WAN core use different tunneling mechanisms, such as SRv6 in the data center and MPLS in the DC WAN core. It is important for DC WAN network operators to have flexibility in the choice of tunneling mechanisms across domains.

3. Use Cases for Inter-domain Intent-based Transport

The use cases for inter-domain intent-based packet transport described in this section are intended to provide motivation for the requirements that follow.

3.1. Inter-domain Data Sovereignty

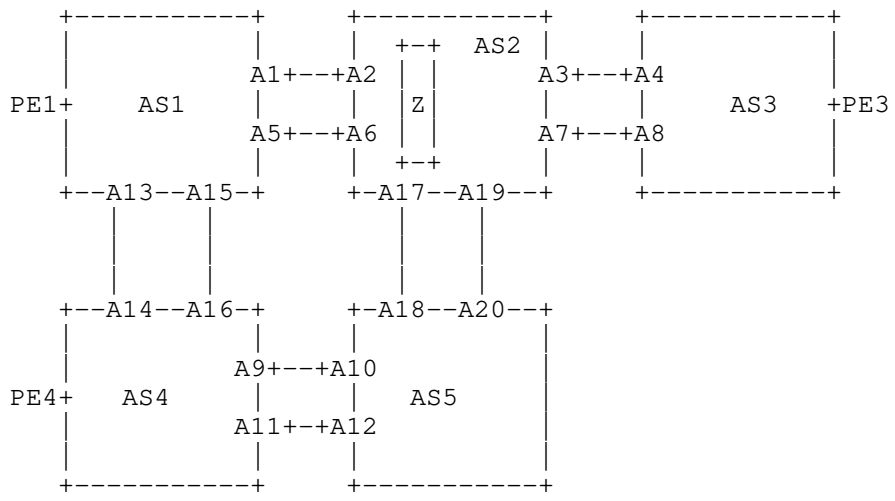


Figure 4: Multi domain Network

Figure Figure 4 depicts a WAN with multiple ASes. Each AS is resides serves a continent (e.g., Asia). Certain traffic from PE1 (in AS1) to PE3 (in AS3) must not traverse country Z in AS2. However, all paths from AS1 to AS3 traverse AS 2. The inter-domain solution should provide end-to-end path creation that traverses AS 2 but avoids country Z.

3.2. Inter-domain Low-Latency Services

Service provider networks running L2 and L3VPNs carry traffic for particular VPNs on low-latency paths that traverse multiple domains.

3.3. Network Mergers

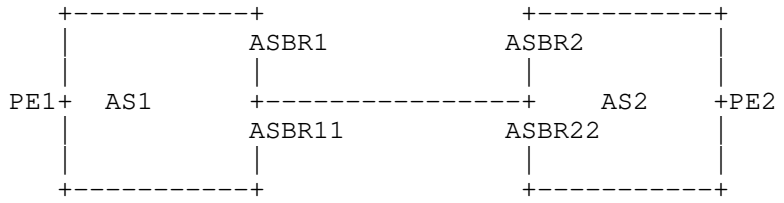


Figure 5: Network Mergers

In diagram Figure 5 above, AS1 and AS2 which were previously under independent administration, merge to come under a single administration. The network operator may decide to merge the two domains into single AS which would need bigger operational effort. Network operators may also retain the two ASes and build end-to-end paths across the two Ases. In this case, the paths in AS1 and AS2 corresponding to the same intent may use different representations in the two ASes. In some cases, organizations may continue to use option A or option B [RFC4364] style interconnectivity in which case the inter-domain solution should satisfy intent of the path on inter-domain links for the service prefixes. In other cases, organizations may prefer to use option C style connectivity from PE1 to PE2. In this case, an inter-domain solution should provide effective mechanisms to translate intent across domains without requiring renumbering of the intent representation.

3.4. Inter-domain Service Function Chaining

[RFC7665] defines service function chaining as an ordered set of service functions and automated steering of traffic through this set of service functions. There could be a variety of service functions such as firewalls, parental control, CGNAT etc. In 5G networks these functions may be completely virtualized or could be a mix of virtualized functions and physical appliances. It is required that the inter-domain solution caters to the service function chaining requirements. The service functions may be virtualized and spread across different data centers attached to different domains.

3.5. AS Confederation

BGP confederation allows the division of a public AS in multiple sub-AS, usually with private identifiers. From outside, the confederation is seen as a single and common AS, the public one. BGP sessions are maintained among sub-AS. In the internals of the confederation, each sub-AS can be configured and run autonomously, even though some BGP parameters (like e.g. LOCAL_PREF or MED) are preserved across sub-AS. Thus, it can be of interest to define end-to-end paths of specific characteristics in terms of SLOs across the sub-AS as well as internally to each sub-AS.

3.6. Inter-domain Multicast Use cases

Multicast services such as IPTV and multicast VPN also need to be supported across a multi-domain service provider network.

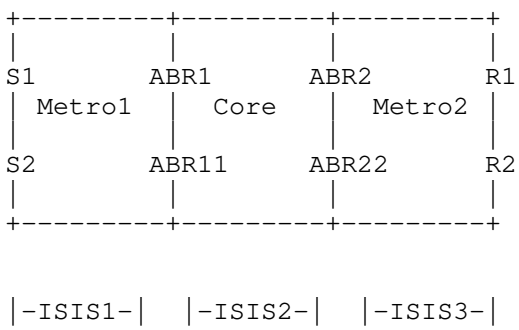


Figure 6: Multicast usecases

Figure 6 shows a simplified multi-domain network supporting multicast. Multicast sources S1 and S2 lie in a different domain from the receivers R1 and R2. Using multiple IGP domains presents a problem for the establishment of multicast replication trees. Typically, a multicast receiver does a reverse path forwarding (RPF) lookup for a multicast source. One solution is to leak the routes for multicast sources across the IGP domains. However, this can compromise the scaling properties of the multi-domain architecture. A distributed inter-domain solution should accommodate a mixture of existing and newer technologies to better facilitate coexistence and migration. A distributed solution should avoid leaking RPF routes into the IGP domains.

4. Requirements

The requirements described in this document are mostly applicable to network under a single administrative domain that are organized into multiple network domains. The requirements are also applicable to multi-AS networks with closely cooperating administration.

4.1. AS and IGP Domain Models

This section describes three different ways that multi-domain networks are organized today. The requirements in subsequent sections are applicable to all three types of multi-domain networks described below.

4.1.1. Multiple ASes connected with E-BGP

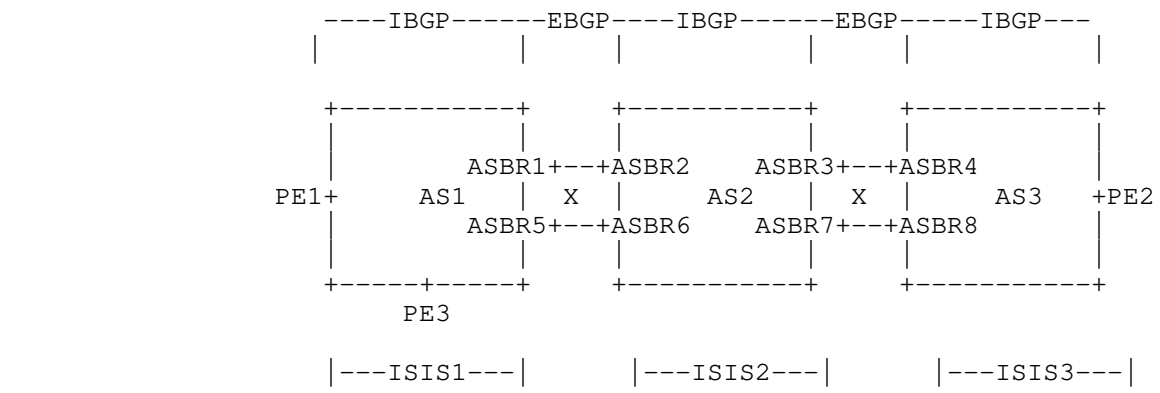


Figure 7: Multiple ASes connected with E-BGP

The above diagram Figure 7 shows three different ASes (AS1, AS2 and AS3.) ASBR1 to ASBR8 are border nodes between the ASes. A given ASBR runs E-BGP sessions with the ASBRs in adjacent ASes. The ASBR also runs I-BGP sessions with other ASBRs in the same AS. Route reflectors can also be used to achieve this full mesh of I-BGP information exchange.Similar scenario applies when considering BGP confederations [RFC5065].

4.1.2. Single AS multiple IGP domains

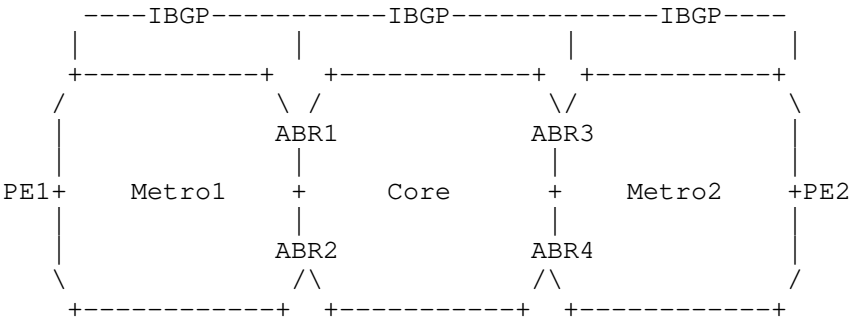


Figure 8: Single AS with Multiple IGP domains

The above diagram Figure 8 shows three different IGP domains, Metro1, Core, and Metro2. The three IGP domains may be realized with

multiple levels in ISIS or multiple areas in OSPF. They can also be realized using separate IGP instances.

This single-AS network uses I-BGP sessions. ABRs and PEs achieve a full mesh of I-BGP information sharing by configuring the ABRs as inline route reflectors.

4.1.3. Single AS, Multiple IGP domains with no common border node

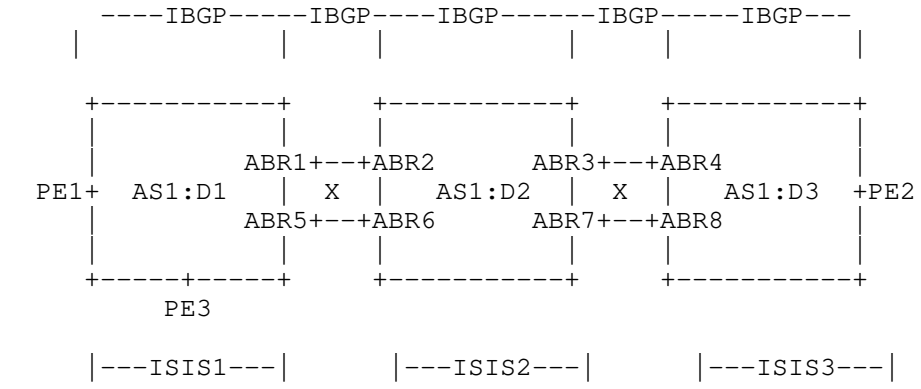


Figure 9: Single AS multiple IGP domains with no common Border node

The above diagram Figure 9 shows a single AS1 with three different IGP domains, D1, D2, and D3. ABR1 to ABR8 are border nodes for the IGP domains and they participate in only one IGP domain.

This single-AS network uses I-BGP sessions. ABRs and PEs achieve a full mesh of I-BGP information sharing by configuring the ABRs as inline route reflectors.

4.2. Transport tunneling Requirements

4.2.1. Unicast tunneling Requirements

The inter-domain solution should support the following unicast tunneling mechanisms:

SR-MPLS tunnels with IPv4 underlay

SR-MPLS tunnels with IPv6 underlay

SR-MPLS tunnels with dual stack underlay

SRv6 tunneling end-to-end

Segment routing TE tunnels and color-only policies as described in [I-D.ietf-idr-segment-routing-te-policy] (both SR-MPLS and SRv6)

Flex-algo [I-D.ietf-lsr-flex-algo] (both SR-MPLS and SRv6)

Pure IP fabric (incapable of supporting MPLS or SRv6 tunneling mechanisms)

RSVP and LDP based tunnels

The inter-domain solution should support the ability to have different domains running different unicast tunneling mechanisms.

The solution should support inter-domain paths that fulfil a common intent using different unicast tunneling mechanisms in different domains.

4.2.2. Multicast tunneling Requirements

The inter-domain solution should support the following multicast tunneling mechanisms:

All of the unicast tunneling mechanisms described in Section 4.2.1 should be supported for multicast service for the purpose of ingress replication.

SR-P2MP as defined in [I-D.voyer-pim-sr-p2mp-policy]

PIM based multicast

RSVP-P2MP and mLDP [RFC6388] based tunnels

BGP based multicast (hop-by-hop or controller-driven, for native IP, labelled, or SRv6 forwarding planes)

The inter-domain solution should support the ability to have different domains running different multicast tunneling mechanisms and should not require to leak RPF routes into IGP domains.

The solution should support inter-domain paths that fulfil a common intent using different multicast tunneling mechanisms in different domains.

4.3. Inter-domain SLA Requirements

This section discusses the end-to-end constraints that intent-based inter-domain path may have to adhere to. The requirements described in this section are applicable to the three types of AS and IGP domain partitioning described in Section 4.1.

4.3.1. Latency, Delay Variation, and Link Loss Constraints

Link delay, delay variation and link loss values can be advertised within a domain using the IGP as described in [RFC8570]. Within an IGP domain, minimum latency, minimum delay variation, and minimum link loss paths can be built using flex-algo [I-D.ietf-lsr-flex-algo]. The end-to-end low latency, low delay variation, or low link loss path requires accumulated metrics for latency, delay variation, and link loss.

The solution should allow the creation of inter-domain paths with low values of latency as calculated over the end-to-end path. It is not necessary that the solution produce the absolute minimum end-to-end latency, delay variation, or link loss path. However, the solution should provide the ability to balance scalability with optimality.

Best path selection at any intermediate border node should be allowed.

The inter-domain solution should allow advertising multiple paths end-to-end and compare the accumulated metric across all of the paths at the ingress.

4.3.2. Bandwidth Constraints

A distributed solution should support the creation of inter-domain paths using intra-domain bandwidth guaranteed paths.

A distributed solution may support optimized path placement with end-to-end bandwidth guarantees.

4.3.3. Link Inclusion/Exclusion Constraints

The links are associated with link-affinity or admin-groups. The link-affinity can be used to indicate a characteristic of a link, such as capacity, encryption, geography, etc. The inter-domain solution should support the creation of paths across different domains that satisfy link inclusion/exclusion constraints. The link constraints should also be satisfied for inter-domain links, such as those between ASBRs.

4.3.4. Node Inclusion/Exclusion Constraints

Creating an inter-domain path that includes or excludes a certain set of nodes in each domain should be supported. The inter-domain solution should be independent of the mechanisms used to achieve the node inclusion/exclusion constraints within a domain. For example, an RSVP-based domain may use link affinities to achieve node exclusion constraints, while an SR-based domain may use flex-algo, which natively supports excluding nodes.

4.3.5. Domain Inclusion/Exclusion Constraints

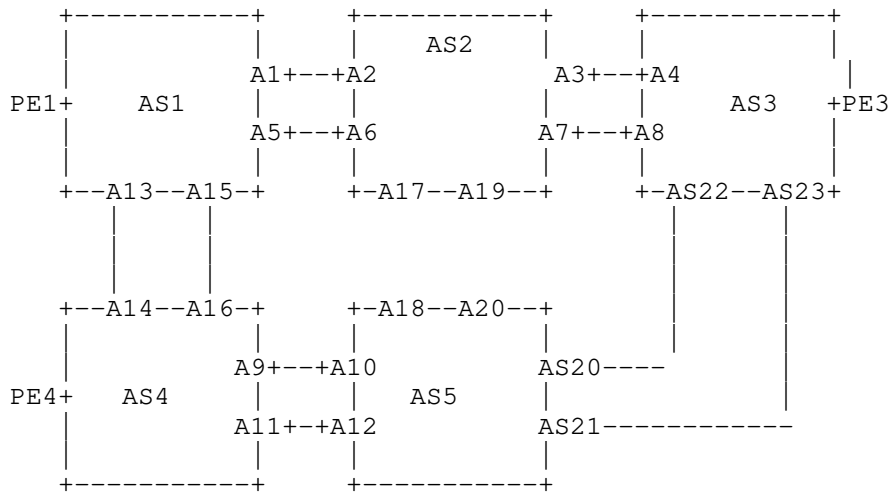


Figure 10: Multi-domain Network

Diagram Figure 10 shows a multi-domain, multi-AS network with the possibility for AS-diverse paths. The inter-domain solution should support creation of end-to-end paths that includes/excludes a certain domain entirely. For example, a network operator should be able to use the solution to create a path from PE1 to PE3 that automatically avoids passing through nodes belonging to AS2.

4.3.6. Diverse Paths

The solution should support the creation of node and link-diverse inter-domain paths.

The intra-domain portion of the end-to-end paths should make use of existing mechanisms for computing and instantiating diverse paths within a domain.

Inter-domain links (such as those connecting ASBRs) should also be taken into account for diverse inter-domain paths.

The solution should support SRLG-aware inter-domain diverse paths.

4.3.7. Constraint applicability to a subset of domains

Use cases such as data sovereignty described in Section 3.1 require that the paths with certain constraints are applicable to only a subset of domains. In domains where a constraint is not applicable, the end-to-end path should not create any state on the internal nodes.

4.3.8. Service function chaining

Support the case where the set of service functions to be applied are deployed in single domain.

Support the case where the set of service functions to be applied are deployed across multiple domains.

Support virtualized service functions as well as service functions based on physical appliances.

Support the movement of a virtualized service function from one location to another.

Support high availability for service functions.

4.4. Multicast specific requirements

Many of the requirements above are applicable to multicast traffic as well. Some requirements need to be refined with respect to multicast. Multicast also has some unique requirements not shared by unicast. These requirements will be covered in a future version of this document.

4.5. Interoperate with BGP-LU

Seamless MPLS architecture is widely deployed and BGP-LU [RFC3107] is used to connect different domains. The inter-domain solution for intent-based paths should be interoperable with BGP-LU.

4.6. Merger and Migration Requirements

4.6.1. Option A and Option B Usecases

Options A and B require additional state on border nodes, so they are typically less scalable than option C. However, options A and B can be advantageous when it is necessary to do filtering or policing on border nodes. When option A or B is deployed, the solution should still meet the SLA requirements described in Section 4.3.

4.6.2. Inter-Domain Intent Translation

In cases where two network domains previously under different administrations merge to come under a single administration, it may be preferable to use option C connectivity between the domains. The paths that fulfill the same intent may be represented using different conventions in each domain. The inter-domain solution should support efficient translation of intent from one representation to another.

4.6.3. Native Support for Best Effort Paths

The inter-domain solution for intent-based paths should also provide the ability to create end-to-end best effort paths with accumulated IGP metric across the domains. A deployment should not require two different mechanisms to be deployed for best effort and intent-based paths.

4.6.4. Interoperate with Other tunneling Mechanisms

As described in Section 4.2.1 and Section 3.6 the inter-domain solution should support one domain having one type of tunneling mechanism and another domain having another type of tunneling mechanism. The different tunneling mechanisms may completely differ in control plane and data plane operations (e.g. SRv6 and MPLS.) The inter-domain solution should provide interoperability between various tunneling mechanisms and provide the ability to create end-to-end intent-based paths.

4.7. Scalability Requirements

The inter-domain solution should be able to support up to 1 million nodes.

The inter-domain solution should facilitate the use of access nodes with low RIB/FIB and low CPU capabilities.

The inter-domain solution should facilitate the use of access nodes with low label stacking capability.

The inter-domain solution should allow for a scalable response to network events. An individual node should only need to respond to a limited subset of network events.

Service routes on the border nodes should be minimized.

Non-MPLS versions of the inter-domain solution should support summarization of prefixes in order achieve scalability.

The inter-domain solution should facilitate filtering in order to ensure the access nodes need to receive and process only the endpoint prefixes that the access node needs to send traffic to.

The inter-domain solution should minimize state on the border nodes in order to reduce label and FIB resource consumption on border nodes.

4.8. Availability Requirements

Traffic should be Fast Reroute (FRR) protected against link, node, and SRLG failures within a domain.

Traffic should be FRR protected against border node failures.

Traffic should be FRR protected against inter-domain link failures.

Traffic should be FRR protected against egress node and egress link failures.

4.9. Operations and Automation Requirements

Each domain should be independent and should not depend on the transport technology in another domain. This allows for more flexible evolution of the network.

Basic MPLS OAM mechanisms described in [RFC8029] should be supported for MPLS based solutions.

End-to-end ping and traceroute procedures should be supported.

The ability to validate the path inside each domain should be supported.

Statistics for inter-domain intent-based paths should be supported on a per path basis on the ingress and egress PE nodes as well as border nodes.

The choice of transport tunnels that make up the inter-domain path should be derived automatically from the intent that the path fulfills.

The intent defined as color in the SR-TE architecture [I-D.ietf-idr-segment-routing-te-policy] should map automatically for all controller to router protocols such as BGP-SR-TE [I-D.ietf-idr-segment-routing-te-policy], PCEP-SR [I-D.ietf-pce-segment-routing-policy-cp], and NETCONF.

The intent should be mapped automatically from flex-algo number [I-D.ietf-lsr-flex-algo].

When access devices have CPU and memory constraints, it is useful to be able to filter prefix advertisements using policies as described in Section 4.7 For large networks it is operationally a tedious and erroneous process to manage this. The inter-domain solution should facilitate filtering the advertisements automatically, based on the service prefixes it receives from end-points.

4.10. Service Mapping Requirements

The above requirements focus on the service independent aspects of inter-domain intent-based paths. In order for different services to effectively use these paths, flexible service mapping is required. The sections below summarize the requirements needed to achieve flexible service mapping.

4.10.1. Traffic service mapping

Automated steering of traffic onto transport paths based on communities carried in the service prefix advertisements should be supported.

Steering of traffic on to transport paths based on the DSCP value carried in IPv4/IPv6 packets should be supported.

Traffic steering based on EXP bits in the MPLS header should be supported.

Traffic steering based on 5-tuple packet filter should be supported. Source address, destination address, source port, destination port and protocol fields should be allowed.

All the above traffic steering mechanisms should be supported for all common types of service traffic, including L2 VPN and L3 VPN traffic and global internet traffic.

When a path that fulfills the desired intent is not available, fallback to a path that fulfills a secondary intent should be supported.

When a path that fulfills the desired intent is not available, fallback to a best-effort path should be supported.

When a path that fulfills the desired intent is not available, the option of not using a fallback path (i,e. dropping the traffic) should be supported.

4.10.2. 1 to N service mapping

The core domain is expected to have more traffic engineering constraints as compared to metros. The ability to map the services to appropriate transport tunnels at service attachment points should be supported.

4.11. Interaction with Other Approaches

This document focuses on use cases and requirements that may benefit from a distributed solution. Many of these same use cases and requirements can be addressed with centralized approaches or other distributed TE solutions. One example of a centralized approach is described in "Interconnecting Millions of Endpoints with Segment Routing" ([RFC8604]).

Distributed and centralized approaches have inherent tradeoffs. Some networks may use a single approach. Other networks may choose to use both distributed and centralized approaches to get the benefits of both. A distributed inter-domain solution should support the requirements below:

Support scenarios where some traffic uses paths created using a centralized approach, and other traffic uses paths created using the distributed solution.

Support scenarios where part of the distributed inter-domain path is created using a centralized approach.

Support scenarios where traffic uses a centralized inter-domain solution for primary traffic, and uses a distributed inter-domain solution as a backup.

The distributed solution should not have any inherent dependencies on centralized approaches.

The distributed solution should co-exist with other distributed TE solutions.

5. Backward Compatibility

6. Security Considerations

TBD

7. IANA Considerations

8. Acknowledgements

Many thanks to Kireeti Kompella, Ron Bonica, Krzysztof Szarcowitz, Srihari Sangli, Julian Lucek, Ram Santhanakrishnan, for discussions and inputs. Thanks to Colby Barth, John Scudder, Joel Halpern for review and comments.

9. Contributors

1.Kaliraj Vairavakkalai

Juniper Networks

kaliraj@juniper.net

2. Jeffrey Zhang

Juniper Networks

zzhang@juniper.net

10. References

10.1. Normative References

[I-D.hegde-rtgwg-egress-protection-sr-networks]
Hegde, S., Lin, W., and S. Peng, "Egress Protection for Segment Routing (SR) networks", draft-hegde-rtgwg-egress-protection-sr-networks-01 (work in progress), November 2020.

[I-D.ietf-idr-performance-routing]
Xu, X., Hegde, S., Talaulikar, K., Boucadair, M., and C. Jacquenet, "Performance-based BGP Routing Mechanism", draft-ietf-idr-performance-routing-03 (work in progress), December 2020.

[I-D.kaliraj-idr-bgp-classful-transport-planes]
Vairavakkalai, K., Venkataraman, N., Rajagopalan, B., Mishra, G., Khaddam, M., and X. Xu, "BGP Classful Transport Planes", draft-kaliraj-idr-bgp-classful-transport-planes-06 (work in progress), January 2021.

[I-D.zzhang-bess-bgp-multicast]
Zhang, Z., Giuliano, L., Patel, K., Wijnands, I., mishra, m., and A. Gulko, "BGP Based Multicast", draft-zzhang-bess-bgp-multicast-03 (work in progress), October 2019.

[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>.

[RFC3107] Rekhter, Y. and E. Rosen, "Carrying Label Information in BGP-4", RFC 3107, DOI 10.17487/RFC3107, May 2001, <https://www.rfc-editor.org/info/rfc3107>.

[RFC8669] Previdi, S., Filsfils, C., Lindem, A., Ed., Sreekantiah, A., and H. Gredler, "Segment Routing Prefix Segment Identifier Extensions for BGP", RFC 8669, DOI 10.17487/RFC8669, December 2019, <https://www.rfc-editor.org/info/rfc8669>.

10.2. Informative References

[I-D.hegde-spring-node-protection-for-sr-te-paths]
Hegde, S., Bowers, C., Litkowski, S., Xu, X., and F. Xu, "Node Protection for SR-TE Paths", draft-hegde-spring-node-protection-for-sr-te-paths-07 (work in progress), July 2020.

[I-D.ietf-idr-link-bandwidth]
Mohapatra, P. and R. Fernando, "BGP Link Bandwidth Extended Community", draft-ietf-idr-link-bandwidth-07 (work in progress), March 2018.

[I-D.ietf-idr-segment-routing-te-policy]
Previdi, S., Filsfils, C., Talaulikar, K., Mattes, P., Rosen, E., Jain, D., and S. Lin, "Advertising Segment Routing Policies in BGP", draft-ietf-idr-segment-routing-te-policy-11 (work in progress), November 2020.

[I-D.ietf-idr-tunnel-encaps]
Patel, K., Velde, G., Sangli, S., and J. Scudder, "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-21 (work in progress), January 2021.

[I-D.ietf-lsr-flex-algo]
Psenak, P., Hegde, S., Filsfils, C., Talaulikar, K., and A. Gulko, "IGP Flexible Algorithm", draft-ietf-lsr-flex-algo-13 (work in progress), October 2020.

[I-D.ietf-mpls-seamless-mpls]
Leymann, N., Decraene, B., Filsfils, C., Konstantynowicz, M., and D. Steinberg, "Seamless MPLS Architecture", draft-ietf-mpls-seamless-mpls-07 (work in progress), June 2014.

[I-D.ietf-pce-segment-routing-policy-cp]
Koldychev, M., Sivabalan, S., Barth, C., Peng, S., and H. Bidgoli, "PCEP extension to support Segment Routing Policy Candidate Paths", draft-ietf-pce-segment-routing-policy-cp-02 (work in progress), January 2021.

[I-D.ietf-rtgwg-segment-routing-ti-lfa]
Litkowski, S., Bashandy, A., Filsfils, C., Decraene, B., and D. Voyer, "Topology Independent Fast Reroute using Segment Routing", draft-ietf-rtgwg-segment-routing-ti-lfa-05 (work in progress), November 2020.

