Network Working Group
Internet Draft
Intended status: Informational
Expires: January 13, 2021

L. Dunbar
J. Guichard
Futurewei
Ali Sajassi
Cisco
J. Drake
Juniper
B. Najem
Bell Canada
Ayan Barnerjee
D. Carrel
Cisco

July 13, 2020

BGP Usage for SDWAN Overlay Networks
draft-dunbar-bess-bgp-sdwan-usage-08

Abstract

The document describes three distinct SDWAN scenarios and discusses the applicability of BGP for each of those scenarios. The goal of the document is to demonstrate how BGP-based control plane is used for large scale SDWAN overlay networks with little manual intervention.

SDWAN edge nodes are commonly interconnected by multiple underlay networks which can be owned and managed by different network providers.

Copyright Notice

Table of Contents

1. Introduction

Here are some key characteristics of "SDWAN" networks:

- Augment of transport, which refers to utilizing overlay paths over different underlay networks. Very often there are multiple parallel overlay paths between any two SDWAN edges, some of which are private networks over which traffic can traverse with or without encryption, others require encryption, e.g. over untrusted public networks.
- Enable direct Internet access from remote sites, instead hauling all traffic to Corporate HQ for centralized policy control.
- Some traffic are routed based on application IDs instead of based on destination IP addresses.
- The Application Routing can also be based on specific performance criteria (e.g. packets delay, packet loos, jitter) to provide better application performance by choosing the right underlay that meets or exceeds the specified criteria.

[Net2Cloud-Problem] describes the network related problems that enterprises face to connect enterprises' branch offices to dynamic workloads in different Cloud DCs, including using SDWAN to aggregate

multiple paths provided by different service providers to achieve better performance and to accomplish application ID based forwarding.

Even though SDWAN has been positioned as a flexible way to reach dynamic workloads in third party Cloud data centers over different underlay networks, scaling becomes a major issue when there are hundreds or thousands of nodes to be interconnected by an SDWAN overlay networks.

BGP is widely used by underlay networks. This document describes using BGP for edge nodes to exchange information across the SDWAN overlay networks.

2. Conventions used in this document

Cloud DC: Third party data centers that host applications and workloads owned by different organizations or tenants.

Controller: Used interchangeably with SDWAN controller to manage SDWAN overlay path creation/deletion and monitor the path conditions between sites.

CPE: Customer Premise Equipment

CPE-Based VPN: Virtual Private Secure network formed among CPEs. This is to differentiate from more commonly used PE-based VPNs [RFC 4364].

Homogeneous SDWAN: A type of SDWAN network in which all traffic to/from the SDWAN edge nodes has to be encrypted regardless of underlay networks. For lack of better terminology, we call this Homogeneous SDWAN throughout this document.

ISP: Internet Service Provider

NSP: Network Service Provider. NSP usually provides more advanced network services, such as MPLS VPN, private leased lines, or managed Secure WAN connections, many

times within a private trusted domain, whereas an ISP usually provides plain internet services over public untrusted domains.

PE: Provider Edge

SDWAN Edge Node: an edge node, which can be physical or virtual, maps the attached clients' traffic to the wide area network (WAN) overlay tunnels.

SDWAN: Software Defined Wide Area Network. In this document, "SDWAN" refers to the solutions of pooling WAN bandwidth from multiple underlay networks to get better WAN bandwidth management, visibility & control. When the underlay networks are private, traffic can traverse without additional encryption; when the underlay networks are public, such as the Internet, some traffic may need to be encrypted when traversing through (depending on user provided policies).

SDWAN IPsec SA: IPsec Security Association between two SDWAN ports or nodes.

SDWAN over Hybrid Networks: SDWAN over Hybrid Networks typically have edge nodes utilizing bandwidth resources from multiple service providers. In Hybrid SDWAN network, packets over private networks can go natively without encryption and are encrypted over the untrusted network, such as the public Internet.

WAN Port: A Port or Interface facing an ISP or Network Service Provider (NSP), with address (usually public routable address) allocated by the ISP or the NSP.

C-PE: SDWAN Edge node, which can be CPE for customer managed SDWAN, or PE that is for provider managed SDWAN services).

ZTP: Zero Touch Provisioning

3. Use Case Scenario Description and Requirements

SDWAN networks can have different topologies and have different traffic patterns. To make it easier for the focused discussion in subsequent drafts on SDWAN control plane and data plane, this section describes several SDWAN scenarios that may have different impact on their corresponding control planes & data planes.

3.1. Requirements

3.1.1. Supporting Multiple SDWAN Segmentations

The term "network segmentation", a.k.a. SDWAN instances, is referring to the process of dividing the network into logical sub-networks using isolation techniques on a forwarding device such as a switch, router, or firewall. For a homogeneous network, such as MPLS VPN or Layer 2 network, VRF or VLAN are used to achieve the network segmentation.

As SDWAN is an overlay network arching over multiple types of networks, MPLS L2VPN/L3VPN or pure L2 underlay can continue using the VRF, VN-ID or VLAN to differentiate SDWAN network segmentations. For public internet, the IPsec inner encapsulation header can carry the SDWAN Instance Identifier to differentiate the packets belonging to different SDWAN instances.

BGP already has the capability to differentiate control packets for different network instances. When using BGP for SDWAN, the SDWAN segmentations can be differentiated by the SDWAN Target ID in the BGP Extended Community in the same way as VPN instances being represented by the Route Target. Same as Route Target, need to use a different name to differentiate from VPN if a CPE supports traditional VPN with multiple VRFs and supports multiple SDWAN Segmentations (instances). The actual SDWAN Target ID encoding is proposed by [SDWAN-EDGE-Discovery].

3.1.2. Client Service Requirement

Client interface of SDWAN nodes can be IP or Ethernet based.

For Ethernet based client interfaces, SDWAN edge should support VLAN-based service interfaces (EVI100), VLAN bundle service interfaces (EVI200), or VLAN-Aware bundling service interfaces. EVPN service requirements are applicable to the Client traffic, as described in the Section 3.1 of RFC8388.

For IP based client interfaces, L3VPN service requirements are applicable.

3.1.3. Application Flow Based Segmentation

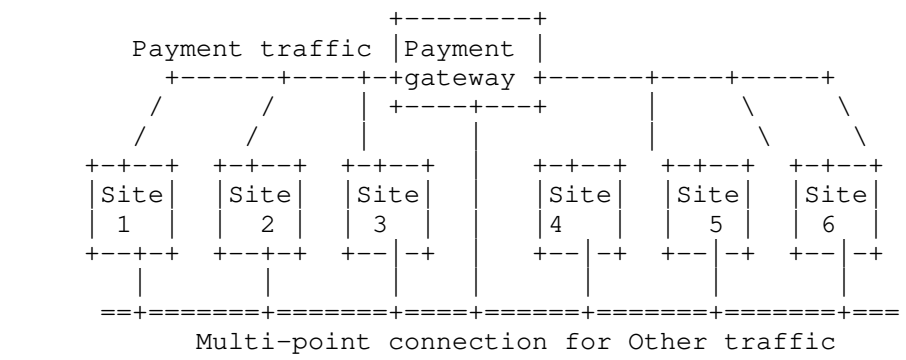
Application Flow based Segmentation, also known as SDWAN Traffic Segmentation, enables the separation of the traffic based on the business and the security needs for different users' groups and/or application requirements. Each user group and/or applications may need different isolated topology and/or policies to fulfill the business requirements.

The Application Flow based Segmentation concept is analogous to VLAN (in L2 network) and VRF (in L3 network).

One can think about the Application Flow based Segmentation as a feature that can be provided or enabled on a single SDWAN service (or domain) to a single Subscriber. Each SDWAN Service can have one or more overlay Segments to support the business requirement; each Segment has its own policy, topology and application/user groups. Applications/users' group can belong to more than one Segment.

For example, a retail business requires the point-of-sales (PoS) application in all stores to be isolated from other applications AND routed only to the payment processing entity at a hub site (i.e. hub and spoke); however, the same retail business requires the other applications to be routed to all sites (i.e. multipoint-to multipoint) AND isolated from the PoS application.

In the figure below, the traffic from the PoS application follows a Tree topology, whereas other traffic can be multipoint-to-multipoint topology.



Another example is an enterprise who wants to isolate the traffic for each department and have different topology and policy for different department; the HR department may need to access certain applications that are NOT accessible by the engineering department. In addition, the contractors may have a limited access to the enterprise resources.

3.1.4. Zero Touch Provisioning

Unlike traditional EVPN or L3VPN whose PEs are deployed for long term, SDWAN edge nodes (virtual or physical) deployment at a specific location can be ephemeral. Therefore, Zero Touch Provisioning (ZTP), or Plug and Play, is a common requirement for SDWAN. When an SDWAN edge is physically installed at a location or instantiated on a VM in a Cloud DC, ZTP automates follow-up steps, including updates to the OS, software version, and configuration prior to connection. From network control perspective, ZTP includes the following:

- Upon power up, an SDWAN node can establish transport layer secure connection (such as TLS, SSL, etc.) to its controller whose address can be burned or preconfigured on the device.

- The SDWAN Controller can designate a Local Network Controller in the proximity of the SDWAN node; the Local Network Controller manages and monitor the communication policies of the edge node.

3.1.5. Constrained Propagation of SDWAN Edge Properties

One SDWAN edge node may only be authorized to communicate with a small number of other SDWAN edge nodes. Under this circumstance, the property of the SDWAN edge node cannot be propagated to any other nodes who are not authorized to communicate. But a remote SDWAN edge node upon powering up might not have the proper policies to know who the authorized peers are. Therefore, it is very essential for SDWAN deployment have a central point to distribute the properties of each SDWAN edge node to its authorized peers.

BGP is well suited for this purpose. RFC 4684 has specified the procedure to constrain the distribution of BGP UPDATE to only a subset of SDWAN edges. Basically, each edge node informs the Route Reflector (RR) [RFC4456] on its interested SDWAN instances. The RR only propagates the BGP UPDATE for the relevant SDWAN instances to the edge.

Usually the connection between a SDWAN edge node and its RR is over insecure network. Therefore, upon power up, a SDWAN node needs to establish a secure transport layer connection (TLS, SSL, etc.) to its designated RR. The BGP UPDATE messages need to be sent over the secure channel (TLS, SSL, etc.) to the RR.

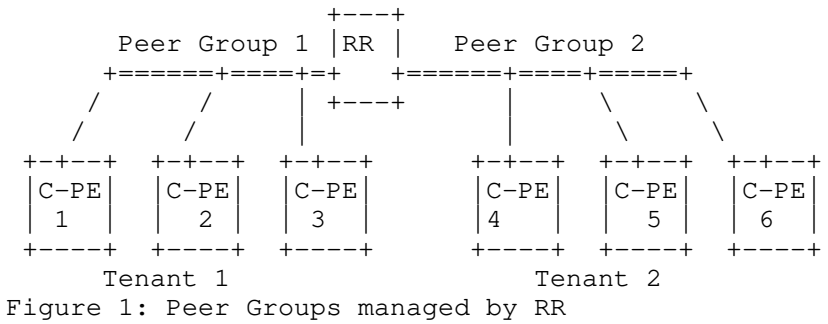


Figure 1: Peer Groups managed by RR

Tenant separation is achieved by the SDWAN instance identification represented in control plane and data plane, respectively.

3.2. Scenarios #1: Homogeneous WAN

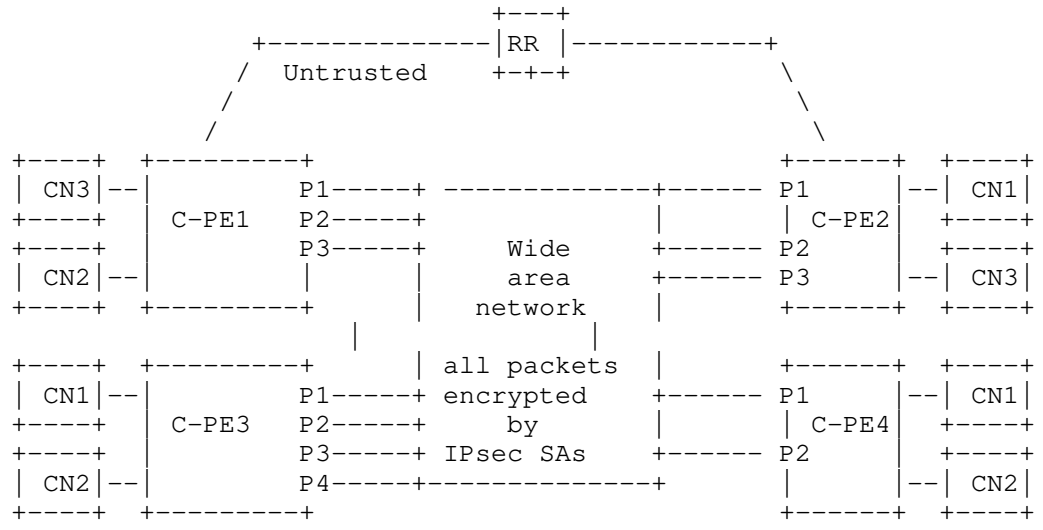
This is referring to a type of SDWAN network with edge nodes encrypting all traffic over WAN to other edge nodes, regardless of whether the underlay is private or public. For lack of better terminology, we call this Homogeneous SDWAN throughout this document.

Some typical scenarios for the use of a Homogeneous SDWAN network are as follows:

- A small branch office connecting to its HQ offices via the Internet. All sensitive traffic to/from this small branch office has to be encrypted, which is usually achieved using IPsec SAs.

- A store in a shopping mall may need to securely connect to its applications in one or more Cloud DCs via the Internet. A common way of achieving this is to establish IPsec SAs to the Cloud DC gateway to carry the sensitive data to/from the store.

As described in [SECURE-EVPN], the granularity of the IPsec SAs for Homogeneous SDWAN can be per site, per subnet, per tenant, or per address. Once the IPsec SA is established for a specific subnet/tenant/site, all traffic to/from the subnets/tenants/site are encrypted.



CN: Client Networks, which is same as Tenant Networks used by NVo3

Figure 2: Homogeneous SDWAN

One of the key properties of homogeneous SDWAN is that the SDWAN Local Network Controller (RR)is connected to C-PEs via untrusted public network, therefore, requiring secure connection between RR and C-PEs (TLS, DTLS, etc.).

Homogeneous SDWAN has some similarity to commonly deployed IPsec VPN, albeit the IPsec VPN is usually point-to-point among a small number of nodes and with heavy manual configuration for IPsec between nodes, whereas an SDWAN network can have a large number of edge nodes with an SDWAN controller to manage requiring zero touch provisioning upon powering up.

Existing Private VPNs (e.g. MPLS based) can use homogeneous SDWAN to extend over public network to remote sites to which the VPN operator does not own or lease infrastructural connectivity, as described in [SECURE-EVPN] and [SECURE-L3VPN]

3.3. Scenario #2: CPE based SDWAN over Hybrid WAN Underlay

In this scenario, SDWAN edge nodes (a.k.a. C-PEs) have some WAN ports connected to PEs of Private VPNs over which packets can be forwarded natively without encryption, and some WAN ports connected to the public Internet over which sensitive traffic have to be encrypted (usually by IPsec SA).

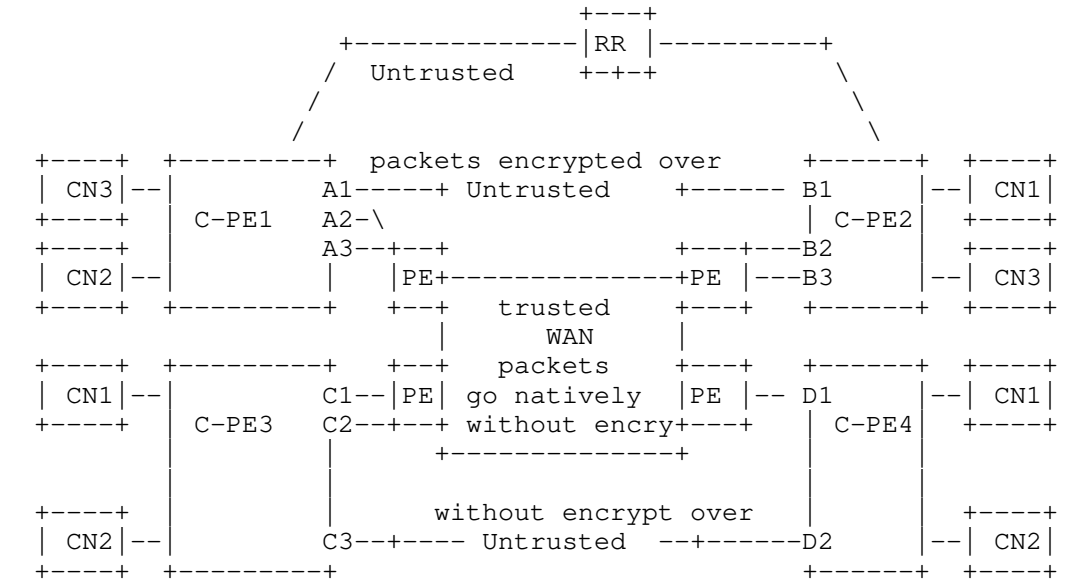
In this scenario, the SDWAN edge nodes' egress WAN ports are all IP/Ethernet based, either egress to PEs of the VPNs or egress to the public Internet. Even if the VPN is a MPLS network, the VPN's PEs have IP/Ethernet links to the SDWAN edge (C-PEs). Throughout this document, this scenario is also called CPE based SDWAN over Hybrid Networks.

Even though IPsec SA can secure the packets traversing the Internet, it does not offer the premium SLA commonly offered by Private VPNs, especially over long distance. Clients need to have policies to specify criteria for flows only traversing private VPNs or traversing either as long as encrypted when over the Internet. For example, client can have those polices for the flows:

1. A policy or criteria for sending the flows over a private network without encryption (for better performance),
2. A policy or criteria for sending the flows over any networks as long as the packets of the flows are encrypted when traversing untrusted networks, or
3. A policy of not needing encryption at all.

If a flow traversing multiple segments, such as A<->B<->C<->D, has either Policy 2 or 3 above, the flow can traverse different underlays in different network segments, such as over Private network underlay between A<->B without encryption, or over the public internet between B<->C in an IPsec SA.

As shown in the figure below, C-PE-1 has two different types of interfaces (A1 to Internet and A2 & A3 to VPN). The C-PEs' loopback addresses and addresses attached to C-PEs may or may not be visible to the ISPs/NSPs. The addresses for the WAN ports can have addresses allocated by service providers or dynamically assigned (e.g. by DHCP). One WAN port shown in the figure below (e.g. A1, A2, A3 etc.) is a logical representation of potential multiple physical ports on the C-PEs.



CN: Client Network

Figure 3: Hybrid SDWAN

Some key characteristics of a Hybrid SDWAN overlay network are as follows:

- one C-PE may be connected to different ISPs/NSPs, with some of its WAN ports addresses being assigned by different ISPs/NSPs.

- The WAN ports connected to PEs of trusted private networks (e.g. MPLS VPN) hand off IP/Ethernet packets, just like today's CPE that do not handle MPLS packets and do not participate in the underlay VPN networks' control plane. Traffic can flow natively without encryption when be forwarded out through those WAN ports for better performance.

- The WAN ports connected to untrusted networks, e.g. the Internet, requires sensitive traffic to be encrypted, i.e. encrypted by IPsec SA.

- An SDWAN local Network Controller (RR) is connected to C-PEs via the untrusted public network, therefore, requiring secure connection between RR and C-PEs via TLS, DTLS, etc.

- The SDWAN nodes' [loopback] addresses might not be routable nor visible in the underlay ISP/NSP networks. Routes & services attached to SDWAN edges at the SDWAN overlay layer are in different address spaces than the underlay networks.

- There could be multiple SDWAN devices sharing a common property, such as a geographic location. Some applications over SDWAN may need to traverse specific geographic locations for various reasons, such as to comply with regulatory rules, to utilize specific value added services, or others.

- The underlay path selection between sites can be a local decision. Some policies allow one service from C-PE1 -> C-PE2 -> C-PE3 using one ISP/NSP underlay in the first segment (C-PE1 -> C-PE2) and using a different ISP/NSP in the second segment (C-PE2-> CPE3).

- Services may not be congruent, i.e. the packets from A-> B may traverse one underlay network, and the packets from B -> A may traverse a different underlay.

- Different services, routes, or VLANs attached to SDWAN nodes can be aggregated over one underlay path; same service/routes/VLAN can spread over multiple SDWAN underlays at different times depending on the policies specified for the service. For example, one tenant's packets to HQ need to be encrypted when sent over the Internet or have to be sent over private networks, while the same

tenant's packets to Facebook can be sent over the Internet without encryption.

3.4. Scenario #3: Private VPN PE based SDWAN

This scenario refers to existing VPN (e.g. MPLS based VPN, such as EVPN or IPVPN) adding extra ports facing untrusted public networks allowing PEs to offload some low priority traffic to ports facing public networks when the VPN MPLS paths are congested. Throughout this document, this scenario is also called Internet Offload for Private VPN, or PE based SDWAN.

In this scenario, the packets offloaded to untrusted public network must be encrypted.

PE based SDWAN can be used by VPN service providers to temporarily increase bandwidth between sites when they are not sure if the demand will sustain for long period of time or as a temporary solution before the permanent infrastructure is built or leased.

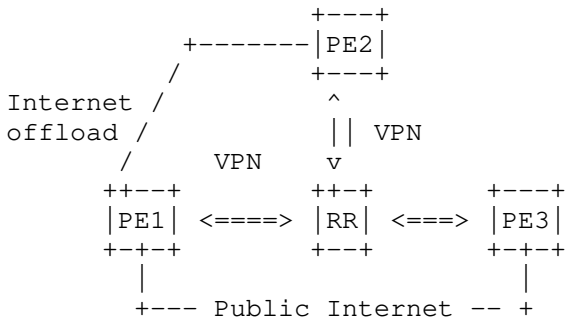


Figure 4: Additional Internet paths added to the VPN

Here are some key properties for PE based SDWAN:

- For MPLS based VPN, PEs continue having MPLS encapsulation handoff to existing paths.

- The BGP RR is connected to PEs in the same way as VPN, i.e. via the trusted network.
- For the added Internet ports, PEs have IP packets handoff, i.e. sending and receiving IP data frames. Internally, PEs can have the option to encapsulate the MPLS payload in IP, as specified by RFC4023.
- The ports facing public internet might get IP addresses assigned by ISPs, which may not be in the same address domain as PEs'.
- Ports facing public internet are not as secure as the ports facing private infrastructure. There could be spoofing, or DDOS attacks to the ports facing public internet. Extra consideration must be given when injecting the new routes learned from public network into VRFs.
- Even though packets are encrypted over public internet, the performance SLA is not guaranteed over public internet. Therefore, clients may have policies only allowing some flows to be offloaded to internet path.

4. BGP Walk Through

4.1. BGP Walk Through for Homogeneous SDWAN

In the figure below, packets destined towards multiple routes attached to the C-PE2 can be carried by one IPsec tunnel. Then one BGP UPDATE can be announced by C-PE2 to its RR.

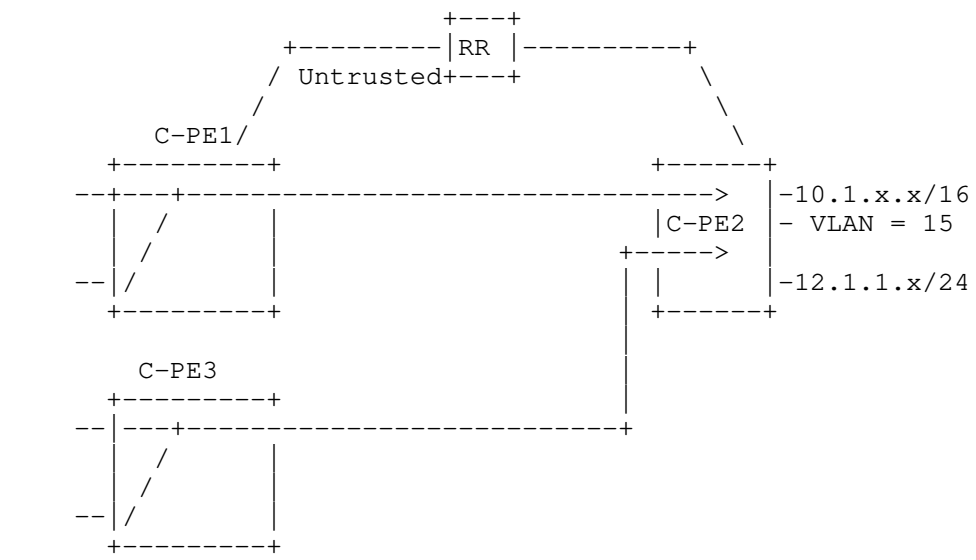


Figure 5: Homogeneous SDWAN

The BGP UPDATE Message from C-PE2 to RR should have the client routes encoded in the MP-NLRI Path Attribute and the IPsec Tunnel associated information encoded in the Tunnel-Encap Path Attributes as described in the [SECURE-EVPN]:

- MP-NLRI Path Attribute: to indicate multiple routes attached to the C-PE2:
  10.1.x.x/16
  VLAN #15
  12.1.1.x/24
- Tunnel-Encap Path Attribute: to describe the IPsec attributes for routes encoded in the NLRI Path Attribute:
  IPsec attributes for remote nodes to establish the IPsec tunnel to C-PE2.

If different client routes attached to C-PE2 needs to be reached by separate IPsec tunnels, then multiple BGP UPDATE messages need to be sent to the remote nodes via RR. If C-PE2 doesn't have the policy on authorized peers for the specific client routes, RR needs to check the client routes policies to propagate the BGP UPDATE messages to the remote authorized edge nodes.

There could be policies governing the topologies of a client's different routes attached to an edge node. For example, VLAN #25 and

route 22.1.1.x/24 could be the Payment Applications described in the Section 3.1.2 that can only communicate with Payment Gateway attached to C-PE3. If C-PEs don't have the policy to govern the communication peers, RR can take over the responsibility of only send BGP UPDATE to the authorized peers.

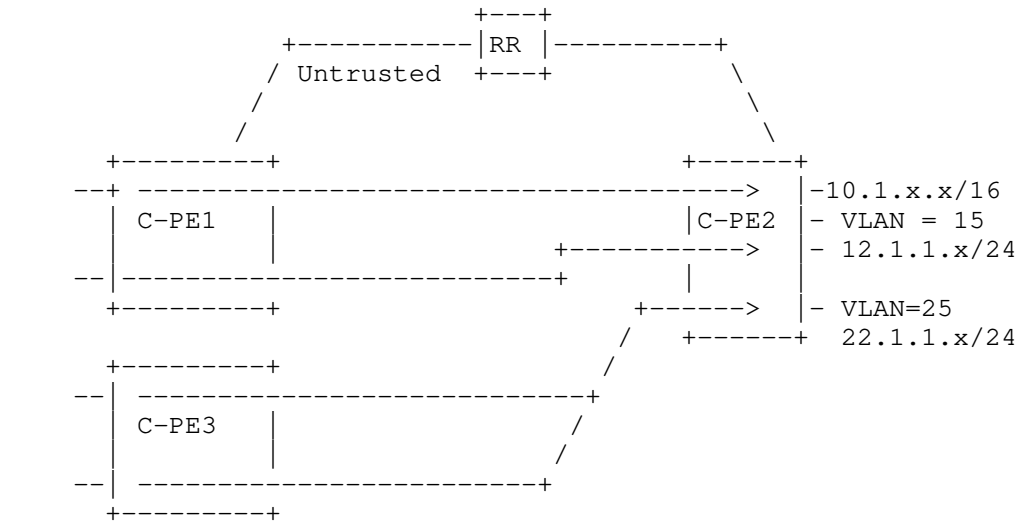


Figure 6: (see *.pdf for more accurate figure)

UPDATE 1:
- MP-NLRI Path Attribute:
  10.1.x.x/16
  VLAN #15
  12.1.1.x/24

- Tunnel-Encap Path Attribute:
  IPsec SA attributes for IPsec tunnels to C-PE2 from any node for reaching 10.1.x.x/16, VLAN #15, and 12.1.1.x/24.

UPDATE 2 (only sent to C-PE3)
- MP-NLRI Path Attribute:
  VLAN #25
  22.1.1.x/24

- Tunnel-Encap:

IPsec SA attributes for IPsec tunnels to C-PE2 from C-PE3 for reaching VLAN #25 and subnet 22.1.1./24.

4.2. BGP Walk Through for Application Flow Based Segmentation

If the applications are assigned with unique IP addresses, the Application Flow based Segmentation described in Section 3.1.2 can be achieved by advertising different BGP UPDATE messages to different nodes. In the Figure below, the following BGP Updates can be advertised to ensure that Payment Application only communicates with the Payment Gateway:

BGP UPDATE #1 from C-PE2 to RR for the P2P topology that is only propagated to Payment GW node:

- MP-NLRI Path Attribute:
  - 30.1.1.x/24
- Tunnel Encap Path Attribute
  - IPsec Attributes for PaymentGW ->C-PE2

BGP UPDATE #2 from C-PE2 to RR for the routes to be reached by C-PE1 and C-PE2:

- MP-NLRI Path Attribute:
  - 10.1.x.x
  - 12.4.x.x
- Tunnel-Encap Path Attribute:
  - Any node to C-PE2

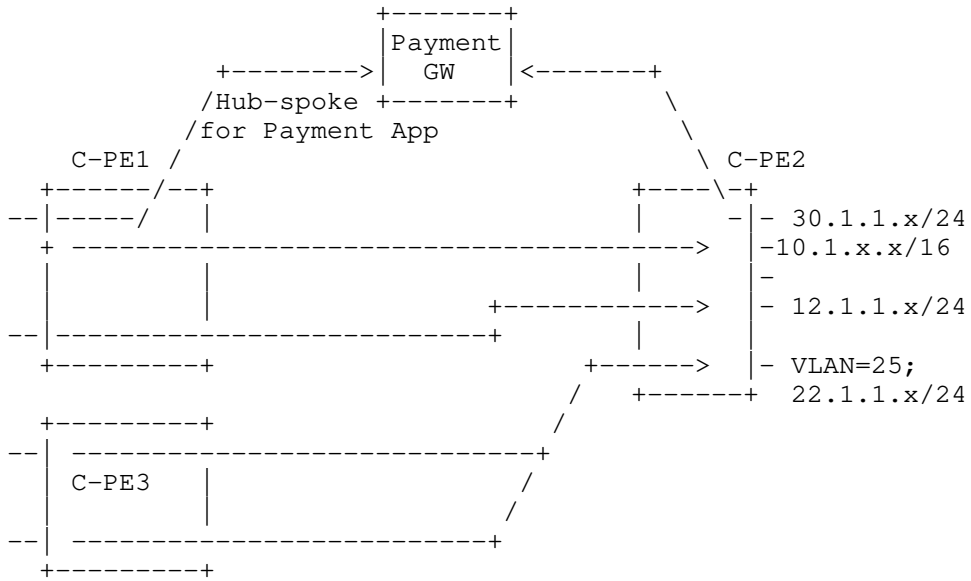


Figure 7: Application Based SDWAN Segmentation

4.3. Client Service Provisioning Model

The provisioning tasks described in Section 4 of RFC8388 are the same for the SDWAN client traffic. When client traffic is multi-homed to two (or more) C-PEs, the Non-Service-Specific parameters need to be provisioned per the Section 4.1.1 of RFC8388.

Since some SDWAN nodes are ephemeral and have small number of IP subnets or VLANs attached to their client ports, it is recommended to have default and simplified Service-specific parameters for each client port, remotely managed by the SDWAN Network Controller via the secure channel (TLS/DTLS) between the controller and the C-PEs.

4.4. WAN Ports Provisioning Model

Since the deployment of PEs to MPLS VPN are for relatively long term, the common provisioning procedure for PE's WAN ports is via CLI.

A SDWAN node deployment can be ephemeral and its location can be in remote locations, manual provisioning for its WAN ports is not acceptable. In addition, a SDWAN WAN port's IP address can be dynamically assigned or using private addresses. Therefore, it is necessary to have a separate control protocol; something like NHRP did for ATM, for a SDWAN node to register its WAN property to its controller dynamically.

Unlike a PE to MPLS based VPN where its WAN ports are homogeneously facing MPLS private network and all traffic are egressed in MPLS data frames through its WAN ports, the WAN ports of a SDWAN node can be connected to a PE of VPN with Ethernet/IP, MPLS private network directly via MPLS headers, or the public Internet.

For Scenario #1 described in Section 3.2, the WAN ports can face public internet or VPN.

For Scenario #2 described in Section 3.3, WAN ports are either configured as connecting to PEs of VPN where traffic can be sent as IP/Ethernet without encryption, or configured as connecting to public Internet that requires encryption for packets egress out.

For Scenario #3 described in Section 3.4, the WAN ports are either configured as VPN egress ports (hand off MPLS data frames), or as connecting to the public internet that requires MPLS in IP in IPsec encapsulation.

4.5. Why BGP as Control Plane for SDWAN?

For a small sized SDWAN network, traditional hub & spoke model using NHRP or DSVPN/DMVPN with a hub node (or controller) managing SDWAN node WAN ports mapping (e.g. local & public addresses and tunnel identifiers mapping) can work reasonably well. However, for a large SDWAN network, say more than 100 nodes with different types of topologies, the traditional approach becomes very messy, complex and error prone.

Here are some of the compelling reasons of using BGP instead of extending NHRP/DSVPN/DMVPN. (Same as the reasons quoted by LSVR on why using BGP):

- BGP has the built-in capability to constrain the propagation of SDWAN edge node properties to a small number of edge nodes [RFC4684].

- RR already has the capability to apply policies to communications among peers.

- BGP is widely deployed as sole protocol (see RFC 7938)

- Robust and simple implementation

- Wide acceptance - minimal learning

- Reliable transport

- Guaranteed in-order delivery

- Incremental updates

- Incremental updates upon session restart

- No flooding and selective filtering

5. SDWAN Traffic Forwarding Walk Through

BGP based EVPN control plane are still applicable to routes attached to the client ports of SDWAN nodes. Section 5 of RFC8388 describes the BGP EVPN NLRI Usage for various routes of client traffic. The procedures described in the Section 6 of RFC8388 are same for the SDWAN client traffic.

The only additional consideration for SDWAN is to control how traffic egress the SDWAN edge node to various WAN ports.

5.1. SDWAN Network Startup Procedures

A SDWAN network can add or delete SDWAN edge nodes on regular basis depending on user requests.

- For Scenario #1: a SDWAN edge node in a shopping mall or Cloud DC can be added or removed on demand. The Zero Touch Provisioning described in 3.1.2 are required for the node startup.
- For Scenario #2: this can be Data Centers or enterprises upgrading their CPEs to add extra bandwidth via public internet in addition to VPN services that they already purchased. Before the node powers up

or upgraded, there should be links connected to the PEs of a provider VPNs.

- For Scenario #3, the Internet facing WAN ports are added to (or removed from) existing VPN PEs.

5.2. Packet Walk-Through for Scenario #1

Upon power up, a SDWAN node can learn client routes from the Client facing ports, in the same way as EVPN described in RFC8388. Controller facilitates the IPsec SA establishment and rekey management as described in [SECURE-EVPN]. Controller manages how client's routes are associated with individual IPSec SA.

[SECURE-EVPN] describes a solution for SDWAN Scenario #1. It utilizes the BGP RR to facilitate the key and policy exchange among PE devices to create private pair-wise IPsec Security Associations without IKEv2 point-to-point signaling or any other direct peer-to-peer session establishment messages.

When C-PEs do not support MPLS, the approaches described by RFC8365 can be used, with addition of IPsec encrypting the IP packets when sending packets over the Black Interfaces.

5.3. Packet Walk-Through for Scenario #2

In this scenario, C-PEs have some WAN ports connected to the public internet and some WAN ports with direct connect to PEs of trusted VPN. The C-PEs in Scenario #2 have the plain IP/Ethernet data frames egress to the PEs of the VPN, encrypted data frames egress the WAN ports facing the public Internet.

Users specify the policy or criteria on which flows can only egress WAN ports facing the trusted VPN without encryption, which can egress the WAN ports facing the public Internet with encryption, or which can egress WAN ports facing the public Internet without encryption.

The internet facing WAN ports can face potential DDoS attacks, additional anti-DDoS mechanism has to be enabled on those WAN ports and the Control Plane should not learn routes from the Public Network facing WAN ports.

For the Scenario #2, if a client route can be reached by MPLS VPN and IPsec Tunnel via public network, the BGP UPDATE for the client

route should indicate all available tunnels in the Tunnel Path Attribute of the BGP NLRI.

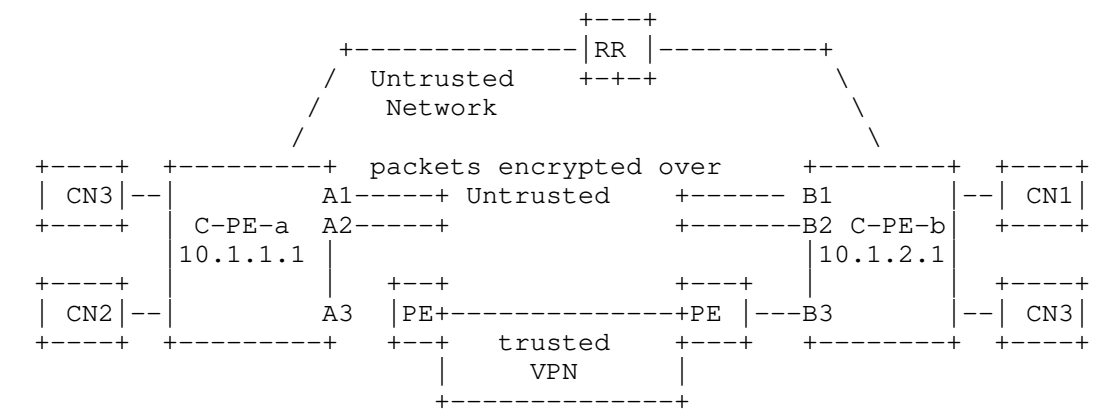


Figure 8: SDWAN Scenario #2

For example, if the CN1 route can be reached by both VPN and Public internet, the CN1's BGP route UPDATE should include the following:

- MP-NLRI Path Attribute:

  CN1

- Tunnel-Encap Path Attribute:

  Tunnel 1: MPLS-in-GRE encapsulation
  With the MPLS-in-GRE Sub-TLV specified by Tunnel-Encap;

  Tunnel 2: IPsec-GRE encapsulation
  With the IPsec Sub-TLVs specified by the [SECURE-EVPN] and [BGP-EDGE-DISCOVERY]

There could be multiple IPsec SA tunnels terminated at the edge node loopback address or terminated at WAN ports. For the Scenario #2, there can be policies to determine which IPsec SA tunnels that the client route can be carried. When a client route can be carried by multiple IPsec SA tunnels terminated by two different WAN ports, multiple Tunnel Path Attributes with different Tunnel-end-point Sub-TLVs need to be included in the NLRI of the BGP UPDATE for the client route.

5.4. Packet Walk-Through for Scenario #3

The behavior described in [SECURE-L3VPN] applies to this scenario.

[SECURE-L3VPN] describes how to extend the RFC4364 VPN to allow some PEs being connected to other PEs via public networks. In this scenario, the PEs is the SDWAN Edge nodes. [SECURE-L3VPN] introduces the concept of RED Interface & Black Interface on those PEs. RED interfaces face the VPN over which packets can be forwarded natively without encryption. Black Interfaces face public network over which only IPsec-protected packets are forwarded. [SECURE-L3VPN] assumes PEs terminate MPLS packets, and use MPLS over IPsec when sending over the Black Interfaces.

The C-PEs not only have RED interfaces facing clients but also have RED interface facing MPLS backbone, with additional BLACK interfaces facing the untrusted public networks for the WAN side. The C-PEs cannot mix the routes learned from the Black Interfaces with the Routes from RED Interfaces. The routes learned from core-facing RED interfaces are for underlay and cannot be mixed with the routes learned over access-facing RED interfaces that are for overlay. Furthermore, the routes learned over core-facing interfaces (both RED and BLACK) can be shared in the same GLOBAL route table.

There may be some added risks of the packets from the ports facing the Internet. Therefore, special consideration has to be given to the routes from WAN ports facing the Internet. RFC4364 describes using an RD to create different routes for reaching same system. A similar approach can be considered to force packets received from the Internet facing ports to go through special security functions before being sent over to the VPN backbone WAN ports.

6. Manageability Considerations

SDWAN overlay networks utilize the SDWAN controller to facilitate route distribution, central configurations, and others. SDWAN Edge nodes need to advertise the attached routes to their controller (i.e. RR in BGP case).

7. Security Considerations

Having WAN ports facing the public Internet introduces the following security risks:

1) Potential DDoS attack to the C-PEs with ports facing internet.

2) Potential risk of provider VPN network being injected with illegal traffic coming from the public Internet WAN ports on the C-PEs.

8. IANA Considerations

None

9. References

9.1. Normative References

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

[RFC4364] E. rosen, Y. Rekhter, "BGP/MPLS IP Virtual Private networks (VPNs)", Feb 2006.

[RFC7296] C. Kaufman, et al, "Internet Key Exchange Protocol Version 2 (IKEv2)", Oct 2014.

[RFC7432] A. Sajassi, et al, "BGP MPLS-Based Ethernet VPN", Feb 2015.

[RFC8365] A. Sajassi, et al, "A network Virtualization Overlay Solution Using Ethernet VPN (EVPN)", March 2018.

9.2. Informative References

[RFC8192] S. Hares, et al, "Interface to Network Security Functions (I2NSF) Problem Statement and Use Cases", July 2017

[RFC5521] P. Mohapatra, E. Rosen, "The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", April 2009.

[BGP-SDWAN-Port] L. Dunbar, H. Wang, W. Hao, "BGP Extension for SDWAN Overlay Networks", draft-dunbar-idr-bgp-sdwan-overlay-ext-03, work-in-progress, Nov 2018.

[Net2Cloud-Gap] L. Dunbar, A. Malis, C. Jacquenet, "Gap Analysis of Interconnecting Underlay with Cloud Overlay", draft-dm-net2cloud-gap-analysis-02, work in progress, Oct. 2018.