[I-D.ietf-spring-segment-routing-policy]
Filsfils, C., Talaulikar, K., Voyer, D., Bogdanov, A., and P. Mattes, "Segment Routing Policy Architecture", draft-ietf-spring-segment-routing-policy-09 (work in progress), November 2020.

[I-D.ietf-spring-sr-service-programming]
Clad, F., Xu, X., Filsfils, C., daniel.bernier@bell.ca, d., Li, C., Decraene, B., Ma, S., Yadlapalli, C., Henderickx, W., and S. Salsano, "Service Programming with Segment Routing", draft-ietf-spring-sr-service-programming-03 (work in progress), September 2020.

[I-D.ietf-spring-srv6-network-programming]
Filsfils, C., Camarillo, P., Leddy, J., Voyer, D., Matsushima, S., and Z. Li, "SRv6 Network Programming", draft-ietf-spring-srv6-network-programming-28 (work in progress), December 2020.

[I-D.saad-sr-fa-link]
Saad, T., Beeram, V., Barth, C., and S. Sivabalan, "Segment-Routing over Forwarding Adjacency Links", draft-saad-sr-fa-link-02 (work in progress), July 2020.

[I-D.voyer-pim-sr-p2mp-policy]
Voyer, D., Filsfils, C., Parekh, R., Bidgoli, H., and Z. Zhang, "Segment Routing Point-to-Multipoint Policy", draft-voyer-pim-sr-p2mp-policy-02 (work in progress), July 2020.

[RFC1997] Chandra, R., Traina, P., and T. Li, "BGP Communities Attribute", RFC 1997, DOI 10.17487/RFC1997, August 1996, <https://www.rfc-editor.org/info/rfc1997>.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <https://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, <https://www.rfc-editor.org/info/rfc4684>.

[RFC5065] Traina, P., McPherson, D., and J. Scudder, "Autonomous System Confederations for BGP", RFC 5065, DOI 10.17487/RFC5065, August 2007, <https://www.rfc-editor.org/info/rfc5065>.

[RFC5357] Hedayat, K., Krzanowski, R., Morton, A., Yum, K., and J. Babiarz, "A Two-Way Active Measurement Protocol (TWAMP)", RFC 5357, DOI 10.17487/RFC5357, October 2008, <https://www.rfc-editor.org/info/rfc5357>.

[RFC6388] Wijnands, IJ., Ed., Minei, I., Ed., Kompella, K., and B. Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", RFC 6388, DOI 10.17487/RFC6388, November 2011, <https://www.rfc-editor.org/info/rfc6388>.

[RFC7311] Mohapatra, P., Fernando, R., Rosen, E., and J. Uttaro, "The Accumulated IGP Metric Attribute for BGP", RFC 7311, DOI 10.17487/RFC7311, August 2014, <https://www.rfc-editor.org/info/rfc7311>.

[RFC7471] Giacalone, S., Ward, D., Drake, J., Atlas, A., and S. Previdi, "OSPF Traffic Engineering (TE) Metric Extensions", RFC 7471, DOI 10.17487/RFC7471, March 2015, <https://www.rfc-editor.org/info/rfc7471>.

[RFC7510] Xu, X., Sheth, N., Yong, L., Callon, R., and D. Black, "Encapsulating MPLS in UDP", RFC 7510, DOI 10.17487/RFC7510, April 2015, <https://www.rfc-editor.org/info/rfc7510>.

[RFC7665] Halpern, J., Ed. and C. Pignataro, Ed., "Service Function Chaining (SFC) Architecture", RFC 7665, DOI 10.17487/RFC7665, October 2015, <https://www.rfc-editor.org/info/rfc7665>.

[RFC8029] Kompella, K., Swallow, G., Pignataro, C., Ed., Kumar, N., Aldrin, S., and M. Chen, "Detecting Multiprotocol Label Switched (MPLS) Data-Plane Failures", RFC 8029, DOI 10.17487/RFC8029, March 2017, <https://www.rfc-editor.org/info/rfc8029>.

[RFC8287] Kumar, N., Ed., Pignataro, C., Ed., Swallow, G., Akiya, N., Kini, S., and M. Chen, "Label Switched Path (LSP) Ping/Traceroute for Segment Routing (SR) IGP-Prefix and IGP-Adjacency Segment Identifiers (SIDs) with MPLS Data Planes", RFC 8287, DOI 10.17487/RFC8287, December 2017, <https://www.rfc-editor.org/info/rfc8287>.

[RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <https://www.rfc-editor.org/info/rfc8402>.

[RFC8570] Ginsberg, L., Ed., Previdi, S., Ed., Giacalone, S., Ward, D., Drake, J., and Q. Wu, "IS-IS Traffic Engineering (TE) Metric Extensions", RFC 8570, DOI 10.17487/RFC8570, March 2019, <https://www.rfc-editor.org/info/rfc8570>.

[RFC8604] Filsfils, C., Ed., Previdi, S., Dawra, G., Ed., Henderickx, W., and D. Cooper, "Interconnecting Millions of Endpoints with Segment Routing", RFC 8604, DOI 10.17487/RFC8604, June 2019, <https://www.rfc-editor.org/info/rfc8604>.

[RFC8679] Shen, Y., Jeganathan, M., Decraene, B., Gredler, H., Michel, C., and H. Chen, "MPLS Egress Protection Framework", RFC 8679, DOI 10.17487/RFC8679, December 2019, <https://www.rfc-editor.org/info/rfc8679>.

[TS.23.501-3GPP]
3rd Generation Partnership Project (3GPP), "System Architecture for 5G System; Stage 2, 3GPP TS 23.501 v16.4.0", March 2020.

Authors' Addresses

Shraddha Hegde
Juniper Networks Inc.
Exora Business Park
Bangalore, KA 560103
India

Email: shraddha@juniper.net

Chris Bowers
Juniper Networks Inc.

Email: cbowers@juniper.net

Xiaohu Xu
Alibaba Inc.
Beijing
China

Email: xiaohu.xxh@alibaba-inc.com

Arkadiy Gulko
EdwardJones

Email: arkadiy.gulko@edwardjones.com

Alex Bogdanov
Google Inc.

Email: bogdanov@google.com

James Uttaro
ATT

Email: ju1738@att.com

Luay Jalil
Verizon

Email: luay.jalil@verizon.com

Mazen Khaddam
Cox communications

Email: mazen.khaddam@cox.com

Andrew Alston
Liquid Telecom

Email: andrew.alston@liquidtelecom.com

Luis M. Contreras
Telefonica
Ronda de la Comunicacion, s/n
Sur-3 building, 3rd floor
Madrid 28050
Spain

Email: luismiguel.contrerasmurillo@telefonica.com
URI: http://lmcontreras.com/

INTERNET-DRAFT
Intended status: Proposed Standard

V. Govindan
M. Mudigonda
A. Sajassi
Cisco Systems
G. Mirsky
ZTE
D. Eastlake
Futurewei Technologies

Expires: August 21, 2021
February 22, 2021

EVPN Network Layer Fault Management
draft-ietf-bess-evpn-bfd-03

Abstract

This document specifies proactive, in-band network layer OAM mechanisms to detect loss of continuity and miss-connection faults that affect unicast and multi-destination paths (used by Broadcast, Unknown Unicast, and Multicast traffic) in an Ethernet VPN (EVPN) network. The mechanisms specified in the draft are based on the widely adopted Bidirectional Forwarding Detection (BFD) protocol.

Table of Contents

1. Introduction.....3
1.1 Terminology.....3

2. Scope of this Document.....5
3. Motivation for Running BFD at the EVPN Network Layer....6
4. Fault Detection for Unicast Traffic.....7

5. Fault Detection for BUM Traffic.....8
5.1 Ingress Replication.....8
5.2 P2MP Tunnels (Label Switched Multicast).....8

6. BFD Packet Encapsulation.....10
6.1 MPLS Encapsulation.....10
6.1.1 Unicast MPLS Encapsulation.....10
6.1.2 MPLS Ingress Replication (MP2P).....11
6.1.3 MPLS LSM (Label Switched Multicast, P2MP).....12
6.2 VXLAN Encapsulation.....12
6.2.1 Unicast VXLAN Encapsulation.....12
6.2.2 VXLAN Ingress Replication (MP2P).....14
6.2.3 VXLAN LSM (Label Switched Multicast, P2MP).....14

7. Scalability Considerations.....15

8. IANA Considerations.....16
8.1 Pseudowire Associated Channel Type.....16
8.2 MAC Addresses.....16
8.3 BFD Discriminator Attribute Type.....16

9. Security Considerations.....17

Acknowledgements.....17

Normative References.....18
Informative References.....20

Authors' Addresses.....21

1. Introduction

[ietf-bess-evpn-oam-req-frmwk] outlines the OAM requirements of Ethernet VPN networks (EVPN [RFC7432]). This document specifies mechanisms for proactive fault detection at the network (overlay) layer of EVPN. The mechanisms proposed in the draft use the widely adopted Bidirectional Forwarding Detection (BFD [RFC5880]) protocol.

EVPN fault detection mechanisms need to consider unicast traffic separately from Broadcast, Unknown Unicast, and Multicast (BUM) traffic since they map to different Forwarding Equivalency Classes (FECs) in EVPN. Hence this document proposes somewhat different fault detection mechanisms depending on the type of traffic and the type of tunnel used as follows:

- Using BFD [RFC5880] for unicast traffic and BUM traffic via MP2P tunnels.
- Using BFD Multipoint Active Tails [RFC8563] [mirsky-mpls-p2mp-bfd] for BUM traffic via a P2MP tunnel.

Packet loss and packet delay measurement are out of scope for this document.

1.1 Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

The following acronyms are used in this document.

BFD - Bidirectional Forwarding Detection [RFC5880]

BUM - Broadcast, Unknown Unicast, and Multicast

CC - Continuity Check

CV - Connectivity Verification

EVI - EVPN Instance

EVPN - Ethernet VPN [RFC7432]

FEC - Forwarding Equivalency Class

GAL - Generic Associated Channel Label [RFC5586]

LSM - Label Switched Multicast (P2MP)

LSP - Label Switched Path

MP2MP - Multi-Point-to-Multi-Point

MP2P - Multi-Point-to-Point

OAM - Operations, Administration, and Maintenance

P2MP - Point-to-Multi-Point (LSM)

P2P - Point to Point.

PE - Provider Edge

VXLAN - Virtual eXtensible Local Area Network (VXLAN) [RFC7348]

2. Scope of this Document

This document specifies BFD based mechanisms for proactive fault detection for EVPN as specified in [RFC7432] and also for EVPN using VXLAN encapsulation [RFC8365]. It covers the following:

- Unicast traffic using Point-to-Point (P2P) and Multi-Point-to-Point (MP2P) tunnels.
- BUM traffic using Multi-Point-to-Point (MP2P) tunnels (ingress replication).
- BUM traffic using Point-to-Multi-Point (P2MP) tunnels (Label Switched Multicast (LSM)).
- MPLS and VXLAN encapsulation.

This document does not discuss BFD mechanisms for:

- EVPN variants like PBB-EVPN [RFC7623]. It is intended to address this in future versions.
- Integrated Routing and Bridging (IRB) solution based on EVPN [ietf-bess-evpn-inter-subnet-forwarding]. It is intended to address this in future versions.
- EVPN using other encapsulations such as NVGRE or MPLS over GRE [RFC8365].
- BUM traffic using MP2MP tunnels.

This document specifies procedures for BFD asynchronous mode. BFD demand mode is outside the scope of this specification except as it is used in [RFC8563]. The use of the Echo function is outside the scope of this specification.

3. Motivation for Running BFD at the EVPN Network Layer

The choice of running BFD at the network layer of the OAM model for EVPN [ietf-bess-evpn-oam-req-frmwk] was made after considering the following:

- In addition to detecting link failures in the EVPN network, BFD sessions at the network layer can be used to monitor the successful setup, such as label programming, of MP2P and P2MP EVPN tunnels transporting Unicast and BUM traffic. The scope of reachability detection covers the ingress and the egress EVPN PE (Provider Edge) nodes and the network connecting them.

- Monitoring a representative set of path(s) or a particular path among multiple paths available between two EVPN PE nodes could be done by exercising entropy mechanisms such as entropy labels, when they are used, or VXLAN source ports. However, paths that cannot be realized by entropy variations cannot be monitored. The fault monitoring requirements outlined by [ietf-bess-evpn-oam-req-frmwk] are addressed by the mechanisms proposed by this draft.

BFD testing between EVPN PE nodes does not guarantee that the EVPN service is functioning. (This can be monitored at the service level, that is CE to CE.) For example, an egress EVPN-PE could understand EVPN labeling received but could switch data to an incorrect interface. However, BFD testing in the EVPN Network Layer does provide additional confidence that data transported using those tunnels will reach the expected egress node. When BFD testing in the EVPN overlay fails, that can be used as an indication of a Loss-of-Connectivity defect in the EVPN underlay that would cause EVPN service failure.

4. Fault Detection for Unicast Traffic

The mechanisms specified in BFD for MPLS LSPs [RFC5884] [RFC7726] are, except as otherwise provided herein, applied to test the handling of unicast EVPN traffic. When discriminators are required for de-multiplexing the BFD sessions, they can be advertised through BGP using the BFD Discriminator Attribute [ietf-bess-mvpn-fast-failover]. Discriminators are needed for MPLS since the label stack does not contain enough information to identify the sender of the packet.

The usage of MPLS entropy labels or various VXLAN source ports takes care of the requirement to monitor various paths of the multi-path server layer network [RFC6790]. Each unique realizable path between the participating PE routers MAY be monitored separately when such entropy is used. At least one path of multi-path connectivity between two PE routers MUST be tracked with BFD, but in that case the granularity of fault-detection will be coarser.

To support unicast fault management to a PE node, that PE MUST allocate or be configured with a BFD discriminator to be used for the BFD messages to it. By default, it advertises this discriminator with BGP using the BFD Discriminator Attribute [ietf-bess-mvpn-fast-failover] with BFD Mode TBD4 in an EVPN MAC/IP Advertisement Route [RFC7432] and extracts its peer's discriminator from such an attribute. however, these discriminators MAY be exchanged out-of-band or through some other mechanism outside the scope of this document.

If configured to do so, once a PE knows a unicast route and discriminator for another PE, it endeavors to bring UP and maintain a BFD session to that other PE. Once the BFD session is UP, the ends of the BFD session MUST NOT change the local discriminator values of the BFD Control packets they generate, unless they first bring down the session as specified in [RFC5884]. The BFD session is brought down if a PE is no longer configured to maintain it or if a route and discriminator are no longer available.

5. Fault Detection for BUM Traffic

Section 5.1 below discusses BUM traffic fault detection for MP2P tunnels using ingress replication and Section 5.2 discusses such fault detection for P2MP tunnels.

5.1 Ingress Replication

Ingress replication uses separate MP2P tunnels for transporting BUM traffic from the ingress PE (head) to a set of one or more egress PEs (tails). The fault detection mechanism specified by this document takes advantage of the fact that the head makes a unique copy for each tail.

Another key aspect to be considered in EVPN is the advertisement of the Inclusive Multicast Ethernet Tag Route [RFC7432]. The BUM traffic flows from a head node to a particular tail only after the head receives such inclusive multicast route from the tail. This route contains the BUM EVPN MPLS label (downstream allocated) corresponding to the MP2P tunnel for MPLS encapsulation and contains the IP address of the PE originating the inclusive multicast route for use in VXLAN encapsulation. It also contains a BFD Discriminator Attribute [ietf-bess-mvpn-fast-failover] with BFD Mode TDB4 giving the BFD discriminator that will be used by the tail. This is the P2P mode since a P2P BFD session is used in the MP2P case.

There MAY exist multiple BFD sessions between a head PE and an individual tail due to (1) the usage of MPLS entropy labels [RFC6790] or VXLAN source ports for an inclusive multicast FEC and (2) due to multiple MP2P tunnels indicated by different tail labels or IP addresses for MPLS or VXLAN. If configured to do so, once a PE knows a multicast route and discriminator for another PE it endeavors to bring UP and maintain a BFD session to that other PE. Once a BFD session for a path is UP, the ends of the BFD session MUST NOT change the local discriminator values of the BFD Control packets they generate, unless they first bring down the session as specified in [RFC5884]. The BFD session is brought down if a PE is no longer configured to maintain it or if a route and discriminator are no longer available.

5.2 P2MP Tunnels (Label Switched Multicast)

Fault detection for BUM traffic distributed using a P2MP tunnel uses BFD Multipoint Active Tails in one of the three methods providing head notification depending on configuration. Sections 5.2.2 and 5.2.3 of [RFC8563] describe two of these methods ("Head Notification

and Tail Solicitation with Multipoint Polling" and "Head Notification with Composite Polling"). The third method ("Head Notification without Polling") is touched on in Section 5.2.1 of [RFC8563] and fully specified in [mirsky-mpls-p2mp-bfd]. All three of these modes assume the existence of a unicast path from each tail to the head. In addition, Head Notification with Composite Polling assumes a head to tail unicast path.

The BUM traffic flows from a head node to the tails after the head receives an Inclusive Multicast Tag Route [RFC7432]. This contains the BUM EVPN MPLS label (upstream allocated) corresponding to the P2MP tunnel for MPLS encapsulation. It also contains a BFD Discriminator Attribute [ietf-bess-mvpn-fast-failover] with BFD Mode 1 and with a Source IP Address TLV giving the address associated with the MultiPoint Head of the P2MP session. This BFD discriminator advertised by a tail in the inclusive multicast route or otherwise configured at or communicated to the head MUST be used in any reverse unicast traffic so the head can determine which tail is responding. If configured to do so, once a PE knows a P2MP multicast route and needed discriminators, it brings UP and maintains a BFD active tails session to the tails. The BFD session is brought down if a PE is no longer configured to maintain it or if the multicast route and discriminators are no longer available.

For MPLS encapsulation of the head to tails BFD, Label Switched Multicast is used. For VXLAN encapsulation, BFD is delivered to the tails through underlay multicast using an outer multicast IP address.

6. BFD Packet Encapsulation

The sections below discuss the MPLS and VXLAN encapsulations of BFD for EVPN network layer fault management.

6.1 MPLS Encapsulation

This section describes use of the Generic Associated Channel Label (GAL) for BFD encapsulation in MPLS based EVPN network layer fault management.

6.1.1 Unicast MPLS Encapsulation

As shown in Figure 1, the packet initially contains the following labels: LSP label (transport), the optional entropy label, the EVPN Unicast label, and then the Generic Associated Channel label with the G-ACh type set to TBD1. The G-ACh payload of the packet MUST contain the destination L2 header (in overlay space) followed by the IP header that encapsulates the BFD packet. The MAC address of the inner packet is used to validate the <EVI, MAC> in the receiving node.

- The destination MAC MUST be the dedicated unicast MAC TBD3 (see Section 8) or the MAC address of the destination PE.
- The destination IP address MUST be 127.0.0.1/32 for IPv4 [RFC1812] or ::1/128 for IPv6 [RFC4291].
- The destination IP port MUST be 3784 [RFC5881].
- The source IP port MUST be in the range 49152 through 65535.
- The discriminator values for BFD are obtained through BGP as discussed in Section 4.

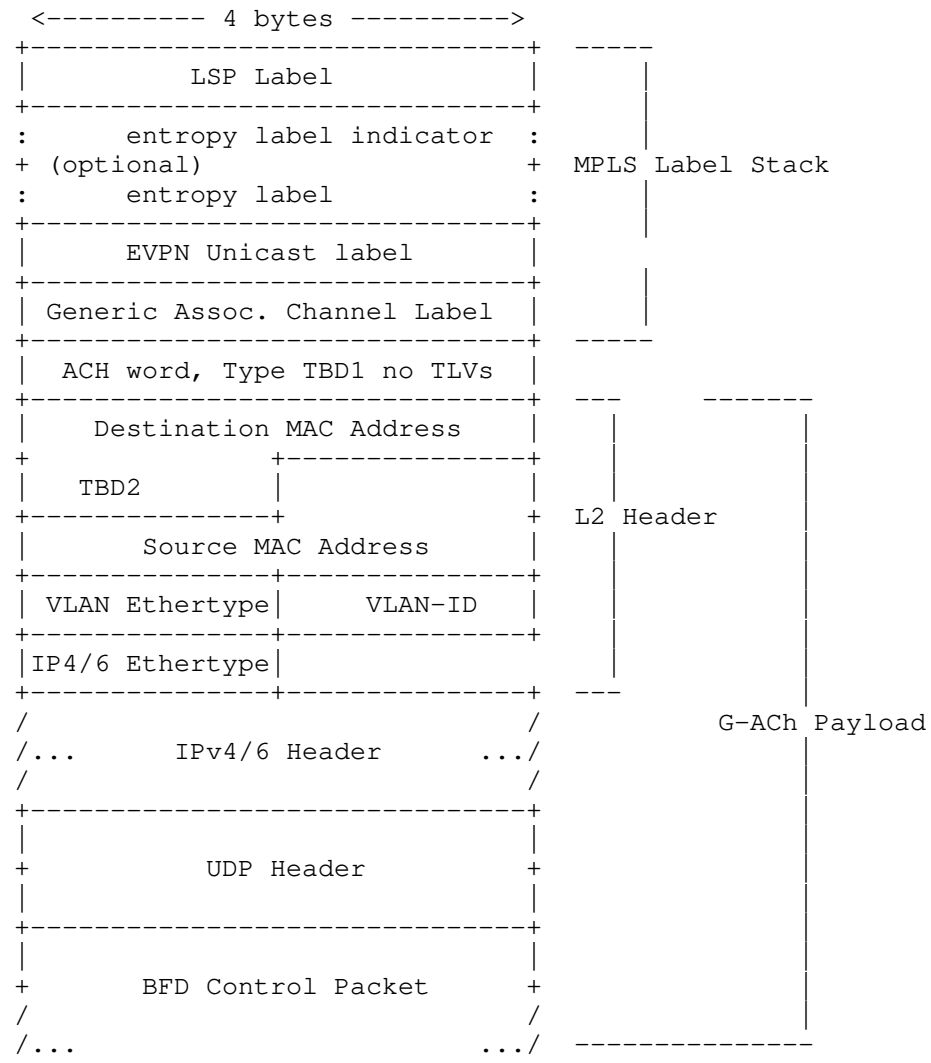


Figure 1. MPLS Unicast Encapsulation

6.1.2 MPLS Ingress Replication (MP2P)

The packet initially contains the following labels: LSP label (transport), the optional entropy label, the BUM label, and the split horizon label [RFC7432] (where applicable). The G-ACh type is set to TBD1. The G-ACh payload of the packet is as described in Section 6.1.1 except that the destination MAC address is the dedicated multicast MAC TBD2.

6.1.3 MPLS LSM (Label Switched Multicast, P2MP)

The encapsulation is the same as in Section 6.1.2 for ingress replication except that the transport label identifies the P2MP tunnel, in effect the set of tail PEs, rather than identifying a single destination PE at the end of an MP2P tunnel.

6.2 VXLAN Encapsulation

This section describes the use of the VXLAN [RFC7348] [RFC8365] for BFD encapsulation in VXLAN based EVPN fault management.

6.2.1 Unicast VXLAN Encapsulation

Figure 2 below shows the unicast VXLAN encapsulation. The outer and inner IP headers have a unicast source IP address of the BFD message source and a destination IP address of the BFD message destination

The destination UDP port MUST be 3784 [RFC5881]. The source port MUST be in the range 49152 through 65535. If the BFD source has multiple IP addresses, entropy MAY be further obtained by using any of those addresses assuming the source is prepared for responses directed to the IP address used.

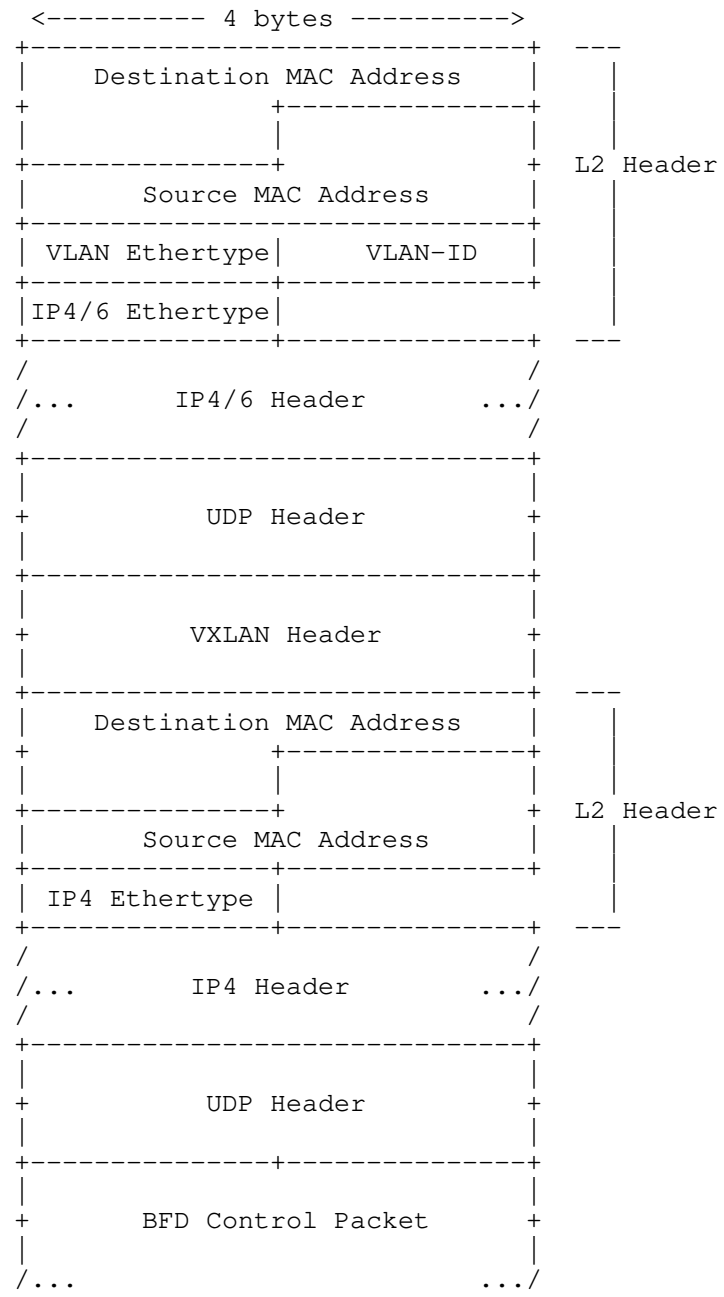


Figure 2. VXLAN Unicast Encapsulation

6.2.2 VXLAN Ingress Replication (MP2P)

The BFD packet construction is as given in Section 6.2.1 except as follows:

(1) The destination IP address used by the BFD message source is that advertised by the destination PE in its Inclusive Multicast EVPN route for the MP2P tunnel in question; and

(2) The Your BFD discriminator used is the one advertised by the BFD destination using BGP as discussed in Section 5.1 for the MP2P tunnel.

6.2.3 VXLAN P2MP

The VXLAN encapsulation for the head-to-tails BFD packets uses the multicast destination IP corresponding to the VXLAN VNI.

The destination port MUST be 3784. For entropy purposes, the source port can vary but MUST be in the range 49152 through 65535 [RFC5881]. If the head PE has multiple IP addresses, entropy MAY be further obtained by using any of those addresses.

The Your BFD discriminator is the value distributed for this multicast fault management purpose as discussed in Section 5.2.

7. Scalability Considerations

The mechanisms proposed by this draft could affect the packet load on the network and its elements especially when supporting configurations involving a large number of EVIs. The option of slowing down or speeding up BFD timer values can be used by an administrator or a network management entity to maintain the overhead incurred due to fault monitoring at an acceptable level.

8. IANA Considerations

The following IANA Actions are requested.

8.1 Pseudowire Associated Channel Type

IANA is requested to assign a channel type from the "Pseudowire Associated Channel Types" registry in [RFC4385] as follows.

Value	Description	Reference
TBD1	BFD-EVPN OAM	[this document]

8.2 MAC Addresses

IANA is requested to assign parallel multicast and unicast MAC addresses under the IANA OUI [0x01005E900101 and 0x00005E900101 suggested] as follows:

IANA Multicast 48-bit MAC Addresses

Address	Usage	Reference
TBD2	EVPN Network Layer OAM	[this document]

IANA Unicast 48-bit MAC Addresses

Address	Usage	Reference
TBD3	EVPN Network Layer OAM	[this document]

8.3 BFD Discriminator Attribute Type

IANA is requested to assign a value from the IETF Review range in the BFD Mode sub-registry on the Border Gateway Protocol Parameters Registry web page as follows:

Value	Description	Reference
TBD4	P2P BFD Session	[this document]

9. Security Considerations

Security considerations discussed in [RFC5880], [RFC5883], and [RFC8029] apply.

MPLS security considerations [RFC5920] apply to BFD Control packets encapsulated in a MPLS label stack. When BPD Control packets are routed, the authentication considerations discussed in [RFC5883] should be followed.

VXLAN BFD security considerations in [RFC8971] apply to BFD packets encapsulate in VXLAN.

Acknowledgements

The authors wish to thank the following for their comments and suggestions:

Mach Chen

Normative References

[ietf-bess-mvpn-fast-failover] Morin, T., Kebler, R., Mirsky, G., "Multicast VPN fast upstream failover", draft-ietf-bess-mvpn-fast-failover (in RFC Editor's queue), February 2019.

[mirsky-mpls-p2mp-bfd] G. Mirsky, S. Mishra, "BFD for Multipoint Networks over Point-to-Multi-Point MPLS LSP", draft-mirsky-mpls-p2mp-bfd (work in progress), November 2020.

[RFC1812] Baker, F., Ed., "Requirements for IP Version 4 Routers", RFC 1812, DOI 10.17487/RFC1812, June 1995, <https://www.rfc-editor.org/info/rfc1812>.

[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>.

[RFC4291] Hinden, R. and S. Deering, "IP Version 6 Addressing Architecture", RFC 4291, DOI 10.17487/RFC4291, February 2006, <https://www.rfc-editor.org/info/rfc4291>.

[RFC4385] Bryant, S., Swallow, G., Martini, L., and D. McPherson, "Pseudowire Emulation Edge-to-Edge (PWE3) Control Word for Use over an MPLS PSN", RFC 4385, DOI 10.17487/RFC4385, February 2006, <http://www.rfc-editor.org/info/rfc4385>.

[RFC5586] Bocci, M., Ed., Vigoureux, M., Ed., and S. Bryant, Ed., "MPLS Generic Associated Channel", RFC 5586, DOI 10.17487/RFC5586, June 2009, <https://www.rfc-editor.org/info/rfc5586>.

[RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, DOI 10.17487/RFC5880, June 2010, <http://www.rfc-editor.org/info/rfc5880>.

[RFC5881] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD) for IPv4 and IPv6 (Single Hop)", RFC 5881, DOI 10.17487/RFC5881, June 2010, <https://www.rfc-editor.org/info/rfc5881>.

[RFC5883] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD) for Multihop Paths", RFC 5883, DOI 10.17487/RFC5883, June 2010, <https://www.rfc-editor.org/info/rfc5883>.

[RFC5884] Aggarwal, R., Kompella, K., Nadeau, T., and G. Swallow, "Bidirectional Forwarding Detection (BFD) for MPLS Label Switched Paths (LSPs)", RFC 5884, DOI 10.17487/RFC5884, June 2010, <https://www.rfc-editor.org/info/rfc5884>.

[RFC6790] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", RFC 6790, DOI 10.17487/RFC6790, November 2012, <http://www.rfc-editor.org/info/rfc6790>.

[RFC7348] Mahalingam, M., Dutt, D., Duda, K., Agarwal, P., Kreeger, L., Sridhar, T., Bursell, M., and C. Wright, "Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", RFC 7348, DOI 10.17487/RFC7348, August 2014, <https://www.rfc-editor.org/info/rfc7348>.

[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>.

[RFC7726] Govindan, V., Rajaraman, K., Mirsky, G., Akiya, N., and S. Aldrin, "Clarifying Procedures for Establishing BFD Sessions for MPLS Label Switched Paths (LSPs)", RFC 7726, DOI 10.17487/RFC7726, January 2016, <https://www.rfc-editor.org/info/rfc7726>.

[RFC8029] Kompella, K., Swallow, G., Pignataro, C., Ed., Kumar, N., Aldrin, S., and M. Chen, "Detecting Multiprotocol Label Switched (MPLS) Data-Plane Failures", RFC 8029, DOI 10.17487/RFC8029, March 2017, <https://www.rfc-editor.org/info/rfc8029>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

[RFC8365] Sajassi, A., Ed., Drake, J., Ed., Bitar, N., Shekhar, R., Uttaro, J., and W. Henderickx, "A Network Virtualization Overlay Solution Using Ethernet VPN (EVPN)", RFC 8365, DOI 10.17487/RFC8365, March 2018, <https://www.rfc-editor.org/info/rfc8365>.

[RFC8563] Katz, D., Ward, D., Pallagatti, S., Ed., and G. Mirsky, Ed., "Bidirectional Forwarding Detection (BFD) Multipoint Active Tails", RFC 8563, DOI 10.17487/RFC8563, April 2019, <https://www.rfc-editor.org/info/rfc8563>.

Informative References

[ietf-bess-evpn-inter-subnet-forwarding] Sajassi, A., Salam, S., Thoria, S., Rekhter, Y., Drake, J., Yong, L., and L. Dunbar, "Integrated Routing and Bridging in EVPN", draft-ietf-bess-evpn-inter-subnet-forwarding-13 (work in progress), February 2021.

[ietf-bess-evpn-oam-req-frmwk] Salam, S., Sajassi, A., Aldrin, S., J. Drake, and D. Eastlake, "EVPN Operations, Administration and Maintenance Requirements and Framework", draft-ietf-bess-evpn-oam-req-frmwk-04, work in progress, July 2019.

[RFC5920] Fang, L., Ed., "Security Framework for MPLS and GMPLS Networks", RFC 5920, DOI 10.17487/RFC5920, July 2010, <https://www.rfc-editor.org/info/rfc5920>.

[RFC7623] Sajassi, A., Ed., Salam, S., Bitar, N., Isaac, A., and W. Henderickx, "Provider Backbone Bridging Combined with Ethernet VPN (PBB-EVPN)", RFC 7623, DOI 10.17487/RFC7623, September 2015, <http://www.rfc-editor.org/info/rfc7623>.

[RFC8971] Pallagatti, S., Ed., Mirsky, G., Ed., Paragiri, S., Govindan, V., and M. Mudigonda, "Bidirectional Forwarding Detection (BFD) for Virtual eXtensible Local Area Network (VXLAN)", RFC 8971, DOI 10.17487/RFC8971, December 2020, <https://www.rfc-editor.org/info/rfc8971>.

Authors' Addresses

Vengada Prasad Govindan
Cisco Systems

Email: venggovi@cisco.com

Mudigonda Mallik
Cisco Systems

Email: mmudigon@cisco.com

Ali Sajassi
Cisco Systems
170 West Tasman Drive
San Jose, CA 95134, USA

Email: sajassi@cisco.com

Gregory Mirsky
ZTE Corp.

Email: gregimirsky@gmail.com

Donald Eastlake, 3rd
Futurewei Technologies
2386 Panoramic Circle
Apopka, FL 32703 USA

Phone: +1-508-333-2270
Email: d3e3e3@gmail.com

BESS Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 26, 2021

P. Brissette, Ed.
A. Sajassi
Cisco Systems
B. Wen
Comcast
E. Leyton
Verizon Wireless
J. Rabadan
Nokia
L. Burdet
S. Thoria
Cisco Systems

February 22, 2021

EVPN multi-homing port-active load-balancing
draft-ietf-bess-evpn-mh-pa-01

Abstract

The Multi-Chassis Link Aggregation Group (MC-LAG) technology enables the establishment of a logical link-aggregation connection with a redundant group of independent nodes. The purpose of multi-chassis LAG is to provide a solution to achieve higher network availability, while providing different modes of sharing/balancing of traffic. EVPN standard defines EVPN based MC-LAG with single-active and all-active multi-homing load-balancing mode. The current draft expands on existing redundancy mechanisms supported by EVPN and introduces support of port-active load-balancing mode. In the current document, port-active load-balancing mode is also referred to as per interface active/standby.

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 https://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 26, 2021.

Table of Contents

1. Introduction 2
1.1. Requirements Language 4
2. Multi-Chassis Ethernet Bundles 4
3. Port-active load-balancing procedure 4
4. Algorithm to elect per port-active PE 5
4.1. Capability Flag 5
4.2. Modulo-based Designated Forwarder Algorithm 6
4.3. HRW Algorithm 6
4.4. Preferred-DF Algorithm 6
5. Convergence considerations 6
5.1. Primary / Backup per Ethernet-Segment 7
5.2. Backward Compatibility 7
6. Applicability 7
7. Overall Advantages 8
8. Security Considerations 8
9. IANA Considerations 8
10. References 9
10.1. Normative References 9
10.2. Informative References 9
Authors' Addresses 10

1. Introduction

EVPN, as per [RFC7432], provides all-active per flow load balancing for multi-homing. It also defines single-active with service carving mode, where one of the PEs, in redundancy relationship, is active per service.

While these two multi-homing scenarios are most widely utilized in data center and service provider access networks, there are scenarios where active-standby per interface multi-homing redundancy is useful

and required. The main consideration for this mode of redundancy is the determinism of traffic forwarding through a specific interface rather than statistical per flow load balancing across multiple PEs providing multi-homing. The determinism provided by active-standby per interface is also required for certain QOS features to work. While using this mode, customers also expect minimized convergence during failures. A new term of load-balancing mode, port-active load- balancing is then defined.

This draft describes how that new redundancy mode can be supported via EVPN

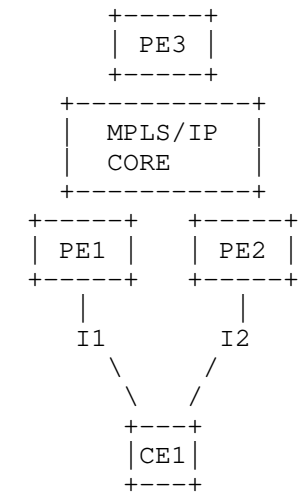


Figure 1: MC-LAG Topology

Figure 1 shows a MC-LAG multi-homing topology where PE1 and PE2 are part of the same redundancy group providing multi-homing to CE1 via interfaces I1 and I2. Interfaces I1 and I2 are Bundle-Ethernet interfaces running LACP protocol. The core, shown as IP or MPLS enabled, provides wide range of L2 and L3 services. MC-LAG multi-homing functionality is decoupled from those services in the core and it focuses on providing multi-homing to CE. With per-port active/ standby redundancy, only one of the two interface I1 or I2 would be in forwarding, the other interface will be in standby. This also implies that all services on the active interface are in active mode and all services on the standby interface operate in standby mode.

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 [RFC2119].

2. Multi-Chassis Ethernet Bundles

When a CE is multi-homed to a set of PE nodes using the [802.1AX] Link Aggregation Control Protocol (LACP), the PEs must act as if they were a single LACP speaker for the Ethernet links to form a bundle, and operate as a Link Aggregation Group (LAG). To achieve this, the PEs connected to the same multi-homed CE must synchronize LACP configuration and operational data among them. InterChassis Communicated-based Protocol (ICCP) has been used for that purpose. EVPN LAG simplifies greatly that solution. Along with the simplification comes few assumptions:

- CE device connected to Multi-homing PEs may has a single LAG with all its active links i.e. Links in the Ethernet Bundle operate in all-active load-balancing mode.

- Same LACP parameters MUST be configured on peering PEs such as system id, port priority and port key.

Any discrepancies from this list is left for future study. Furthermore, mis-configuration and mis-wiring detection across peering PEs are also left for further study.

3. Port-active load-balancing procedure

Following steps describe the proposed procedure with EVPN LAG to support port-active load-balancing mode:

a. The Ethernet-Segment Identifier (ESI) MUST be assigned per access interface as described in [RFC7432], which may be auto derived or manually assigned. Access interface MAY be a Layer-2 or Layer3 interface. The usage of ESI over L3 interfce is newly described in this document.

b. Ethernet-Segment MUST be configured in port-active load-balancing mode on peering PEs for specific access interface

c. Peering PEs MAY exchange only Ethernet-Segment route (Route Type-4) when ESI is configured on a Layer3 interface.

d. PEs in the redundancy group leverage the DF election defined in [RFC8584] to determine which PE keeps the port in active mode and

which one(s) keep it in standby mode. While the DF election defined in [RFC8584] is per [ES, Ethernet Tag] granularity, for port-active mode of multi-homing, the DF election is done per ES. The details of this algorithm are described in Section 4.

e. DF router MUST keep corresponding access interface in up and forwarding active state for that Ethernet-Segment

f. Non-DF routers MAY bring and keep peering access interface attached to it in operational down state. If the interface is running LACP protocol, then the non-DF PE MAY also set the LACP state to OOS (Out of Sync) as opposed to interface state down. This allows for better convergence on standby to active transition.

g. For EVPN-VPWS service, the usage of primary/backup bits of EVPN Layer2 attributes extended community [RFC8214] is highly recommended to achieve better convergence.

4. Algorithm to elect per port-active PE

The ES routes, running in port-active load-balancing mode, are advertised with a new capability in the DF Election Extended Community as defined in [RFC8584]. Moreover, the ES associated to the port leverages existing procedure of single-active, and signals single-active bit along with Ethernet-AD per-ES route. Finally, as in [RFC7432], the ESI-label based split-horizon procedures should be used to avoid transient echo'ed packets when L2 circuits are involved.

4.1. Capability Flag

[RFC8584] defines a DF Election extended community, and a Bitmap field to encode "capabilities" to use with the DF election algorithm in the DF algorithm field. Bitmap (2 octets) is extended by the following value:

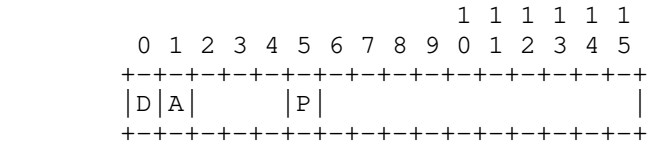


Figure 2: Amended Bitmap field in the DF Election Extended Community

Bit 0: 'Don't Preempt' bit, as explained in [PREF-DF].

Bit 1: AC-Influenced DF Election, as explained in [RFC8584].

Bit 5: (corresponds to Bit 25 of the DF Election Extended Community and it is defined by this document): P bit or 'Port Mode' bit (P hereafter), determines that the DF-Algorithm should be modified to consider the port only and not the Ethernet Tags.

4.2. Modulo-based Designated Forwarder Algorithm

The default DF Election algorithm, or modulus-based algorithm as in [RFC7432] and updated by [RFC8584], is used here, at the granularity of ES only. Given the fact, ES-Import RT community inherits from ESI only byte 1-6, many deployments differentiate ESI within these bytes only. For Modulo calculation, bytes [3-6] are used to determine the designated forwarder using Modulo-based DF assignment.

4.3. HRW Algorithm

Highest Random Weight (HRW) algorithm defined in [RFC8584] MAY also be used and signalled, and modified to operate at the granularity of ES rather than per [ES, VLAN].

[RFC8584] describes computing a 32 bit CRC over the concatenation of Ethernet Tag and ESI. For port-active load-balancing mode, the Ethernet Tag is simply removed from the CRC computation.

4.4. Preferred-DF Algorithm

When the new capability 'Port-Mode' is signalled, the algorithm is modified to consider the port only and not any associated Ethernet Tags. Furthermore, the "port-based" capability MUST be compatible with the 'DP' capability (for non-revertive). The AC-DF bit MUST be set to zero. When an AC (sub-interface) goes down, it does not influence the DF election.

5. Convergence considerations

To improve the convergence, upon failure and recovery, when port-active load-balancing mode is used, some advanced synchronization between peering PEs may be required. Port-active is challenging in a sense that the "standby" port is in down state. It takes some time to bring a "standby" port in up-state and settle the network. For IRB and L3 services, ARP / ND cache may be synchronized. Moreover, associated VRF tables may also be synchronized. For L2 services, MAC table synchronization may be considered.

Finally, for Bundle-Ethernet interface where LACP is running the ability to set the "standby" port in "out-of-sync" state aka "warm-standby" can be leveraged.

5.1. Primary / Backup per Ethernet-Segment

The L2 Info Extended Community MAY be advertised in Ethernet A-D per ES routes for fast convergence. Only the P and B bits are relevant to this specification. When advertised, the L2 Info Extended Community SHALL have only P or B bits set and all other bits must be zero. MTU must also be zero. Remote PE receiving optional L2 Info Extended Community on Ethernet A-D per ES routes SHALL consider only P and B bits. P and B bits received on Ethernet A-D per EVI routes per [RFC8214] are overridden.

5.2. Backward Compatibility

Implementations that comply with [RFC7432] or [RFC8214] only (i.e., implementations that predate this specification) will not advertise the L2 Info Extended Community in Ethernet A-D per ES routes. That means that all remote PEs in the ES will not receive P and B bit per ES and will continue to receive and honour the P and B bits Ethernet A-D per EVI routes. Similarly, an implementation that complies with [RFC7432] or [RFC8214] only and that receives a L2 Info Extended Community will ignore it and will continue to use the default path resolution algorithm.

6. Applicability

A common deployment is to provide L2 or L3 service on the PEs providing multi-homing. The services could be any L2 EVPN such as EVPN VPWS, EVPN [RFC7432], etc. L3 service could be in VPN context [RFC4364] or in global routing context. When a PE provides first hop routing, EVPN IRB could also be deployed on the PEs. The mechanism defined in this draft is used between the PEs providing the L2 and/or L3 service, when the requirement is to use per port active.

A possible alternate solution is the one described in this draft is MC-LAG with ICCP [RFC7275] active-standby redundancy. However, ICCP requires LDP to be enabled as a transport of ICCP messages. There are many scenarios where LDP is not required e.g. deployments with VXLAN or SRv6. The solution defined in this draft with EVPN does not mandate the need to use LDP or ICCP and is independent of the underlay encapsulation.

7. Overall Advantages

The use of port-active multi-homing brings the following benefits to EVPN networks:

a. Open standards based per interface single-active redundancy mechanism that eliminates the need to run ICCP and LDP.

b. Agnostic of underlay technology (MPLS, VXLAN, SRv6) and associated services (L2, L3, Bridging, E-LINE, etc).

c. Provides a way to enable deterministic QOS over MC-LAG attachment circuits.

d. Fully compliant with [RFC7432], does not require any new protocol enhancement to existing EVPN RFCs.

e. Can leverage various DF election algorithms e.g. modulo, HRW, etc.

f. Replaces legacy MC-LAG ICCP-based solution, and offers following additional benefits:

g.

* Efficiently supports 1+N redundancy mode (with EVPN using BGP RR) where as ICCP requires full mesh of LDP sessions among PEs in redundancy group.

* Fast convergence with mass-withdraw is possible with EVPN, no equivalent in ICCP

h. Customers want per interface single-active redundancy, but don't want to enable LDP (e.g. they may be running VXLAN or SRv6 in the network). Currently there is no alternative to this.

8. Security Considerations

The same Security Considerations described in [RFC7432] are valid for this document.

9. IANA Considerations

This document solicits the allocation of the following values:

o Bit 5 in the [RFC8584] DF Election Capabilities registry, with name "P" (port mode load-balancing) Capability" for port-active ES.

10. References

10.1. 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>.

[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>.

[RFC8214] Boutros, S., Sajassi, A., Salam, S., Drake, J., and J. Rabadan, "Virtual Private Wire Service Support in Ethernet VPN", RFC 8214, DOI 10.17487/RFC8214, August 2017, <https://www.rfc-editor.org/info/rfc8214>.

[RFC8584] Rabadan, J., Ed., Mohanty, S., Ed., Sajassi, A., Drake, J., Nagaraj, K., and S. Sathappan, "Framework for Ethernet VPN Designated Forwarder Election Extensibility", RFC 8584, DOI 10.17487/RFC8584, April 2019, <https://www.rfc-editor.org/info/rfc8584>.

10.2. Informative References

[PREF-DF] Rabadan, J., "Preference-based EVPN DF Election", 2020.

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

[RFC7275] Martini, L., Salam, S., Sajassi, A., Bocci, M., Matsushima, S., and T. Nadeau, "Inter-Chassis Communication Protocol for Layer 2 Virtual Private Network (L2VPN) Provider Edge (PE) Redundancy", RFC 7275, DOI 10.17487/RFC7275, June 2014, <https://www.rfc-editor.org/info/rfc7275>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

Authors' Addresses

Patrice Brissette (editor)
Cisco Systems
Ottawa, ON
Canada

Email: pbrisset@cisco.com

Ali Sajassi
Cisco Systems
USA

Email: sajassi@cisco.com

Bin Wen
Comcast
USA

Email: Bin_Wen@comcast.com

Edward Leyton
Verizon Wireless
USA

Email: edward.leyton@verizonwireless.com

Jorge Rabadan
Nokia
USA

Email: jorge.rabadan@nokia.com

Luc Andre Burdet
Cisco Systems
Canada

Email: lburdet@cisco.com

Samir Thoria
Cisco Systems
USA

Email: sthoria@cisco.com

Network Working Group
Internet-Draft
Intended status: Experimental
Expires: February 19, 2021

K. Vairavakkalai
M. Jeyananth
Juniper Networks, Inc.
August 18, 2020

BGP signalled private MPLS-labels
draft-kaliraj-bess-bgp-sig-private-mpls-labels-01

Abstract

The MPLS-forwarding-layer in a core network is a shared resource. The MPLS FIB at nodes in this layer contains labels that are dynamically allocated and locally significant at that node.

For some usecases like upstream-label-allocation, it is useful to be able to create virtual private MPLS-forwarding-layers over this shared MPLS-forwarding-layer. This allows installing deterministic private label-values in the private-FIBs created at nodes participating in this private MPLS forwarding-layer, while preserving the "locally significant" nature of the underlying shared 'public' MPLS-forwarding-layer.

This specification describes the procedures to create such virtual private MPLS-forwarding layers (private MPLS-planes) using a new BGP family. And gives a few example use-cases on how this private forwarding-layers can be used.

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].

This Internet-Draft will expire on February 19, 2021.

Table of Contents

1. Introduction 3
2. Motivation 3
3. Constructs and building blocks 4
3.1. Context Protocol Nexthop Address 4
3.2. MPLS context FIB 4
3.3. Context Label 5
3.4. Roles of nodes in a MPLS-plane 5
3.4.1. Edge-nodes (PLER) 5
3.4.2. Transit-nodes (PLSR) 5
3.5. Sending traffic into the MPLS plane 5
4. Terminology 6
5. BGP families, routes and encoding 7
5.1. New address-families 7
5.1.1. AFI: MPLS, SAFI: 128 7
5.1.2. AFI: MPLS, SAFI: 1 8
5.2. Routes and Operational procedures 8
5.2.1. "Context-Nexthop" discovery route 8
5.2.2. "Private Label" routes 8
6. Example of Usecases 10
6.1. Mezanine transport layer in a Seamless-MPLS network . . . 10
6.2. Service Forwarding Helper usecase 11
6.3. Standard BGP API to a MPLS network's forwarding-plane . . 12
6.4. Traffic engineering and Security advantages 12
7. IANA Considerations 12
8. Security Considerations 13
9. Acknowledgements 13
10. Normative References 13
Authors' Addresses 13

1. Introduction

The MPLS-forwarding-layer in a core network is a shared resource. The MPLS FIB at nodes in this layer contains labels that are dynamically allocated and locally significant at that node.

For some usecases like upstream-label-allocation, it is useful to be able to create virtual private MPLS-forwarding-layers over this shared MPLS-forwarding-layer. This allows installing deterministic private label-values in the private-FIBs in this private forwarding-layer, while preserving the "locally significant" nature of the underlying shared 'public' MPLS-forwarding-layer.

It can be noted that, mechanism described in this document is nothing but a [RFC4364] style BGP VPN where the FEC is MPLS-Label, instead of IP-prefix. This document defines new address-families (AFI: MPLS, SAFI: VPN-Unicast, Unicast) and associated signaling mechanisms to create and use MPLS forwarding-contexts in a network. The concepts of MPLS-Context-tables and upstream allocation are described in [RFC5331].

BGP speakers participating in the private MPLS FIB layer create instances of "MPLS forwarding-context" FIBs, which are identified using a "Context-Protocol-Nexthop (CPNH)". A Context-label MAY be advertised in conjunction with the Context Protocol Nexthop (CPNH) using new BGP address-family to other speakers.

2. Motivation

A provider's core network consists of a global-domain (default forwarding-tables in P and PE nodes) that is shared by all tenants in the network and may also contain multiple private user-domains (e.g. VRF route tables).

The global MPLS forwarding-layer can be viewed as the collection of all default MPLS forwarding-tables. This global MPLS Fib layer contains labels locally significant to each node. The "local-significance of labels" gives the nodes freedom to participate in MPLS-forwarding with whatever label-ranges they can support in forwarding hardware.

In emerging usecases some applications using the MPLS-network may benefit from a "static labels" view of the MPLS-network. In some other usecases, a standard mechanism to do Upstream label-allocation is beneficial.

It is desirable to leave the global MPLS FIB layer intact, and build private MPLS FIB-layers on top of it to achieve these requirements.

The private-MPLS-FIBs can then be used by the applications as desired. The private MPLS-FIBs need to be created only at the nodes in the network where predictable label-values (external label allocation) is desired. E.g. P-routers that need to act as a "Detour-nodes" or "Service-Forwarding-Helpers" that need to mirror service-labels.

In other words, provisioning of these private MPLS-FIBs can be gradual and can co-exist with nodes not supporting the feature described in this document. These private-MPLS-FIBs can be stitched together using either the Context-labels over the existing shared MPLS-network tunnels, or 'private' context-interfaces - to form the "private MPLS-FIB layer".

An application can then install the routes with desired label-values in the private forwarding-contexts with desired forwarding-semantics.

3. Constructs and building blocks

The building-blocks that construct a private MPLS plane are described in this section.

3.1. Context Protocol Nexthop Address

A private MPLS plane (just "MPLS plane" here-after) is identified by an IP-address called Context Protocol Nexthop (CPNH). This address is unique in the core-network, like any other loopback address.

A loopback-address uniquely identifies a specific node in the network, and we call it Global Protocol Nexthop (GPNH) in this document. The CPNH address uniquely identifies a "MPLS-plane".

Each node that has forwarding-context for a MPLS-plane MUST be configured with the same CPNH but a different RD, such that the RD:CPNH will uniquely identify that node in the MPLS-plane.

3.2. MPLS context FIB

An instance of a MPLS forwarding-table at a node in the private MPLS-plane. This Private MPLS FIB contains the private-label routes.

A node can have context-FIB for multiple MPLS-planes. The same label-value can have a different forwarding-semantic in each MPLS-plane. Thus the applications using that MPLS-plane get a deterministic label-value independent of other applications using other MPLS-planes.

The terms "private MPLS FIB-layer" and "private MPLS-plane" are used interchangeably in this document.

3.3. Context Label

A context-label is a non-reserved dynamically allocated label, that is installed in the global MPLS FIB, and points to a MPLS-Context-FIB. The Context-Label have forwarding semantics as follows in the global MPLS-FIB:

Context-Label -> Pop and Lookup in MPLS-Context-Fib

Advertising the "Context-Label in conjunction with the GPNH" tells the network how to reach a "RD:CPNH".

3.4. Roles of nodes in a MPLS-plane

The node roles in a MPLS-plane can be classified into "edge nodes" (call them PLER) or "transit-nodes" (call them PLSR).

3.4.1. Edge-nodes (PLER)

Private Label Edge-routers (PLER) have MPLS context-FIB that belong to the MPLS-plane. They advertise the presence of this context-FIB, and private-label routes from this FIB, using new BGP AFI/SAFI described in this document.

3.4.2. Transit-nodes (PLSR)

Private Label Transit-nodes do label-swap forwarding for the Context-Labels they see in the Context-Protocol-Nexthop advertisement routes going thru them. They basically stitch/extend the label switched path to a RD:CPNH when they re-advertise the CPNH routes with nexthop-self.