[SDWAN-EDGE-Discovery] L. Dunbar, S. Hares, R. Raszuk, K. Majumdar, "BGP UPDATE for SDWAN Edge Discovery", draft-dunbar-idr-sdwan-edge-discovery-00, work-in-progress, July 2020.

[VPN-over-Internet] E. Rosen, "Provide Secure Layer L3VPNs over Public Infrastructure", draft-rosen-bess-secure-l3vpn-00, work-in-progress, July 2018

[DMVPN] Dynamic Multi-point VPN: https://www.cisco.com/c/en/us/products/security/dynamic-multipoint-vpn-dmvpn/index.html

[DSVPN] Dynamic Smart VPN: http://forum.huawei.com/enterprise/en/thread-390771-1-1.html

[SECURE-EVPN] A. Sajassi, et al, "Secure EVPN", draft-sajassi-bess-secure-evpn-01, Work-in-progress, March 2019.

[SECURE-L3VPN] E. Rosen, R. Bonica, "Secure Layer L3VPN over Public Infrastructure", draft-rosen-bess-secure-l3vpn-00, Work-in-progress, June 2018.

[ITU-T-X1036] ITU-T Recommendation X.1036, "Framework for creation, storage, distribution and enforcement of policies for network security", Nov 2007.

[Net2Cloud-Problem] L. Dunbar and A. Malis, "Seamless Interconnect Underlay to Cloud Overlay Problem Statement", draft-dm-net2cloud-problem-statement-02, June 2018

[Net2Cloud-gap] L. Dunbar, A. Malis, and C. Jacquenet, "Gap Analysis of Interconnecting Underlay with Cloud Overlay", draft-dm-net2cloud-gap-analysis-02, work-in-progress, Aug 2018.

[Tunnel-Encap] E. Rosen, et al "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-10, Aug 2018.

10. Acknowledgments

Acknowledgements to Jim Guichard, John Scudder, Darren Dukes, Andy Malis and Donald Eastlake for their review and contributions.

This document was prepared using 2-Word-v2.0.template.dot.

Authors' Addresses

Linda Dunbar
Futurewei
Email: ldunbar@futurewei.com

James Guichard
Futurewei
Email: james.n.guichard@futurewei.com

Ali Sajassi
Cisco
Email: sajassi@cisco.com

John Drake
Juniper
Email: jdrake@juniper.net

Basil Najem
Bell Canada
Email: basil.najem@bell.ca

David Carrel
Cisco
Email: carrel@cisco.com

Ayan Banerjee
Cisco
Email: ayabaner@cisco.com

BESS
Internet-Draft
Intended status: Standards Track
Expires: December 12, 2020

Z. Zhang
L. Giuliano
Juniper Networks
K. Patel
Arrcus
I. Wijnands
M. Mishra
Cisco Systems
A. Gulko
Refinitiv

June 10, 2020

BGP Based Multicast
draft-ietf-bess-bgp-multicast-02

Abstract

This document specifies a BGP address family and related procedures that allow BGP to be used for setting up multicast distribution trees. This document also specifies procedures that enable BGP to be used for multicast source discovery, and for showing interest in receiving particular multicast flows. Taken together, these procedures allow BGP to be used as a replacement for other multicast routing protocols, such as PIM or mLDP. The BGP procedures specified here are based on the BGP multicast procedures that were originally designed for use by providers of Multicast Virtual Private Network service.

Requirements Language

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

Status of This Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at 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 December 12, 2020.

Table of Contents

1. Introduction

1.1. Motivation

This section provides some motivation for BGP signaling for native and labeled multicast. One target deployment would be a Data Center that requires multicast but uses BGP as its only routing protocol [RFC7938]. In such a deployment, it would be desirable to support multicast by extending the deployed routing protocol, without requiring the deployment of tree building protocols such as PIM, mLDP, RSVP-TE P2MP, and without requiring an IGP.

Additionally, compared to PIM, BGP based signaling has several advantage as described in the following section, and may be desired in non-DC deployment scenarios as well.

1.1.1. Native/unlabeled Multicast

Protocol Independent Multicast (PIM) has been the prevailing multicast protocol for many years. Despite its success, it has two drawbacks:

- The ASM model, which is prevalent, introduces complexity in the following areas: source discovery procedures, need for Rendezvous Points (RPs) and group-to-RP mappings, need to switch between RP-rooted trees and source-rooted trees, etc.

- Periodical protocol state refreshes due to soft state nature.

PIM-SSM removes much of the complexity of PIM-ASM by moving source discovery to the application layer. However, for various reasons, many legacy applications and devices still rely upon network-based source discovery. PIM-Port (PIM over Reliable Transport) solves the soft state issue, though its deployment has also been limited for two reasons:

- o It does not remove the ASM complexities.
- o In many of the scenarios where reliable transport is deemed important, BGP-based multicast (e.g. BGP-MVPN) has been used instead of PORT.

Partly because of the above mentioned problems, some Data Center operators have been avoiding deploying multicast in their networks.

BGP-MVPN [RFC6514] uses BGP to signal VPN customer multicast state over provider networks. It removes the above mentioned problems from the SP environment, and the deployment experiences have been encouraging. While RFC 6514 makes it possible for an SP to provide MVPN service without running PIM on its backbone, that RFC still assumes that PIM (or mLDP) runs on the PE-CE links. [draft-ietf-bess-mvpn-pe-ce] adapts the concept of BGP-MVPN to PE-CE links so that the use of PIM on the PE-CE links can be eliminated (though the PIM-ASM complexities still remains in the customer network), and this document extends it further to general topologies, so that they can be run on any router, as a replacement for PIM or mLDP.

With that, PIM can be completely eliminated from the network. PIM soft state is replaced by BGP hard state. For ASM, source specific trees are set up directly after simpler source discovery (data driven on FHRs and control driven elsewhere), all based on BGP. All the complexities related to source discovery and shared/source tree switch are also eliminated. Additionally, the trees can be setup with MPLS labels, with just minor enhancements in the signaling.

1.1.2. Labeled Multicast

There could be two forms of labeled multicast signaled by BGP. The first one is labeled (x,g) multicast where 'x' stands for either 's' or '*'. Basically, it is for BGP-signaled multicast tree as described in previous section but with labels. The second one is for mLDP tunnels with BGP signaling in part or whole through a BGP domain.

For both cases, BGP is used because other label distribution mechanisms like mLDP may not be desired by some operators. For example, a DC operator may prefer to have a BGP-only deployment.

1.2. Overview

1.2.1. (x,g) Multicast

PIM-like functionality is provided, using BGP-based join/prune signaling and BGP-based source discovery for ASM. The BGP-based join signaling supports both labeled multicast and IP multicast.

The same RPF procedures as in PIM are used for each router to determine the RPF neighbor for a particular source or RPA (in case of Bidirectional Tree). Except in the Bidirectional Tree case and a special case described in Section 1.2.1.2, no (*,G) join is used - LHR routers discover the sources for ASM and then join towards the sources directly. Data driven mechanisms like PIM Assert is replaced by control driven mechanisms (Section 1.2.4).

The joins are carried in BGP Updates with MCAST-TREE SAFI and S-PMSI/ Leaf A-D routes defined in this document. The updates are targeted at the upstream neighbor by use of Route Targets. There are three benefits of using S-PMSI/Leaf routes for this purpose: a) when the routes go through RRs, we have to distinguish different routes based on upstream router and downstream router. This leads to Leaf routes. b) for labeled bidirectional trees, we need to signal "upstream fec". S-PMSI suits this very well. c) we may want to allow the option of setting up trees or parts of a tree from the root/upstream towards leaves/downstream and S-PMSI suits that very well.

If the BGP updates carry labels (via Tunnel Encapsulation Attribute [I-D.ietf-idr-tunnel-encaps]), then (s,g) multicast traffic can use the labels. This is very similar to mLDP Inband Signaling [RFC6826], except that there are no corresponding "mLDP tunnels" for the PIM trees. Similar to mLDP, labeled traffic on transit LANs are point to point. Of course, traffic sent to receivers on a LAN by a LHR is native multicast.

For labeled bidirectional (*,g) trees, downstream traffic (away from the RPA) can be forwarded as in the (s,g) case. For upstream traffic (towards RPA), the upstream neighbor needs to advertise a label for its downstream neighbors. The same label that the upstream neighbor advertises to its upstream is the same one that it advertises to its downstreams, using an S-PMSI A-D route.

1.2.1.1. Source Discovery for ASM

This document does not support ASM via shared trees (aka RP Tree, or RPT) with one exception discussed in the next section. Instead, FHRs, LHRs, and optionally RRs work together to propagate/discover

source information via control plane and LHRs join source specific Shortest Path Trees (SPT) directly.

A FHR originates Source Active A-D routes upon discovering sources for particular flows and advertise them to its peers. It is desired that the SA routes only reach LHRs that are interested in receiving the traffic. To achieve that, the SA routes carry an IPv4 or IPv6 address specific Route Target. The Global Administrator field is set the group address of the flow, and the Local Administrator field is set to 0 or a pre-assigned domain-wide unique value that identifies a VPN. An LHR advertises Route Target Membership routes, with the Route Target field in the NLRI set according to the groups it wants to receive traffic for, as how a FHR encode the Route Target in its Source Active routes. The propagation of the SA routes is subject to cooperative export filtering as specified in [RFC4684] and referred to as RTC mechanism in this document. That way, the LHR only receives Source Active routes for groups that it is interested in.

Typically, a set of RRs are used and they maintains all Source Active routes but only distribute to interested LHRs on demand (upon receiving corresponding Route Target Membership routes, which are triggered on LHRs when they receive IGMP/MLD membership routes). The rest of the document assumes that RRs are used, even though that is not required.

1.2.1.2. ASM Shared-tree-only Mode

It may be desired that only a shared tree is used to distribute all traffic for a particular ASM group from its RP to all LHRs, as described in Section 4.1 "PIM Shared Tree Forwarding" of [RFC7438]. This will significantly cut down the number of trees and works out very well in certain deployment scenarios. For example, all the sources could be connected to the RP, or clustered close to RP. In the latter case, either the path from FHRs to the RP do not intersect the shared tree so native forwarding can be used between the FHRs and the RP, or other means outside of this document could be used to forward traffic from FHRs to the RP.

For native forwarding from FHRs to the RP, SA routes may be used to announce the sources so that the RP can join source specific trees to pull traffic, but the group specific Route Target is not needed. The LHRs do not advertise the group specific Route Target Membership routes as they do not need the SA routes.

To establish the shared tree, (*,g) Leaf A-D routes are used as in the bidirectional tree case, though no forwarding state is established to forward traffic from downstream neighbors.

1.2.1.3. Integration with BGP-MVPN

For each VPN, the Source Active routes distribution in that VPN do not have to involve PEs at all unless there are sources/receivers directly connected to some PEs and they are independent of MVPN SA routes. For example, FHRs and LHRs establish BGP sessions with RRs of that particular VPN for the purpose of SA distribution.

After source discovery, BGP multicast signaling is done from LHRs towards the sources. When the signaling reaches an egress PE, BGP-MVPN signaling takes over, as if a PIM (s,g) join/prune was received on the PE-CE interface. When the BGP-MVPN signaling reaches the ingress PE, BGP multicast signaling as specified in this document takes over, similar to how BGP-MVPN triggers PIM (s,g) join/prune on PE-CE interfaces.

1.2.2. BGP Inband Signaling for mLDP Tunnel

Part of an (or the whole) mLDP tunnel can also be signaled via BGP and seamlessly integrated with the rest of mLDP tunnel signaled natively via mLDP. All the procedures are similar to mLDP except that the signaling is done via BGP. The mLDP FEC is encoded as the BGP NLRI, with MCAST-TREE SAFI and S-PMSI/Leaf A-D Routes for C-multicast mLDP defined in this document. The Leaf A-D routes correspond to mLDP Label Mapping messages, and the S-PMSI A-D routes are used to signal upstream FEC for MP2MP mLDP tunnels, similar to the bidirection (*,g) case.

1.2.3. BGP Sessions

In order for two BGP speakers to exchange MCAST-TREE NLRI, they must use BGP Capabilities Advertisement [RFC5492] to ensure that they both are capable of properly processing the MCAST-TREE NLRI. This is done as specified in [RFC4760], by using a capability code 1 (multiprotocol BGP) with an AFI of IPv4 (1) or IPv6 (2) and a SAFI of MCAST-TREE with a value to be assigned by IANA.

How the BGP peer sessions are provisioned, whether EBGP or IBGP, whether statically, automatically (e.g., based on IGP neighbor discovery), or programmably via an external controller, is outside the scope of this document.

In case of IBGP, it could be that every router peering with Route Reflectors, or hop by hop IBGP sessions could be used to exchange MCAST-TREE NLRIs for joins. In the latter case, unless desired otherwise for reasons outside of the scope of this document, the hop by hop IBGP sessions SHOULD only be used to exchange MCAST-TREE NLRIs.

When multihop BGP is used, a router advertises its local interface addresses, for the same purposes that the Address List TLV in LDP serves. This is achieved by advertising the interface address as host prefixes with IPv4/v6 Address Specific ECs corresponding to the router's local addresses used for its BGP sessions (Section 2.1.5).

Because the BGP Capability Advertisement is only between two peers, when the sessions are only via RRs, a router needs another way to determine if its neighbor is capable of signaling multicast via BGP. The interface address advertisement can be used for that purpose - the inclusion of a Session Address EC indicates that the BGP speaker identified in the EC supports the C-Multicast NLRI.

FHRs and LHRs may also establish BGP sessions to some Route Reflectors for source discovery purpose (Section 1.2.1.1).

With the traditional PIM, the FHRs and LHRs refer to the PIM DRs on the source or receiver networks. With BGP based multicast, PIM may not be running at all, and the FHRs and LHRs refer to the IGMP/MLD queriers, or the DF elected per [I-D.wijnands-bier-mld-lan-election]. Alternatively, if it is known that a network only has senders then no IGMP/MLD or DF election is needed - any router may generate SA routes. That will not cause any issue other than redundant SA routes being originated.

1.2.4. LAN and Parallel Links

There could be parallel links between two BGP peers. A single multi-hop session, whether IBGP or EBGP, between loopback addresses may be used. Except for LAN interfaces in case of unlabeled (x,g) unidirectional trees (note that transit LAN interface is not supported for BGP signaled (*,g) bidirectional tree and for mLDP tunnels, traffic on transit LAN is point to point between neighbors), any link between the two peers can be automatically used by a downstream peer to receive traffic from the upstream peer, and it is for the upstream peer to decide which link to use. If one of the links goes down, the upstream peer switches to a different link and there is no change needed on the downstream peer.

For unlabeled (x,g) unidirectional trees, the upstream peer MAY prefer LAN interfaces to send traffic, since multiple downstream peers may be reached simultaneously, or it may make a decision based on local policy, e.g., for load balancing purpose. Because different downstream peers might choose different upstream peers for RPF, when an upstream peer decides to use a LAN interface to send traffic, it originates an S-PMSI A-D route indicating that one or more LAN interface will be used. The route carries Route Targets specific to the LANs so that all the peers on the LANs import the route. If more

than one router originate the route specifying the same LAN for the same (s,g) or (*,g) flow, then assert procedure based on the S-PMSI A-D routes happens and assert losers will stop sending traffic to the LAN.

1.2.5. Transition

A network currently running PIM can be incrementally transitioned to BGP based multicast. At any time, a router supporting BGP based multicast can use PIM with some neighbors (upstream or downstream) and BGP with some other neighbors. PIM and BGP MUST not be used simultaneously between two neighbors for multicast purpose, and routers connected to the same LAN MUST be transitioned during the same maintenance window.

In case of PIM-SSM, any router can be transitioned at any time (except on a LAN). It may receive source tree joins from a mixed set of BGP and PIM downstream neighbors and send source tree joins to its upstream neighbor using either PIM or BGP signaling.

In case of PIM-ASM, the RPs are first upgraded to support BGP based multicast. They learn sources either via PIM procedures from PIM FHRs, or via Source Active A-D routes from BGP FHRs. In the former case, the RPs can originate proxy Source Active A-D routes. There may be a mixed set of RPs/RRs - some capable of both traditional PIM RP functionalities while some only redistribute SA routes.

Then any routers can be transitioned incrementally. A transitioned LHR router will pull Source Active A-D routes from the RPs/RRs when they receive IGMP/MLD (*,G) joins for ASM groups, and may send either PIM (s,g) joins or BGP Source Tree Join routes. A transitioned transit router may receive (*,g) PIM joins but only send source tree joins after pulling Source Active A-D routes from RPs/RRs.

Similarly, a network currently running mLDP can be incrementally transitioned to BGP signaling. Without the complication of ASM, any router can be transitioned at any time, even without the restriction of coordinated transition on a LAN. It may receive mixed mLDP label mapping or BGP updates from different downstream neighbors, and may exchange either mLDP label mapping or BGP updates with its upstream neighbors, depending on if the neighbor is using BGP based signaling or not.

1.2.6. Inter-region Multicast

An end-to-end multicast tree or P2MP tunnel may span multiple regions, where a region could be an IGP area (or even a sub-area) or an Autonomous System (AS). There are several situations to consider.

1.2.6.1. Same BGP Signaling Inline across a Region

With inline signaling, the multicast tree/tunnel is signaled through the region and internal routers in the region maintain corresponding per-tree/tunnel state.

If all routers in the region have route towards the source/root of the tree/tunnel then there is nothing different from the intra-region case. On the other hand, if internal routers do not have route towards the source/root, e.g. BGP-LU is used as in Seamless MPLS, the internal routers need to do RPF towards an upstream Regional Border Router (RBR). To signal the RBR information to an internal upstream router, the Leaf A-D Route carries a new BGP Extended Community referred to as Multicast RPF Address EC, similar to PIM RPF Vector [RFC5496] and mLDP Recursive FEC [RFC6512].

1.2.6.2. Different Signaling Inline across a Region

Just like that part of a PIM multicast tree can be signaled as an mLDP P2MP/MP2MP tunnel with mLDP Inband Signaling [RFC6826], BGP-signaled (*/s, g) multicast tree can be signaled with mLDP Inband Signaling or even with PIM across the region, and a BGP-signaled p2mp tunnel can be signaled with mLDP across the region. A RBR will stitch the upstream portion (e.g BGP-signaled) to downstream portion (e.g mLDP-signaled).

Depending on whether internal routers have route towards the source/root, PIM RPF Vector or mLDP Recursive FEC may be used.

1.2.6.3. Overlay Signaling Over a Region

With overlay signaling, a downstream RBR signals via BGP to its upstream RBR over the region (whether via a RR or not) and the internal routers do not maintain the state of the (overlay) tree/tunnel. The upstream RBR tunnels packets to the downstream RBR, just as in the intra-region case when two routers on the tree/tunnel are not directly connected. For example, when BGP-LU is used as in Seamless MPLS, a downstream RBR determines that the route towards the source/root has a BGP Next Hop towards a BGP speaker capable of multicast signaling via BGP as specified in this document, so it signals to that BGP speaker (via a RR or not).

Suppose an upstream RBR receives the signaling for the same tree/tunnel from several downstream RBRs. It could use Ingress Replication to replicate packets directly to those downstream RBRs, or it could use underlay P2MP tunnels instead.

In the latter case, the upstream RBR advertises an S-PMSI A-D route with a Provider Tunnel Attribute (PTA) specifying the underlay tunnel. This is very much like the "mLDP Over Targeted Sessions" [RFC7060] or BGP-MVPN [RFC6514]. If the mapping between overlay tree/tunnel and underlay tunnel is one-to-one, the MPLS Label field in the PTA is set to 0 or otherwise set to a Domain-wide Common Block (DCB) label [I-D.ietf-bess-mvpn-evpn-aggregation-label] or an upstream-assigned label corresponding to the overlay tree/tunnel.

The underlay tunnel, whether P2P to individual downstream RBRs or P2MP to the set of downstream RBRs, can be of any type including Segment Routing (SR) [RFC8402] policies [I-D.ietf-spring-segment-routing-policy] [I-D.voyer-pim-sr-p2mp-policy].

1.2.6.4. Controller Based Signaling

[I-D.ietf-bess-bgp-multicast-controller] specifies the procedures for a controller to signal multicast forwarding state to each router on a multicast tree based on the controller's computation. Depending on deployment scenarios, in inter-region cases it is possible that the hop-by-hop signaling specified in this document and the controller based signaling may be used in different regions.

Consider a situation where an ABR is connected three regions A, B, and C, where hop-by-hop signaling is used in A and B, while controller based signaling is used in C.

For a particular multicast tree, A is the upstream region, while B and C are two downtream regions. The ABR receives a Leaf A-D route from region B and a Leaf A-D route from C's controller, and sends a Leaf A-D route to its upstream router in A.

For a different tree, C is the upstream region while A and B are downstream. The ABR receives two Leaf A-D routes for the tree from regions A and B, and one Leaf A-D route from C's controller. Note that the ABR needs to signal to the controller that it is a leaf of the tree (because of the Leaf A-D routes received from regions A and B).

For both cases, the ABR stitches together different segments in different regions by creating forwarding state based on the Leaf A-D routes (optionally based on the S-PMSI A-D routes in region A and B in addition.)