PLSRs dont have context-FIBs. PLSRs dont have Context Protocol-Nexthop. Because they dont have Private label routes to originate.

However a node in the network can play both roles, of PLER and PLSR.

3.5. Sending traffic into the MPLS plane

MPLS-traffic arriving with private-labels hits the correct private MPLS-FIB by virtue of either arriving on a "private network-interface" that is attached to the FIB, or arriving on a shared network-interface with a "Context-label".

To send data traffic into this private MPLS FIB-layer, the application MUST use as handle either a "Context-label" advertised by a node or a "Private-interface" owned by the application at the node.

The Context-Label is the only label-value the application needs to learn from the network (PLER node it is connected to), to be able to use the private MPLS-plane. The application can decide the value of the labels to be programmed in the private MPLS-FIBs.

Once the packet enters the private MPLS plane at an edge-node (PLER), the node will forward the packet to the next node (PLSR or PLER), by pushing the Context-label advertised by that next-node, and the transport-label to reach that node's GPNH. This will repeat until the packet reaches the private MPLS-FIB that originated that private MPLS-label.

At each PLER in the MPLS-plane, the private-label value remains the same, and points towards the same resource attached to the MPLS-plane. This allows the applications using the MPLS-network a static-labels view of the resourses attached to the private MPLS-plane.

At each PLSR in the MPLS-plane, the context-label value will change (be swapped in forwarding), but is transparent to the application.

4. Terminology

P-router : A Provider core router, also called a LSR

LSR : Label Switch Router (pure transport node speaking LDP, RSVP etc)

PLSR: a transit node in a private MPLS-plane. It has a forwarding-context for private-labels.

PLER: an edge node in a private MPLS-plane. It has a forwarding-context for private-labels.

Detour-router : A P-router that is used as a loose-hop in a traffic-engineered path

PE-router : Provider Edge router, that hosts a service (Internet, L3VPN etc)

SE-router : Service Edge router. Same as PE.

SFH-router : Service Forwarding Helper. A node helping an SE-router with service-traffic forwarding, using Service-routes mirrored by the SE.

MPLS FIB : MPLS Forwarding table

Global MPLS FIB : Global MPLS Forwarding table, to which shared-interfaces are connected

Private MPLS FIB : Private MPLS Forwarding table, to which private-interfaces are connected

Private MPLS FIB Layer : The group of Private MPLS FIBs in the network, connected together via Context-Labels

Context-Label : Locally-significant Non-reserved label pointing to a private MPLS FIB

Context nexthop IP-address (CPNH) : An IP-address that identifies the "Private MPLS FIB Layer". RD:CPNH identifies a Private MPLS FIB at a node.

Global nexthop IP-address (GPNH) : Global Protocol Nexthop address. E.g. a loopback address used as transport tunnel end-point.

5. BGP families, routes and encoding

This section describes the new constructs defined by this document.

5.1. New address-families

This document defines a new AFI: "MPLS". And two new address-families.

5.1.1. AFI: MPLS, SAFI: 128

This address-family is used to exchange private label-routes into private MPLS-FIBs at routers that are connected using a common network-interface.

Routes in this family contain Route-Target extended-community identifying the private-FIB-Layer (VPN) the route belongs to. This address-family also advertises the Context-Label that the receiving router uses to access the private MPLS-FIB. The Context-Label is required when the connecting-interface is a shared common interface that terminates into the global MPLS FIB. The Context-Label installed in the global MPLS-FIB points to the private MPLS-FIB.

5.1.2. AFI: MPLS, SAFI: 1

This address-family is used to exchange private label-routes in private MPLS-FIBs to routers that are connected using a private network-interface.

Because the interface is private, and terminates directly into the private MPLS-FIB, a Context-Label is not required to access the private MPLS-FIB.

5.2. Routes and Operational procedures

5.2.1. "Context-Nexthop" discovery route

The Context-NH discovery route is a [BGP-CT] family route that carries CPNH in the "Prefix" portion of the NLRI. And the Context-Label is carried in the "Label" field in the [RFC8277] format NLRI.

This route is advertised with the following path-attributes:

- BGP Nexthop attribute (code 14, MP_REACH) carrying GPNH address.
- Route-Target extended community, identifying the private FIB-layer

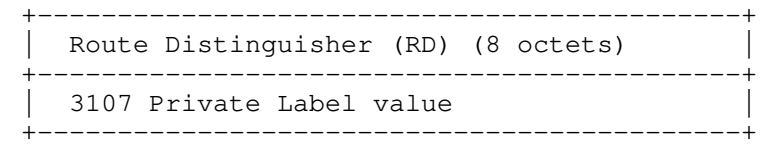
The "Context-Nexthop discovery route" is originated by each speaker who acts as a PLER. The "RD:Context-nexthop" uniquely identifies the private-FIB at the speaker. The "Context-nexthop address" uniquely identifies the private-FIB-layer.

A speaker readvertising a Context-Nexthop discovery-route MUST follow the mechanisms described in [BGP-CT]. Specifically when re-advertising with "next-hop self" MUST allocate a new Label with a forwarding semantic of "Swap Received-Context-Label, Forward to Received-GPNH". This extends reachability to the CPNH across tunnel domains.

5.2.2. "Private Label" routes

The Private Label routes are carried in the new address-family "MPLS VpnUnicast" defined in this document.

NLRI Label Prefix (Private Label route)

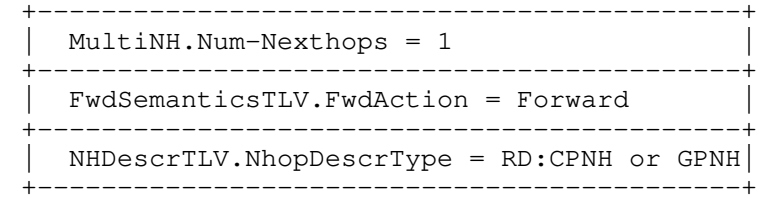


Private-Label-Value: The (upstream assigned) label value

Attributes on this route:

- BGP Nexthop attribute (code 14, MP_REACH) carrying a GPNH address. (OR)
- The Multi-nexthop attribute [MULTI-NH] with forwarding-semantic:
  - "Forward to RD:CPNH"
- Route-Target extended-community, identifying the private FIB-layer

MultiNexthop BGP-attribute (Private Label route)



A speaker MAY readvertise a private-label-route without changing the Nexthop (RD:CPNH) carried in it, if the speaker is a pure PLSR.

If it does alter the nexthop to SelfRD:CPNH, it SHOULD act as a PLER, and for e.g. originate a "Context-Nexthop discovery route" for prefix "SelfRD:CPNH".

Even if the speaker sets nexthop-address to Self because of regular BGP readvertisement-rules, Label Prefix MUST NOT be altered, and the received NLRI "RD:Private-Label1" MUST be re-advertised as-is. Such

that value of label "Private-Label1" doesn't change while the packet traverses multiple nodes in the private-MPLS-FIB-layer.

The Route-target attached to the route is the one identifying the private MPLS FIB layer (VPN). The Private-label routes resolve over the Context-nexthop route that belong to the same VPN.

A node receiving a "Private-Label route" RD:L1 MUST install the label L1 in the private MPLS Forwarding-context idenfied by the Route-Target attached to the route.

The label route MUST be installed with forwarding-semantic as specified in the received Multi-nexthop attribute. As an example, a Detour node MAY receive the private-label-route with a forwarding-semantic of "Forward to RD:CPNH" operation. And an Egress node MAY receive a private-label-route with a forwarding-semantic pointing to a resource it houses. Note that such a Private-label BGP-route MAY be received from external-application also.

5.2.2.1. Resolving received Private Label-routes

A node receiving a "Context-nexthop discovery route" MUST be capable of using either the CPNH or the RD:CPNH carried in the NLRI, to resolve other routes received with this CPNH address or RD:CPNH in the "Nexthop-attributes".

The receiver of a private-label route MUST recursively resolve the received nexthop (RD:CPNH) over the Context-Nexthop discovery-route for prefix "RD:CPNH" to determine the label stack "Context-Label, Transport-Label" to push, so that the MPLS packet with private-label reaches the private MPLS FIB originating the route.

If a node receives multiple "Context-nexthop discovery route" for a CPNH, it SHOULD run path-selection after stripping the RD, to find the closest ingress to the private-MPLS-plane identified by the CPNH. This best path SHOULD be used to resolve a received private-label-route.

6. Example of Usecases

6.1. Mezanine transport layer in a Seamless-MPLS network

Typically service-routes in a MPLS network bind to the following entities that identify point-of-presence of a service:

- Protocol Nexthop - PE loopback address (GPNH)

- Service Label - PE advertised locally signifcant label that identifies the service

In this model, whenever a PE is taken out of service the GPNH changes, and Service-Label changes - which causes maintenance a heavy convergence event. Because the service-routes with massive-scale need to be readvertised with new service-label or PE-address.

An alternate model could be: to advertise the Service-routes with a protocol-nexthop of CPNH (without RD), with a forwarding-semantic of:

- "Push <Private-Label>, and Forward to CPNH"

This model fully decouples the service-layer from the transport-layer identifiers, by making the Service-routes refer to the CPNH and Private-Labels. Thus the underlying transport-layer can change (nodes representing a Private-label can be added or removed) without any changes to the service-routes. Which present good scaling properties for the network.

This model also allows anycast traffic forwarding to any resource in the network. Multiple PEs can advertise the same Private-Label to identify a specific service (e.g. peering with an AS) they are offering.

Once the service-route traffic enters the private-FIB-layer, at the closest entry-point determined by path-selection of CPNH auto-discovery routes; then the Private-Labels (with pre-determined values) pushed will determine the loose hop path taken by the traffic and also the destination-resource.

6.2. Service Forwarding Helper usecase

In a virtualized environment a Service-PE node (that comprises of a vCP and multiple vFPs) can mirror MPLS labels (GL1) in its global MPLS-FIB to a private forwarding context at an upstream node (SFH) with information on which vFPs are optimal exit-points for that label. Such that the SFH can optimally forward traffic to GL1 to the right vFPs, thus avoiding intra fabric traffic hops.

To do this, the service-PE advertises a private-label route with RD:GL1 to the SFH node. The route is advertised with a Multi-nexthop attribute with one or more legs that have a "Forward to SEPx" semantics. Where SEPx is one of many exit-points at the Service-PE node.

6.3. Standard BGP API to a MPLS network's forwarding-plane

This mechanism facilitates predictable (external-allocator determined) label-values, using a standard BGP-family as the API. It gives the external applications a separate MPLS-FIB to play with, totally separate from other applications.

This also avoids vendor specific-API dependencies for external-allocators (controller softwares), and vice-versa.

This mechanism also increases the overal MPLS label-space available in the network, because it creates per-app label-forwarding-contexts (namespaces), instead of reserving/splitting the global MPLS FIB among various applications.

6.4. Traffic engineering and Security advantages

- Ability of ingress to steer mpls-traffic thru specific detour loose-hop nodes using predictable-labels' stack.
- Provide label-spoofing protection at edge-nodes - by virtue of using separate mpls-forwarding-contexts
- Allow private-MPLS label usage to spread across multiple-domains/ AS and work seamlessly with existing technologies like Inter-AS VPN option C.

7. IANA Considerations

This document makes following requests of IANA.

New BGP AFI code:

- <TBD> for "MPLS"

Which will be used to create new BGP AFI-SAFI pairs:

- MPLS Uni(SAFI:1),
- MPLS VpnUni(SAFI:128)

.

New NLRI Route-types for these AFI SAFIs:

- Type 1: Context-Nexthop-Discovery-route.
- Type 2: Private-Label route

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

8. Security Considerations

Using separate mpls-forwarding-contexts for separate applications and stitching them into separate MPLS-planes increases the security attributes of the MPLS network.

9. Acknowledgements

The authors thank Jeffrey (Zhaohui) Zhang, Ron Bonica, Jeff Haas and John Scudder for the valuable discussions.

10. Normative References

[BGP-CT] Vairavakkalai, K., "BGP Classful Transport Planes", July 2020, <https://tools.ietf.org/html/draft-kaliraj-idr-bgp-classful-transport-planes-01#section-8>.

[MULTI-NH] Vairavakkalai, K., "BGP MultiNexthop attribute", June 2017, <https://tools.ietf.org/html/draft-kaliraj-idr-multinexthop-attribute-00>.

[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>.

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

[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>.

[RFC8277] Rosen, E., "Using BGP to Bind MPLS Labels to Address Prefixes", RFC 8277, DOI 10.17487/RFC8277, October 2017, <https://www.rfc-editor.org/info/rfc8277>.

Authors' Addresses

Kaliraj Vairavakkalai
Juniper Networks, Inc.
1133 Innovation Way,
Sunnyvale, CA 94089
US

Email: kaliraj@juniper.net

Minto Jeyananth
Juniper Networks, Inc.
1133 Innovation Way,
Sunnyvale, CA 94089
US

Email: minto@juniper.net

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 19, 2021

K. Vairavakkalai
N. Venkataraman
B. Rajagopalan
Juniper Networks, Inc.
G. Mishra
Verizon Communications Inc.
M. Khaddam
Cox Communications Inc.
X. Xu
Alibaba Inc.
R. Szarecki
Google.

February 15, 2021

BGP Classful Transport Planes
draft-kaliraj-idr-bgp-classful-transport-planes-07

Abstract

This document specifies a mechanism, referred to as "service mapping", to express association of overlay routes with underlay routes satisfying a certain SLA, using BGP. The document describes a framework for classifying underlay routes into transport classes, and mapping service routes to specific transport class.

The "Transport class" construct maps to a desired SLA, and can be used to realize the "Topology Slice" in 5G Network slicing architecture.

This document specifies BGP protocol procedures that enable dissimination of such service mapping information that may span multiple co-operating administrative domains. These domains may be administetered by the same provider or closely co-ordinating provider networks.

It makes it possible to advertise multiple tunnels to the same destination address, thus avoiding need of multiple loopbacks on the egress node.

A new BGP transport layer address family (SAFI 76) is defined for this purpose that uses RFC-4364 technology and follows RFC-8277 NLRI encoding. This new address family is called "BGP Classful Transport", aka BGP CT.

It carries transport prefixes across tunnel domain boundaries (e.g. in Inter-AS Option-C networks), parallel to BGP LU (SAFI 4) . It dissiminates "Transport class" information for the transport prefixes

across the participating domains, which is not possible with BGP LU. This makes the end-to-end network a "Transport Class" aware tunneled network.

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 https://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, 2021.

Copyright Notice

Table of Contents

1. Introduction 3
2. Terminology 5
3. Transport Class 6
4. "Transport Class" Route Target Extended Community 7
5. Transport RIB 9
6. Transport Routing Instance 9
7. Nexthop Resolution Scheme 9
8. BGP Classful Transport Family NLRI 10
9. Comparison with other families using RFC-8277 encoding . . . 10
10. Protocol Procedures 11
11. Scaling considerations 15
11.1. Avoiding unintended spread of CT routes across domains. 15
11.2. Constrained distribution of PNHs to SNs (On Demand Nexthop) 15
11.3. Limiting scope of visibility of PE loopback as PNHs . . 16
12. OAM considerations 17
13. Applicability to Network Slicing 18
14. Illustration of procedures with example topology 18
14.1. Topology 18
14.2. Service Layer route exchange 20
14.3. Transport Layer route propagation 20
14.4. Data plane view 22
14.4.1. Steady state 22
14.4.2. Absorbing failure of primary path 23
15. IANA Considerations 24
15.1. New BGP SAFI 24
15.2. New Format for BGP Extended Community 24
15.2.1. Existing registries to be modified 24
15.2.2. New registries to be created 25
15.3. MPLS OAM code points 26
16. Security Considerations 27
17. Acknowledgements 27
18. References 27
18.1. Normative References 27
18.2. URIs 29
Authors' Addresses 29

1. Introduction

To facilitate service mapping, the tunnels in a network can be grouped by the purpose they serve into a "Transport Class". The tunnels could be created using any signaling protocol, such as LDP, RSVP, BGP LU or SPRING. The tunnels could also use native IP or IPv6, as long as they can carry MPLS payload. Tunnels may exist between different pair of end points. Multiple tunnels may exist between the same pair of end points.

Thus, a Transport Class consists of tunnels created by various protocols that satisfy the properties of the class. For example, a "Gold" transport class may consist of tunnels that traverse the shortest path with fast re-route protection, a "Silver" transport class may hold tunnels that traverse shortest paths without protection, a "To NbrAS Foo" transport class may hold tunnels that exit to neighboring AS Foo, and so on.

The extensions specified in this document can be used to create a BGP transport tunnel that potentially spans domains, while preserving its Transport Class. Examples of domain are Autonomous System (AS), or IGP area. Within each domain, there is a second level underlay tunnel used by BGP to cross the domain. The second level underlay tunnels could be hetrogeneous: Each domain may use a different type of tunnel (e.g. MPLS, IP, GRE), or use a differnet signaling protocol. A domain boundary is demarcated by a rewrite of BGP nexthop to 'self' while re-advertising tunnel routes in BGP. Examples of domain boundary are inter-AS links and inter-region ABRs. The path uses MPLS label-switching when crossing domain boundary and uses the native intra-AS tunnel of the desired transport class when traversing within a domain.

Overlay routes carry sufficient indication of the Transport Classes they should be encapsulated over, in form of BGP community called the "Mapping community". Based on the mapping community, "route resolution" procedure on the ingress node selects from the corresponding Transport Class an appropriate tunnel whose destination matches (LPM) the nexthop of the overlay route. If the overlay route is carried in BGP, the protocol nexthop (or, PNH) is generally carried as an attribute of the route.

The PNH of the overlay route is also referred to as "service endpoint" (SEP). The service endpoint may exist in the same domain as the service ingress node or lie in a different domain, adjacent or non-adjacent. In the former case, reachability to the SEP is provided by an intra-domain tunneling protocol, and in the latter case, reachability to the SEP is via BGP transport families.

In this architecture, the intra-domain transport protocols (e.g. RSVP, SRTE) are also "Transport Class aware", and they publish ingress routes in Transport RIB associated with the Transport Class, at the tunnel ingress node. These routes are then redistributed into BGP CT to be advertised to adjacent domains. It is outside the scope of this document how exactly the transport protocols are made transport class aware, though configuration on the tunnel ingress node is a simple mechanism to achieve it.

This document describes mechanisms to:

Model a "Transport Class" as "Transport RIB" on a router, consisting of tunnel ingress routes of a certain class.

Enable service routes to resolve over an intended Transport Class by virtue of carrying the appropriate "Mapping community". Which results in using the corresponding Transport RIB for finding nexthop reachability.

Advertise tunnel ingress routes in a Transport RIB via BGP without any path hiding, using BGP VPN technology and Add-path. Such that overlay routes in the receiving domains can also resolve over tunnels of associated Transport Class.

Provide a way for co-operating domains to reconcile any differences in extended community namespaces, and interoperate between different transport signaling protocols in each domain.

In this document we focus mainly on MPLS as the intra-domain transport tunnel forwarding, but the mechanisms described here would work in similar manner for non-MPLS (e.g. IP, GRE, UDP) transport tunnel forwarding technologies too.

This document assumes MPLS forwarding when crossing domain boundaries, as that is the defacto standard in deployed networks today. But mechanisms specified in this document can also support different forwarding technologies (e.g. SRv6). A future document may describe such adaptations, it is out of scope of this document.

The document Seamless Segment Routing [Seamless-SR] describes various use cases and applications of procedures described in this document.

2. Terminology

LSP: Label Switched Path.

TE : Traffic Engineering.

SN : Service Node.

BN : Border Node.

TN : Transport Node, P-router.

BGP-VPN : VPNs built using RFC4364 mechanisms.

RT : Route-Target extended community.

RD : Route-Distinguisher.

PNH : Protocol-Nexthop address carried in a BGP Update message.

SEP : Service End point, the PNH of a Service route.

LPM : Longest Prefix Match.

Service Family : BGP address family used for advertising routes for "data traffic", as opposed to tunnels.

Transport Family : BGP address family used for advertising tunnels, which are in turn used by service routes for resolution.

Transport Tunnel : A tunnel over which a service may place traffic. These tunnels can be GRE, UDP, LDP, RSVP, or SR-TE.

Tunnel Domain : A domain of the network containing SN and BN, under a single administrative control that has a tunnel between SN and BN. An end-to-end tunnel spanning several adjacent tunnel domains can be created by "stitching" them together using labels.

Transport Class : A group of transport tunnels offering the same type of service.

Transport Class RT : A Route-Target extended community used to identify a specific Transport Class.

Transport RIB : At the SN and BN, a Transport Class has an associted Transport RIB that holds its tunnel routes.

Transport Plane : An end to end plane comprising of transport tunnels belonging to same transport class. Tunnels of same transport class are stitched together by BGP route readvertisements with nexthop-self, to span across domain boundaries using Label-Swap forwarding mechanism similar to Inter-AS option-b.

Mapping Community : BGP Community/Extended-community on a service route, that maps it to resolve over a Transport Class.

3. Transport Class

A Transport Class is defined as a set of transport tunnels that share certain characteristics useful for underlay selection.

On the wire, a transport class is represented as the Transport Class RT, which is a new Route-Target extended community.

A Transport Class is configured at SN and BN, along with attributes like RD and Route-Target. Creation of a Transport Class instantiates

the associated Transport RIB and a Transport routing instance to contain them all.

The operator may configure a SN/BN to classify a tunnel into an appropriate Transport Class, which causes the tunnel's ingress routes to be installed in the corresponding Transport RIB. At a BN, these tunnel routes may then be advertised into BGP CT.

Alternatively, a router receiving the transport routes in BGP with appropriate signaling information can associate those ingress routes to the appropriate Transport Class. E.g. for Classful Transport family (SAFI 76) routes, the Transport Class RT indicates the Transport Class. For BGP LU family(SAFI 4) routes, import processing based on Communities or inter-AS source-peer may be used to place the route in the desired Transport Class.

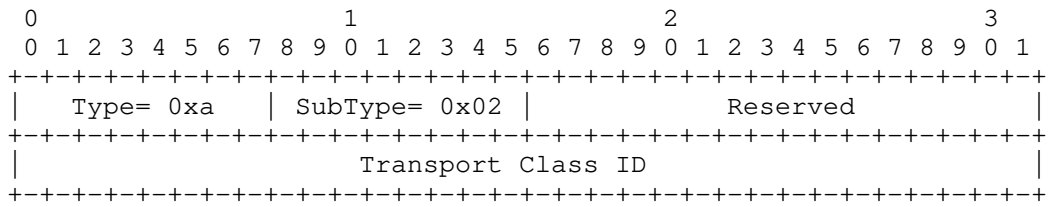
When the ingress route is received via SRTE [SRTE], which encodes the Transport Class as an integer 'Color' in the NLRI as "Color:Endpoint", the 'Color' is mapped to a Transport Class during import processing. SRTE ingress route for 'Endpoint' is installed in that transport class. The SRTE route when advertised out to BGP speakers will then be advertised in Classful Transport family with Transport Class RT and a new label. The MPLS swap route thus installed for the new label will pop the label and deliver decapsulated traffic into the path determined by SRTE route.

4. "Transport Class" Route Target Extended Community

This document defines a new type of Route Target, called "Transport Class" Route Target Extended Community.

"Transport Class" Route Target extended community is a transitive extended community EXT-COMM [RFC4360] of extended-type, with a new Format (Type high = 0xa) and SubType as 0x2 (Route Target).

This new Route Target Format has the following encoding:



"Transport Class" Route Target Extended Community

Type: 2 octets

Type field contains value 0xa.

SubType: 2 octets

Subtype field contain 0x2. This indicates 'Route Target'.

Transport Class ID: 4 octets

The least significant 32-bits of the value field contain the "Transport Class" identifier, which is a 32-bit integer.

The remaining 2 octets after SubType field are Reserved, they MUST be set to zero by originator, and ignored, left unaltered by receiver.

The "Transport class" Route Target Extended community follows the mechanisms for VPN route import, export as specified in BGP-VPN [RFC4364], and Route Target Contrain mechanisms as specified in VPN-RTC [RFC4684]

A BGP speaker that implements RT Constraint VPN-RTC [RFC4684] MUST apply the RT Constraint procedures to the "Transport class" Route Target Extended community as-well.

The Transport Class Route Target Extended community is carried on Classful Transport family routes, and allows associating them with appropriate Transport RIBs at receiving BGP speakers.

Use of the Transport Class Route Target Extended community with a new Type code avoids conflicts with any VPN Route Target assignments already in use for service families.

5. Transport RIB

A Transport RIB is a routing-only RIB that is not installed in forwarding path. However, the routes in this RIB are used to resolve reachability of overlay routes' PNH. Transport RIB is created when the Transport Class it represents is configured.

Overlay routes that want to use a specific Transport Class confine the scope of nexthop resolution to the set of routes contained in the corresponding Transport RIB. This Transport RIB is the "Routing Table" referred in Section 9.1.2.1 RFC4271 [1]

Routes in a Transport RIB are exported out in 'Classful Transport' address family.

6. Transport Routing Instance

A BGP VPN routing instance that is a container for the Transport RIB. It imports, and exports routes in this RIB with Transport Class RT. Tunnel destination addresses in this routing instance's context come from the "provider namespace". This is different from user VRFs for e.g., which contain prefixes in "customer namespace"

The Transport Routing instance uses the RD and RT configured for the Transport Class.

7. Nexthop Resolution Scheme

An implementation may provide an option for the service route to resolve over less preferred Transport Classes, should the resolution over preferred, or "primary" Transport Class fail.

To accomplish this, the set of service routes may be associated with a user-configured "resolution scheme", which consists of the primary Transport Class, and optionally, an ordered list of fallback Transport Classes.

A community called as "Mapping Community" is configured for a "resolution scheme". A Mapping community maps to exactly one resolution scheme. A resolution scheme comprises of one primary transport class and optionally one or more fallback transport classes.

A BGP route is associated with a resolution scheme during import processing. The first community on the route that matches a mapping community of a locally configured resolution scheme is considered the effective mapping community for the route. The resolution scheme thus found is used when resolving the route's PNH. If a route

contains more than one mapping community, it indicates that the route considers these multiple mapping communities as equivalent. So the first community that maps to a resolution scheme is chosen.

A transport route received in BGP Classful Transport family SHOULD use a resolution scheme that contains the primary Transport Class without any fallback to best effort tunnels. The primary Transport Class is identified by the Transport Class RT carried on the route. Thus Transport Class RT serves as the Mapping Community for Classful Transport routes.

A service route received in a BGP service family MAY map to a resolution scheme that contains the primary Transport Class identified by the mapping community on the route, and a fallback to best effort tunnels transport class. The primary Transport Class is identified by the Mapping community carried on the route. For e.g. the Extended Color community may serve as the Mapping Community for service routes. Color:0:<n> MAY map to a resolution scheme that has primary transport class <n>, and a fallback to best-effort transport class.

8. BGP Classful Transport Family NLRI

The Classful Transport family will use the existing AFI of IPv4 or IPv6, and a new SAFI 76 "Classful Transport" that will apply to both IPv4 and IPv6 AFIs.

The "Classful Transport" SAFI NLRI itself is encoded as specified in https://tools.ietf.org/html/rfc8277#section-2 [RFC8277].

When AFI is IPv4 the "Prefix" portion of Classful Transport family NLRI consists of an 8-byte RD followed by an IPv4 prefix. When AFI is IPv6 the "Prefix" consists of an 8-byte RD followed by an IPv6 prefix.

Attributes on a Classful Transport route include the Transport Class Route-Target extended community, which is used to leak the route into the right Transport RIBs on SNs and BNs in the network.

9. Comparison with other families using RFC-8277 encoding

SAFI 128 (Inet-VPN) is a RF8277 encoded family that carries service prefixes in the NLRI, where the prefixes come from the customer namespaces, and are contexualized into separate user virtual service RIBs called VRFs, using RFC4364 procedures.

SAFI 4 (BGP LU) is a RFC8277 encoded family that carries transport prefixes in the NLRI, where the prefixes come from the provider namespace.