2. Specification

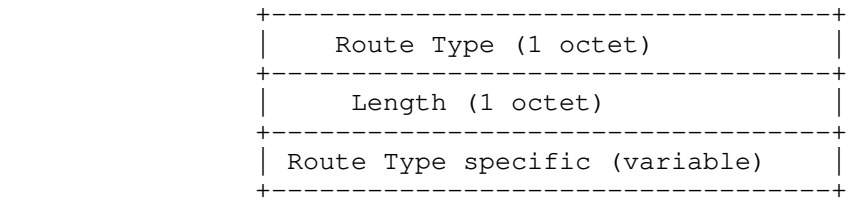
2.1. BGP NLRIs and Attributes

The BGP Multiprotocol Extensions [RFC4760] allow BGP to carry routes from multiple different "AFI/SAFIs". This document defines a new SAFI known as a MCAST-TREE SAFI with a value to be assigned by the IANA. This SAFI is used along with the AFI of IPv4 (1) or IPv6 (2).

The MCAST-TREE NLRI defined below is carried in the BGP UPDATE messages [RFC4271] using the BGP multiprotocol extensions [RFC4760] with a AFI of IPv4 (1) or IPv6 (2) assigned by IANA and a MCAST-TREE SAFI with a value to be assigned by the IANA.

The Next hop field of MP_REACH_NLRI attribute SHALL be interpreted as an IPv4 address whenever the length of the Next Hop address is 4 octets, and as an IPv6 address whenever the length of the Next Hop is address is 16 octets.

The NLRI field in the MP_REACH_NLRI and MP_UNREACH_NLRI is a prefix with a maximum length of 12 octets for IPv4 AFI and 36 octets for IPv6 AFI. The following is the format of the MCAST-TREE NLRI:



The Route Type field defines encoding of the rest of the MCAST-TREE NLRI. (Route Type specific MCAST-TREE NLRI).

The Length field indicates the length in octets of the Route Type specific field of MCAST-TREE NLRI.

The following new route types are defined:

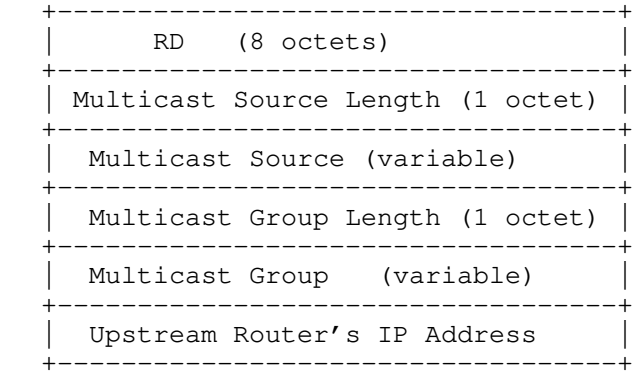
```
      3 -  S-PMSI A-D Route for (x,g)
      4 -  Leaf A-D Route
      5 -  Source Active A-D Route
   0x43 -  S-PMSI A-D Route for C-multicast mLDP
```

Except for the Source Active A-D routes, the routes are to be consumed by targeted upstream/downstream neighbors, and are not propagated further. This can be achieved by outbound filtering based on the RTs that lead to the importation of the routes.

The Type-3/4 routes MAY carry a Tunnel Encapsulation Attribute (TEA) [I-D.ietf-idr-tunnel-encaps]. The Type-0x43 route MUST carry a TEA. When used for mLDP, the Type-4 route MUST carry a TEA. Only the MPLS tunnel type for the TEA is considered. Others are outside the scope of this document.

2.1.1. S-PMSI A-D Route

Similar to defined in RFC 6514, an S-PMSI A-D Route Type specific MCAST-TREE NLRI consists of the following:



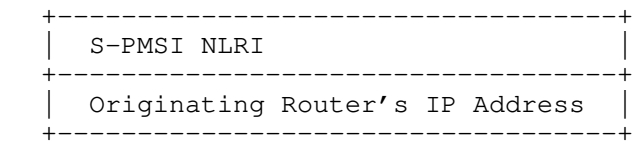
If the Multicast Source (or Group) field contains an IPv4 address, then the value of the Multicast Source (or Group) Length field is 32. If the Multicast Source (or Group) field contains an IPv6 address, then the value of the Multicast Source (or Group) Length field is 128.

Usage of other values of the Multicast Source Length and Multicast Group Length fields is outside the scope of this document.

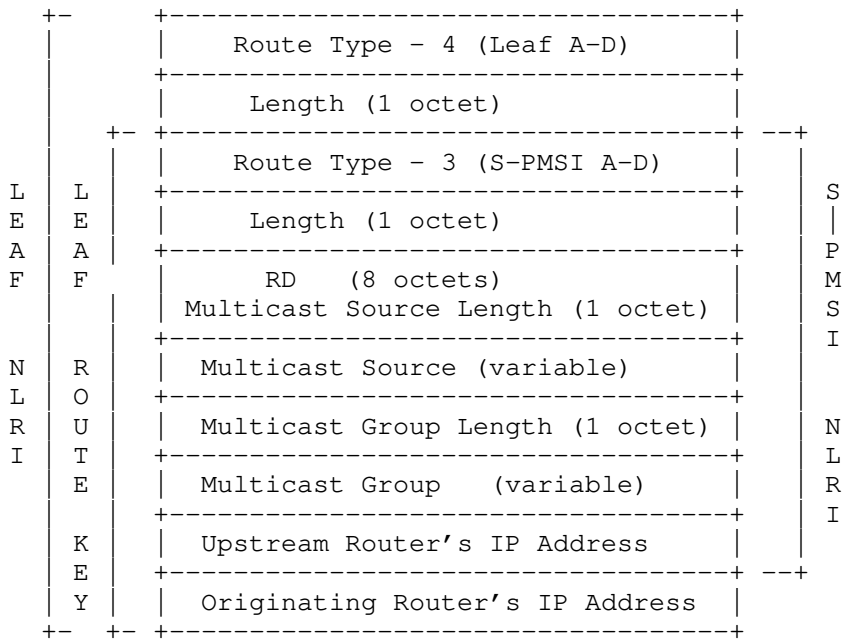
There are two usages for S-PMSI A-D route. They're described in Section 2.2.5 and Section 2.2.6 respectively.

2.1.2. Leaf A-D Route

Similar to the Leaf A-D route in [RFC6514], a MCAST-TREE Leaf A-D route's route key includes the corresponding S-PMSI NLRI, plus the Originating Router's IP Addr. The difference is that there is no RD.



For example, the entire NLRI of a Leaf A-D route for (x,g) tree is as following:

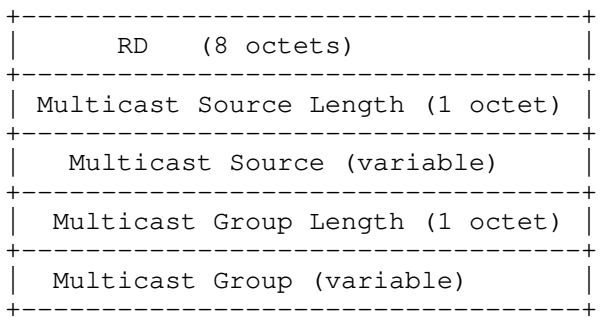


Even though the MCAST-TREE Leaf A-D route is unsolicited, unlike the Leaf A-D route for GTM in [RFC7524], it is encoded as if a corresponding S-PMSI A-D route had been received.

When used for signaling mLDP tunnels, even though the Leaf A-D route is unsolicited, unlike the "Route-type 0x44 Leaf A-D route for C-multicast mLDP" as in [RFC7441], it is Route-type 4 and encoded as if a corresponding S-PMSI A-D route had been received.

2.1.3. Source Active A-D Route

Similar to defined in RFC 6514, a Source Active A-D Route Type specific MCAST NLRI consists of the following:



The definition of the source/length and group/length fields are the same as in the S-PMSI A-D routes.

Usage of Source Active A-D routes is described in Section 1.2.1.1.

2.1.4. S-PMSI A-D Route for C-multicast mLDP

The route is used to signal upstream FEC for an MP2MP mLDP tunnel. The route key include the mLDP FEC and the Upstream Router's IP Address field. The encoding is similar to the same route in [RFC7441].

2.1.5. Session Address Extended Community

For two BGP speakers to determine if they are directly connected, each will advertise their local interface addresses, with an Session Address Extended Community. This is an IPv4/IPv6 Address Specific EC with the Global Admin Field set to the local address used for its multihop sessions and the Local Admin Field set to the prefix length corresponding to the interface's network mask.

For example, if a router has two interfaces with address 10.10.10.1/24 and 10.12.0.1/16 respectively (notice the different network mask), and a loopback address 11.11.11.1/32 that is used for BGP sessions, then it will advertise prefix 10.10.10.1/32 with a Session Address EC 11.11.11.1:24 and 10.12.0.1/32 with a Session Address EC 11.11.11.1:16. If it also uses another loopback address 11.11.11.11/32 for other BGP sessions, then the routes will additionally carry Session Address EC 11.11.11.11:24 and 11.11.11.11:16 respectively.

This achieves what the Address List TLV in LDP Address Messages achieves, and can also be used to indicate that a router supports the BGP multicast signaling procedures specified in this document.

Only those interface addresses that will be used as resolved nexthops in the RIB need to be advertised with the Session Address EC. For example, the RPF lookup may say that the resolved nexthop address is A1, so the router needs to find out the corresponding BGP speaker with address A1 through the (interface address, session address) mapping built according to the interface address NLRI with the Session Address EC. For comparison with LDP, this is done via the (interface address, session address) mapping that is built by the LDP Address Messages.

2.1.6. Multicast RPF Address Extended Community

This is an IP or IPv6 Address Specific EC with the Global Admin Field set to the address of the upstream RBR and the Local Admin Field set to 0.

2.2. Procedures

2.2.1. Source Discovery for ASM

When a FHR first receives a multicast packet addressed to an ASM group, it originates a Source Active route. It carries a IP/IPv6 Address Specific RT, with the Global Admin Field set to the group address and the Local Admin Field set to 0. The route is advertised to its peers, who will re-advertise further based on the RTC mechanisms. Note that typically the route is advertised only to the RRs.

The FHRs withdraws the Source Active route after a certain amount of time since it last received a packet of an (s,g) flow. The amount of time to wait is a local matter.

2.2.2. Originating Tree Join Routes

Note that in this document, tree join routes are S-PMSI/Leaf A-D routes.

2.2.2.1. (x,g) Multicast Tree

When a router learns from IGMP/MLD or a downstream PIM/BGP peer that it needs to join a particular (s,g) tree, it determines the RPF nexthop address wrt the source, following the same RPF procedures as defined for PIM. It further finds the BGP router that advertised the nexthop address as one of its local addresses.

If the RPF neighbor supports MCAST-TREE SAFI, this router originates a Leaf A-D route. Although it is unsolicited, it is constructed as if there was a corresponding S-PMSI A-D route. The Upstream Router's

IP Address field is set to the RPF neighbor's session address (learnt via the EC attached to the host route for the RPF nexthop address). An Address Specific RT corresponding to the session address is attached to the route, with the Global Administrative Field set to the session address and the local administrative field set to 0 or a pre-assigned domain-wide unique value that identifies a VPN.

Similarly, when a router learns that it needs to join a bi-directional tree for a particular group, it determines the RPF neighbor wrt the RPA. If the neighbor supports MCAST-TREE SAFI, it originates a Leaf A-D Route and advertises the route to the RPF neighbor (in case of EBGP or hop-by-hop IBGP), or one or more RRs.

When a router first learns that it needs to receive traffic for an ASM group, either because of a local (*,g) IGMP/MLD report or a downstream PIM (*,g) join, it originates a RTC route with the NLRI's AS field set to its AS number and the Route Target field set to an address based RT, with the Global Administrator field set to group address and the Local Administrator field set to 0 or a pre-assigned domain-wide unique value that identifies a VPN. The route is advertised to its peers (most practically some RRs), so that the router can receive matching Source Active A-D routes. Upon the receiving of the Source Active A-D routes, the router originates Leaf A-D routes as described above, as long as it still needs to receive traffic for the flows (i.e., the corresponding IGMP/MLD membership exists or join from downstream PIM/BGP neighbor exists).

When a Leaf A-D route is originated by this router, it sets up corresponding forwarding state such that the expected incoming interface list includes all non-LAN interfaces directly connecting to the upstream neighbor. LAN interfaces are added upon receiving corresponding S-PMSI A-D route (Section 2.2.5.2). If the upstream neighbor is not directly connected, tunnels may be used - details to be included in future revisions.

When the upstream neighbor changes, the previously advertised Leaf A-D route is withdrawn. If there is a new upstream neighbor, a new Leaf A-D route is originated, corresponding to the new neighbor. Because NLRIs are different for the old and new Leaf A-D routes, make-before-break as well as MoFRR [RFC7431] can be achieved.

2.2.2.2. BGP Inband Signaling for mLDP Tunnel

The same mLDP procedures as defined in [RFC6388] are followed, except that where a label mapping message is sent in [RFC6388], a Leaf A-D route is sent if the the upstream neighbor supports BGP based signaling.

2.2.3. Receiving Tree Join Routes

A router (auto-)configures Import RTs matching itself so that it can import tree join routes from their peers. Note that in this document, tree join routes are S-PMSI/Leaf A-D routes.

When a router receives a tree join route and imports it, it determines if it needs to originate its own corresponding route and advertise further upstream wrt the source/RPA or mLDP tunnel root. If this router is the FHR or is on the RPL or is the tunnel root, then it does not need to. Otherwise the procedures in Section 2.2.2 are followed.

Additionally, the router sets up its corresponding forwarding state such that traffic will be sent to the downstream neighbor, and received from the downstream neighbor in case of bidirectional tree/ tunnel. If the downstream neighbor is not directly connected, tunnels may be used - details to be included in future revisions.

2.2.4. Withdrawl of Tree Join Routes

For a particular tree or tunnel, if a downstream neighbor withdraws its Leaf A-D route, the neighbor is removed from the corresponding forwarding state. If all downstream neighbors withdraw their tree join routes and this router no longer has local receivers, it withdraws the tree join routes that it previously originated.

As mentioned earlier, when the upstream neighbor changes, the previously advertised Leaf A-D route is also withdrawn. The corresponding incoming interfaces are also removed from the corresponding forwarding state.

2.2.5. LAN procedures for (x,g) Unidirectional Tree

For a unidirectional (x,g) multicast tree, if there is a LAN interface connecting to the downstream neighbor, it MAY be preferred over non-LAN interfaces, but an S-PMSI A-D route MUST be originated to facilitate the analog of the Assert process (Section 2.2.5.1).

2.2.5.1. Originating S-PMSI A-D Routes

If this router chooses to use a LAN interface to send traffic to its neighbors for a particular (s,g) or (*,g) flow, it MUST announce that by originating a corresponding S-PMSI A-D route. The Tunnel Type in the PMSI Tunnel Attribute (PTA) is set to 0 (no tunnel information Present). The LAN interface is identified by an IP address specific RT, with the Global Administrative Field set to the LAN interface's address prefix and the Local Administrative Field set to the prefix

length. The RT also serves the purpose of restricting the importing of the route by all routers on the LAN. An operator MUST ensure that RTs encoded as above are not used for other purposes. Practically that should not be unreasonable.

If multiple LAN interfaces are to be used (to reach different sets of neighbors), then the route will include multiple RTs, one for each used LAN interface as described above.

The S-PMSI A-D routes may also be used to announce tunnels that could be used to send traffic to downstream neighbors that are not directly connected. Details may be added in future revisions.

2.2.5.2. Receiving S-PMSI A-D Routes

A router (auto-)configures an Import RT for each of its LAN interfaces over which BGP is used for multicast signaling. The construction of the RT is described in the previous section.

When a router R1 imports an S-PMSI A-D route for flow (x,g) from router R2, R1 checks to see if it also originating an S-PMSI A-D route with the same NLRI except the Upstream Router's IP Address field. When a router R1 originates an S-PMSI A-D route, it checks to see if it also has installed an S-PMSI A-D route, from some other router R2, with the same NLRI except the Upstream Router's IP Address field. In either case, R1 checks to see if the two routes have an RT in common and the RT is encoded as in Section 2.2.5.1. If so, then there is a LAN attached to both R1 and R2, and both routers are prepared to send (S,G) traffic onto that LAN. This kicks off the assert procedure to elect a winner - the one with the highest Upstream Router's IP Address in the NLRI wins. An assert loser will not include the corresponding LAN interface in its outgoing interface list, but it keeps the S-PMSI A-D route that it originates.

If this router does not have a matching S-PMSI route of its own with some common RTs, and the originator of the received S-PMSI route is a chosen upstream neighbor for the corresponding flow, then this router updates its forwarding state to include the LAN interface in the incoming interface list. When the last S-PMSI route with a RT matching the LAN is withdrawn later, the LAN interface is removed from the incoming interface list.

Note that a downstream router on the LAN does not participate in the assert procedure. It adds/keeps the LAN interface in the expected incoming interfaces as long as its chosen upstream peer originates the S-PMSI AD route. It does not switch to the assert winner as its upstream. An assert loser MAY keep sending joins upstream based on

local policy even if it has no other downstream neighbors (this could be used for fast switch over in case the assert winner would fail).

2.2.6. Distributing Label for Upstream Traffic for Bidirectional Tree/ Tunnel

For MP2MP mLDP tunnels or labeled (*,g) bidirectional trees, an upstream router needs to advertise a label to all its downstream neighbors so that the downstream neighbors can send traffic to itself.

For MP2MP mLDP tunnels, the same procedures for mLDP are followed except that instead of MP2MP-U Label Mapping messages, S-PMSI A-D Routes for C-Multicast mLDP are used.

For labeled (*,g) bidirectional trees, for a Leaf A-D route received from a downstream neighbor, a corresponding S-PMSI A-D route is sent back to the downstream router.

In both cases, a single S-PMSI A-D route is originated for each tree from this router, but with multiple RTs (one for each downstream neighbor on the tree). A TEA specifies a label allocated by the upstream router for its downstream neighbors to send traffic with. Note that this is still a "downstream allocated" label (the upstream router is "downstream" from traffic direction point of view).

The S-PMSI routes do not carry a PTA, unless a P2MP tunnel is used to reach downstream neighbors. Such use case is out of scope of this document for now and may be specified in the future.

3. IANA Considerations

This document requests IANA to assign a new BGP SAFI value for the MCAST-TREE SAFI.

This document requests IANA to create a new "BGP MCAST-TREE Route Types" registry, referencing this document. The following initial values are defined:

```
      0~2 -  Reserved
      3   -  S-PMSI A-D Route for (x,g)
      4   -  Leaf A-D Route
      5   -  Source Active A-D Route
   0x43   -  S-PMSI A-D Route for C-multicast mLDP
```

This document requests IANA to assign two Sub-type values from Transitive IPv4-Address-Specific Extended Community Sub-types

Registry for Session Address EC and Multicast RPF Address EC respectively.

This document requests IANA to assign two Type values from Transitive IPv6-Address-Specific Extended Community Types Registry for Session Address EC and Multicast RPF Address EC respectively.

4. Security Considerations

This document does not introduce new security risks.

5. Acknowledgements

The authors thank Marco Rodrigues for his initial idea/ask of using BGP for multicast signaling beyond MVPN. We thank Eric Rosen for his questions, suggestions, and help finding solutions to some issues. We also thank Luay Jalil and James Uttaro for their comments and support for the work.

6. References

6.1. Normative References

[I-D.ietf-idr-tunnel-encaps]
Patel, K., Velde, G., and S. Ramachandra, "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-15 (work in progress), December 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>.

[RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, DOI 10.17487/RFC4601, August 2006, <https://www.rfc-editor.org/info/rfc4601>.

[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>.

[RFC5015] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", RFC 5015, DOI 10.17487/RFC5015, October 2007, <https://www.rfc-editor.org/info/rfc5015>.

[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>.

[RFC7441] Wijnands, IJ., Rosen, E., and U. Joorde, "Encoding Multipoint LDP (mLDP) Forwarding Equivalence Classes (FECs) in the NLRI of BGP MCAST-VPN Routes", RFC 7441, DOI 10.17487/RFC7441, January 2015, <https://www.rfc-editor.org/info/rfc7441>.

6.2. Informative References

[I-D.ietf-bess-bgp-multicast-controller]
Zhang, Z., Raszuk, R., Pacella, D., and A. Gulko, "Controller Based BGP Multicast Signaling", draft-ietf-bess-bgp-multicast-controller-01 (work in progress), June 2020.

[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-03 (work in progress), October 2019.

[I-D.ietf-bess-mvpn-pe-ce]
Patel, K., Rosen, E., and Y. Rekhter, "BGP as an MVPN PE-CE Protocol", draft-ietf-bess-mvpn-pe-ce-01 (work in progress), October 2015.

[I-D.ietf-spring-segment-routing-policy]
Filsfils, C., Sivabalan, S., Voyer, D., Bogdanov, A., and P. Mattes, "Segment Routing Policy Architecture", draft-ietf-spring-segment-routing-policy-07 (work in progress), May 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-01 (work in progress), April 2020.

[I-D.wijnands-bier-mld-lan-election]
Wijnands, I., Pfister, P., and Z. Zhang, "Generic Multicast Router Election on LAN's", draft-wijnands-bier-mld-lan-election-01 (work in progress), July 2016.

[RFC5496] Wijnands, IJ., Boers, A., and E. Rosen, "The Reverse Path Forwarding (RPF) Vector TLV", RFC 5496, DOI 10.17487/RFC5496, March 2009, <https://www.rfc-editor.org/info/rfc5496>.

[RFC6512] Wijnands, IJ., Rosen, E., Napierala, M., and N. Leymann, "Using Multipoint LDP When the Backbone Has No Route to the Root", RFC 6512, DOI 10.17487/RFC6512, February 2012, <https://www.rfc-editor.org/info/rfc6512>.

[RFC6826] Wijnands, IJ., Ed., Eckert, T., Leymann, N., and M. Napierala, "Multipoint LDP In-Band Signaling for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", RFC 6826, DOI 10.17487/RFC6826, January 2013, <https://www.rfc-editor.org/info/rfc6826>.

[RFC7060] Napierala, M., Rosen, E., and IJ. Wijnands, "Using LDP Multipoint Extensions on Targeted LDP Sessions", RFC 7060, DOI 10.17487/RFC7060, November 2013, <https://www.rfc-editor.org/info/rfc7060>.

[RFC7431] Karan, A., Filsfils, C., Wijnands, IJ., Ed., and B. Decraene, "Multicast-Only Fast Reroute", RFC 7431, DOI 10.17487/RFC7431, August 2015, <https://www.rfc-editor.org/info/rfc7431>.

[RFC7938] Lapukhov, P., Premji, A., and J. Mitchell, Ed., "Use of BGP for Routing in Large-Scale Data Centers", RFC 7938, DOI 10.17487/RFC7938, August 2016, <https://www.rfc-editor.org/info/rfc7938>.

[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>.

Authors' Addresses

Zhaohui Zhang
Juniper Networks

EMail: zzhang@juniper.net

Lenny Giuliano
Juniper Networks

EMail: lenny@juniper.net

Keyur Patel
Arrcus

EMail: keyur@arrcus.com

IJsbrand Wijnands
Cisco Systems

EMail: ice@cisco.com

Mankamana Mishra
Cisco Systems

EMail: mankamis@cisco.com

Arkadiy Gulko
Refinitiv

EMail: arkadiy.gulko@refinitiv.com

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

Z. Zhang
Juniper Networks
R. Raszuk
Bloomberg LP
D. Pacella
Verizon
A. Gulko
Refinitiv

September 22, 2020

Controller Based BGP Multicast Signaling
draft-ietf-bess-bgp-multicast-controller-05

Abstract

This document specifies a way that one or more centralized controllers can use BGP to set up a multicast distribution tree in a network. In the case of labeled tree, the labels are assigned by the controllers either from the controllers' local label spaces, or from a common Segment Routing Global Block (SRGB), or from each routers Segment Routing Local Block (SRLB) that the controllers learn. In case of labeled unidirectional tree and label allocation from the common SRGB or from the controllers' local spaces, a single common label can be used for all routers on the tree to send and receive traffic with. Since the controllers calculate the trees, they can use sophisticated algorithms and constraints to achieve traffic engineering.

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.

Status of This Memo

Table of Contents

1. Overview

1.1. Introduction

[I-D.ietf-bess-bgp-multicast] describes a way to use BGP as a replacement signaling for PIM [RFC7761] or mLDP [RFC6388]. The BGP-based multicast signaling described there provides a mechanism for setting up both (s,g)/(*,g) multicast trees (as PIM does, but optionally with labels) and labeled (MPLS) multicast tunnels (as mLDP does). Each router on a tree performs essentially the same procedures as it would perform if using PIM or mLDP, but all the inter-router signaling is done using BGP.

These procedures allow the routers to set up a separate tree for each individual multicast (x,g) flow where the 'x' could be either 's' or '*', but they also allow the routers to set up trees that are used for more than one flow. In the latter case, the trees are often referred to as "multicast tunnels" or "multipoint tunnels", and specifically in this document they are mLDP tunnels (except that they are set up with BGP signaling). While it actually does not have to be restricted to mLDP tunnels, mLDP FEC is conveniently borrowed to identify the tunnel. In the rest of the document, the term tree and tunnel are used interchangeably.

The trees/tunnels are set up using the "receiver-initiated join" technique of PIM/mLDP, hop by hop from downstream routers towards the root. The BGP messages are either sent hop by hop between downstream routers and their upstream neighbors, or can be reflected by Route Reflectors (RRs).

As an alternative to each hop independently determining its upstream router and signaling upstream towards the root (following PIM/mLDP model), the entire tree can be calculated by a centralized controller, and the signaling can be entirely done from the controller, using the same BGP messages as defined in [I-D.ietf-bess-bgp-multicast]. For that, some additional procedures and optimizations are specified in this document.

While it is outside the scope of this document, signaling from the controllers could be done via other means as well, like Netconf or any other SDN methods.

1.2. Resilience

Each router could establish direct BGP sessions with one or more controllers, or it could establish BGP sessions with RRs who in turn peer with controllers. For the same tree/tunnel, each controller may independently calculate the tree/tunnel and signal the routers on the tree/tunnel using MCAST-TREE Leaf A-D routes [I-D.ietf-bess-bgp-multicast]. How the tree/tunnel roots/leaves are discovered and how the calculation is done are outside the scope of this document.

On each router, BGP route selection rules will lead to one controller's route for the tree/tunnel being selected as the active route and used for setting up forwarding state. As long as all the routers on a tree/tunnel consistently pick the same controller's routes for the tree/tunnel, the setup should be consistent. If the tree/tunnel is labeled, different labels will be used from different controllers so there is no traffic loop issue even if the routers do not consistently select the same controlle's routes. In the unlabeled case, to ensure the consistency the selection SHOULD be solely based on the identifier of the controller, which could be carried in an Address Specific Extended Community (EC).

Another consistency issue is when a bidirectional tree/tunnel needs to be re-routed. Because this is no longer triggered hop-by-hop from downstream to upstream, it is possible that the upstream change happens before the downstream, causing traffic loop. In the unlabeled case, there is no good solution (other than that the controller issues upstream change only after it gets acknowledgement from downstream). In the labeled case, as long as a new label is used there should be no problem.

Besides the traffic loop issue, there could be transient traffic loss before both the upstream and downstream's forwarding state are updated. This could be mitigated if the upstream keep sending traffic on the old path (in addition to the new path) and the downstream keep accepting traffic on the old path (but not on the new path) for some time. It is a local matter when for the downstream to switch to the new path - it could be data driven (e.g., after traffic arrives on the new path) or timer driven.

For each tree, multiple disjoint instances could be calculated and signaled for live-live protection. Different labels are used for different instances, so that the leaves can differentiate incoming

traffic on different instances. As far as transit routers are concerned, the instances are just independent. Note that the two instances are not expected to share common transit routers (it is otherwise outside the scope of this document/revision).

1.3. Signaling

Each router only receives Leaf A-D routes from the controllers but does not originate or re-advertise S-PMSI/Leaf A-D routes. The re-advertisement of a received route can be blocked based on the fact that a configured import RT matches the RT of the route, which indicates that this router is the target and consumer of the route hence it should not be re-advertised further. The routes includes the forwarding information in the form of Tunnel Encapsulation Attributes (TEA) [I-D.ietf-idr-tunnel-encaps], with enhancements specified in this document.

Suppose that for a particular tree, there are two downstream routers D1 and D2 for a particular upstream router U. A controller C may send two Leaf A-D routes to U, as if the two routes were originated by D1 and D2 but reflected by the controller. Alternatively, C could just send one route to U, with the Upstream Router's IP Address field set to U's IP address and the TEA specifying both the two downstreams and its upstream (see Section 2.1.4). In this case, the Originating Router's Address field of the Leaf A-D route is set to the controller's address. Note that for a TEA attached to a unicast NLRI, only one of the tunnels in a TEA is used for forwarding a particular packet, while all the tunnels in a TEA are used to reach multiple endpoints when it is attached to a multicast NLRI.

Note that, in case of labeled trees, the (x,g) or mLDP FEC signaling is actually not needed to transit routers but only needed on tunnel root/leaves. However, for consistency, the same signaling is used to all routers.

1.4. Label Allocation

In the case of labeled multicast signaled hop by hop towards the root, whether it's (x,g) multicast or "mLDP" tunnel, labels are assigned by a downstream router and advertised to its upstream router (from traffic direction point of view). In the case of controller based signaling, routers do not originate tree join (S-PMSI/Leaf A-D) routes anymore, so the controllers have to assign labels on behalf of routers, and there are three options for label assignment:

- From each router's SRLB that the controller learns
- From the common SRGB that the controller learns

o From the controller's local label space

Assignment from each router's SRLB is no different from each router assigning labels from its own local label space in the hop-by-hop signaling case. The assignments for a router is independent of assignments for another router, even for the same tree.

Assignment from the controller's local label space is upstream-assigned [RFC5331]. It is used if the controller does not learn the common SRGB or each router's SRLB. Assignment from the SRGB [RFC8402] is only meaningful if all SRGBs are the same and a single common label is used for all the routers on a tree in case of unidirectional tree/tunnel (Section 1.4.1). Otherwise, assignment from SRLB is preferred.

The choice of which of the options to use depends on many factors. An operator may want to use a single common label per tree for ease of monitoring and debugging, but that requires explicit RPF checking and either SRGB or upstream assigned labels, which may not be supported due to either the software or hardware limitations (e.g. label imposition/disposition limits). In an SR network, assignment from the common SRGB if it's required to use a single common label per unidirectional tree, or otherwise assignment from SRLB is a good choice because it does not require support for context label spaces.

1.4.1. Using a Common per-tree Label for All Routers

MPLS labels only have local significance. For an LSP that goes through a series of routers, each router allocates a label independently and it swaps the incoming label (that it advertised to its upstream) to an outgoing label (that it received from its downstream) when it forwards a labeled packet. Even if the incoming and outgoing labels happen to be the same on a particular router, that is just incidental.

With Segment Routing, it is becoming a common practice that all routers use the same SRGB so that a SID maps to the same label on all routers. This makes it easier for operators to monitor and debug their network. The same concept applies to multicast trees as well - a common per-tree label is used for a router to receive traffic from its upstream neighbor and replicate traffic to all its downstream neighbor.

However, a common per-tree label can only be used for unidirectional trees. Additionally, it requires each router to do explicit RPF check, so that only packets from its expected upstream neighbor are accepted. Otherwise, traffic loop may form during topology changes, because the forwarding state update is no longer ordered.

Traditionally, p2mp mpls forwarding does not require explicit RPF check as a downstream router advertises a label only to its upstream router and all traffic with that incoming label is presumed to be from the upstream router and accepted. When a downstream router switches to a different upstream router a different label will be advertised, so it can determine if traffic is from its expected upstream neighbor purely based on the label. Now with a single common label used for all routers on a tree to send and receive traffic with, a router can no longer determine if the traffic is from its expected neighbor just based on that common tree label. Therefore, explicit RPF check is needed. Instead of interface based RPF checking as in PIM case, neighbor based RPF checking is used - a label identifying the upstream neighbor precedes the tree label and the receiving router checks if that preceding neighbor label matches its expected upstream neighbor. Notice that this is similar to what's described in Section "9.1.1 Discarding Packets from Wrong PE" of RFC 6513 (an egress PE discards traffic sent from a wrong ingress PE). The only difference is one is used for label based forwarding and the other is used for (s,g) based forwarding. [note: for bidirectional trees, we may be able to use two labels per tree - one for upstream traffic and one for downstream traffic. This needs further verification].

Both the common per-tree label and the neighbor label are allocated either from the common SRGB or from the controller's local label space. In the latter case, an additional label identifying the controller's label space is needed, as described in the following section.

1.4.2. Upstream-assignment from Controller's Local Label Space

In this case in the multicast packet's label stack the tree label and upstream neighbor label (if used in case of single common-label per tree) are preceded by a downstream-assigned "context label". The context label identifies a context-specific label space (the controller's local label space), and the upstream-assigned label that follows it is looked up in that space.

This specification requires that, in case of upstream-assignment from a controller's local label space, each router D to assign, corresponding to each controller C, a context label that identifies the upstream-assigned label space used by that controller. This label, call it Lc-D, is communicated by D to C via BGP-LS [RFC 7752].

Suppose a controller is setting up unidirectional tree T. It assigns that tree the label Lt, and assigns label Lu to identify router U which is the upstream of router D on tree T. C needs to tell U: "to send a packet on the given tree/tunnel, one of the things you have to

do is push Lt onto the packet's label stack, then push Lu, then push Lc-D onto the packet's label stack, then unicast the packet to D". Controller C also needs to inform router D of the correspondence between <Lc-D, Lu, Lt> and tree T.

To achieve that, when C sends a Leaf A-D route, for each tunnel in the TEA, it includes a label stack Sub-TLV [I-D.ietf-idr-tunnel-encaps], with the outer label being the context label Lc-D (received by the controller from the corresponding downstream), the next label being the upstream neighbor label Lu, and the inner label being the label Lt assigned by the controller for the tree. The router receiving the route will use the label stacks to send traffic to its downstreams.

For C to signal the expected label stack for D to receive traffic with, we overload a tunnel TLV in the TEA of the Leaf A-D route sent to D - if the tunnel TLV has a RPF sub-TLV (Section 2.1.4), then it indicates that this is actually for receiving traffic from the upstream.

1.5. Determining Root/Leaves

For the controller to calculate a tree, it needs to determine the root and leaves of the tree. This may be based on provisioning (static or dynamically programmed), or based on BGP signaling using the BGP multicast messages defined in [I-D.ietf-bess-bgp-multicast], as described in the following two sections.

In both cases, the BGP updates are targeted at the controller, via an address specific Route Target with Global Administration Field set to the controller's address and the Local Administration Field set to 0, or a value pre-assigned to identify a VPN.

1.5.1. PIM-SSM/Bidir or mLDP

In this case, the PIM Last Hop Routers (LHRs) with interested receivers or mLDP tunnel leaves encode a Leaf A-D route with the Upstream Router's IP Address field set to the controller's address and the Originating Router's IP Address set to the address of the LHR or the P2MP tunnel leaf. The encoded PIM SSM source or mLDP FEC provides root information and the Originating Router's IP Address provides leaves information.

1.5.2. PIM ASM

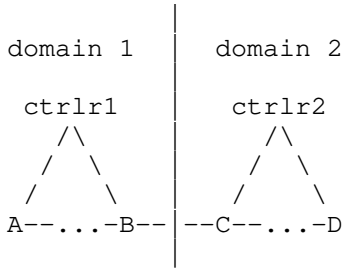
In this case, the First Hop Routers (FHRs) originate Source Active routes which provides root information, and the LHRs originate Leaf

A-D routes, encoded as in the PIM-SSM case except that it is (*,G) instead of (S,G). The Leaf A-D routes provide leaf information.

1.6. Multiple Domains

An end to end multicast tree may span multiple routing domains, and the setup of the tree in each domain may be done differently as specified in [I-D.ietf-bess-bgp-multicast]. This section discusses a few aspects specific to controller signaling.

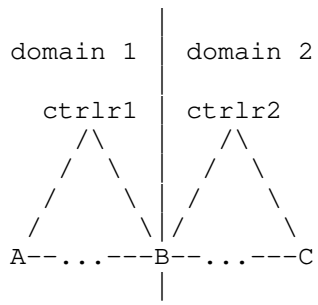
Consider two adjacent domains each with its own controller in the following configuration where router B is an upstream node of C for a multicast tree:



In the case of native (un-labeled) IP multicast, nothing special is needed. Controller 1 signals B to send traffic out of B-C link while Controller 2 signals C to accept traffic on the B-C link.

In the case of labeled IP multicast or mLDP tunnel, the controllers may be able to coordinate their actions such that Controller 1 signals B to send traffic out of B-C link with label X while Controller 2 signals C to accept traffic with the same label X on the B-C link. If the coordination is not possible, then C needs to use hop-by-hop BGP signaling to signal towards B, as specified in [I-D.ietf-bess-bgp-multicast].

The configuration could also be as following, where router B borders both domain 1 and domain 2 and is controlled by both controllers:



As discussed in Section 1.2, when B receives signaling from both Controller 1 and Controller 2, only one of the routes would be selected as the best route and used for programming the forwarding state of the corresponding segment. For B to stitch the two segments together, it is expected for B to know by provisioning that it is a border router so that B will look for the other segment (represented by the signaling from the other controller) and stitch the two together.

1.7. SR-P2MP

[I-D.voyer-pim-sr-p2mp-policy] describes an architecture to construct a Point-to-Multipoint (P2MP) tree to deliver Multi-point services in a Segment Routing domain. An SR P2MP tree is constructed by stitching together a set of Replication Segments that are specified in [I-D.voyer-spring-sr-replication-segment]. An SR Point-to-Multipoint (SR P2MP) Policy is used to define and instantiate a P2MP tree which is computed by a controller.

An SR P2MP tree is no different from an mLDP tunnel in MPLS forwarding plane. The difference is in control plane - instead of hop-by-hop mLDP signaling from leaves towards the root, to set up SR P2MP trees controllers program forwarding state (referred to as Replication Segments) to the root, leaves, and intermediate replication points using Netconf, PCEP, BGP or any other reasonable signaling/programming methods.

Procedures in this document can be used for controllers to set up SR P2MP trees with just an additional S-PMSI route type.

If/once the SR Replication Segment is extended to bi-redirectional, and SR MP2MP is introduced, the same procedures in this document would apply to SR MP2MP as well.

2. Specification

2.1. Enhancements to TEA

This document specifies two new Tunnel Types and four new sub-TLVs. The type codes will be assigned by IANA from the "BGP Tunnel Encapsulation Attribute Tunnel Types".

2.1.1. Any-Encapsulation Tunnel

When a multicast packet needs to be sent from an upstream node to a downstream node, it may not matter how it is sent - natively when the two nodes are directly connected or tunneled otherwise. In case of tunneling, it may not matter what kind of tunnel is used - MPLS, GRE, IPinIP, or whatever.

To support this, an "Any-Encapsulation" tunnel type is defined. This tunnel MUST have a Tunnel Endpoint Sub-TLV and SHOULD NOT have any other Sub-TLVs. The Tunnel Endpoint Sub-TLV specifies an IP address, which could be any of the following:

- An interface's local address - when a packet needs to sent out of the corresponding interface natively. On a LAN multicast MAC address MUST be used.

- A directly connected neighbor's interface address - when a packet needs to unicast to the address natively.

- An address that is not directly connected - when a packet needs to be tunneled to the address (any tunnel type/instance can be used).

2.1.2. Load-balancing Tunnel

Consider that a multicast packet needs to be sent to a downstream node, which could be reached via four paths P1~P4. If it does not matter which of path is taken, an "Any-Encapsulation" tunnel with the Tunnel Endpoint Sub-TLV specifying the downstream node's loopback address works well. If the controller wants to specify that only P1~P2 should be used, then a "Load-balancing" tunnel needs to be used, listing P1 and P2 as member tunnels of the "Load-balancing" tunnel.

A load-balancing tunnel has one "Member Tunnels" Sub-TLV defined in this document. The Sub-TLV is a list of tunnels, each specifying a way to reach the downstream. A packet will be sent out of one of the tunnels listed in the Member Tunnels Sub-TLV of the load-balancing tunnel.

2.1.3. Receiving MPLS Label Stack

While [I-D.ietf-bess-bgp-multicast] uses S-PMSI A-D routes to signal forwarding information for MP2MP upstream traffic, when controller signaling is used, a single Leaf A-D route is used for both upstream and downstream traffic. Since different upstream and downstream labels need to be used, a new "Receiving MPLS Label Stack" of type TBD is added as a tunnel sub-TLV in addition to the existing MPLS Label Stack sub-TLV. Other than type difference, the two are the encoded the same way.

The Receiving MPLS Label Stack sub-TLV is added to each downstream tunnel in the TEA of Leaf A-D route for an MP2MP tunnel to specify the forwarding information for upstream traffic from the corresponding downstream node. A label stack instead of a single label is used because of the need for neighbor based RPF check, as further explained in the following section.

The Receiving MPLS Label Stack sub-TLV is also used for downstream traffic from the upstream for both P2MP and MP2MP, as specified below.

2.1.4. RPF Sub-TLV

The RPF sub-TLV has a type to be allocated by IANA and a one-octet length. The length is 0 currently, but if necessary in the future, sub-sub-TLVs could be placed in its value part. If the RPF sub-TLV appears in a tunnel, it indicates that the "tunnel" is for the upstream node instead of a downstream node. The tunnel contains an Receiving MPLS Label Stack sub-TLV for downstream traffic from the upstream node, and in case of MP2MP it also contains a regular MPLS Label Stack sub-TLV for upstream traffic to the upstream node.

The inner most label in the Receiving MPLS Label Stack is the incoming label identifying the tree (for comparison the inner most label for a regular MPLS Label Stack is the outgoing label). If the Receiving MPLS Label Stack sub-TLVe has more than one labels, the second inner most label in the stack identifies the expected upstream neighbor and explicit RPF checking needs to be set up for the tree label accordingly.

2.1.5. Tree Label Stack sub-TLV

The MPLS Label Stack sub-TLV can be used to specify the complete label stack used to send traffic, with the stack including both a transport label (stack) and label(s) that identify the (tree, neighbor) to the downstream node. There are cases where the controller only wants to specify the tree-identifying labels but

leave the transport details to the router itself. For example, the router could locally determine a transport label (stack) and combine with the tree-identifying labels signaled from the controller to get the complete outgoing label stack.