SAFI 76 (Classful Transport) is a RFC8277 encoded family that carries transport prefixes in the NLRI, where the prefixes come from the provider namespace, but are contexualized into separate Transport RIBs, using RFC4364 procedures.

It is worth noting that SAFI 128 has been used to carry transport prefixes in "L3VPN Inter-AS Carrier's carrier" scenario, where BGP LU/LDP prefixes in CsC VRF are advertised in SAFI 128 towards the remote-end baby carrier.

In this document a new AFI/SAFI is used instead of reusing SAFI 128 to carry these transport routes, because it is operationally advantageous to segregate transport and service prefixes into separate address families, RIBs. E.g. It allows to safely enable "per-prefix" label allocation scheme for Classful Transport prefixes without affecting SAFI 128 service prefixes which may have huge scale. "per prefix" label allocation scheme keeps the routing churn local during topology changes.

A new family also facilitates having a different readvertisement path of the transport family routes in a network than the service route readvertisement path. viz. Service routes (Inet-VPN) are exchanged over an EBGP multihop session between Autonomous systems with nexthop unchanged; whereas Classful Transport routes are readvertised over EBGP single hop sessions with "nexthop-self" rewrite over inter-AS links.

The Classful Transport family is similar in vein to BGP LU, in that it carries transport prefixes. The only difference is, it also carries in Route Target an indication of which Transport Class the transport prefix belongs to, and uses RD to disambiguate multiple instances of the same transport prefix in a BGP Update.

10. Protocol Procedures

This section summarizes the procedures followed by various nodes speaking Classful Transport family

Preparing the network for deploying Classful Transport planes

- Operator decides on the Transport Classes that exist in the network, and allocates a Route-Target to identify each Transport Class.

Operator configures Transport Classes on the SNs and BNs in the network with unique Route-Distinguishers and Route-Targets.

Implementations may provide automatic generation and assignment of RD, RT values for a transport routing instance; they MAY also provide a way to manually override the automatic mechanism, in order to deal with any conflicts that may arise with existing RD, RT values in the different network domains participating in a deployment.

Origination of Classful Transport route:

At the ingress node of the tunnel's home domain, the tunneling protocols install routes in the Transport RIB associated with the Transport Class the tunnel belongs to.

The ingress node then advertises this tunnel destination into BGP as a Classful Transport family route with NLRI RD:TunnelEndpoint, attaching a 'Transport Class' Route Target that identifies the Transport Class. This BGP CT route is advertised to EBGP peers and IBGP peers which are RR-clients. This route MUST NOT be advertised to the IBGP peers who are not RR-clients.

Alternatively, the egress node of the tunnel i.e. the tunnel endpoint can originate the same BGP Classful Transport route, with NLRI RD:TunnelEndpoint and PNH TunnelEndpoint, which will resolve over the tunnel route at the ingress node. When the tunnel is up, the Classful Transport BGP route will become usable and get re-advertised.

Unique RD SHOULD be used by the originator of a Classful Transport route to disambiguate the multiple BGP advertisements for a transport end point.

Ingress node receiving Classful Transport route

On receiving a BGP Classful Transport route with a PNH that is not directly connected, e.g. an IBGP-route, a mapping community on the route (the Transport Class RT) indicates which Transport Class this route maps to. The routes in the associated Transport RIB are used to resolve the received PNH. If there does not exist a route in the Transport RIB matching the PNH, the Classful Transport route is considered unusable, and MUST NOT be re-advertised further.

Border node readvertising Classful Transport route with nexthop self:

The BN allocates an MPLS label to advertise upstream in Classful Transport NLRI. The BN also installs an MPLS swap-route for that label that swaps the incoming label with a label received from the downstream BGP speaker, or pops the incoming label. And then pushes received traffic to the transport tunnel or direct interface that the Classful Transport route's PNH resolved over.

The label SHOULD be allocated with "per-prefix" label allocation semantics. The prefix used as key is formed by stripping RD from the BGP CT NLRI prefix. This helps in avoiding BGP CT route churn through out the CT network when a failure happens in a domain. The failure is not propagated further than the BN closest to the failure.

The value of advertised MPLS label is locally significant, and is dynamic by default. The BN may provide option to allocate a value from a statically carved out range. This can be achieved using locally configured export policy, or via mechanisms described in BGP Prefix-SID [RFC8669].

Border node receiving Classful Transport route on EBGP :

If the route is received with PNH that is known to be directly connected, e.g. EBGP single-hop peering address, the directly connected interface is checked for MPLS forwarding capability. No other nexthop resolution process is performed, as the inter-AS link can be used for any Transport Class.

If the inter-AS links should honor Transport Class, then the BN SHOULD follow procedures of an Ingress node described above, and perform nexthop resolution process. The interface routes SHOULD be installed in the Transport RIB belonging to the associated Transport Class.

Avoiding path-hiding through Route Reflectors

When multiple BNs exist that advertise a RDn:PEn prefix to RRs, the RRs may hide all but one of the BNs, unless ADDPATH [RFC7911] is used for the Classful Transport family. This is similar to L3VPN option-B scenarios. Hence ADDPATH SHOULD be used for Classful Transport family, to avoid path-hiding through RRs.

Avoiding loop between Route Reflectors in forwarding path

Pair of redundant ABRs acting as RR with nexthop-self may chose each other as best path instead of the upstream ASBR, causing a traffic forwarding loop.

Implementations SHOULD provide a way to alter the tie-breaking rule specified in BGP RR [RFC4456] to tie-break on CLUSTER_LIST step before ROUTER-ID step, when performing path selection for BGP CT routes. RFC4456 considers pure RR which is not in forwarding path. When RR is in forwarding path and reflects routes with nexthop-self, which is the case for ABR BNs in a BGP transport network, this rule may cause loops. This document suggests the following modification to the BGP Decision Process Tie Breaking rules (Sect. 9.1.2.2, [RFC4271]) when doing path selection for BGP CT family routes:

The following rule SHOULD be inserted between Steps e) and f): a BGP Speaker SHOULD prefer a route with the shorter CLUSTER_LIST length. The CLUSTER_LIST length is zero if a route does not carry the CLUSTER_LIST attribute.

Some deployment considerations can also help in avoiding this problem:

IGP metric should be assigned such that "ABR to redundant ABR" cost is inferior than "ABR to upstream ASBR" cost.

Tunnels belonging to special Transport classes SHOULD NOT be provisioned between ABR to ABRs. This will ensure that the route received from an ABR with nexthop-self will not be usable at a redundant ABR.

This avoids possibility of such loops altogether, irrespective of whether the path selection modification mentioned above is implemented.

Ingress node receiving service route with mapping community

Service routes received with mapping community resolve using Transport RIBs determined by the resolution scheme. If the resolution process does not find an usable Classful Transport route or tunnel route in any of the Transport RIBs, the service route MUST be considered unusable for forwarding purpose.

Coordinating between domains using different community namespaces.

Cooperating option-C domains may sometimes not agree on RT, RD, Mapping-community or Transport Route Target values because of differences in community namespaces; e.g. during network mergers or renumbering for expansion. Such deployments may deploy mechanisms to map and rewrite the Route-target values on domain boundaries, using per ASBR import policies. This is no different than any other BGP VPN family. Mechanisms employed in inter-AS

VPN deployments may be used with the Classful Transport family also.

The resolution schemes SHOULD allow association with multiple mapping communities. This helps with renumbering, network mergers, or transitions.

Though RD can also be rewritten on domain boundaries, deploying unique RDs is strongly RECOMMENDED, because it helps in trouble shooting by uniquely identifying originator of a route, and avoids path-hiding.

This document defines a new format of Route-Target extended-community to carry Transport Class, this avoids collision with regular Route Target namespace used by service routes.

11. Scaling considerations

11.1. Avoiding unintended spread of CT routes across domains.

RFC8212 [RFC8212] suggests BGP speakers require explicit configuration of both BGP Import and Export Policies for any EBGP sessions, in order to receive or send routes on EBGP sessions.

It is recommended to follow this for BGP CT routes. It will prohibit unintended advertisement of transport routes through out the BGP CT transport domain which may span multiple AS. This will conserve usage of MPLS label and nexthop resources in the network. An ASBR of a domain can be provisioned to allow routes with only the Transport targets that are required by SNs in the domain.

11.2. Constrained distribution of PNHs to SNs (On Demand Nexthop)

This section describes how the number of Protocol Nexthops advertised to a SN or BN can be constrained using BGP Classsful Transport and VPN RTC [RFC4684]

An egress SN MAY advertise BGP CT route for RD:eSN with two Route Targets: transport-target:0:<TC> and a RT carrying <eSN>:<TC>. Where TC is the Transport Class identifier, and eSN is the IP-address used by SN as BGP nexthop in it’s service route advertisements.

transport-target:0:<TC> is the new type of route target (Transport Class RT) defined in this document. It is carried in BGP extended community attribute (BGP attribute code 16).

The RT carrying <eSN>:<TC> MAY be an IP-address specific regular RT (BGP attribute code 16), IPv6-address specific RT (BGP attribute code 25), or a Wide-communities based RT (BGP attribute code 34) as described in RTC-Ext [RTC-Ext]

An ingress SN MAY import BGP CT routes with Route Target carrying: <eSN>:<TC>. The ingress SN MAY learn the eSN values either by configuration, or it MAY discover them from the BGP nexthop field in the BGP VPN service routes received from eSN. A BGP ingress SN receiving a BGP service route with nexthop of eSN SHOULD generate a RTC/Extended-RTC route for Route Target prefix <Origin ASN>:<eSN>/[80|176] in order to learn BGP CT transport routes to reach eSN. This allows constrained distribution of the transport routes to the PNHs actually required by iSN.

When path of route propogation of BGP CT routes is same as the RTC routes, a BN would learn the RTC routes advertised by ingress SNs and propagate further. This will allow constraining distribution of BGP CT routes for a PNH to only the necessary BNs in the network, closer to the egress SN.

This mechanism provides "On Demand Nexthop" of BGP CT routes, which help with scaling of MPLS forwarding state at SN and BN.

But the amount of state carried in RTC family may become proportional to number of PNHs in the network. To strike a balance, the RTC route advertisements for <Origin ASN>:<eSN>/[80|176] MAY be confined to the BNs in home region of ingress-SN, or the BNs of a super core.

Such a BN in the core of the network SHOULD import BGP CT routes with Transport Class Route Target: 0:<TC>, and generate a RTC route for <Origin ASN>:0:<TC>/96, while not propagating the more specific RTC requests for specific PNHs. This will let the BN learn transport routes to all eSN nodes. But confine their propagation to ingress-SNs.

11.3. Limiting scope of visibility of PE loopback as PNHs

It may be even more desirable to limit the number of PNHs that are globaly visible in the network. This is possible using mechanism described in MPLS Namespaces [MPLS-NAMESPACES]

Such that advertisement of PE loopback addresses as next-hop in BGP service routes is confined to the region they belong to. An anycast IP-address called "Context Protocol Nexthop Address" abstracts the PEs in a region from other regions in the network, swapping the PE scoped service label with a CPNH scoped private namespace label.

This provides much greater advantage in terms of scaling and convergence. Changes to implement this feature are required only on the region's BNs and RR.

12. OAM considerations

Standard MPLS OAM procedures specified in [RFC8029] also apply to BGP Classful Transport.

The 'Target FEC Stack' sub-TLV for IPv4 Classful Transport has a Sub-Type of [TBD], and a length of 13. The Value field consists of the RD advertised with the Classful Transport prefix, the IPv4 prefix (with trailing 0 bits to make 32 bits in all), and a prefix length, encoded as follows:

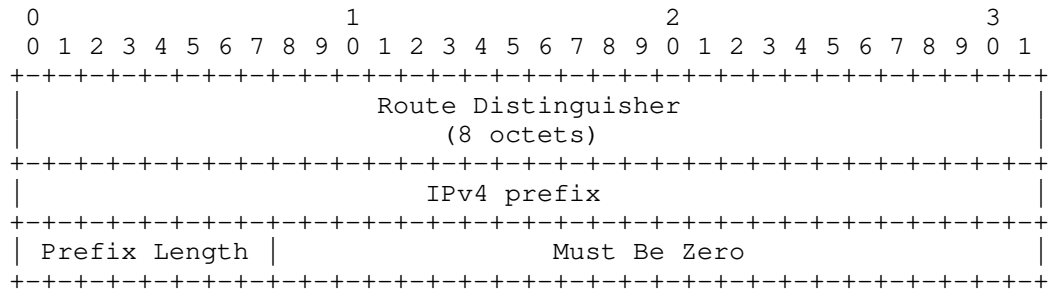


Figure 1: Classful Transport IPv4 FEC

The 'Target FEC Stack' sub-TLV for IPv6 Classful Transport has a Sub-Type of [TBD], and a length of 25. The Value field consists of the RD advertised with the Classful Transport prefix, the IPv6 prefix (with trailing 0 bits to make 128 bits in all), and a prefix length, encoded as follows:

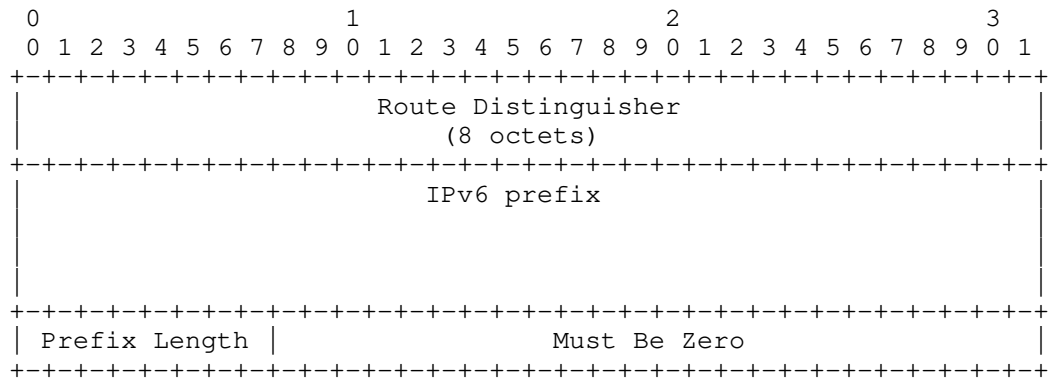


Figure 2: Classful Transport IPv6 FEC

13. Applicability to Network Slicing

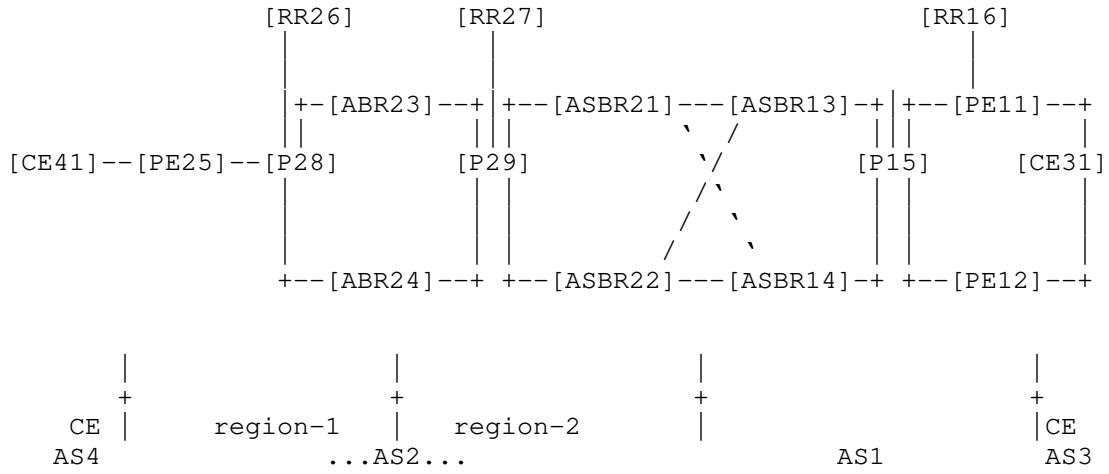
In Network Slicing, the Transport Slice Controller (TSC) sets up the Topology (e.g. RSVP, SR-TE tunnels with desired characteristics) and resources (e.g. polices/shapers) in a transport network to create a Transport slice. The Transport class construct described in this document represents the "Topology Slice" portion of this equation.

The TSC can use the Transport Class Identifier (Color value) to provision a transport tunnel in a specific Topology Slice.

Further, Network slice controller can use the Mapping community on the service route to map traffic to the desired Transport slice.

14. Illustration of procedures with example topology

14.1. Topology



This example shows a provider network that comprises of two Autonomous systems, AS1, AS2. They are serving customers AS3, AS4 respectively. Traffic direction being described is CE41 to CE31. CE31 may request a specific SLA, e.g. Gold for this traffic, when traversing these provider networks.

AS2 is further divided into two regions. So there are three tunnel domains in provider space. AS1 uses ISIS Flex-Algo intra-domain tunnels, whereas AS2 uses RSVP intra-domain tunnels.

The network has two Transport classes: Gold with transport class id 100, Bronze with transport class id 200. These transport classes are provisioned at the PEs and the Border nodes (ABRs, ASBRs) in the network.

Following tunnels exist for Gold transport class.

PE25_to_ABR23_gold - RSVP tunnel

PE25_to_ABR24_gold - RSVP tunnel

ABR23_to_ASBR22_gold - RSVP tunnel

ASBR13_to_PE11_gold - ISIS FlexAlgo tunnel

ASBR14_to_PE11_gold - ISIS FlexAlgo tunnel

Following tunnels exist for Bronze transport class.

PE25_to_ABR23_bronze - RSVP tunnel

ABR23_to_ASBR21_bronze - RSVP tunnel

ABR23_to_ASBR22_bronze - RSVP tunnel

ABR24_to_ASBR21_bronze - RSVP tunnel

ASBR13_to_PE12_bronze - ISIS FlexAlgo tunnel

ASBR14_to_PE11_bronze - ISIS FlexAlgo tunnel

These tunnels are either provisioned or auto-discovered to belong to transport class 100 or 200.

14.2. Service Layer route exchange

Service nodes PE11, PE12 negotiate service families (SAFI 1, 128) on the BGP session with RR16. Service helpers RR16, RR26 have multihop EBGP session to exchange service routes between the two AS. Similarly PE25 negotiates service families with RR26.

Forwarding happens using service routes at service nodes PE25, PE11, PE12 only. Routes received from CEs are not present in any other nodes' FIB in the network.

CE31 advertises a route for example prefix 31.31.31.31 with nexthop self to PE11, PE12. CE31 can attach a mapping community Color:0:100 on this route, to indicate its request for Gold SLA. Or, PE11 can attach the same using locally configured policies. Let us assume CE31 is getting VPN service from PE25.

The 31.31.31.31 route is readvertised in SAFI 128 by PE11 with nexthop self (1.1.1.1) and label V-L1, to RR16 with the mapping community Color:0:100 attached. This SAFI 128 route reaches PE25 via RR16, RR26 with the nexthop unchanged, as PE11 and label V-L1. Now PE25 can resolve the PNH 1.1.1.1 using transport routes received in BGP CT or BGP LU.

The IP FIB at PE25 will have a route for 31.31.31.31 with a nexthop thus found, that points to a Gold tunnel in ingress domain.

14.3. Transport Layer route propagation

ASBR13 negotiates BGP CT family with transport ASBRs ASBR21, ASBR22. They negotiate BGP CT family with RR27 in region 2. ABR23, ABR24 negotiate BGP CT family with RR27 in region 2 and RR26 in region 1. PE25 receives BGP CT routes from RR26. BGP LU family is also

negotiated on these sessions alongside BGP CT family. BGP LU carries "best effort" transport class routes, BGP CT carries gold, bronze transport class routes.

ASBR13 is provisioned with transport class 100, RD value 1.1.1.3:10 and a transport route target 0:100. And a Transport class 200 with RD value 1.1.1.3:20, and transport route target 0:200.

Similarly, these transoprt classes are also configured on ASBRs, ABRs and PEs, with same transport route target, but unique RDs.

Ingress route for ASBR13_to_PE11_gold is advertised by ASBR13 in BGP CT family to ASBRs ASBR21, ASBR22. This route is sent with a NLRI containing RD prefix 1.1.1.3:10:1.1.1.1, Label B-L1 and a route target extended community transport-target:0:100. MPLS swap route is installed at ASBR13 for B-L1 with a nexthop pointing to ASBR13_to_PE11_gold tunnel.

Ingress route for ASBR13_to_PE11_bronze is advertised by ASBR13 in BGP CT family to ASBRs ASBR21, ASBR22. This route is sent with a NLRI containing RD prefix 1.1.1.3:20:1.1.1.1, Label B-L2 and a route target extended community transport-target:0:200. MPLS swap route is installed at ASBR13 for label B-L2 with a nexthop pointing to ASBR13_to_PE11_bronze tunnel

ASBR21 receives BGP CT route 1.1.1.3:10:1.1.1.1 over the single hop EBGP sesion, and readvertises with nexthop self (loopback adderss 2.2.2.1) to RR27, advertising a new label B-L3. MPLS swap route is installed for label B-L3 at ASBR21 to swap to received label B-L1 and forwards to ASBR13. RR27 readvertises this BGP CT route to ABR23, ABR24.

ASBR22 receives BGP CT route 1.1.1.3:10:1.1.1.1 over the single hop EBGP sesion, and readvertises with nexthop self (loopback adderss 2.2.2.2) to RR27, advertising a new label B-L4. MPLS swap route is installed for label B-L4 at ASBR21 to swap to received label B-L2 and forwards to ASBR13. RR27 readvertises this BGP CT route to ABR23, ABR24.

Addpath is enabled for BGP CT family on the sessions between RR27 and ASBRs, ABRs. Such that routes for 1.1.1.3:10:1.1.1.1 with the nexthops ASBR21 and ASBR22 are reflected to ABR23, ABR24 without any path hiding. Thus giving ABR23 visibiity of both available nexthops for Gold SLA.

ABR23 receives the route with nexthop 2.2.2.1, label B-L3 from RR27. The route target "transport-target:0:100" on this route acts as mapping community, and instructs ABR23 to strictly resolve the

nexthop using transport class 100 routes only. ABR23 is unable to find a route for 2.2.2.1 with transport class 100. Thus it considers this route unusable and does not propagate it further. This prunes ASBR21 from Gold SLA tunneled path.

ABR23 also receives the route with nexthop 2.2.2.2, label B-L4 from RR27. The route target "transport-target:0:100" on this route acts as mapping community, and instructs ABR23 to strictly resolve the nexthop using transport class 100 routes only. ABR23 successfully resolves the nexthop to point to ABR23_to_ASBR22_gold tunnel. ABR23 readvertises this route with nexthop self (loopback address 2.2.2.3) and a new label B-L5 to RR26. Swap route for B-L5 is installed by ABR23 to swap to label B-L4, and forward into ABR23_to_ASBR22_gold tunnel.

RR26 reflects the route from ABR23 to PE25. PE25 receives the BGP CT route for prefix 1.1.1.3:10:1.1.1.1 with label B-L5, nexthop 2.2.2.3 and transport-target:0:100 from RR26. And it similarly resolves the nexthop 2.2.2.3 over transport class 100, pushing labels associated with PE25_to_ABR23_gold tunnel.

In this manner, the Gold transport LSP "ASBR13_to_PE11_gold" in egress-domain is extended by BGP CT until the ingress-node PE25 in ingress domain, to create an end-to-end Gold SLA path. MPLS swap routes are installed at ASBR13, ASBR22 and ABR23, when propagating the PE11 BGP CT Gold transport class route 1.1.1.3:10:1.1.1.1 with nexthop self towards PE25.

The BGP CT LSP thus formed, originates in PE25, and terminates in ASBR13, traversing over the Gold underlay LSPs in each domain. ASBR13 uses UHP to stitch the BGP CT LSP into the "ASBR13_to_PE11_gold" LSP to traverse the last domain, thus satisfying Gold SLA end-to-end.

When PE25 receives service route with nexthop 1.1.1.1 and mapping community Color:0:100, it resolves over this BGP CT route 1.1.1.3:10:1.1.1.1. Thus pushing label B-L5, and pushing as top label the labels associated with PE25_to_ABR23_gold tunnel.

14.4. Data plane view

14.4.1. Steady state

This section describes how the data plane looks like in steady state.

CE41 transmits an IP packet with destination as 31.31.31.31. On receiving this packet PE25 performs a lookup in the IP FIB associated with the CE41 interface. This lookup yeids the service route that

pushes the VPN service label V-L1, BGP CT label B-L5, and labels for PE25_to_ABR23_gold tunnel. Thus PE25 encapsulates the IP packet in MPLS packet with label V-L1(innermost), B-L5, and top label as PE25_to_ABR23_gold tunnel. This MPLS packet is thus transmitted to ABR23 using Gold SLA.

ABR23 decapsulates the packet received on PE25_to_ABR23_gold tunnel as required, and finds the MPLS packet with label B-L5. It performs lookup for label B-L5 in the global MPLS FIB. This yields the route that swaps label B-L5 with label B-L4, and pushes top label provided by ABR23_to_ASBR22_gold tunnel. Thus ABR23 transmits the MPLS packet with label B-L4 to ASBR22, on a tunnel that satisfies Gold SLA.

ASBR22 similarly performs a lookup for label B-L4 in global MPLS FIB, finds the route that swaps label B-L4 with label B-L2, and forwards to ASBR13 over the directly connected MPLS enabled interface. This interface is a common resource not dedicated to any specific transport class, in this example.

ASBR13 receives the MPLS packet with label B-L2, and performs a lookup in MPLS FIB, finds the route that pops label B-L2, and pushes labels associated with ASBR13_to_PE11_gold tunnel. This transmits the MPLS packet with VPN label V-L1 to PE11, using a tunnel that preserves Gold SLA in AS 1.

PE11 receives the MPLS packet with V-L1, and performs VPN forwarding. Thus transmitting the original IP payload from CE41 to CE31. The payload has traversed path satisfying Gold SLA end-to-end.

14.4.2. Absorbing failure of primary path

This section describes how the data plane reacts when gold path experiences a failure.

Let us assume tunnel ABR23_to_ASBR22_gold goes down, such that now end-to-end Gold path does not exist in the network. This makes the BGP CT route for RD prefix 1.1.1.1:10:1.1.1.1 unusable at ABR23. This makes ABR23 send a BGP withdrawal for 1.1.1.1:10:1.1.1.1 to RR26, which then withdraws the prefix from PE25.

Withdrawal for 1.1.1.1:10:1.1.1.1 allows PE25 to react to the loss of gold path to 1.1.1.1. Let us assume PE25 is provisioned to use best-effort transport class as the backup path. This withdrawal of BGP CT route allows PE25 to adjust the nexthop of the VPN Service-route to push the labels provided by the BGP LU route. That repairs the traffic to go via best effort path. PE25 can also be provisioned to use Bronze transport class as the backup path. The repair will happen in similar manner in that case as-well.

Traffic repair to absorb the failure happens at ingress node PE25, in a service prefix scale independent manner. This is called PIC (Prefix scale Independent Convergence). The repair time will be proportional to time taken for withdrawing the BGP CT route.

15. IANA Considerations

This document makes following requests of IANA.

15.1. New BGP SAFI

New BGP SAFI code for "Classful Transport". Value 76.

This will be used to create new AFI,SAFI pairs for IPv4, IPv6 Classful Transport families. viz:

- "Inet, Classful Transport". AFI/SAFI = "1/76" for carrying IPv4 Classful Transport prefixes.
- "Inet6, Classful Transport". AFI/SAFI = "2/76" for carrying IPv6 Classful Transport prefixes.

15.2. New Format for BGP Extended Community

Please assign a new Format (Type high = 0xa) of extended community EXT-COMM [RFC4360] called "Transport Class" from the following registries:

the "BGP Transitive Extended Community Types" registry, and

the "BGP Non-Transitive Extended Community Types" registry.

Please assign the same low-order six bits for both allocations.

This document uses this new Format with subtype 0x2 (route target), as a transitive extended community.

The Route Target thus formed is called "Transport Class" route target extended community.

Taking reference of RFC7153 [RFC7153] , following requests are made:

15.2.1. Existing registries to be modified

15.2.1.1. Registries for the "Type" Field

15.2.1.1.1. Transitive Types

This registry contains values of the high-order octet (the "Type" field) of a Transitive Extended Community.