For that purpose, a new Tree Label Stack sub-TLV is defined, with a one-octet length field. The value field contains a label stack with the same encoding as value part of the MPLS Label Stack sub-TLV, but the sub-TLV has a different type. A stack is specified because it may take up to three labels (see Section 1.4):

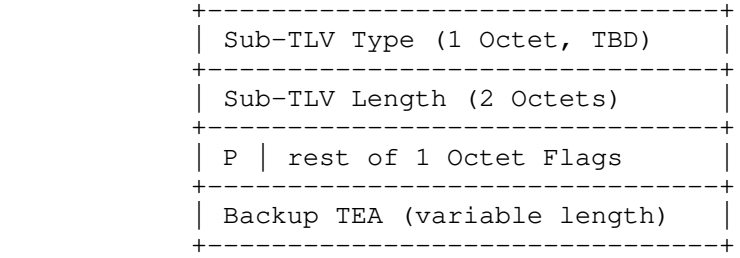
- o If different nodes use different labels (allocated from the common SRGB or the node's SRLB) for a (tree, neighbor) tuple, only a single label is in the stack. This is similar to current mLDP hop by hop signaling case.

- o If different nodes use the same tree label, then an additional neighbor-identifying label is needed in front of the tree label.

- o For the previous bullet, if the neighbor-identifying label is allocated from the controller's local label space, then an additional context label is needed in front of the neighbor label.

2.1.6. Backup Tunnel sub-TLV

The Backup Tunnel sub-TLV is used to specify the backup paths for the tunnel. The length is two-octet. The value part encodes a one-octet flags field and a variable length Tunnel Encapsulation Attribute. If the tunnel goes down, traffic that is normally sent out of the tunnel is fast rerouted to the tunnels listed in the encoded TEA.



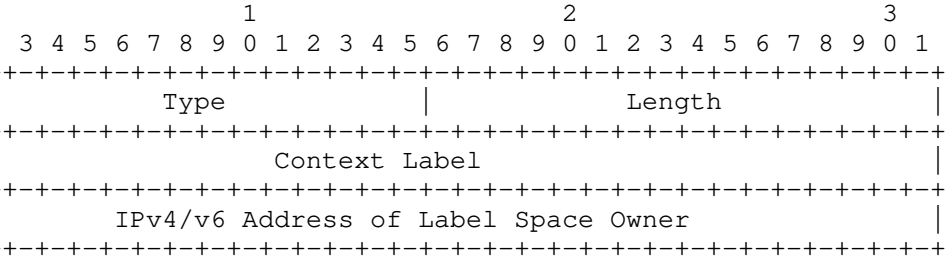
The backup tunnels can be going to the same or different nodes reached by the original tunnel.

If the tunnel carries a RPF sub-TLV and a Backup Tunnel sub-TLV, then both traffic arriving on the original tunnel and on the tunnels encoded in the Backup Tunnel sub-TLV's TEA can be accepted, if the Parallel (P-)bit in the flags field is set. If the P-bit is not set,

then traffic arriving on the backup tunnel is accepted only if router has switched to receiving on the backup tunnel (this is the equivalent of PIM/mLDP MoFRR).

2.2. Context Label TLV in BGP-LS Node Attribute

For a router to signal the context label that it assigns for a controller (or any label allocator that assigns labels - from its local label space -- that will be received by this router), a new BGP-LS Node Attribute TLV is defined:



The Length field implies the type of the address. Multiple Context Label TLVs may be included in a Node Attribute, one for each label space owner.

An as example, a controller with address 11.11.11.11 allocates label 200 from its own label space, and router A assigns label 100 to identify this controller's label space. The router includes the Context Label TLV (100, 11.11.11.11) in its BGP-LS Node Attribute and the controller instructs router B to send traffic to router A with a label stack (100, 200), and router A uses label 100 to determine the Label FIB in which to look up label 200.

2.3. SR P2MP Signaling

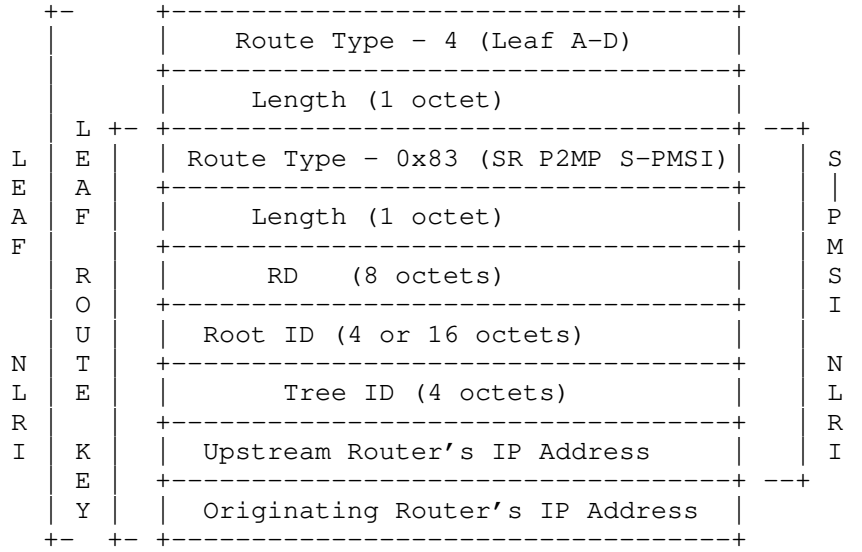
An SR P2MP policy for an SR P2MP tree is identified by a (Root, Tree-id) tuple. It has a set of leaves and set of Candidate Paths (CPs). The policy is instantiated on the root of the tree, with corresponding Replication Segments - identified by (Root, Tree-id, Tree-Node-id) - instantiated on the tree nodes (root, leaves, and intermediate replication points). The Candidate Path is implicitly identified by the Route Distinguisher.

2.3.1. S-PMSI A-D Route for SR P2MP

With BGP signaled IP multicast trees and mLDP tunnels, the tree/ tunnel identification is encoded in the NLRI of S-PMSI A-D routes and corresponding Leaf A-D routes. The signaling sets up forwarding

state on each node of the tree, so the NLRI also contains the identification of the node in the "Upstream Router's IP Address" field.

For SR P2MP, forwarding state are represented as Replication Segments and are signaled from controllers to tree nodes. A Replication Segment is identified in a new type of S-PMSI A-D route and corresponding Leaf A-D route (note that the "Leaf" term here does not refer to tree leaves):

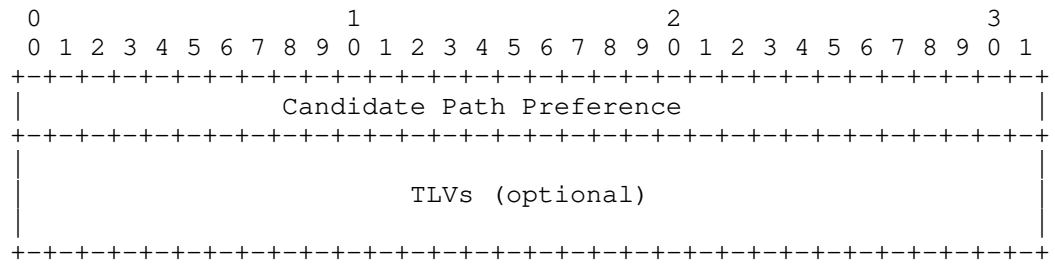


Leaf A-D route for SR Replication Segment

2.3.2. BGP Community Container for SR P2MP Policy

The Leaf A-D route for Replication Segments signaled to the root is also used to signal (parts of) the SR P2MP Policy - the policy name, the set of leaves (optional, for informational purpose), preference of the CP and other information are all encoded in a newly defined BGP Community Container (BCC) [I-D.ietf-idr-wide-bgp-communities] called SR P2MP Policy BCC.

The SR P2MP Policy BCC has a BGP Community Container type to be assigned by IANA. It is composed of a fixed 4-octet Candidate Path Preference value, optionally followed by TLVs.



BGP Community Container for SR P2MP Policy

One optional TLV is to enclose the following optional Atoms TLVs that are already defined in [I-D.ietf-idr-wide-bgp-communities]:

- An IPv4 or IPv6 Prefix list - for the set of leaves
- A UTF-8 string - for the policy name

If more information for the policy are needed, more Atoms TLVs or SR P2MP Policy BCC specific TLVs can be defined.

The root receives one Leaf A-D route for each Candidate Path of the policy. Only one of the routes need to, though more than one MAY include the above listed optional Atom TLVs in the SR P2MP Policy BCC.

2.3.3. SR Policy Tunnel Type

The Tunnel Encapsulation Attribute (TEA) attached to Leaf A-D routes encodes all replication branch information. For example, if an SR explicit path is to be used to reach a particular downstream node, the TEA will include a tunnel that lists the entire label stack for that SR path, plus the label that identifies the SR P2MP tree to the downstream node.

That SR path may have been installed on the node as a unicast SR policy with a corresponding Binding SID. In stead of listing the entire label stack in an MPLS tunnel in the TEA, a different tunnel, SR Policy Tunnel [I-D.ietf-idr-segment-routing-te-policy], can be used as an alternative. The tunnel includes a Binding SID sub-TLV, an optional endpoint sub-TLV that identifies the downstream node, and an optional one-segment segment list that identifies to the downstream node the SR P2MP tree. When a node receives the Leaf A-D route with the TEA that contains an SR Policy Tunnel without a RPF sub-TLV, the Binding SID is used to locate corresponding outgoing segment lists used to reach the downstream node; the tree-identifying

segment from the optional one-segment segment list is added to to outgoing segment lists mapped from the binding SID to form the entire segment list used to send traffic to downstream node.

Note that, the SR Policy Tunnel is initially defined to instantiate an SR policy. For that use case it provides information associated with the policy, e.g., Binding SID, preference, and segment lists. The receiving node installs that policy and establishes the mapping from the Binding SID to the outgoing segments. The use of SR Policy Tunnel in this document is to refer to a pre-installed SR policy so the preference and segment lists are not used.

If a tunnel in the TEA carries a RPF sub-TLV, it is for the upstream node. The tunnel may be an MPLS tunnel in case of SR MPLS, and the Receiving MPLS Label Stack sub-TLV specifies the incoming label stack that identifies the tree and optionally the upstream neighbor. Alternatively, for both SR-MPLS and SRv6 an SR Policy Tunnel with the RPF sub-TLV can be used, in which the Binding SID sub-TLV is the SID for the tree.

If the node is the root and a Binding SID is allocated by the controller, the Binding SID is signaled to the root in a TEA tunnel with a RPF sub-TLV as above but without a destination sub-TLV.

3. Procedures

Details to be added. The general idea is described in the introduction section.

4. Security Considerations

This document does not introduce new security risks.

5. IANA Considerations

This document makes the following IANA requests:

- Assign "Any-Encapsulation" and "Load-balancing" tunnel types from the "BGP Tunnel Encapsulation Attribute Tunnel Types" registry

- Assign "Member Tunnels", "Receiving MPLS Label Stack", "Tree Label Stack" and "RPF" sub-TLV types from the "BGP Tunnel Encapsulation Attribute Sub-TLVs" registry. The "Member Tunnels" sub-TLV has a two-octet value length (so the type should be in the 128-255 range), while the "Receiving MPLS Label Stack", "Tree Label" and "RPF" sub-TLV has a one-octet value length.

- Assign "Context Label TLV" type from the "BGP-LS Node Descriptor, Link Descriptor, Prefix Descriptor, and Attribute TLVs" registry.

- Assign "S-PMSI A-D Route for SR P2MP" route type from the "BGP MCAST-TREE Route Types" registry, with a suggested value of 0x83.

- Assign a new BGP Community Container type "SR P2MP Policy", and to create an "SR P2MP Policy Community Container TLV Registry", with an initial entry for "TLV for Atoms".

6. Acknowledgements

The authors Eric Rosen for his questions, suggestions, and help finding solutions to some issues like the neighbor based explicit RPF checking. The authors also thank Lenny Giuliano, Sanoj Vivekanandan and IJsbrand Wijnands for their review and comments.

7. References

7.1. Normative References

[I-D.ietf-bess-bgp-multicast]
Zhang, Z., Giuliano, L., Patel, K., Wijnands, I., mishra, m., and A. Gulko, "BGP Based Multicast", draft-ietf-bess-bgp-multicast-02 (work in progress), June 2020.

[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-09 (work in progress), May 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-17 (work in progress), July 2020.

[I-D.ietf-idr-wide-bgp-communities]
Raszuk, R., Haas, J., Lange, A., Decraene, B., Amante, S., and P. Jakma, "BGP Community Container Attribute", draft-ietf-idr-wide-bgp-communities-05 (work in progress), July 2018.

[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.

[I-D.voyer-spring-sr-replication-segment]
Voyer, D., Filsfils, C., Parekh, R., Bidgoli, H., and Z. Zhang, "SR Replication Segment for Multi-point Service Delivery", draft-voyer-spring-sr-replication-segment-04 (work in progress), July 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>.

[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>.

7.2. Informative References

[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>.

[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>.

[RFC7761] Fenner, B., Handley, M., Holbrook, H., Kouvelas, I., Parekh, R., Zhang, Z., and L. Zheng, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", STD 83, RFC 7761, DOI 10.17487/RFC7761, March 2016, <https://www.rfc-editor.org/info/rfc7761>.

[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>.

Authors' Addresses

Zhaohui Zhang
Juniper Networks

EMail: zzhang@juniper.net

Robert Raszuk
Bloomberg LP

EMail: robert@raszuk.net

Dante Pacella
Verizon

EMail: dante.j.pacella@verizon.com

Arkadiy Gulko
Refinitiv

EMail: arkadiy.gulko@refinitiv.com

BESS WorkGroup
Internet-Draft
Intended status: Standards Track
Expires: January 28, 2021

N. Malhotra, Ed.
A. Sajassi
Cisco Systems
J. Rabadan
Nokia
J. Drake
Juniper
A. Lingala
ATT
S. Thoria
Cisco Systems
July 27, 2020

Weighted Multi-Path Procedures for EVPN All-Active Multi-Homing
draft-ietf-bess-evpn-unequal-lb-06

Abstract

In an EVPN-IRB based network overlay, EVPN all-active multi-homing enables multi-homing for a CE device connected to two or more PEs via a LAG, such that bridged and routed traffic from remote PEs can be equally load balanced (ECMPed) across the multi-homing PEs. This document defines extensions to EVPN procedures to optimally handle unequal access bandwidth distribution across a set of multi-homing PEs in order to:

- provide greater flexibility, with respect to adding or removing individual PE-CE links within the access LAG.
- handle PE-CE LAG member link failures that can result in unequal PE-CE access bandwidth across a set of multi-homing PEs.

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 January 28, 2021.

Table of Contents

1. Requirements Language and Terminology

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

"Local PE" in the context of an ESI refers to a provider edge switch OR router that physically hosts the ESI.

"Remote PE" in the context of an ESI refers to a provider edge switch OR router in an EVPN overlay, whose overlay reachability to the ESI is via the Local PE.

- BW: Band-Width
- LAG: Link Aggregation Group
- ES: Ethernet Segment
- vES: Virtual Ethernet Segment
- EVI: Ethernet virtual Instance, this is a mac-vrf.
- IMET: Inclusive Multicast Route
- DF: Designated Forwarder
- BDF: Backup Designated Forwarder
- DCI: Data Center Interconnect Router

2. Introduction

In an EVPN-IRB based network overlay, with a CE multi-homed via a EVPN all-active multi-homing, bridged and routed traffic from remote PEs can be equally load balanced (ECMPed) across the multi-homing PEs:

- ECMP Load-balancing for bridged unicast traffic is enabled via
- aliasing and mass-withdraw procedures detailed in RFC 7432.
- ECMP Load-balancing for routed unicast traffic is enabled via existing L3 ECMP mechanisms.
- Load-sharing of bridged BUM traffic on local ports is enabled via EVPN DF election procedure detailed in RFC 7432

All of the above load balancing and DF election procedures implicitly assume equal bandwidth distribution between the CE and the set of multi-homing PEs. Essentially, with this assumption of equal "access" bandwidth distribution across all PEs, ALL remote traffic is equally load balanced across the multi-homing PEs. This assumption of equal access bandwidth distribution can be restrictive with respect to adding / removing links in a multi-homed LAG interface and may also be easily broken on individual link failures. A solution to handle unequal access bandwidth distribution across a set of multi-homing EVPN PEs is proposed in this document. Primary motivation behind this proposal is to enable greater flexibility with respect to adding / removing member PE-CE links, as needed and to optimally handle PE-CE link failures.

2.1. PE-CE Link Provisioning

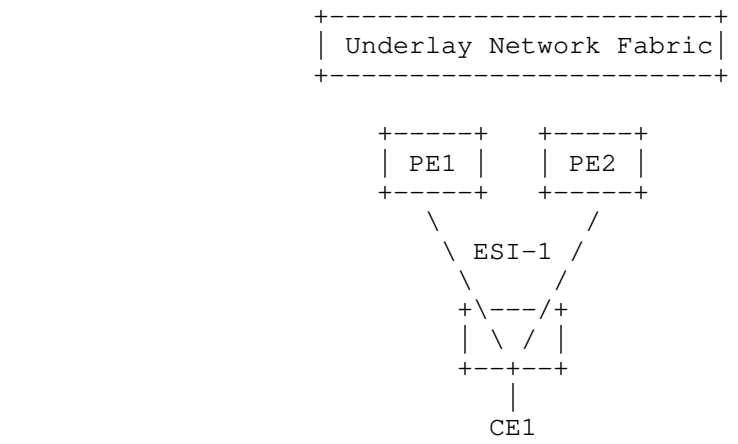


Figure 1

Consider CE1 that is dual-homed to PE1 and PE2 via EVPN all-active multi-homing with single member links of equal bandwidth to each PE (aka, equal access bandwidth distribution across PE1 and PE2). If the provider wants to increase link bandwidth to CE1, it must add a link to both PE1 and PE2 in order to maintain equal access bandwidth distribution and inter-work with EVPN ECMP load balancing. In other words, for a dual-homed CE, total number of CE links must be provisioned in multiples of 2 (2, 4, 6, and so on). For a triple-homed CE, number of CE links must be provisioned in multiples of three (3, 6, 9, and so on). To generalize, for a CE that is multi-homed to "n" PEs, number of PE-CE physical links provisioned must be an integral multiple of "n". This is restrictive in case of dual-homing and very quickly becomes prohibitive in case of multi-homing.

Instead, a provider may wish to increase PE-CE bandwidth OR number of links in any link increments. As an example, for CE1 dual-homed to PE1 and PE2 in all-active mode, provider may wish to add a third link to only PE1 to increase total bandwidth for this CE by 50%, rather than being required to increase access bandwidth by 100% by adding a link to each of the two PEs. While existing EVPN based all-active load balancing procedures do not necessarily preclude such asymmetric access bandwidth distribution among the PEs providing redundancy, it may result in unexpected traffic loss due to congestion in the access interface towards CE. This traffic loss is due to the fact that PE1 and PE2 will continue to be treated as equal cost paths at remote PEs, and as a result may attract approximately equal amount of CE1 destined traffic, even when PE2 only has half the bandwidth to CE1 as PE1. This may lead to congestion and traffic loss on the PE2-CE1 link. If bandwidth distribution to CE1 across PE1 and PE2 is 2:1, traffic from remote hosts must also be load balanced across PE1 and PE2 in 2:1 manner.

2.2. PE-CE Link Failures

More importantly, unequal PE-CE bandwidth distribution described above may occur during regular operation following a link failure, even when PE-CE links were provisioned to provide equal bandwidth distribution across multi-homing PEs.

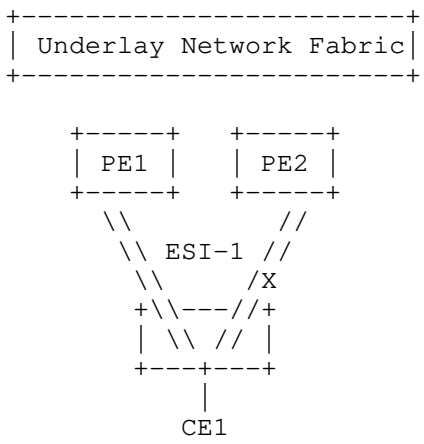


Figure 2

Consider a CE1 that is multi-homed to PE1 and PE2 via a LAG with two member links to each PE. On a PE2-CE1 physical link failure, LAG represented by an Ethernet Segment ESI-1 on PE2 stays up, however, its bandwidth is cut in half. With existing ECMP procedures, both PE1 and PE2 may continue to attract equal amount of traffic from

remote PEs, even when PE1 has double the bandwidth to CE1. If bandwidth distribution to CE1 across PE1 and PE2 is 2:1, traffic from remote hosts must also be load balanced across PE1 and PE2 in 2:1 manner to avoid unexpected congestion and traffic loss on PE2-CE1 links within the LAG. As an alternative, min-link on LAGs is sometimes used to bring down the LAG interface on member link failures. This however results in loss of available bandwidth in the network, and is not ideal.