Registry Name: BGP Transitive Extended Community Types

TYPE VALUE	NAME
+ 0x0a	Transitive Transport Class Extended Community (Sub-Types are defined in the "Transitive Transport Class Extended Community Sub-Types" registry)

15.2.1.1.2. Non-Transitive Types

This registry contains values of the high-order octet (the "Type" field) of a Non-transitive Extended Community.

Registry Name: BGP Non-Transitive Extended Community Types

TYPE VALUE	NAME
+ 0x4a	Non-Transitive Transport Class Extended Community (Sub-Types are defined in the "Non-Transitive Transport Class Extended Community Sub-Types" registry)

15.2.2. New registries to be created

15.2.2.1. Transitive "Transport Class" Extended Community Sub-Types Registry

This registry contains values of the second octet (the "Sub-Type" field) of an extended community when the value of the first octet (the "Type" field) is 0x07.

Registry Name: Transitive Transport Class Extended Community Sub-Types

RANGE	REGISTRATION PROCEDURE
0x00-0xBF	First Come First Served
0xC0-0xFF	IETF Review

SUB-TYPE VALUE	NAME
0x02	Route Target

15.2.2.2. Non-Transitive "Transport Class" Extended Community Sub-Types Registry

This registry contains values of the second octet (the "Sub-Type" field) of an extended community when the value of the first octet (the "Type" field) is 0x47.

Registry Name: Non-Transitive Transport Class Extended Community Sub-Types

RANGE	REGISTRATION PROCEDURE
0x00-0xBF	First Come First Served
0xC0-0xFF	IETF Review

SUB-TYPE VALUE	NAME
0x02	Route Target

15.3. MPLS OAM code points

The following two code points are sought for Target FEC Stack sub-TLVs:

- IPv4 BGP Classful Transport
- IPv6 BGP Classful Transport

16. Security Considerations

Mechanisms described in this document carry Transport routes in a new BGP address family. That minimizes possibility of these routes leaking outside the expected domain or mixing with service routes.

When redistributing between SAFI 4 and SAFI 76 Classful Transport routes, there is a possibility of SAFI 4 routes mixing with SAFI 1 service routes. To avoid such scenarios, it is RECOMMENDED that implementations support keeping SAFI 4 routes in a separate transport RIB, distinct from service RIB that contain SAFI 1 service routes.

17. Acknowledgements

The authors thank Jeff Haas, John Scudder, Navaneetha Krishnan, Ravi M R, Chandrasekar Ramachandran, Shradha Hegde, Richard Roberts, Krzysztof Szarkowicz, John E Drake, Srihari Sangli, Vijay Kestur, Santosh Kolenchery, Robert Raszuk, Ahmed Darwish for the valuable discussions and review comments.

The decision to not reuse SAFI 128 and create a new address-family to carry these transport-routes was based on suggestion made by Richard Roberts and Krzysztof Szarkowicz.

18. References

18.1. Normative References

[MPLS-NAMESPACES] Vairavakkalai, Ed., "Private MPLS-label namespaces", 08 2020, <https://tools.ietf.org/html/draft-kaliraj-bess-bgp-sig-private-mpls-labels-01#section-6.1>.

[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>.

[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, <https://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, <https://www.rfc-editor.org/info/rfc4360>.

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

[RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, DOI 10.17487/RFC4456, April 2006, <https://www.rfc-editor.org/info/rfc4456>.

[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, <https://www.rfc-editor.org/info/rfc4684>.

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

[RFC7911] Walton, D., Retana, A., Chen, E., and J. Scudder, "Advertisement of Multiple Paths in BGP", RFC 7911, DOI 10.17487/RFC7911, July 2016, <https://www.rfc-editor.org/info/rfc7911>.

[RFC8029] Kompella, K., Swallow, G., Pignataro, C., Ed., Kumar, N., Aldrin, S., and M. Chen, "Detecting Multiprotocol Label Switched (MPLS) Data-Plane Failures", RFC 8029, DOI 10.17487/RFC8029, March 2017, <https://www.rfc-editor.org/info/rfc8029>.

[RFC8212] Mauch, J., Snijders, J., and G. Hankins, "Default External BGP (EBGP) Route Propagation Behavior without Policies", RFC 8212, DOI 10.17487/RFC8212, July 2017, <https://www.rfc-editor.org/info/rfc8212>.

[RFC8277] Rosen, E., "Using BGP to Bind MPLS Labels to Address Prefixes", RFC 8277, DOI 10.17487/RFC8277, October 2017, <https://www.rfc-editor.org/info/rfc8277>.

[RFC8669] Previdi, S., Filsfils, C., Lindem, A., Ed., Sreekantiah, A., and H. Gredler, "Segment Routing Prefix Segment Identifier Extensions for BGP", RFC 8669, DOI 10.17487/RFC8669, December 2019, <https://www.rfc-editor.org/info/rfc8669>.

[RTC-Ext] Zhang, Z., Ed., "Route Target Constrain Extension", 07 2020, <https://tools.ietf.org/html/draft-zzhang-idr-bgp-rt-constrains-extension-00#section-2>.

[Seamless-SR] Hegde, Ed., "Seamless Segment Routing", 11 2020, <https://datatracker.ietf.org/doc/html/draft-hegde-spring-mpls-seamless-sr-03>.

[SRTE] Previdi, S., Ed., "Advertising Segment Routing Policies in BGP", 11 2019, <https://tools.ietf.org/html/draft-ietf-idr-segment-routing-te-policy-08>.

18.2. URIs

[1] https://www.rfc-editor.org/rfc/rfc4271#section-9.1.2.1

Authors' Addresses

Kaliraj Vairavakkalai
Juniper Networks, Inc.
1133 Innovation Way,
Sunnyvale, CA 94089
US

Email: kaliraj@juniper.net

Natrajan Venkataraman
Juniper Networks, Inc.
1133 Innovation Way,
Sunnyvale, CA 94089
US

Email: natv@juniper.net

Balaji Rajagopalan
Juniper Networks, Inc.
Electra, Exora Business Park˜Marathahalli - Sarjapur Outer Ring Road,
Bangalore, KA 560103
India

Email: balajir@juniper.net

Gyan Mishra
Verizon Communications Inc.
13101 Columbia Pike
Silver Spring, MD 20904
USA

Email: gyan.s.mishra@verizon.com

Mazen Khaddam
Cox Communications Inc.
Atlanta, GA
USA

Email: mazen.khaddam@cox.com

Xiaohu Xu
Alibaba Inc.
Beijing
China

Email: xiaohu.xxh@alibaba-inc.com

Rafal Jan Szarecki
Google.
1160 N Mathilda Ave, Bldg 5,
Sunnyvale,, CA 94089
USA

Email: szarecki@google.com

INTERNET-DRAFT
Intended Status: Informational
Expires: July 29, 2021

L. Krattiger, Ed.
A. Sajassi, Ed.
S. Thoria
Cisco Systems

J. Rabadan
Nokia

J. Drake
Juniper

January 25, 2021

EVPN Interoperability Modes
draft-krattiger-evpn-modes-interop-03

Abstract

Ethernet VPN (EVPN) provides different functional modes in the area of Service Interface, Integrated Route and Bridge (IRB) and IRB Core connectivity. This document specifies how the different EVPN functional modes and types can interoperate with each other. This document doesn't aim to redefine the existing functional modes but extend them for interoperability.

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 3
1.1 Requirements Language 3
2. Valid Combinations for Interoperability 3
3. Service Interface Interoperability 5
3.1. VLAN-Aware Bundle and VLAN-Based 5
3.1.1. VLAN-Aware Bundle Service PE 6
3.1.2. VLAN-Based Service PE 7
3.2. Service Interface Interop Mode of Operation 7
4. Interoperability for different IRB Types 8
4.1. Asymmetric IRB and Symmetric IRB 8
4.1.1. Asymmetric IRB PE 10
4.1.2. Symmetric IRB PE 10
4.2. IRB Interop Mode of Operation 11
5. Interoperability for different IRB Core Connectivity Modes . . 12
5.1. Interface-Less and Interface-Ful Unnumbered IRB 12
5.1.1. Interface-Less PE 15
5.1.2. Interface-Ful Unnumbered IRB 15
5.2. Tunnel Encapsulation Consideration 17
6. Security Considerations 17
7. IANA Considerations 17
8. References 18
8.1. Normative References 18
8.2. Informative References 18
9. Conclusion 19
9.1. Demonstration of Applicability 19
9.1.1. Service Interface Interoperability 19
9.1.2. IRB Types 19
9.1.3. IRB Core Connectivity Types 20

10. Authors' Addresses 20

1. Introduction

Ethernet VPN (EVPN) provides different functional modes in the area of Service Interface, Integrated Route and Bridge (IRB) and IRB connection model. It is understood that the different modes are defined with different use-cases in mind. Even with the specific use-cases and the resulting mode definition, the aim of interoperability is critical.
The following EVPN modes are considered for interoperability. It is limited to most pertinent interop modes as oppose to all permutations. In the future if other modes are identified, it will be addressed in future revisions.

- For Service Interfaces, the VLAN Aware Bundle and VLAN Based types.

- In Integrated Routing and Bridging (IRB) the Asymmetric IRB and Symmetric IRB type.

- Within the IRB connectivity types, interface-less and the interface-ful Unnumbered IRB.

1.1 Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

2. Valid Combinations for Interoperability

The tables below provide an overview of the valid combinations for interoperability described in this Internet-Draft.

For the Service Interface Types as described in [RFC7432] section 6 and [RFC8365] section 5.1.2. Interoperability considerations are provided for the VLAN-Based Service interface ([RFC7432], section 6.1) and the VLAN-Aware Bundle Service Interface type ([RFC7432] section 6.3). The VLAN Bundle Service Interface ([RFC7432] section 6.2) is not considered at this time.

Table 1 represent the considered Service Interface Types interoperability:

	VLAN-Based	VLAN Bundle	VLAN-Aware Bundle
VLAN-Based	YES	NO	YES
VLAN Bundle	NO	YES	NO
VLAN-Aware Bundle	YES	NO	YES

Table 1

In regards to Integrated Route and Bridge (IRB), two different modes are defined in [EVPN-INTERSUBNET], with section 5 describing Symmetric IRB and section 6 Asymmetric IRB:

> The interoperability considerations for Asymmetric IRB and Symmetric IRB mode are represented within this Internet-Draft.

For the IRB Core Connectivity, from all the available modes as described in [EVPN-PREFIX], considered for interoperability is the interface-less mode (section 4.4.1) in conjunction with only one of the interface-ful modes, namely interface-ful IP-VRF-to-IP-VRF with Unnumbered SBD IRB (section 4.4.3). With the implementation proximation between the two interface-ful modes, considerations for interoperability between interface-less and interface-ful Numbered are currently not considered. Similarly, the interoperability between the two interface-ful modes is currently not being considered, given the already close relation and to limit permutations. Future revisions of this Internet-Draft might address further variations of interoperability.

Table 2 represent the considered IRB Core Connectivity interoperability.

	Interface-Less	Interface-Ful Numbered IRB	Interface-Ful Unnumbered IRB
Interface-Less	YES	NO	YES
Interface-Ful Numbered IRB	NO	YES	NO
Interface-Ful Unnumbered IRB	YES	NO	YES

Table 2

3. Service Interface Interoperability

3.1. VLAN-Aware Bundle and VLAN-Based

[RFC7432] section 6 describes three different Service Interface Types. The two modes in focus for interoperability are namely the VLAN-Based Service Interface as defined in [RFC7432] section 6.1 and the VLAN-Aware Bundle Service Interface as defined in [RFC7432] section 6.3. The VLAN Bundle Service Interface is not considered.

The VLAN-Based Service Interface defines an EVPN instance consisting of only a single broadcast domain or "Single Broadcast Domain per EVI" as described in [RFC8365] section 5.1.2 Option 1. In this mode, individual BGP Route Distinguisher (RD) and Route Target (RT) are required for each EVI. Each EVI corresponds to a single MAC-VRF identified by the RT, which provides the advantage of an BGP RT constraint mechanisms in order to limit the propagation and import of routes to only the PE that are interested. With VLAN-Based, the MAC-VRF corresponds to only a single bridge table. The VLAN-Based Service Interface uses the EVPN MAC/IP Advertisement route ([RFC7432], section 7.2) with the MUST requirement of the Ethernet Tag ID being set to zero.

Differently, the VLAN-Aware Bundle Service Interface follows a bundling of multiple broadcast domains, with each having its own bridge table, into a single EVI. This refers to the definition of "Multiple Broadcast Domain per EVI" as described in [RFC8365] section 5.1.2 Option 2. The advantage of this model is that it doesn't require the provisioning of an RD/RT per broadcast domain, which is a moot point when VLAN-Base uses auto-derivation of RD/RT. With VLAN-Aware Bundle Service, RT Constraint, as defined in [RFC4684], does not help to reduce the dissemination of routes for a BD to the PEs attached to that BD. This is given by the nature of the bundle service where the RT is not sufficient to identify the MAC-VRF and corresponding bridge table. The differences between the two modes of Service Interfaces, namely VLAN-Based and VLAN-Aware Bundle Service Interface, lie in the definition of the Ethernet Tag field in the EVPN routes. While VLAN-Based Service Interface defines the EtherTag as "must be set to zero", the VLAN-Aware Bundle Service interface uses the VID within the EtherTag to identify the bridge table within the MAC-VRF. These two requirements are orthogonal and as a result make the interoperability of the two types mutually exclusive, an interoperability is not achievable (Figure 1).

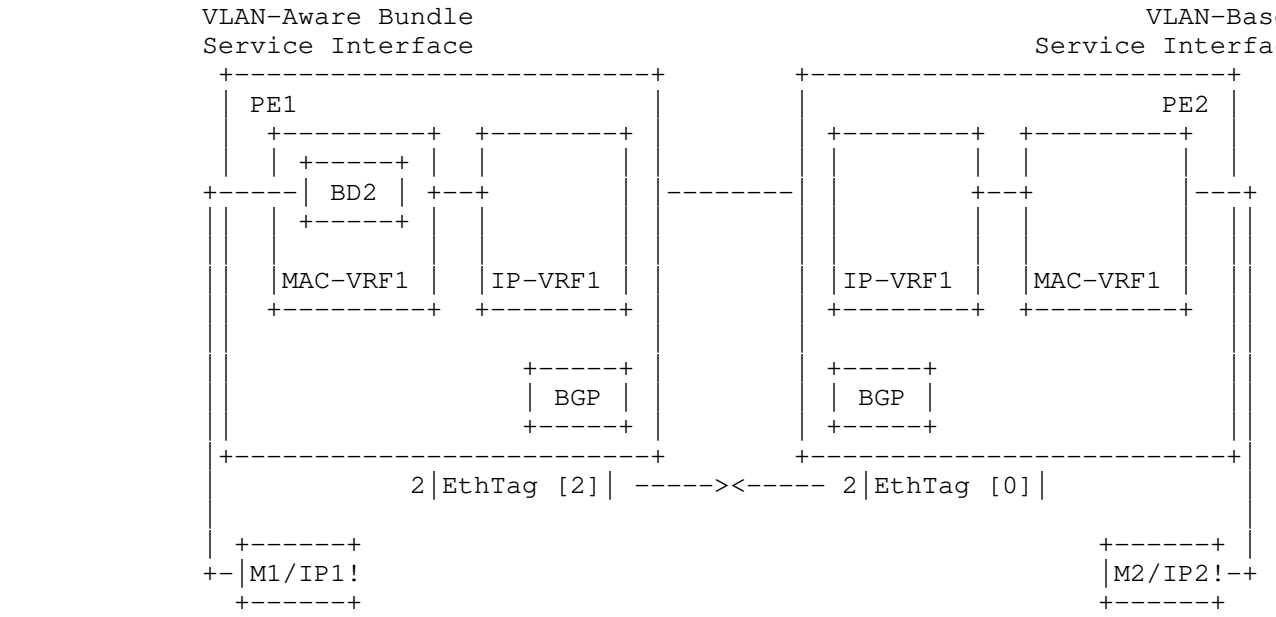


Figure 1: Interop of different Service Interface Types

As illustrated in Figure 1, the MAC/IP routes exchanged by PE1 and PE2 contain Ethernet Tags 2 and 0 respectively. The receiving PE will not process these routes and will normally discard them (treat-as-withdraw)."

By extending the requirements currently present, an interoperability is achievable. The adjustment would be as follows.

3.1.1. VLAN-Aware Bundle Service PE

In case of VLAN Aware Bundle Service Interface on the receiving PE and with the consideration of VLAN Based Service Interface on the advertising PE:

- MUST Operate in Single Broadcast Domain per EVI.

- Multiple Broadcast Domain per EVI case is not considered.

- MUST allow to send and receive zero EtherTag.

- The import of routes is performed based on the import policy (route-target).

- With single bridge table per MAC-VRF, additional evaluation of the EtherTag field is not required; the bridge table is sufficiently defined by the import route-target.

- No Change to data-plane operation, the MPLS label identifies MAC-VRF + bridge-table, or the VNI identifies the MAC-VRF + the bridge-table.

3.1.2. VLAN-Based Service PE

- Operates in Single Broadcast Domain per EVI.

In case of VLAN Based Service Interface on the receiving PE and with the consideration of VLAN Based Service Interface on the advertising PE:

- Operates in Single Broadcast Domain per EVI.

- MUST allow receiving of non-zero EtherTag.

- No Change in control-plane operation, the EVI import policy (route-target) identifies the broadcast domain (bridge-table) within a MAC-VRF.

- No Change to data-plane operation, the MPLS label identifies MAC-VRF + bridge-table, or the VNI identifies the MAC-VRF + the bridge-table.

While the expansion introduces additional configuration requirement for the VLAN-Aware Bundle Service Interface, it also allows for broader interoperability in the eventuality of Vendor "A" only implemented VLAN-Based while Vendor "B" only implemented VLAN-Aware Bundle Service Interface.

3.2. Service Interface Interop Mode of Operation

When Service Interface interoperability is required, a given PE should follow this section’s procedures for all its broadcast domains (BDs) and not just the BDs that need interoperability.

For those BDs where interoperability between VLAN-Aware Bundle and VLAN-Based Service Interface is needed, ignoring the presence of the EVPN routes Ethernet Tag ID on the PEs supporting VLAN-Based mode is not enough. Each PE needs to clearly signal what mode it supports, so that all the PEs attached to the same EVI can understand in what mode the EVI operates.

Consider a scenario where PE1 is attached to the BD range BD1-10 and it operates in VLAN-Aware mode, whereas PE2 is attached to the BD range BD7-20 and operates in VLAN-Based mode. Interoperability is required for the intersecting BDs, I.e., BD7-10.

For PE1, this means BD7-10 need to be separated into a dedicated MAC-VRF each. EVPN routes for each of these four MAC-VRFs MUST be advertised by PE1 with an Ethernet Tag ID of zero. In this way, PE1 indicates the use of VLAN-Based mode for those BDs. On reception, PE1 imports the BD7-10 routes based on the Route Target and ignoring the Ethernet Tag ID, as the Route Target alone is sufficient to identify the correct MAC-VRF and Bridge Table. The remaining BDs on PE1 (range BD1-6) continue operating in VLAN-Aware Bundle mode.

In the same example, other PEs attached to BD1-6 must still process the received Ethernet Tag ID in the EVPN routes from PE1, so that they can identify the correct Bridge Table in a given MAC-VRF.

PE2 operates in VLAN-Based mode for BD7-20, as per [RFC7432] and [RFC8365]. PE2's EVPN route advertisements for BD7-20 will include individual Route Targets per BD and an Ethernet Tag ID of zero. On reception, PE2 identifies the MAC-VRF and Bridge Table solely based on the Route Target.

4. Interoperability for different IRB Types

4.1. Asymmetric IRB and Symmetric IRB

The differences in the two inter-subnet forwarding modes, namely Asymmetric IRB and Symmetric IRB, are beyond just the information difference in the control-plane from an EVPN Route Type 2 perspective. The two IRB modes have significant differences in inter-subnet forwarding behavior and as a result different operation during label imposition or encapsulation.

With the Asymmetric IRB mode, the ingress PE performs a "bridge-and-route" operation while the egress PE follows a "bridge-only" approach. Differently, the forwarding behavior in Symmetric IRB mode performs a "bridge-and-route" operation on the ingress PE followed by a "route-and bridge" operation at the egress PE. The significance in difference is not only in the forwarding behavior itself but also around how the respective EVPN attribute are used for driving the inter-subnet operation. More specifically, in the case of inter-subnet forwarding with Asymmetric IRB, MPLS Label1 is used towards the egress PE to specify the MAC-VRF and respective Bridge-Domain for forwarding. In inter-subnet forwarding with Symmetric IRB, MPLS Label2 associated with the IP-VRF is used for the inter-subnet

forwarding operation towards egress PE.

The respective forwarding behaviors are described in [EVPN-INTERSUBNET]. The following steps are required to ensure the interoperability between the Asymmetric and Symmetric IRB modes.

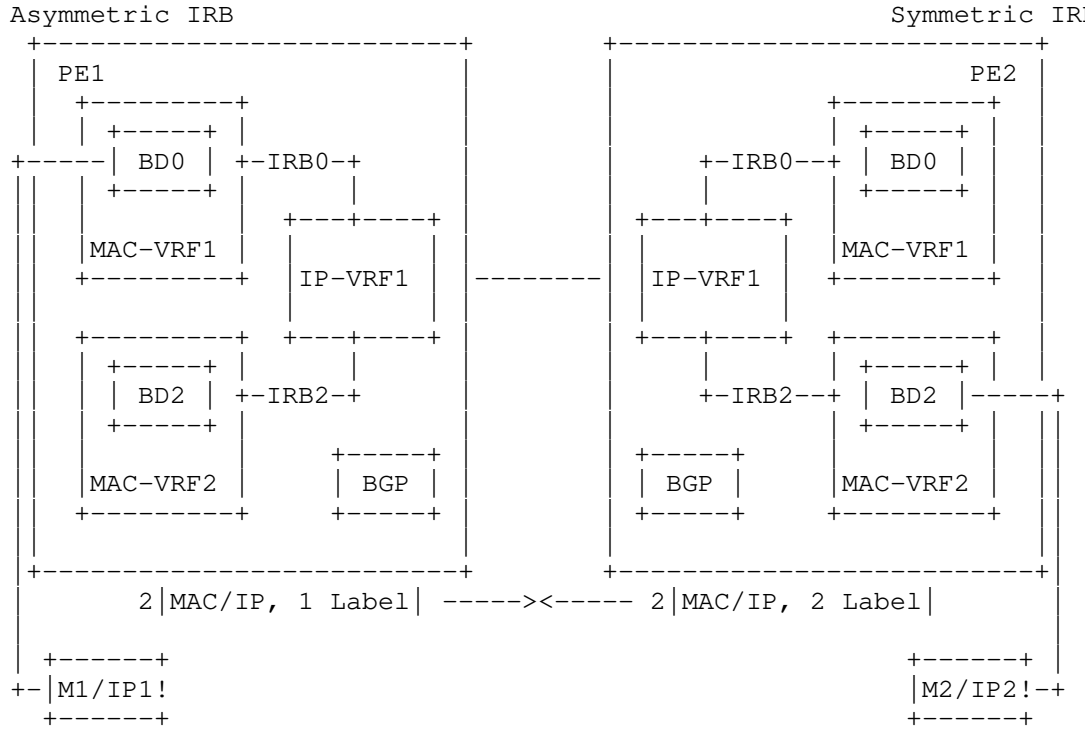


Figure 2: Asymmetric IRB and Symmetric IRB

Figure 2 illustrates the overview of and Asymmetric IRB PE (PE1) and a Symmetric IRB PE (PE2) within a interoperability deployment scenario. Attached to PE1, end-point M1/IP1 is attached to BD0 within MAC-VRF1. Respectively, on PE2 end-point M2/IP2 is connected via attachment circuit to BD2 positioned within MAC-VRF2. IRB0 and IRB2 represent the host-facing IRB interface for inter-subnet communication between the different end-points located in the different IP Subnet. The IRB interfaces for a common MAC-VRF/BD on PE1 and PE2 use the same IP address. With the difference of the IRB modes between PE1 (Asymmetric IRB) and PE2 (Symmetric IRB), there is a difference in the MPLS Label presence as part of the MAC/IP routes exchanged between the PEs. PE1s update contains a single label,

representing MPLS Label1 used for bridging purposes. PE2s advertisement contains two labels, one for bridging and one for routing, as part of the MAC/IP route. While PE1 receives all information necessary from PE2, PE2 is missing information necessary for its routing operation. As a result, Inter-Subnet routing between PE1 and PE2 is not achieved.

By following the current existing forwarding behavior as described in [EVPN-INTERSUBNET], interoperability is theoretically achievable without changes in the control-plane format. Nevertheless, there are steps required that involve predominantly the local behavior of the PE with Symmetric IRB mode.

4.1.1. Asymmetric IRB PE

In case of Asymmetric IRB as the advertising PE and with Symmetric IRB on the receiving PE:

- Asymmetric IRB PE MUST send MAC and IP information with MPLS Label1.

In case of Symmetric IRB as the advertising PE and with Asymmetric IRB on the receiving PE:

- Asymmetric IRB PE MUST be able to ignore MPLS Label2.

4.1.2. Symmetric IRB PE

In case of Symmetric IRB as the advertising PE and with Asymmetric IRB on the receiving PE:

- Symmetric IRB PE has no additional requirements.

In case of Asymmetric IRB as the advertising PE and with Symmetric IRB on the receiving PE:

- Symmetric IRB PE requires to add the host-binding information, MAC and IP, and associates them to the adjacency (ARP/ND) table facing the PE with Asymmetric IRB; this is in addition of adding the MAC address into the MAC-VRF table. Since there is no MPLS Label2 or Route-Target for the IP-VRF, the Host IP is not specifically added to IP-VRF table.

4.2. IRB Interop Mode of Operation

Interoperability between the Asymmetric IRB and Symmetric IRB mode follows specific defined behavior that is predominantly required on the PE that operates in the Symmetric IRB mode. Nevertheless, in support for the interoperability, the PE operating in Asymmetric IRB MUST accommodate the following two minimal requirements (with references to Figure 2): 1) The PE that operates in Asymmetric IRB mode (PE1), MUST send the MAC/IP route including the Host IP address. 2) The PE with Asymmetric IRB (PE1) MUST accept the MAC/IP routes sent from PE2 (Symmetric IRB), while ignoring the additional information of MPLS Label2 and Route-Target of the IP-VRF.

In reference to 1), the PE MUST always send the end-point MAC address, Host IP address and related MPLS Label1 as part of the MAC/IP route towards the PE with Symmetric IRB (PE2). This route will be sent only with MPLS Label1 and the Route-Target of the matching MAC-VRF. In reference to the illustration in Figure 2, PE1 MUST generate and advertise an EVPN MAC/IP route using:

- MAC Length of 48

- MAC Address of M1

- IP Length of 32 / 128

- IP Address of IP1

- Label for MAC-VRF1

- Route-Target of MAC-VRF1

- Next-Hop PE1

For completeness of the requirements and in reference of 2), the MAC/IP route advertised from the PE operating in Symmetric IRB (PE2) is as follow:

- MAC Length of 48

- MAC Address of M2

- IP Length of 32 /128

- IP Address of IP2

- Label for MAC-VRF2, IP-VRF1

- Route-Target of MAC-VRF2, IP-VRF1

- Next-Hop PE2

As defined in 2), the Label and Route-Target information for IP-VRF1 MUST be ignored by PE1 (PE with Asymmetric IRB).

With PE2 operating in Symmetric IRB and with enabled interop mode, the MAC/IP route from PE1 (Asymmetric IRB) is processed in the respective bridging, routing and adjacency table. Based on the Route-Target for MAC-VRF1, the MAC address M1 will be imported into MAC-VRF1 respectively and placed within BD0. In addition, the host-binding information M1/IP1 MUST be installed within PE2s adjacency table. Subsequent, on PE2 the MAC address M1 and the host-binding information (adjacency table entry) of M1/IP1 MUST point towards PE1 as the next-hop. With no presence of the Route-Target for IP-VRF1, the IP address IP1 will not be specifically imported into IP-VRF1 and is not associated with a MPLS Label2. As a result of the interoperability, the additional efficiency provided by Symmetric IRB in regards of preserving adjacency table exhaustion is reduced; this is specifically when communicating with an Asymmetric IRB based egress PE. In contrary, the interop mode allows for communication between the different IRB modes. As a result, in the eventuality that Vendor "A" only provides Asymmetric IRB, while Vendor "B" only has Symmetric IRB available, interoperability for inter-subnet forwarding can be seamlessly achieved. In addition, two further benefits are present by implementing an Asymmetric/Symmetric Co-Existence on the same PE (dual-mode PE).

- A dual-mode PE can seamlessly communicate with PE's that are either in Asymmetric or in Symmetric IRB mode.

- A dual-mode PE can act as Anchor for interconnecting Symmetric IRB and Asymmetric IRB based PE's (deployment restrictions might apply).

5. Interoperability for different IRB Core Connectivity Modes

5.1. Interface-Less and Interface-Ful Unnumbered IRB

The two modes, namely interface-less and interface-ful Unnumbered SBD IRB, are closely related in regards to the information required in the EVPN Route Type 5. While interface-less provides all information for the IP prefix advertisement within the EVPN Route Type 5, in the case of interface-ful Unnumbered SBD IRB, an additional EVPN Route Type 2 is required for the next-hop recursive lookup. From a forwarding behavior, both approaches are similar and follow a symmetric routing approach but are not interoperable. Note as per

[EVPN-PREFIX] the interface-ful Unnumbered SBD IRB is an OPTIONAL mode.

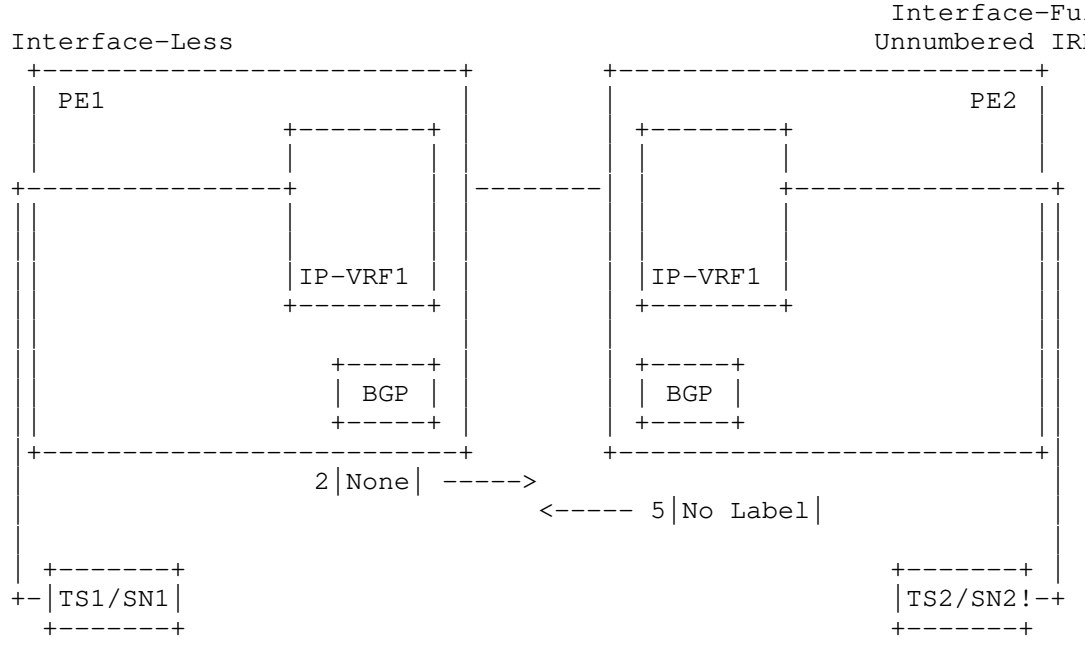


Figure 3: Interoperability of different IRB Core Connectivity Mode (unnumbered)

The illustration in Figure 3 represents the possible deployment scenario between two different Core IRB Connectivity modes. Specifically, PE1 is operating with interface-less Core IRB Mode while PE2 operates with the interface-ful Unnumbered SDB IRB mode; both operate without interoperability capabilities. Attached to PE1 and PE2 respectively, Tenant System 1 (TS1) and Tenant System 2 (TS2) with different IP Subnet are present. TS1 attached on PE1 as well as TS2 attached to PE2 are represented in a common IP-VRF (IP-VRF1), sharing a common Route-Target between the PEs. With the different IRB Core Connectivity modes on PE1 and PE2 respectively, the differences in IP prefix advertisements as described in [EVPN-PREFIX] are present. PE1 advertises only a single EVPN Route Type 5 (IP Prefix Route) for TS1 using the fields following the interface-less mode:

EVPN Route Type 5:

- IP Length of 0 to 32 / 0 to 128

- IP Address of SN1

- Label for IP-VRF1

- GW IP Address set to zero

- Route-Target of IP-VRF1

- Router's MAC Extended Community of PE1

- Next-Hop PE1

Differently, PE2 advertises an EVPN Route Type 2 (MAC/IP Route) next to the EVPN Route Type 5 (IP Prefix Route). The MAC/IP Route supports the requirement for recursive next-hop resolution for the next-hop used in the IP Prefix Route. Below the fields used in the Route Type 5 and respective Route Type 2 according to the interface-ful Unnumbered IRB mode:

EVPN Route Type 5:

- IP Length of 0 to 32 / 0 to 128

- IP Address of SN1

- Label SHOULD be set to 0

- GW IP Address SHOULD be set to "

- Route-Target of IP-VRF1

- Router's MAC Extended Community of PE2

- Next-Hop PE2

EVPN Route Type 2:

- MAC Length of 48

- MAC Address of PE2

- IP Length of 32 / 128

- IP Address of PE2

- Label for IP-VRF1

- Route-Target of IP-VRF1

- Next-Hop PE2

While PE1 is missing the MPLS Label for the IP-VRF from PE2, PE2 is missing the MPLS Label information and the necessary info for the next-hop recursion. As a result, Routing with IP Prefix Advertisement between PE1 and PE2 is not achieved.

By advertising an additional EVPN Route Type 2 from interface-less (PE1) and by advertising the MPLS Label as part of EVPN Route Type 5 from PE2, interoperability is achievable. The specific mode of operation would be as per the following two section and refers to Figure 3 and Figure 4.

5.1.1. Interface-Less PE

In case of interface-less on the advertising PE and with the consideration of interface-ful Unnumbered IRB as the receiving PE:

Shall generate and Advertise EVPN Route Type 2 for every IP-VRF using.

- MAC Length of 48

- MAC Address with "Router MAC"

- IP Length of 32

- IP Address with NVE IP

- Label for IP-VRF

- Route-Target of IP-VRF

- Router-MAC Extended Community

In case of interface-less on the receiving PE and with the consideration of interface-ful Unnumbered IRB as the advertising PE:

- MUST ignore EVPN Route Type 2 with MPLS Label and route-target matching the IP-VRF because there is no MAC-VRF defined matching these information.

5.1.2. Interface-Ful Unnumbered IRB

In case of interface-ful Unnumbered on the advertising PE and with

the consideration of interface-less as the receiving PE:

- Shall advertise MPLS Label for IP-VRF in EVPN Route Type 5 with matching route-target.

In case of interface-ful Unnumbered on the receiving PE and with the consideration of interface-less as the advertising PE:

- No Additions Required.

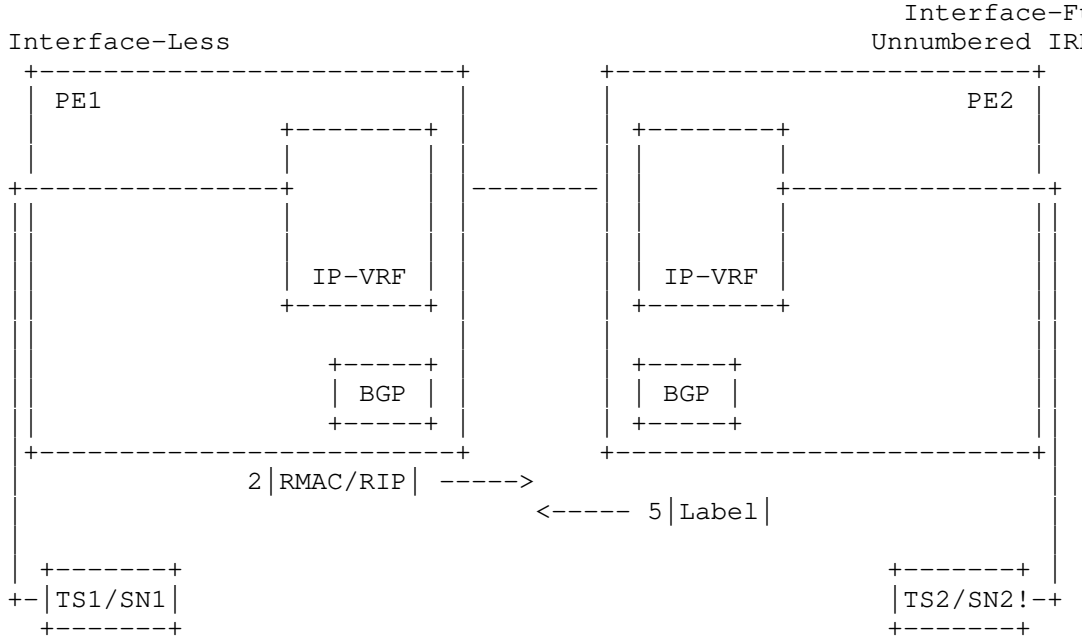


Figure 4: Interop of different IRB Core Connectivity Types (unnumbered)

Illustrated in Figure 4 are the additional requirements for interface-less IRB Core Connectivity mode, specifically the MAC/IP Route (EVPN Route Type 2) necessary for PE2s next-hop recursion. Also, the MPLS Label addition within PE2s IP Prefix route (EVPN Route Type 5) is represented, which is required for interface-ful Unnumbered IRBs advertisement towards an interface-less PE (PE1)

The interop mode introduces additional control-plane advertisements from an Interface-less perspective. This is necessary to allow interface-ful Unnumbered SBD IRB to perform the recursive lookup required. From a EVPN Type 5 perspective between the two types, most

of the fields are already equally defined and populated as per [EVPN-PREFIX]. Exception is the IP-VRF Label, which is required to be added in the interface-ful Unnumbered SBD IRB's EVPN Type 5. In addition, the Interface-less addition allows the Co-Existence of both types on the same PE (dual-mode PE). Such a dual-mode PE can communicate at the same time with PE's that are in Interface-less or in interface-ful Unnumbered SBD IRB mode.

The disadvantage of the additional advertisement has to be put into relation to advantage of successful interoperability where eventualy Vendor "A" only implemented interface-less while Vendor "B" only implemented interface-ful Unnumbered SBD IRB.

5.2. Tunnel Encapsulation Consideration

In regards to IRB core connectivity both solutions, namely interface-less and interface-ful, provide a solution for Layer 3 connectivity among the IP-VRFs. Even as the functional result of both modes is the same, there are important considerations in regards to tunnel encapsulations.

[EVPN-IRB] section 4 considers the choice for the NVO tunnel should be dictated by the tunnel capabilities. For example for the IP-VRF-to-IP-VRF model with interface-less, the NVO tunnel for MPLS needs to be IP NVO and for VXLAN needs to be Ethernet NVO.

With the "IP-VRF-to-IP-VRF" model that is used in interface-ful (numbered or unnumbered), section 4.4.2 or 4.4.3 respectively describe the solution to accommodate Ethernet NVO tunnels (VXLAN or GPE, GENEVE, MPLS with MAC payload) only. In the case of interface-ful unnumbered, the Router-MAC Extended Community is always signaled via EVPN update message, which implies the presence of a MAC payload. IP NVO Tunnel are not applicable to these two use-cases/models

Depending on the use of NVO tunnels, interoperability between interface-les and interface-ful unnumbered requires additional changes on the Tunnel Encapsulation mode. This Internet-Draft considers the usage of a compatible NVO Tunnel mode between a PE operating in interface-les and a PE operating nterface-ful unnumbered mode.

6. Security Considerations

TBD.

7. IANA Considerations

TBD.

8. References

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

[RFC8365] Sajassi, A., Ed., Drake, J., Ed., Bitar, N., Shekhar, R., Uttaro, J., and W. Henderickx, "A Network Virtualization Overlay Solution Using Ethernet VPN (EVPN)", RFC 8365, DOI 10.17487/RFC8365, March 2018, <https://www.rfc-editor.org/info/rfc8365>.

[EVPN-INTERSUBNET] Sajassi et al., "Integrated Routing and Bridging in EVPN", draft-ietf-bess-evpn-inter-subnet-forwarding-11, work in progress, October, 2020.

[EVPN-PREFIX] Rabadan et al., "IP Prefix Advertisement in EVPN", draft-ietf-bess-evpn-prefix-advertisement-11, May 2018.

[KEYWORDS] 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>.

[RFC1776] Crocker, S., "The Address is the Message", RFC 1776, DOI 10.17487/RFC1776, April 1 1995, <http://www.rfc-editor.org/info/rfc1776>.

[TRUTHS] Callon, R., "The Twelve Networking Truths", RFC 1925, DOI 10.17487/RFC1925, April 1 1996, <http://www.rfc-editor.org/info/rfc1925>.

8.2. Informative References

[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, <https://www.rfc-editor.org/info/rfc4684>.

[EANTC] EANTC, "Multi-Vendor Interoperability Test", February 2019, <http://www.eantc.de/fileadmin/eantc/downloads/News/2019/EANTC-MPLSSDNNFV2019-WhitePaper-v1.2.pdf>.

[EVILBIT] Bellovin, S., "The Security Flag in the IPv4 Header", RFC 3514, DOI 10.17487/RFC3514, April 1 2003, <http://www.rfc-editor.org/info/rfc3514>.

[RFC5513] Farrel, A., "IANA Considerations for Three Letter Acronyms", RFC 5513, DOI 10.17487/RFC5513, April 1 2009, <http://www.rfc-editor.org/info/rfc5513>.

[RFC5514] Vyncke, E., "IPv6 over Social Networks", RFC 5514, DOI 10.17487/RFC5514, April 1 2009, <http://www.rfc-editor.org/info/rfc5514>.

9. Conclusion

With minimal additions, the most common EVPN types for Virtual Identifiers to EVI Mapping, Integrated Routing and Bridging and IP Prefix Advertisement can be made interoperable. The aim for interoperability doesn't remove the requirement for optimized types for different use-cases but allows flexibility and basic interoperability.

9.1. Demonstration of Applicability

Cisco, Juniper and Nokia demonstrated successfully the ability of EVPN interoperability modes during EANTCs yearly "Multi-Vendor Interoperability Test". The Whitepaper can be obtained through EANTC with the latest version being available at [EANTC].

9.1.1. Service Interface Interoperability

A proof of the benefit with this interoperability mode has already been demonstrated during EVPN Multi-Vendor interoperability testing and also, in production environments. Specifically, Cisco and Nokia's VLAN-Based Service Interface successful proofed interoperability with Junipers VLAN-Aware Bundle Service Interface.

9.1.2. IRB Types

A proof of the benefit with this interoperability mode has already successfully demonstrated during EVPN Multi-Vendor interoperability

testing. Specifically, Cisco operated in a Hybrid IRB (Dual-Mode) mode while other Vendor operated in an Asymmetric IRB mode. Forwarding was achieved through dynamic detection of the alternate Vendor PE's mode and adjustment to Asymmetric IRB for these specific BDs. Communication for all other BDs continued to be Symmetric IRB.

9.1.3. IRB Core Connectivity Types

A proof of an interoperability mode between interface-less and interface-ful Unnumbered SBD IRB has already been demonstrated in production environments and during EVPN Multi-Vendor interoperability testing. Specifically, Cisco's addition for Interface-less is successfully deployed with Nokia's and Nuage's interface-ful Unnumbered SBD IRB at customers

10. Authors' Addresses

Lukas Krattiger
Cisco
USA
EMail: lkrattig@cisco.com

Ali Sajassi
Cisco
USA
EMail: sajassi@cisco.com

Samir Thoria
Cisco
USA
EMail: sthoria@cisco.com

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

John E. Drake
Juniper
EMail: jdrake@juniper.net

BESS Working Group
Internet-Draft
Intended status: Best Current Practice
Expires: August 26, 2021

G. Mishra
Verizon Inc.
M. Mishra
Cisco Systems
J. Tantsura
Apstra, Inc.
L. Wang
Juniper Networks, Inc.
Q. Yang
Arista Networks
A. Simpson
Nokia
S. Chen
Huawei Technologies

February 22, 2021

IPv4 NLRI with IPv6 Next Hop Use Cases
draft-mishra-bess-ipv4nlri-ipv6nh-use-cases-08

Abstract

As Enterprises and Service Providers upgrade their brown field or green field MPLS/SR core to an IPv6 transport such as MPLS LDPv6, SR-MPLSv6 or SRv6, Multiprotocol BGP (MP-BGP)now plays an important role in the transition of the core from IPv4 to IPv6 being able to continue to support legacy IPv4, VPN-IPv4, and Multicast VPN IPv4 customers.

This document describes the critical use case and OPEX savings of being able to leverage the MP-BGP capability exchange usage as a pure transport allowing both IPv4 and IPv6 to be carried over the same BGP TCP session. By doing so, allows for the elimination of Dual Stacking on the PE-CE connections making the peering IPv6-ONLY to now carry both IPv4 and IPv6 Network Layer Reachability Information (NLRI). This document now provides a solution for IXPs (Internet Exchange points) that are facing IPv4 address depletion at these peering points to use BGP-MP capability exchange defined in [RFC5549] to carry IPv4 (Network Layer Reachability Information) NLRI in an IPv6 next hop using the [RFC5565] softwire mesh framework.

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 https://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 26, 2021.

Table of Contents

1. Introduction 3
2. Requirements Language 6
3. eBGP PE-CE IPv4 and IPv6 NLRI over IPv6 Next Hop Peer Use Case Interop Testing 6
4. RFC 8950 updates to RFC 5549 6
5. Operational Improvements with Single IPv6 transport peer . . 8
6. Operational Considerations 8
7. IANA Considerations 9
8. Security Considerations 9
9. Acknowledgments 9
10. References 9
10.1. Normative References 9
10.2. Informative References 10
Appendix A. IPv4 NLRI IPv6 Next Hop Vendor Testing 12
A.1. Router and Switch Vendors Support and Quality Assurance Engineering Lab Results. 12
A.2. Router and Switch Vendors Interoperability Lab Results. . 12
Authors' Addresses 13

1. Introduction

As Enterprises and Service Providers upgrade their brown field or green field MPLS/SR core to an IPv6 transport such as MPLS LDPv6, SR-MPLSv6 or SRv6, Multiprotocol BGP (MP-BGP)now plays an important role in the transition of the core from IPv4 to IPv6, and being able to continue to support legacy IPv4, VPN-IPv4, and Multicast VPN IPv4 customers.

IXPs (Internet Exchange points) are also facing IPv4 address depletion at their peering points, which are large Layer 2 transit backbones that service providers peer and exchange IPv4 and IPv6 (Network Layer Reachability Information) NLRI. Today these transit exchange points are dual stacked. One proposal to solve this issue is to use [RFC5549] to carry IPv4 (Network Layer Reachability Information) NLRI in an IPv6 next hop and eliminate the IPv4 peering completely using the concept of [RFC5565] softwire mesh framework. So now with the MP-BGP reach capability exchanged over IPv4 AFI over IPv6 next hop peer we can now advertise IPv4(Network Layer Reachability Information) NLRI over IPv6 peering using the [RFC5565] softwire mesh framework.

Multiprotocol BGP (MP-BGP) specifies that the set of usable next-hop address families is determined by the Address Family Identifier (AFI) and the Subsequent Address Family Identifier (SAFI). Historically the AFI/SAFI definitions for the IPv4 address family only have provisions for advertising a Next Hop address that belongs to the IPv4 protocol when advertising IPv4 or VPN-IPv4 Network Layer Reachability Information (NLRI). [RFC5549] specifies the extensions necessary to allow advertising IPv4 NLRI or VPN-IPv4 NLRI with a Next Hop address that belongs to the IPv6 protocol. This comprises an extension of the AFI/SAFI definitions to allow the address of the Next Hop for IPv4 NLRI or VPN-IPv4 NLRI to also belong to the IPv6 Protocol. [RFC5549] defines the encoding of the Next Hop to determine which of the protocols the address actually belongs to, and a new BGP Capability allowing MP-BGP Peers to dynamically discover whether they can exchange IPv4 NLRI and VPN-IPv4 NLRI with an IPv6 Next Hop.

With this new MP-BGP capability exchange allows the BGP peering session to act as a pure transport to allow the session to carry Address Family Identifier (AFI) and the Subsequent Address Family Identifier (SAFI) for both IPv4 and IPv6.

Furthermore, a number of these existing AFI/SAFIs allow the Next Hop to belong to either the IPv4 Network Layer Protocol or the IPv6 Network Layer Protocol, and specify the encoding of the Next Hop information to determine which of the protocols the address actually

belongs to. For example, [RFC4684] allows the Next Hop address to be either IPv4 or IPv6 and states that the Next Hop field address shall be interpreted as an IPv4 address whenever the length of Next Hop address is 4 octets, and as an IPv6 address whenever the length of the Next Hop address is 16 octets.

For example, the AFI/SAFI <25/65> used (as per [RFC6074]) to perform L2VPN auto-discovery, allows advertising NLRI that contains the identifier of a Virtual Private LAN Service (VPLS) instance or that identifies a particular pool of attachment circuits at a given Provider Edge (PE), while the Next Hop field contains the loopback address of a PE. Similarly, the AFI/SAFI <1/132> (defined in [RFC4684]) to advertise Route Target (RT) membership information, allows advertising NLRI that contains such RT membership information, while the Next Hop field contains the address of the advertising router.

There are situations such as those described in [RFC4925] and in [RFC5565] where carriers (or large enterprise networks acting as carrier for their internal resources) may be required to establish connectivity between 'islands' of networks of one address family type across a transit core of a differing address family type. This includes both the case of IPv6 islands across an IPv4 core and the case of IPv4 islands across an IPv6 core. Where Multiprotocol BGP (MP-BGP) is used to advertise the corresponding reachability information, this translates into the requirement for a BGP speaker to advertise Network Layer Reachability Information (NLRI) of a given address family via a Next Hop of a different address family (i.e., IPv6 NLRI with IPv4 Next Hop and IPv4 NLRI with IPv6 Next Hop).

The current AFI/SAFI definitions for the IPv6 address family assume that the Next Hop address belongs to the IPv6 address family type. Specifically, as per [RFC2545] and [RFC8277], when the <AFI/SAFI> is <2/1>, <2/2>, or <2/4>, the Next Hop address is assumed to be of IPv6 type. As per [RFC4659], when the <AFI/SAFI> is <2/128>, the Next Hop address is assumed to be of IPv6-VPN type.

However, [RFC4798] and [RFC4659] specify how an IPv4 address can be encoded inside the Next Hop IPv6 address field when IPv6 NLRI needs to be advertised with an IPv4 Next Hop. [RFC4798] defines how the IPv4-mapped IPv6 address format specified in the IPv6 addressing architecture ([RFC4291]) can be used for that purpose when the <AFI/SAFI> is <2/1>, <2/2>, or <2/4>. [RFC4659] defines how the IPv4-mapped IPv6 address format as well as a null Route Distinguisher can be used for that purpose when the <AFI/SAFI> is <2/128>. Thus, there are existing solutions for the advertisement of IPv6 NLRI with an IPv4 Next Hop.

Similarly, the current AFI/SAFI definitions for advertisement of IPv4 NLRI or VPN-IPv4 NLRI assume that the Next Hop address belongs to the IPv4 address family type. Specifically, as per [RFC4760] and [RFC8277], when the <AFI/SAFI> is <1/1>, <1/2>, or <1/4>, the Next Hop address is assumed to be of IPv4 type. As per [RFC4364], when the <AFI/SAFI> is <1/128>, the Next Hop address is assumed to be of VPN-IPv4 type. As per [RFC6513] and [RFC6514], when the <AFI/SAFI> is <1/129>, the Next Hop address is assumed to be of VPN-IPv4 type. There is clearly no generally applicable method for encoding an IPv6 address inside the IPv4 address field of the Next Hop. Hence, there is currently no specified solution for advertising IPv4 or VPN-IPv4 NLRI with an IPv6 Next Hop.

A new specification for carrying IPv4 Network Layer Reachability Information (NLRI) of a given address family via a Next Hop of a different address family is now defined in [RFC5549], and specifies the extensions necessary to do so. This comprises an extension of the AFI/SAFI definitions to allow the address of the Next Hop for IPv4 NLRI or VPN-IPv4 NLRI to belong to either the IPv4 or the IPv6 protocol, the encoding of the Next Hop information to determine which of the protocols the address actually belongs to, and a new BGP Capability allowing MP-BGP peers to dynamically discover whether they can exchange IPv4 NLRI and VPN- IPv4 NLRI with an IPv6 Next Hop.

With the new extensions defined in [RFC5549] supporting Network Layer Reachability Information (NLRI) and next hop address family mismatch, the BGP peer session can now be treated as a pure transport and carry both IPv4 and IPv6 NLRI at the PE-CE edge over a single IPv6 TCP session. This allows for the elimination of dual stack from the PE-CE peering point, and now allow the peering to be IPv6-ONLY. The elimination of IPv4 on the PE-CE peering points translates into OPEX expenditure savings of point-to-point infrastructure links as well as /31 address space savings and administration and network management of both IPv4 and IPv6 BGP peers. This reduction decreases the number of PE-CE BGP peers by fifty percent, which is a tremendous cost savings for all Enterprises and Service Providers.

While the savings exists at the PE-CE edge, on the core side PE to Route Reflector peering carrying <AFI/SAFI> IPv4 <1/1>, VPN-IPV4 <1/128>, and Multicasat VPN <1/129>, the cost savings nets to a break even to be the same as with an IPV4 Core carrying IPv6 NLRI IPV6 <2/1>, VPN-IPV6 <2/128>, and Multicasat VPN <2/129>. This document also provides a possible solution for IXPs (Internet Exchange points) that are facing IPv4 address depletion at these peering points to use BGP-MP capability exchange defined in [RFC5549] to carry IPv4 (Network Layer Reachability Information) NLRI in an IPv6 next hop using the [RFC5565] softwire mesh framework.

2. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

3. eBGP PE-CE IPv4 and IPv6 NLRI over IPv6 Next Hop Peer Use Case Interop Testing

This particualr use case for external BGP PE-CE edge peering interoperability testing defined in this draft utilizing [RFC8950] next hop encoding to carry both IPv4 and IPv6 NLRI over an IPv6 Next hop encoded peer. Today the IPv4 NLRI and IPv6 NLRI are carried over separate BGP sessions based on the address family of the NLRI being transported. With this drafts use case, the IPv6 NLRI Unicast SAFI along with now the IPv4 NLRI Unicast SAFI, is now being carried by the sinlge transport style IPv6 next hop peer.

This document describes the use case of advertising with IPv4 NLRI over IPv6 Next hop with MP_REACH_NLRI with:

- AFI = 1
- SAFI = 1
- Length of Next Hop Address = 16 or 32
- Next Hop Address = IPv6 address of next hop (potentially followed by the link-local IPv6 address of the next hop). This field is to be constructed as per Section 3 of [RFC2545].

The BGP speaker receiving the advertisement MUST use the Length of Next Hop Address field to determine which network-layer protocol the next hop address belongs to.

Note that this method of using the Length of the Next Hop Address field to determine which network-layer protocol the next hop address belongs to (out of the set of protocols allowed by the AFI/SAFI definition) is the same as used in [RFC4684] and [RFC6074].