2.3. Design Requirement

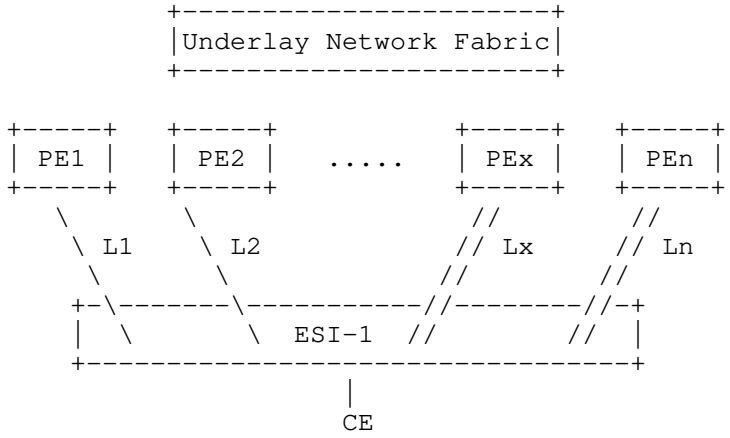


Figure 3

To generalize, if total link bandwidth to a CE is distributed across "n" multi-homing PEs, with Lx being the total bandwidth to PEx across all links, traffic from remote PEs to this CE must be load balanced unequally across [PE1, PE2, ..., PEn] such that, fraction of total unicast and BUM flows destined for CE that are serviced by PEx is:

Lx / [L1+L2+.....+Ln]

The solution proposed below includes extensions to EVPN procedures to achieve the above.

3. Solution Overview

In order to achieve weighted load balancing for overlay unicast traffic, Ethernet A-D per-ES route (EVPN Route Type 1) is leveraged to signal the Ethernet Segment bandwidth to remote PEs. Using Ethernet A-D per-ES route to signal the Ethernet Segment bandwidth provides a mechanism to be able to react to changes in access bandwidth in a service and host independent manner. Remote PEs

computing the MAC path-lists based on global and aliasing Ethernet A-D routes now have the ability to setup weighted load balancing path-lists based on the ESI access bandwidth received from each PE that the ESI is multi-homed to. If Ethernet A-D per-ES route is also leveraged for IP path-list computation, as per [EVPN-IP-ALIASING], it also provides a method to do weighted load balancing for IP routed traffic.

In order to achieve weighted load balancing of overlay BUM traffic, EVPN ES route (Route Type 4) is leveraged to signal the ESI bandwidth to PEs within an ESI's redundancy group to influence per-service DF election. PEs in an ESI redundancy group now have the ability to do service carving in proportion to each PE's relative ESI bandwidth.

Procedures to accomplish this are described in greater detail next.

4. Weighted Unicast Traffic Load-balancing

4.1. Local PE Behavior

A PE that is part of an Ethernet Segment's redundancy group would advertise an additional "link bandwidth" extended community attribute with Ethernet A-D per-ES route (EVPN Route Type 1), that represents total bandwidth of PE's physical links in an Ethernet Segment. BGP link bandwidth extended community defined in [BGP-LINK-BW] is re-used for this purpose.

4.2. EVPN Link Bandwidth Extended Community

A new EVPN Link Bandwidth extended community is defined to signal local ES link bandwidth to remote PEs. This extended-community is defined of type 0x06 (EVPN). IANA is requested to assign a sub-type value of 0x10 for the EVPN Link bandwidth extended community, of type 0x06 (EVPN). EVPN Link Bandwidth extended community is defined as conditional transitive with the following behavior:

- Pass it across eBGP session when next-hop is not rewritten.
- Drop it across eBGP session when next-hop is rewritten.

Link bandwidth extended community described in [BGP-LINK-BW] for layer 3 VPNs was considered for re-use here. This Link bandwidth extended community is however defined in [BGP-LINK-BW] as optional non-transitive. In inter-AS scenarios, link-bandwidth needs to be signaled to an eBGP neighbor when the next-hop is not unchanged. Since it is not possible to change deployed behavior of this extended-community, it was decided to define a new one.

4.3. Remote PE Behavior

A receiving PE SHOULD use per-ES link bandwidth attribute received from each PE to compute a relative weight for each remote PE, per-ES, and then use this relative weight to compute a weighted path-list to be used for load balancing, as opposed to using an ECMP path-list for load balancing across the PE paths. PE Weight and resulting weighted path-list computation at remote PEs is a local matter. An example computation algorithm is shown below to illustrate the idea:

if,

L(x,y) : link bandwidth advertised by PE-x for ESI-y

W(x,y) : normalized weight assigned to PE-x for ESI-y

H(y) : Highest Common Factor (HCF) of [L(1,y), L(2,y), ..., L(n,y)]

then, the normalized weight assigned to PE-x for ESI-y may be computed as follows:

W(x,y) = L(x,y) / H(y)

For a MAC+IP route (EVPN Route Type 2) received with ESI-y, receiving PE may compute MAC and IP forwarding path-list weighted by the above normalized weights.

As an example, for a CE dual-homed to PE-1, PE-2, PE-3 via 2, 1, and 1 GE physical links respectively, as part of a LAG represented by ESI-10:

L(1, 10) = 2000 Mbps

L(2, 10) = 1000 Mbps

L(3, 10) = 1000 Mbps

H(10) = 1000

Normalized weights assigned to each PE for ESI-10 are as follows:

W(1, 10) = 2000 / 1000 = 2.

W(2, 10) = 1000 / 1000 = 1.

W(3, 10) = 1000 / 1000 = 1.

For a remote MAC+IP host route received with ESI-10, forwarding load balancing path-list may now be computed as: [PE-1, PE-1, PE-2, PE-3] instead of [PE-1, PE-2, PE-3]. This now results in load balancing of all traffic destined for ESI-10 across the three multi-homing PEs in proportion to ESI-10 bandwidth at each PE.

Weighted path-list computation must only be done for an ESI if a link bandwidth attribute is received from all of the PE's advertising reachability to that ESI via Ethernet A-D per-ES Route Type 1. In an unlikely event that link bandwidth attribute is not received from one or more subset of PEs, forwarding path-list should be computed using regular ECMP semantics. Note that a default weight cannot be assumed for a PE that does not advertise its link bandwidth as the weight attribute t be used in path-list computation is relative.

5. Weighted BUM Traffic Load-Sharing

Optionally, load sharing of per-service DF role, weighted by individual PE's link-bandwidth share within a multi-homed ES may also be achieved.

In order to do that, a new DF Election Capability [RFC8584] called "BW" (Bandwidth Weighted DF Election) is defined. BW MAY be used along with some DF Election Types, as described in the following sections.

5.1. The BW Capability in the DF Election Extended Community

[RFC8584] defines a new extended community for PEs within a redundancy group to signal and agree on uniform DF Election Type and Capabilities for each ES. This document requests IANA for a bit in the DF Election extended community Bitmap:

Bit 28: BW (Bandwidth Weighted DF Election)

ES routes advertised with the BW bit set will indicate the desire of the advertising PE to consider the link-bandwidth in the DF Election algorithm defined by the value in the "DF Type".

As per [RFC8584], all the PEs in the ES MUST advertise the same Capabilities and DF Type, otherwise the PEs will fall back to Default [RFC7432] DF Election procedure.

The BW Capability MAY be advertised with the following DF Types:

- Type 0: Default DF Election algorithm, as in [RFC7432]
- Type 1: HRW algorithm, as in [RFC8584]

- Type 2: Preference algorithm, as in [EVPN-DF-PREF]

- Type 4: HRW per-multicast flow DF Election, as in [EVPN-PER-MCAST-FLOW-DF]

The following sections describe how the DF Election procedures are modified for the above DF Types when the BW Capability is used.

5.2. BW Capability and Default DF Election algorithm

When all the PEs in the Ethernet Segment (ES) agree to use the BW Capability with DF Type 0, the Default DF Election procedure as defined in [RFC7432] is modified as follows:

- Each PE advertises a "Link Bandwidth" extended community attribute along with the ES route to signal the PE-CE link bandwidth (LBW) for the ES.

- A receiving PE MUST use the ES link bandwidth attribute received from each PE to compute a relative weight for each remote PE.

- The DF Election procedure MUST now use this weighted list of PEs to compute the per-VLAN Designated Forwarder, such that the DF role is distributed in proportion to this normalized weight. As a result, a single PE may have multiple ordinals in the DF candidate PE list and 'N' used in (V mode N) operation as defined in [RFC7432] is modified to be total number of ordinals instead of being total number of PEs.

Considering the same example as in Section 3, the candidate PE list for DF election is:

[PE-1, PE-1, PE-2, PE-3].

The DF for a given VLAN-a on ES-10 is now computed as (VLAN-a % 4). This would result in the DF role being distributed across PE1, PE2, and PE3 in portion to each PE's normalized weight for ES-10.

5.3. BW Capability and HRW DF Election algorithm (Type 1 and 4)

[RFC8584] introduces Highest Random Weight (HRW) algorithm (DF Type 1) for DF election in order to solve potential DF election skew depending on Ethernet tag space distribution. [EVPN-PER-MCAST-FLOW-DF] further extends HRW algorithm for per-multicast flow based hash computations (DF Type 4). This section describes extensions to HRW Algorithm for EVPN DF Election specified in [RFC8584] and in [EVPN-PER-MCAST-FLOW-DF] in order to achieve DF election distribution that is weighted by link bandwidth.

5.3.1. BW Increment

A new variable called "bandwidth increment" is computed for each [PE, ES] advertising the ES link bandwidth attribute as follows:

In the context of an ES,

L(i) = Link bandwidth advertised by PE(i) for this ES

L(min) = lowest link bandwidth advertised across all PEs for this ES

Bandwidth increment, "b(i)" for a given PE(i) advertising a link bandwidth of L(i) is defined as an integer value computed as:

b(i) = L(i) / L(min)

As an example,

with PE(1) = 10, PE(2) = 10, PE(3) = 20

bandwidth increment for each PE would be computed as:

b(1) = 1, b(2) = 1, b(3) = 2

with PE(1) = 10, PE(2) = 10, PE(3) = 10

bandwidth increment for each PE would be computed as:

b(1) = 1, b(2) = 1, b(3) = 1

Note that the bandwidth increment must always be an integer, including, in an unlikely scenario of a PE's link bandwidth not being an exact multiple of L(min). If it computes to a non-integer value (including as a result of link failure), it MUST be rounded down to an integer.

5.3.2. HRW Hash Computations with BW Increment

HRW algorithm as described in [RFC8584] and in [EVPN-PER-MCAST-FLOW-DF] computes a random hash value for each PE(i), where, (0 < i <= N), PE(i) is the PE at ordinal i, and Address(i) is the IP address of PE(i).

For 'N' PEs sharing an Ethernet segment, this results in 'N' candidate hash computations. The PE that has the highest hash value is selected as the DF.

We refer to this hash value as "affinity" in this document. Hash or affinity computation for each PE(i) is extended to be computed one per bandwidth increment associated with PE(i) instead of a single affinity computation per PE(i).

PE(i) with b(i) = j, results in j affinity computations:

affinity(i, x), where 1 < x <= j

This essentially results in number of candidate HRW hash computations for each PE that is directly proportional to that PE's relative bandwidth within an ES and hence gives PE(i) a probability of being DF in proportion to it's relative bandwidth within an ES.

As an example, consider an ES that is multi-homed to two PEs, PE1 and PE2, with equal bandwidth distribution across PE1 and PE2. This would result in a total of two candidate hash computations:

affinity(PE1, 1)

affinity(PE2, 1)

Now, consider a scenario with PE1's link bandwidth as 2x that of PE2. This would result in a total of three candidate hash computations to be used for DF election:

affinity(PE1, 1)

affinity(PE1, 2)

affinity(PE2, 1)

which would give PE1 2/3 probability of getting elected as a DF, in proportion to its relative bandwidth in the ES.

Depending on the chosen HRW hash function, affinity function MUST be extended to include bandwidth increment in the computation.

For e.g.,

affinity function specified in [EVPN-PER-MCAST-FLOW-DF] MAY be extended as follows to incorporate bandwidth increment j:

affinity(S,G,V, ESI, Address(i,j)) =
(1103515245.((1103515245.Address(i).j + 12345) XOR
D(S,G,V,ESI))+12345) (mod 2^31)

affinity or random function specified in [RFC8584] MAY be extended as follows to incorporate bandwidth increment j:

affinity(v, Es, Address(i,j)) = (1103515245((1103515245.Address(i).j + 12345) XOR D(v,Es))+12345)(mod 2^31)

5.4. BW Capability and Preference DF Election algorithm

This section applies to ES'es where all the PEs in the ES agree use the BW Capability with DF Type 2. The BW Capability modifies the Preference DF Election procedure [EVPN-DF-PREF], by adding the LBW value as a tie-breaker as follows:

Section 4.1, bullet (f) in [EVPN-DF-PREF] now considers the LBW value:

f) In case of equal Preference in two or more PEs in the ES, the tie-breakers will be the DP bit, the LBW value and the lowest IP PE in that order. For instance:

- If vES1 parameters were [Pref=500,DP=0,LBW=1000] in PE1 and [Pref=500,DP=1, LBW=2000] in PE2, PE2 would be elected due to the DP bit.

- If vES1 parameters were [Pref=500,DP=0,LBW=1000] in PE1 and [Pref=500,DP=0, LBW=2000] in PE2, PE2 would be elected due to a higher LBW, even if PE1's IP address is lower.

- The LBW exchanged value has no impact on the Non-Revertive option described in [EVPN-DF-PREF].

6. Cost-Benefit Tradeoff on Link Failures

While incorporating link bandwidth into the DF election process provides optimal BUM traffic distribution across the ES links, it also implies that DF elections are re-adjusted on link failures or bandwidth changes. If the operator does not wish to have this level of churn in their DF election, then they should not advertise the BW capability. Not advertising BW capability may result in less than optimal BUM traffic distribution while still retaining the ability to allow a remote ingress PE to do weighted ECMP for its unicast traffic to a set of multi-homed PEs.

7. Real-time Available Bandwidth

PE-CE link bandwidth availability may sometimes vary in real-time disproportionately across PE-CE links within a multi-homed ESI due to various factors such as flow based hashing combined with fat flows

and unbalanced hashing. Reacting to real-time available bandwidth is at this time outside the scope of this document. Procedures described in this document are strictly based on static link bandwidth parameter.

8. Routed EVPN Overlay

An additional use case is possible, such that traffic to an end host in the overlay is always IP routed. In a purely routed overlay such as this:

- o A host MAC is never advertised in EVPN overlay control plane.
- o Host /32 or /128 IP reachability is distributed across the overlay via EVPN route type 5 (RT-5) along with a zero or non-zero ESI.
- o An overlay IP subnet may still be stretched across the underlay fabric, however, intra-subnet traffic across the stretched overlay is never bridged.
- o Both inter-subnet and intra-subnet traffic, in the overlay is IP routed at the EVPN GW.

Please refer to [RFC7814] for more details.

Weighted multi-path procedure described in this document may be used together with procedures described in [EVPN-IP-ALIASING] for this use case. Ethernet A-D per-ES route advertised with Layer 3 VRF RTs would be used to signal ES link bandwidth attribute instead of the Ethernet A-D per-ES route with Layer 2 VRF RTs. All other procedures described earlier in this document would apply as is.

If [EVPN-IP-ALIASING] is not used for routed fast convergence, link bandwidth attribute may still be advertised with IP routes (RT-5) to achieve PE-CE link bandwidth based load balancing as described in this document. In the absence of [EVPN-IP-ALIASING], re-balancing of traffic following changes in PE-CE link bandwidth will require all IP routes from that CE to be re-advertised in a prefix dependent manner.

9. EVPN-IRB Multi-homing With Non-EVPN routing

EVPN-LAG based multi-homing on an IRB gateway may also be deployed together with non-EVPN routing, such as global routing or an L3VPN routing control plane. Key property that differentiates this set of use cases from EVPN IRB use cases discussed earlier is that EVPN control plane is used only to enable LAG interface based multi-homing and NOT as an overlay VPN control plane. EVPN control plane in this case enables:

- DF election via EVPN RT-4 based procedures described in [RFC7432]
- Local MAC sync across multi-homing PEs via EVPN RT-2
- Local ARP and ND sync across multi-homing PEs via EVPN RT-2

Applicability of weighted ECMP procedures proposed in this document to these set of use cases is an area of further consideration.

10. Operational Considerations

None

11. Security Considerations

This document raises no new security issues for EVPN.

12. IANA Considerations

[RFC8584] defines a new extended community for PEs within a redundancy group to signal and agree on uniform DF Election Type and Capabilities for each ES. This document requests IANA for a bit in the DF Election extended community Bitmap:

Bit 28: BW (Bandwidth Weighted DF Election)

A new EVPN Link Bandwidth extended community is defined to signal local ES link bandwidth to remote PEs. This extended-community is defined of type 0x06 (EVPN). IANA is requested to assign a sub-type value of 0x10 for the EVPN Link bandwidth extended community, of type 0x06 (EVPN). EVPN Link Bandwidth extended community is defined as conditional transitive with the following behavior:

- Pass it across eBGP session when next-hop is not rewritten.
- Drop it across eBGP session when next-hop is rewritten.

13. Acknowledgements

Authors would like to thank Satya Mohanty for valuable review and inputs with respect to HRW and weighted HRW algorithm refinements proposed in this document.

14. Contributors

Satya Ranjan Mohanty
Cisco Systems
US
Email: satyamoh@cisco.com

15. Normative References

[BGP-LINK-BW]
Mohapatra, P. and R. Fernando, "BGP Link Bandwidth Extended Community", draft-ietf-idr-link-bandwidth-07 (work in progress), March 2019.

[EVPN-DF-PREF]
Rabadan, J., Sathappan, S., Przygienda, T., Lin, W., Drake, J., Sajassi, A., Mohanty, S., and , "Preference-based EVPN DF Election", draft-ietf-bess-evpn-pref-df-05 (work in progress), December 2019.

[EVPN-IP-ALIASING]
Sajassi, A. and G. Badoni, "L3 Aliasing and Mass Withdrawal Support for EVPN", draft-sajassi-bess-evpn-ip-aliasing-01 (work in progress), March 2020.

[EVPN-PER-MCAST-FLOW-DF]
Sajassi, A., mishra, m., Thoria, S., Rabadan, J., and J. Drake, "Per multicast flow Designated Forwarder Election for EVPN", draft-ietf-bess-evpn-per-mcast-flow-df-election-01 (work in progress), March 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>.

[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>.

[RFC7814] Xu, X., Jacquenet, C., Raszuk, R., Boyes, T., and B. Fee, "Virtual Subnet: A BGP/MPLS IP VPN-Based Subnet Extension Solution", RFC 7814, DOI 10.17487/RFC7814, March 2016, <https://tools.ietf.org/html/rfc7814>.

[RFC8584] Rabadan, J., Ed., Mohanty, R., Sajassi, N., Drake, A., 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>.

Authors' Addresses

Neeraj Malhotra (editor)
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: nmalhotr@cisco.com

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

Email: sajassi@cisco.com

Jorge Rabadan
Nokia
777 E. Middlefield Road
Mountain View, CA 94043
USA

Email: jorge.rabadan@nokia.com

John Drake
Juniper

Email: jdrake@juniper.net

Avinash Lingala
ATT
200 S. Laurel Avenue
Middletown, CA 07748
USA

Email: ar977m@att.com

Samir Thoria
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: sthoria@cisco.com

BESS Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 12, 2021

G. Dawra, Ed.
LinkedIn
C. Filsfils
Cisco Systems
R. Raszuk
Bloomberg LP
B. Decraene
Orange
S. Zhuang
Huawei Technologies
J. Rabadan
Nokia

July 11, 2020

SRv6 BGP based Overlay services
draft-ietf-bess-srv6-services-03

Abstract

This draft defines procedures and messages for SRv6-based BGP services including L3VPN, EVPN and Internet services. It builds on RFC4364 "BGP/MPLS IP Virtual Private Networks (VPNs)" and RFC7432 "BGP MPLS-Based Ethernet VPN".

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 January 12, 2021.

Table of Contents

1. Introduction

SRv6 refers to Segment Routing [RFC8402] instantiated on the IPv6 dataplane [RFC8754].

SRv6 based BGP services refers to the L3 and L2 overlay services with BGP as control plane and SRv6 as dataplane.

SRv6 SID refers to a SRv6 Segment Identifier as defined in [RFC8402].

SRv6 Service SID refers to an SRv6 SID associated with one of the service specific behavior on the advertising Provider Edge (PE) router, such as (but not limited to), END.DT (Table lookup in a VRF) or END.DX (cross-connect to a nexthop) behaviors in the case of L3VPN service as defined in [I-D.ietf-spring-srv6-network-programming].

To provide SRv6 service with best-effort connectivity, the egress PE signals an SRv6 Service SID with the BGP overlay service route. The ingress PE encapsulates the payload in an outer IPv6 header where the destination address is the SRv6 Service SID provided by the egress PE. The underlay between the PEs only need to support plain IPv6 forwarding [RFC8200].

To provide SRv6 service in conjunction with an underlay SLA from the ingress PE to the egress PE, the egress PE colors the overlay service route with a Color extended community [I-D.ietf-idr-segment-routing-te-policy]. The ingress PE encapsulates the payload packet in an outer IPv6 header with the segment list of SR policy associated with the related SLA followed by the SRv6 Service SID associated with the route. The underlay nodes whose SRv6 SID's are part of the segment list MUST support SRv6 data plane.

BGP is used to advertise the reachability of prefixes of a particular service from an egress PE to ingress PE nodes.

This document describes how existing BGP messages between PEs may carry SRv6 Service SIDs as a means to interconnect PEs and form VPNs.