4. RFC 8950 updates to RFC 5549

This section describes the updates to [RFC8950] next hop encoding from [RFC5549]. In [RFC5549] when AFI/SAFI 1/128 is used, the next-hop address is encoded as an IPv6 address with a length of 16 or 32 bytes. To accommodate all existing implementations and bring

consistency with VPNv4oIPv4 and VPNv6oIPv6, this document modifies how the next-hop address is encoded. The next-hop address is now encoded as a VPN-IPv6 address with a length of 24 or 48 bytes [RFC8950] (see Sections 3 and 6.2). This change addresses Erratum ID 5253 (Err5253). As all known and deployed implementations are interoperable today and use the new proposed encoding, the change does not break existing interoperability.

[RFC5549] next hop encoding of MP_REACH_NLRI with:

- AFI = 1
- SAFI = 1, 2, or 4
- Length of Next Hop Address = 16 or 32
- Next Hop Address = IPv6 address of next hop (potentially followed by the link-local IPv6 address of the next hop). This field is to be constructed as per Section 3 of [RFC2545].
- NLRI= NLRI as per current AFI/SAFI definition

It also allows advertising with [RFC4760] of an MP_REACH_NLRI with:

- AFI = 1
- SAFI = 128 or 129
- Length of Next Hop Address = 16 or 32
- NLRI= NLRI as per current AFI/SAFI definition

[RFC8950] next hop encoding of MP_REACH_NLRI with:

- AFI = 1
- SAFI = 1, 2, or 4
- Length of Next Hop Address = 16 or 32
- Next Hop Address = IPv6 address of next hop (potentially followed by the link-local IPv6 address of the next hop). This field is to be constructed as per Section 3 of [RFC2545].
- NLRI= NLRI as per current AFI/SAFI definition

It also allows advertising with [RFC4760] of an MP_REACH_NLRI with:

- AFI = 1
- SAFI = 128 or 129
- Length of Next Hop Address = 24 or 48
- Next Hop Address = VPN-IPv6 address of next hop with an 8-octet RD set to zero (potentially followed by the link-local VPN-IPv6 address of the next hop with an 8-octet RD is set to zero).
- NLRI= NLRI as per current AFI/SAFI definition

5. Operational Improvements with Single IPv6 transport peer

As Enterprises and Service Providers migrate their IPv4 core to an MPLS LDPv6 or SRv6 transport, they must continue to be able to support legacy IPv4 customers. With the new extensions defined in [RFC4760], supporting Network Layer Reachability Information (NLRI) and next hop address family mismatch, the BGP peer session can now be treated as a pure transport and carry both IPv4 and IPv6 NLRI at the PE-CE edge. This paves the way to now eliminate dual stacking on all PE-CE peering points to customers making the peering IPv6 only. With this change all IPv4 and IPv6 Network Layer Reachability Information (NLRI) will now be carried over a single BGP session. This also solves the dual stack issue with IXP (Internet Exchange Points) having to maintain separate peering for both IPv4 and IPv6. From an operations perspective the PE-CE edge peering will be drastically simplified with the elimination of IPv4 peers yielding a reduction of peers by 50 percent. From an operations perspective prior to elimination of IPv4 peers an audit is recommended to identify and IPv4 and IPv6 peering incongruencies that may exist and to rectify prior to elimination of the IPv4 peers. No operational impacts or issues are expected with this change.

6. Operational Considerations

With a sinlge IPv6 Peer carrying both IPv4 and IPv6 NLRI there are some operational considerations in terms of what changes and what does not change.

What does not change with a single IPv6 transport peer carrying IPv4 NLRI and IPv6 NLRI below:

Routing Policy configuration is still separate for IPv4 and IPv6 configured by capability as previously

Layer 1, Layer 2 issues such as 1 way fiber or fiber cut will impact both IPv4 and IPv6 as previously.

If the interface is admin down the IPv6 peer would go down and IPv4 NLRI and IPv6 NLRI would be withdrawn as previously.

What does change with a single IPv6 transport peer carrying IPv4 NLRI and IPv6 NLRI below:

Physical interface is no longer dual stacked. Any change in IPv6 address or DAD state will impact both IPv4 and IPv6 NLRI exchange

Single BFD session for both IPv4 and IPv6 NLRI fate sharing as the session is now tied to the transport which now is only IPv6 address family

Both IPv4 and IPv6 peer now exists under the IPv4 address family configuration

Fate sharing of IPv4 and IPv6 address family from a logical perspective now carried over a single IPv6 peer

7. IANA Considerations

There are not any IANA considerations.

8. Security Considerations

The extensions defined in this document allow BGP to propagate reachability information about IPv6 routes over an MPLS IPv4 core network. As such, no new security issues are raised beyond those that already exist in BGP-4 and use of MP-BGP for IPv6. The security features of BGP and corresponding security policy defined in the ISP domain are applicable. For the inter-AS distribution of IPv6 routes according to case (a) of Section 4 of this document, no new security issues are raised beyond those that already exist in the use of eBGP for IPv6 [RFC2545].

9. Acknowledgments

10. References

10.1. 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>.

[RFC2545] Marques, P. and F. Dupont, "Use of BGP-4 Multiprotocol Extensions for IPv6 Inter-Domain Routing", RFC 2545, DOI 10.17487/RFC2545, March 1999, <https://www.rfc-editor.org/info/rfc2545>.

[RFC4291] Hinden, R. and S. Deering, "IP Version 6 Addressing Architecture", RFC 4291, DOI 10.17487/RFC4291, February 2006, <https://www.rfc-editor.org/info/rfc4291>.

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

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <https://www.rfc-editor.org/info/rfc4760>.

[RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, DOI 10.17487/RFC5492, February 2009, <https://www.rfc-editor.org/info/rfc5492>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

[RFC8277] Rosen, E., "Using BGP to Bind MPLS Labels to Address Prefixes", RFC 8277, DOI 10.17487/RFC8277, October 2017, <https://www.rfc-editor.org/info/rfc8277>.

10.2. Informative References

[I-D.ietf-idr-dynamic-cap] Ramachandra, S. and E. Chen, "Dynamic Capability for BGP-4", draft-ietf-idr-dynamic-cap-14 (work in progress), December 2011.

[RFC4659] De Clercq, J., Ooms, D., Carugi, M., and F. Le Faucheur, "BGP-MPLS IP Virtual Private Network (VPN) Extension for IPv6 VPN", RFC 4659, DOI 10.17487/RFC4659, September 2006, <https://www.rfc-editor.org/info/rfc4659>.

[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, <https://www.rfc-editor.org/info/rfc4684>.

[RFC4798] De Clercq, J., Ooms, D., Prevost, S., and F. Le Faucheur, "Connecting IPv6 Islands over IPv4 MPLS Using IPv6 Provider Edge Routers (6PE)", RFC 4798, DOI 10.17487/RFC4798, February 2007, <https://www.rfc-editor.org/info/rfc4798>.

[RFC4925] Li, X., Ed., Dawkins, S., Ed., Ward, D., Ed., and A. Durand, Ed., "Softwire Problem Statement", RFC 4925, DOI 10.17487/RFC4925, July 2007, <https://www.rfc-editor.org/info/rfc4925>.

[RFC5549] Le Faucheur, F. and E. Rosen, "Advertising IPv4 Network Layer Reachability Information with an IPv6 Next Hop", RFC 5549, DOI 10.17487/RFC5549, May 2009, <https://www.rfc-editor.org/info/rfc5549>.

[RFC5565] Wu, J., Cui, Y., Metz, C., and E. Rosen, "Softwire Mesh Framework", RFC 5565, DOI 10.17487/RFC5565, June 2009, <https://www.rfc-editor.org/info/rfc5565>.

[RFC6074] Rosen, E., Davie, B., Radoaca, V., and W. Luo, "Provisioning, Auto-Discovery, and Signaling in Layer 2 Virtual Private Networks (L2VPNs)", RFC 6074, DOI 10.17487/RFC6074, January 2011, <https://www.rfc-editor.org/info/rfc6074>.

[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>.

[RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 8126, DOI 10.17487/RFC8126, June 2017, <https://www.rfc-editor.org/info/rfc8126>.

[RFC8950] Litkowski, S., Agrawal, S., Ananthamurthy, K., and K. Patel, "Advertising IPv4 Network Layer Reachability Information (NLRI) with an IPv6 Next Hop", RFC 8950, DOI 10.17487/RFC8950, November 2020, <https://www.rfc-editor.org/info/rfc8950>.

Appendix A. IPv4 NLRI IPv6 Next Hop Vendor Testing

IPv4 NLRI with IPv6 Next Hop encoding is supported for all BGP peers both iBGP and eBGP.

This section details the vendor support QA testing of RFC 8950 Next Hop Encoding for "PE-CE eBGP" using GUA (Global Unicast Address), Link Local (LL) peering. This drafts goal is to first ensure that QA testing of all features and functionality works with "eBGP PE-CE" use case single peer carrying both IPv4 NLRI and IPv6 NLRI and that the routing policy features are all still fully functionality do not change.

A.1. Router and Switch Vendors Support and Quality Assurance Engineering Lab Results.

Vendor	PE-CE eBGP GUI	PE-CE eBGP LL	QA Tested
Cisco	***		
Juniper	***		
Nokia/ALU	***		
Arista	***		
Huawei	***		

Table 1: Vendor Support

A.2. Router and Switch Vendors Interoperability Lab Results.

This section details the vendor interoperability testing and support of RFC5549 that all features and functionality works with "eBGP PE-CE" use case with having a single peer carrying both IPv4 NLRI and IPv6 NLRI and that the routing policy features are fully tested for quality assurance.

Vendor	Cisco	Juniper	Nokia/ALU	Arista	Huawei
Cisco	N/A				
Juniper		N/A			
Nokia/ALU			N/A		
Arista				N/A	
Huawei					N/A

Table 2: Vendor Interop

Authors' Addresses

Gyan Mishra
Verizon Inc.

Email: gyan.s.mishra@verizon.com

Mankamana Mishra
Cisco Systems
821 Alder Drive,
MILPITAS CALIFORNIA 95035

Email: mankamis@cisco.com

Jeff Tantsura
Apstra, Inc.

Email: jefftant.ietf@gmail.com

Lili Wang
Juniper Networks, Inc.
10 Technology Park Drive,
Westford MA 01886
US

Email: liliw@juniper.net

Qing Yang
Arista Networks

Email: qyang@arista.com

Adam Simpson
Nokia

Email: adam.1.simpson@nokia.com

Shuanglong Chen
Huawei Technologies

Email: chenshuanglong@huawei.com

BESS Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 13, 2021

W. Wang
A. Wang
China Telecom
H. Wang
Huawei Technologies

March 12, 2021

Layer-3 Accessible EVPN Services
draft-wang-bess-l3-accessible-evpn-04

Abstract

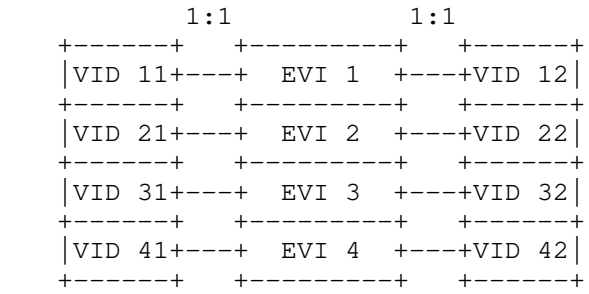
This draft describes layer-3 accessible EVPN service interfaces according to [RFC7432], and proposes a new solution which can simplify the deployment of layer-3 accessible EVPN service. This solution allows each PE in EVPN network to maintain only one IP-VRF.

Table of Contents

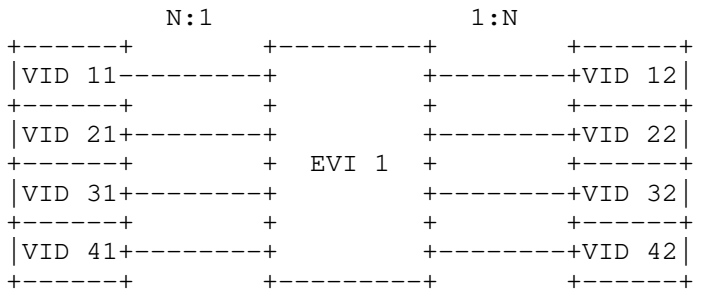
1. Introduction 2
2. Conventions used in this document 4
3. Terminology 4
4. Service Interfaces in layer-3 accessible EVPN 5
5. Solutions of LSI-aware bundle service interface 6
6. Protocol Extensions 8
6.1. Forwarding Plane 8
6.1.1. Extensions to VxLAN 8
6.2. Control Plane 8
7. Security Considerations 9
8. IANA Considerations 9
9. Normative References 9
Authors' Addresses 10

1. Introduction

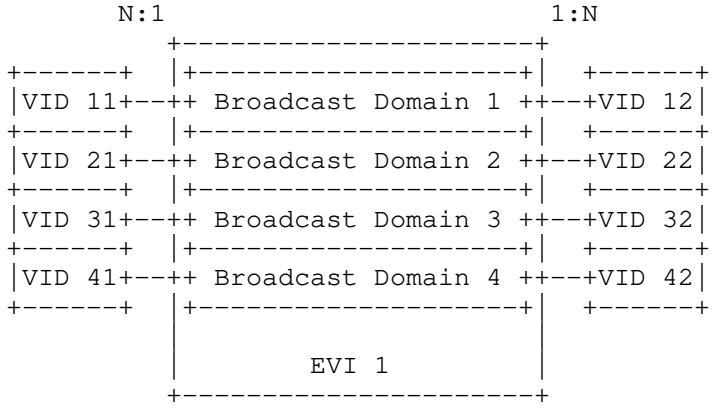
[RFC7432]defines three service interfaces for layer-2 accessible EVPN: VLAN-Based Service Interface, VLAN-Bundle Service Interface and VLAN-Aware Bundle Service Interface. These three types of service interfaces can realize the isolation of layer-2 traffic of customers in different ways, as shown in Figure 1.



VLAN-based Service Interface



VLAN-bundle Service Interface



VLAN-Aware Bundle Service Interface

Figure 1: EVPN Service Interfaces Overview

For VLAN-based service interface, there is a one to one mapping between VID and EVI. Each EVI has a single broadcast domain so that traffic from different customers can be isolated.

For VLAN-bundle service interface, there is a N to one mapping between VID and EVI. Each EVI has a single broadcast domain, but the MAC address MUST be unique that can be used for customer traffic isolation.

For VLAN-aware bundle service interface, there is a N to one mapping between VID and EVI. Each EVI has multiple broadcast domains while the MAC address can overlap. One broadcast domain corresponds to one VID, which can be used to customer traffic isolation.

In the scenarios corresponding to these service interfaces, CE-PE should be placed in the same Layer-2 network. In most of provider network, CE-PE need to cross a Layer-3 network, then the above service interfaces should be extended to adapt to the layer-3 network.

In this draft, we describe three layer-3 accessible interfaces for EVPN, summarize the existing layer-3 accessible EVPN solutions, and propose a new solution which can simplify the depolyment of layer-3 accessible EVPN service.

2. 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] .

3. Terminology

The following terms are defined in this draft:

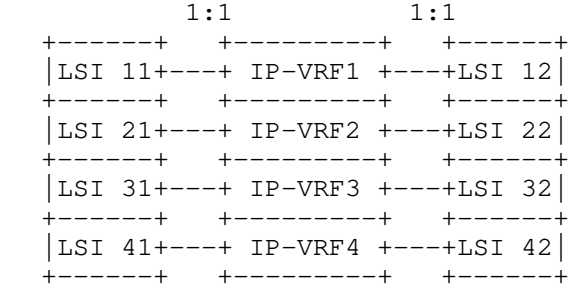
- CE: Client Edge
- PE: Provider Edge
- EVPN: BGP/MPLS Ethernet VPN, defined in [RFC7432]
- VxLAN: Virtual eXtensible Local Area Network, defined in [RFC7348]
- IPSec: Internet Protocol Security, defined in [RFC4301]

4. Service Interfaces in layer-3 accessible EVPN

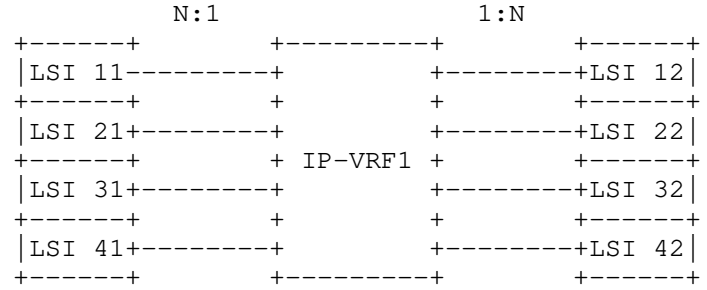
In most of provider network, CE-PE need to cross a Layer-3 network. With this scenario, service interfaces defined in [RFC7432] should be extended to adapt to the layer-3 network. To achieve the traffic isolation, tunnel encapsulation technologies can be used.

We define Logical Session Identifier(LSI) to distinguish the packets from different tunnels, which is related to VNI/SPI. The length of LSI is 16 bits.

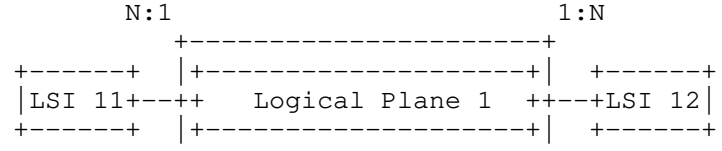
The layer-3 accessible interfaces for EVPN are shown in Figure 2, refer to [RFC7432]



LSI-based Service Interface



LSI-bundle Service Interface



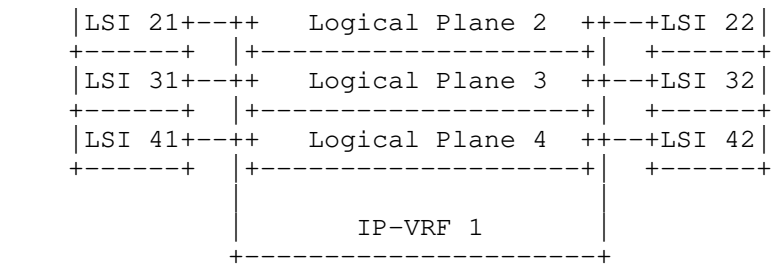


Figure 2: Layer-3 accessible EVPN Service Interfaces Overview

For LSI-based service interface, there is a one to one mapping between LSI and IP-VRF. Each IP-VRF has a single logical plane so that traffic from different customers can be isolated.

For LSI-bundle service interface, there is a N to one mapping between LSI and IP-VRF. Each IP-VRF has a single logical plane, but the IP address MUST be unique that can be used for customer traffic isolation.

For LSI-aware bundle service interface, there is a N to one mapping between LSI and IP-VRF. Each IP-VRF has multiple logical planes while the IP address can overlap. One logical plane corresponds to one LSI, which can be used to customer traffic isolation.

5. Solutions of LSI-aware bundle service interface

Let’s assume a scenario as shown in Figure 3. PE1, PE2 and PE3 are EVPN peers, the customer data transmission between PEs relies on VxLAN. CE1, CE2 and CE3 are connected to the sites of customer for its department A and B.

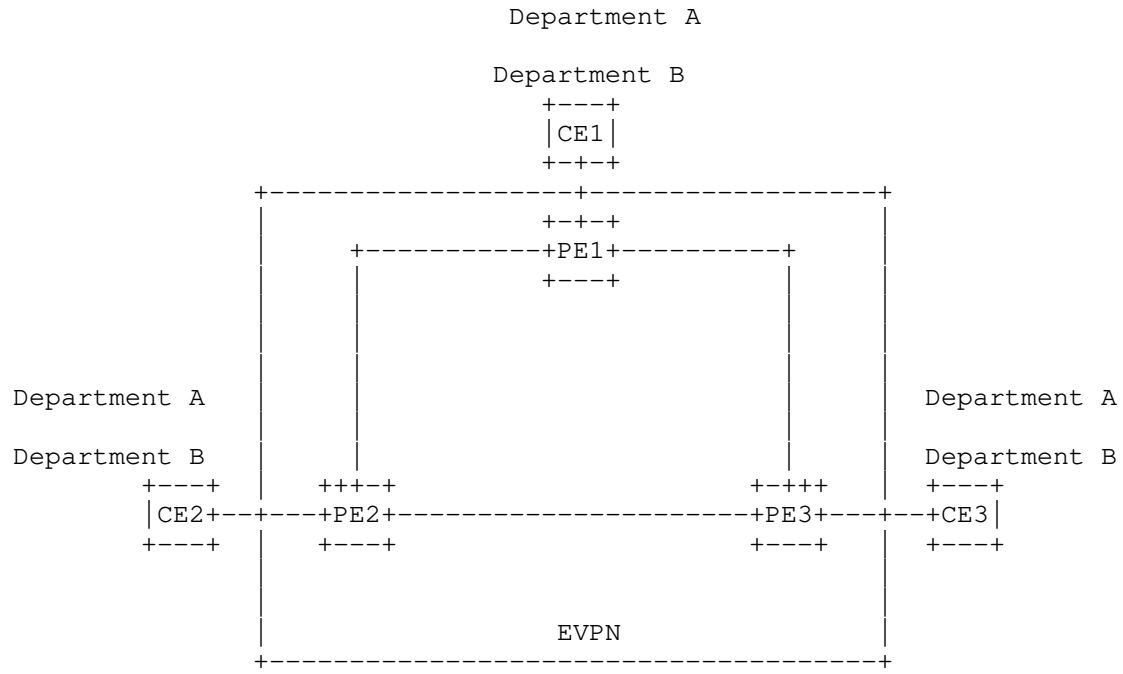


Figure 3: LSI-aware bundle service interface scenario

If each VNI has its own IP-VRF, each PE and CE maintain an IP-VRF for each deployment. In this situation, customer traffic can be isolated by different VNIs, and there is no need for extending control plane/ forwarding plane protocols.

For deployment, we expect a simpler way, such as assign an IP-VRF to each customer, not to each department. That is to say, all VNIs share one IP-VRF on PEs. In this situation, each CE still maintain an IP-VRF for each deployment, but each PE maintains only one VRF for all deployments. In this situation, customer traffic cannot be isolated by VNIs. We propose a solution for this scenario:

- Using LSI information to identify different customer routes / traffic. As described above, LSI can be generated by VNI/SPI, and there is a one to one mapping between LSI and VNI/SPI. PEs should maintain the mapping table of LSI and VNI/SPI, so that they can distinguish different customer routes / traffic. LSI information can be transmitted by using Ethernet Tag ID or a newly defined ESI type.

o TBD (more solutions are welcome).

6. Protocol Extensions

6.1. Forwarding Plane

6.1.1. Extensions to VxLAN

When the forwarding plane uses VxLAN tunnel technologes, we should extend the VxLAN GPE header to carry the LSI information, the extentions to the VxLAN GPE header is shown in Figure 4:

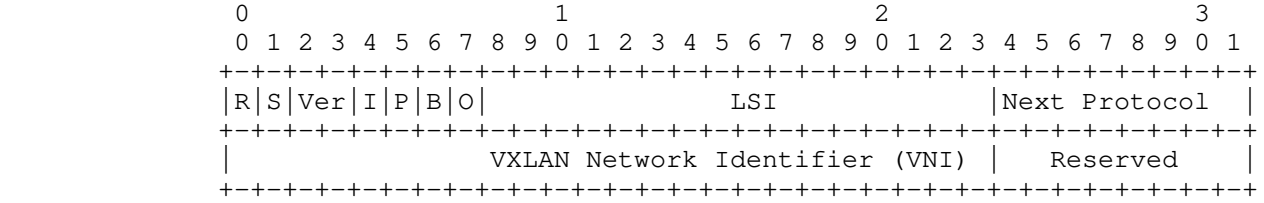


Figure 4: The extentions to VxLAN GPE header

We define a S bit. If S is set to 1, it means the field after O bit contains LSI information.

6.2. Control Plane

We proposed two methods to identify the routes that related to different LSI information:

o Reusing the Ethernet Tag ID. This method requires the update of [I-D.ietf-bess-evpn-prefix-advertisement] (Etherenet Tag ID is set to 0 for route type 5), and may arises some confuse with the original defination of Ethernet Tag ID.

o Using the newly defined ESI type as shown in Figure 5. This method can preserve the original purpose of ESI defination (multi-homing).

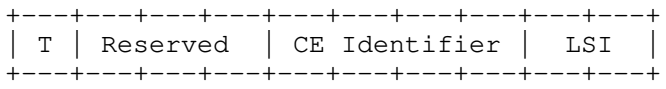


Figure 5: The format of new ESI type

Where:

- T (1 octet): specifys the ESI Type. The recommended value is 0x06.
- CE Identifier (3 octets): the route ID/IPv4 address of CE.
- LSI (2 octets): the LSI information.

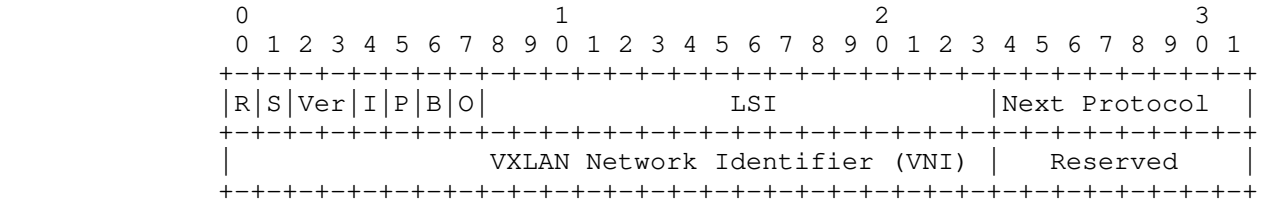
Since the length of LSI is 16 bits, while the length of Ethernet Tag ID and ESI are 80 bits and 32 bits, respectively. We can only use the lower 16 bits of Ethernet Tag ID / ESI field to carry LSI information, the other locations MUST set to 0.

7. Security Considerations

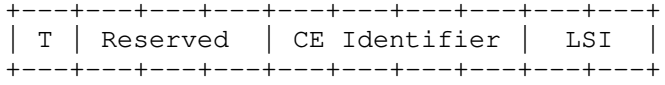
TBD

8. IANA Considerations

This draft extends the VxLAN GPE header, S bit of Flag and LSI field are added:



This draft also define a new ESI type:



9. Normative References

[I-D.ietf-bess-evpn-prefix-advertisement]
Rabadan, J., Henderickx, W., Drake, J., Lin, W., and A. Sajassi, "IP Prefix Advertisement in EVPN", draft-ietf-bess-evpn-prefix-advertisement-11 (work in progress), May 2018.

[I-D.ietf-bess-mvpn-evpn-aggregation-label]
Zhang, Z., Rosen, E., Lin, W., Li, Z., and I. Wijnands, "MVPN/EVPN Tunnel Aggregation with Common Labels", draft-ietf-bess-mvpn-evpn-aggregation-label-05 (work in progress), January 2021.

[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>.

[RFC2890] Dommety, G., "Key and Sequence Number Extensions to GRE", RFC 2890, DOI 10.17487/RFC2890, September 2000, <https://www.rfc-editor.org/info/rfc2890>.

[RFC4301] Kent, S. and K. Seo, "Security Architecture for the Internet Protocol", RFC 4301, DOI 10.17487/RFC4301, December 2005, <https://www.rfc-editor.org/info/rfc4301>.

[RFC7348] Mahalingam, M., Dutt, D., Duda, K., Agarwal, P., Kreeger, L., Sridhar, T., Bursell, M., and C. Wright, "Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", RFC 7348, DOI 10.17487/RFC7348, August 2014, <https://www.rfc-editor.org/info/rfc7348>.

[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>.

Authors' Addresses

Wei Wang
China Telecom
Beiqijia Town, Changping District
Beijing, Beijing 102209
China

Email: weiwang94@foxmail.com

Aijun Wang
China Telecom
Beiqijia Town, Changping District
Beijing, Beijing 102209
China

Email: wangaj3@chinatelecom.cn

Haibo Wang
Huawei Technologies
Huawei Building, No.156 Beiqing Rd.
Beijing, Beijing 100095
China

Email: rainsword.wang@huawei.com