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. SRv6 Services TLVs

This document extends the BGP Prefix-SID attribute [RFC8669] to carry SRv6 SIDs and associated information.

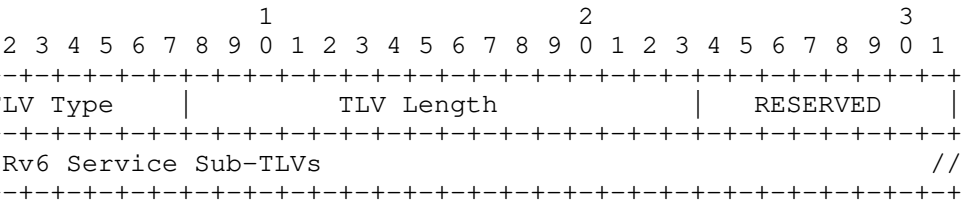
The SRv6 Service TLVs are defined as two new TLVs of the BGP Prefix-SID Attribute to achieve signaling of SRv6 SIDs for L3 and L2 services.

- SRv6 L3 Service TLV: This TLV encodes Service SID information for SRv6 based L3 services. It corresponds to the equivalent functionality provided by an MPLS Label when received with a Layer 3 service route. Some behaviors which MAY be encoded, but not limited to, are End.DX4, End.DT4, End.DX6, End.DT6, etc.

- SRv6 L2 Service TLV: This TLV encodes Service SID information for SRv6 based L2 services. It corresponds to the equivalent functionality provided by an MPLS Label1 for EVPN Route-Types as defined in[RFC7432]. Some behaviors which MAY be encoded, but not limited to, are End.DX2, End.DX2V, End.DT2U, End.DT2M etc.

When an egress PE is enabled for BGP Services over SRv6 data-plane, it MUST signal one or more SRv6 Service SIDs enclosed in SRv6 Service TLV(s) within the BGP Prefix-SID Attribute attached to MP-BGP NLRIs defined in [RFC4760] [RFC4659] [I-D.ietf-bess-rfc5549revision] [RFC7432] [RFC4364] where applicable as described in Section 5 and Section 6.

The following depicts the SRv6 Service TLVs encoded in the BGP Prefix-SID Attribute:



- TLV Type (1 octet): This field is assigned values from the IANA registry "BGP Prefix-SID TLV Types". It is set to 5 for SRv6 L3 Service TLV. It is set to 6 for SRv6 L2 Service TLV.

- TLV Length (2 octets): Specifies the total length of the TLV Value.

- RESERVED (1 octet): This field is reserved; it SHOULD be set to 0 by the sender and MUST be ignored by the receiver.

- SRv6 Service Sub-TLVs (variable): This field contains SRv6 Service related information and is encoded as an unordered list of Sub-TLVs whose format is described below.

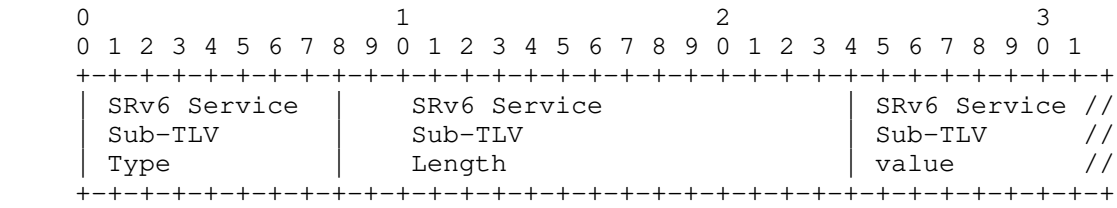
A BGP speaker receiving a route containing BGP Prefix-SID Attribute with one or more SRv6 Service TLVs observes the following rules when advertising the received route to other peers:

- if the nexthop is unchanged during advertisement, the SRv6 Service TLVs, including any unrecognized Types of Sub-TLV and Sub-Sub-TLV, SHOULD be propagated further. In addition, all Reserved fields in the TLV or Sub-TLV or Sub-Sub-TLV MUST be propagated unchanged.

- if the nexthop is changed, the TLVs, Sub-TLVs and Sub-Sub-TLVs SHOULD be updated as appropriate. Any unrecognized received sub-TLVs and Sub-Sub-TLVs MUST be removed.

3. SRv6 Service Sub-TLVs

The format of a single SRv6 Service Sub-TLV is depicted below:



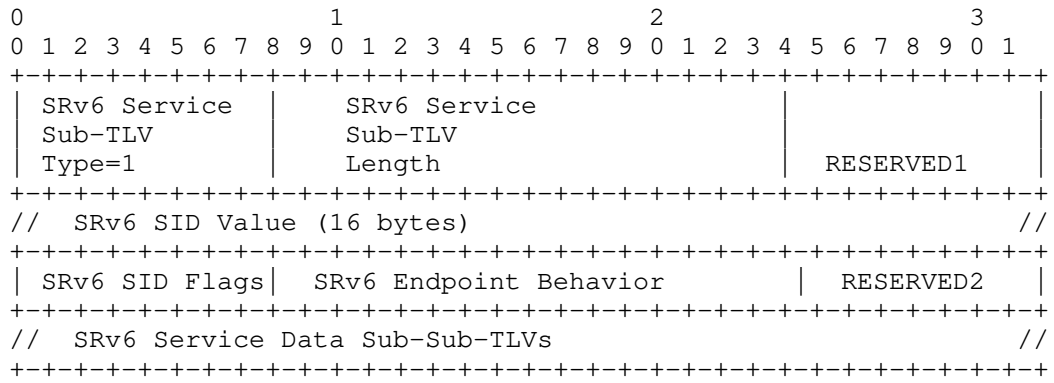
- SRv6 Service Sub-TLV Type (1 octet): Identifies the type of SRv6 service information. It is assigned values from the IANA Registry "SRv6 Service Sub-TLV Types".

- SRv6 Service Sub-TLV Length (2 octets): Specifies the total length of the Sub-TLV Value field.

- SRv6 Service Sub-TLV Value (variable): Contains data specific to the Sub-TLV Type. In addition to fixed length data, it contains other properties of the SRv6 Service encoded as a set of SRv6 Service Data Sub-Sub-TLVs whose format is described in Section 3.2 below.

3.1. SRv6 SID Information Sub-TLV

SRv6 Service Sub-TLV Type 1 is assigned for SRv6 SID Information Sub-TLV. This Sub-TLV contains a single SRv6 SID along with its properties. Its encoding is depicted below:



- SRv6 Service Sub-TLV Type (1 octet): This field is set to 1 to represent SRv6 SID Information Sub-TLV.

- SRv6 Service Sub-TLV Length (2 octets): This field contains the total length of the Value field of the Sub-TLV.

- RESERVED1 (1 octet): SHOULD be set to 0 by the sender and MUST be ignored by the receiver.

- SRv6 SID Value (16 octets): Encodes an SRv6 SID as defined in [I-D.ietf-spring-srv6-network-programming]

- SRv6 SID Flags (1 octet): Encodes SRv6 SID Flags - none are currently defined. SHOULD be set to 0 by sender and MUST be ignored by the receiver.

- SRv6 Endpoint Behavior (2 octets): Encodes SRv6 Endpoint behavior codepoint value from the IANA registry defined in section 9.2 of [I-D.ietf-spring-srv6-network-programming] that is associated with SRv6 SID. The opaque behavior (i.e. value 0xFFFF) or an unrecognized behavior MUST NOT be considered as invalid by the receiver.

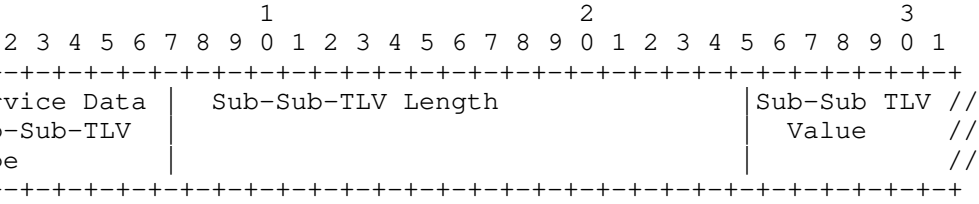
- RESERVED2 (1 octet): SHOULD be set to 0 by the sender and MUST be ignored by the receiver.

- SRv6 Service Data Sub-Sub-TLV Value (variable): Used to advertise properties of the SRv6 SID. It is encoded as a set of SRv6 Service Data Sub-Sub-TLVs.

When multiple SRv6 SID Information Sub-TLVs are present, the ingress PE SHOULD use the SRv6 SID from the first instance of the Sub-TLV. An implementation MAY provide a local policy to override this selection.

3.2. SRv6 Service Data Sub-Sub-TLVs

The format of the SRv6 Service Data Sub-Sub-TLV is depicted below:



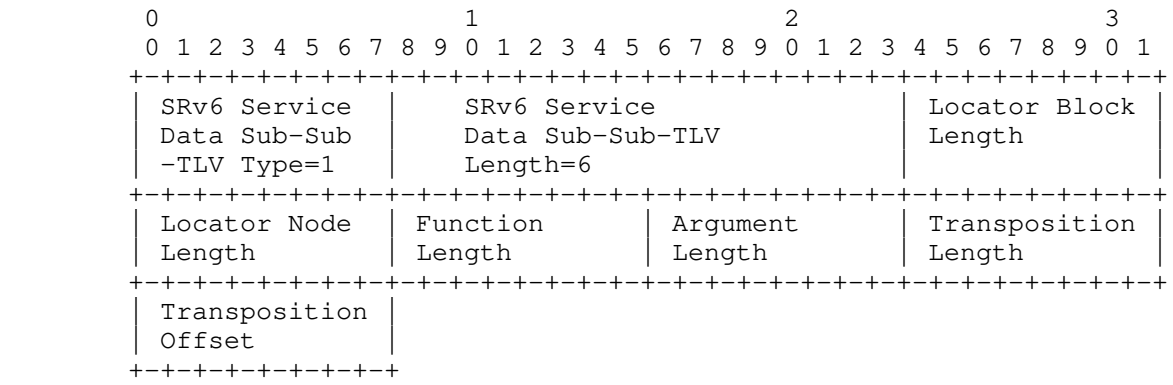
- SRv6 Service Data Sub-Sub-TLV Type (1 octet): Identifies the type of Sub-Sub-TLV. It is assigned values from the IANA Registry "SRv6 Service Data Sub-Sub-TLVs".

- SRv6 Service Data Sub-Sub-TLV Length (2 octets): Specifies the total length of the Sub-Sub-TLV Value field.

- SRv6 Service Data Sub-Sub-TLV Value (variable): Contains data specific to the Sub-Sub-TLV Type.

3.2.1. SRv6 SID Structure Sub-Sub-TLV

SRv6 Service Data Sub-Sub-TLV Type 1 is assigned for SRv6 SID structure Sub-Sub-TLV. SRv6 SID Structure Sub-Sub-TLV is used to advertise the lengths of each individual parts of the SRv6 SID as defined in [I-D.ietf-spring-srv6-network-programming]. It is carried as Sub-Sub-TLV in SRv6 SID Information Sub-TLV



- SRv6 Service Data Sub-Sub-TLV Type (1 octet): This field is set to 1 to represent SRv6 SID Structure Sub-Sub-TLV.

- SRv6 Service Data Sub-Sub-TLV Length (2 octets): This field contains the total length of 6 bytes.

- Locator Block Length (1 octet): Contains length of SRv6 SID locator Block in bits.

- Locator Node Length (1 octet): Contains length of SRv6 SID locator Node in bits.

- Function Length (1 octet): Contains length of SRv6 SID Function in bits.

- Argument Length (1 octet): Contains length of SRv6 SID argument in bits.

- Transposition Length (1 octet): Size in bits for the part of SID that has been transposed (or shifted) into a label field

- Transposition Offset (1 octet): The offset position in bits for the part of SID that has been transposed (or shifted) into a label field.

Section 4 describes mechanisms for signaling of the SRv6 Service SID by transposing a variable part of the SRv6 SID value (function and/or the argument parts) and carrying them in existing label fields to achieve more efficient packing of those service prefix NLRIs in BGP update messages. The SRv6 SID Structure Sub-Sub-TLV contains appropriate length fields when the SRv6 Service SID is signaled in split parts to enable the receiver to put together the SID accurately.

Transposition Offset indicates the bit position and Transposition Length indicates the number of bits that are being taken out of the SRv6 SID value and put into high order bits of label field. The bits that have been shifted out MUST be set to 0 in the SID value.

Transposition Length of 0 indicates nothing is transposed and that the entire SRv6 SID value is encoded in the SID Information sub-TLV. In this case, the Transposition Offset MUST be set to 0.

Since size of label field is 24 bits, only that many bits can be transposed from the SRv6 SID value into it.

As an example, when the entire function part of size 16 of an SRv6 SID is transposed and the sum of the locator block and locator node parts is 64, then the transposition offset would be set to 64 and the transposition length is set to 16.

BGP speakers that do not support this specification may misinterpret, on reception of an SRv6-based BGP service route update, the function and/or argument parts of the SRv6 SID encoded in label field(s) as MPLS label values for MPLS-based services. Implementations supporting this specification SHOULD provide a mechanism to control advertisement of SRv6-based BGP service routes on a per neighbor and per service basis.

Arguments MAY be generally applicable for SIDs of only specific behaviors (e.g. End.DT2M) and therefore the argument length MUST be set to 0 for SIDs where the argument is not applicable.

4. Encoding SRv6 SID information

The SRv6 Service SID(s) for a BGP Service Prefix are carried in the SRv6 Services TLVs of the BGP Prefix-SID Attribute.

For certain types of BGP Services like L3VPN where a per-VRF SID allocation is used (i.e. End.DT4 or End.DT6 behaviors), the same SID is shared across multiple NLRIs thus providing efficient packing. However, for certain other types of BGP Services like EVPN VPWS where a per-PW SID allocation is required (i.e. End.DX2 behavior), each NLRI would have its own unique SID there by resulting in inefficient packing.

To achieve efficient packing, this document allows the encoding of the SRv6 Service SID either as a whole in the SRv6 Services TLVs or the encoding of only the common part of the SRv6 SID (e.g. Locator) in the SRv6 Services TLVs and encoding the variable (e.g. Function and Argument parts) in the existing label fields specific to that service encoding. This later form of encoding is referred to as the

Transposition Scheme where the SRv6 SID Structure Sub-Sub-TLV describes the sizes of the parts of the SRv6 SID and to also indicate offset of variable part along with its length in SRv6 SID value. The use of the Transposition Scheme is RECOMMENDED for the specific service encodings that allow it as described further in Section 5 and Section 6.

As an example, for the EVPN VPWS service prefix described further in Section 6.1.2, the function part of the SRv6 SID is encoded in the MPLS Label field of the NLRI and the SID value in the SRv6 Services TLV carries only the locator part with the SRv6 SID Structure Sub-Sub-TLV. The SRv6 SID Structure sub-sub-TLV defines the lengths of locator block, locator node and function parts (arguments are not applicable for the End.DX2 behavior). Transposition Offset indicates the bit position and Transposition Length indicates the number of bits that are being taken out of the SID and put into label field.

In yet another example, for the EVPN Per-ES A-D route described further in Section 6.1.1, only the argument of the SID needs to be signaled. This argument part of the SRv6 SID MAY be transposed in the ESI Label field of the ESI Label Extended Community and the SID value in the SRv6 Services TLV is set to 0 with the SRv6 SID Structure Sub-Sub-TLV. The SRv6 SID Structure sub-sub-TLV defines the lengths of locator block, locator node, function and argument parts. The offset and length of argument part SID value moved to label field is set in transposition offset and length of SID structure TLV. The receiving router is then able to put together the entire SRv6 Service SID (e.g. for the End.DT2M behavior) placing the label value received in the ESI Label field of the Per-ES A-D route into the correct transposition offset and length in the SRv6 SID with the End.DT2M behavior received for a EVPN Route Type 3 value.

5. BGP based L3 service over SRv6

BGP egress nodes (egress PEs) advertise a set of reachable prefixes. Standard BGP update propagation schemes[RFC4271], which may make use of route reflectors [RFC4456], are used to propagate these prefixes. BGP ingress nodes (ingress PEs) receive these advertisements and may add the prefix to the RIB in an appropriate VRF.

Egress PEs which supports SRv6 based L3 services advertises overlay service prefixes along with a Service SID enclosed in a SRv6 L3 Service TLV within the BGP Prefix-SID Attribute. This TLV serves two purposes - first, it indicates that the egress PE supports SRv6 overlay and the BGP ingress PE receiving this route MUST choose to perform IPv6 encapsulation and optionally insert an SRH [RFC8754] when required; second ,it indicates the value of the Service SID to be used in the encapsulation.

The Service SID thus signaled only has local significance at the egress PE, where it may be allocated or configured on a per-CE or per-VRF basis. In practice, the SID may encode a cross-connect to a specific Address Family table (END.DT) or next-hop/interface (END.DX) as defined in [I-D.ietf-spring-srv6-network-programming].

The SRv6 Service SID SHOULD be routable within the AS of the egress PE and serves the dual purpose of providing reachability between ingress PE and egress PE while also encoding the endpoint behavior.

When the egress PE sets the next-hop to a value that is not covered by the SRv6 Locator from which the SRv6 Service SID is allocated, then the ingress PE SHOULD perform reachability check for the SRv6 Service SID in addition to the BGP next-hop reachability procedures.

At an ingress PE, BGP installs the received prefix in the correct RIB table, recursing via an SR Policy leveraging the received SRv6 Service SID.

Assuming best-effort connectivity to the egress PE, the ingress PE encapsulates the payload in an outer IPv6 header where the destination address is the SRv6 Service SID associated with the related BGP route update.

However, when the received route is colored with an extended color community 'C' and Next-Hop 'N', and the ingress PE has a valid SRv6 Policy (C, N) associated with SID list <S1,S2, S3> [I-D.ietf-spring-segment-routing-policy], then the effective SR Policy is <S1, S2, S3-Service-SID>.

Multiple VPN routes MAY resolve recursively via the same SR Policy.

5.1. IPv4 VPN Over SRv6 Core

The MP_REACH_NLRI for SRv6 core is encoded according to IPv4 VPN Over IPv6 Core defined in [I-D.ietf-bess-rfc5549revision].

Label field of IPv4-VPN NLRI is encoded as specified in [RFC8277] with the Label Value set to the Function part of the SRv6 SID when the Transposition Scheme of encoding (Section 4) is used and otherwise set to Implicit NULL.

SRv6 Service SID is encoded as part of the SRv6 L3 Service TLV. The behavior of the SRv6 SID is entirely up to the originator of the advertisement. In practice, the behavior is End.DX4 or End.DT4.

5.2. IPv6 VPN Over SRv6 Core

The MP_REACH_NLRI for SRv6 core is encoded according to IPv6 VPN over IPv6 Core is defined in [RFC4659].

Label field of the IPv6-VPN NLRI is encoded as specified in [RFC8277] with the Label Value set to the Function part of the SRv6 SID when the Transposition Scheme of encoding (Section 4) is used and otherwise set to Implicit NULL.

SRv6 Service SID is encoded as part of the SRv6 L3 Service TLV. The behavior of the SRv6 SID is entirely up to the originator of the advertisement. In practice, the behavior is End.DX6 or End.DT6.

5.3. Global IPv4 over SRv6 Core

The MP_REACH_NLRI for SRv6 core is encoded according to IPv4 over IPv6 Core is defined in [I-D.ietf-bess-rfc5549revision].

SRv6 Service SID is encoded as part of the SRv6 L3 Service TLV. The behavior of the SRv6 SID is entirely up to the originator of the advertisement. In practice, the behavior is End.DX4 or End.DT4.

5.4. Global IPv6 over SRv6 Core

The MP_REACH_NLRI for SRv6 core is encoded according to [RFC2545]

SRv6 Service SID is encoded as part of the SRv6 L3 Service TLV. The behavior of the SRv6 SID is entirely up to the originator of the advertisement. In practice, the behavior is End.DX6 or End.DT6.

6. BGP based Ethernet VPN (EVPN) over SRv6

[RFC7432] provides an extendable method of building an Ethernet VPN (EVPN) overlay. It primarily focuses on MPLS based EVPNs and [RFC8365] extends to IP based EVPN overlays. [RFC7432] defines Route Types 1, 2 and 3 which carry prefixes and MPLS Label fields; the Label fields have specific use for MPLS encapsulation of EVPN traffic. Route Type 5 carrying MPLS label information (and thus encapsulation information) for EVPN is defined in [I-D.ietf-bess-evpn-prefix-advertisement]. Route Types 6,7 and 8 are defined in [I-D.ietf-bess-evpn-igmp-mld-proxy].

- Ethernet Auto-discovery Route (Route Type 1)
- MAC/IP Advertisement Route (Route Type 2)
- Inclusive Multicast Ethernet Tag Route (Route Type 3)

- Ethernet Segment route (Route Type 4)
- IP prefix route (Route Type 5)
- Selective Multicast Ethernet Tag route (Route Type 6)
- IGMP join sync route (Route Type 7)
- IGMP leave sync route (Route Type 8)

To support SRv6 based EVPN overlays, one or more SRv6 Service SIDs are advertised with Route Type 1,2,3 and 5. The SRv6 Service SID(s) per Route Type are advertised in SRv6 L3/L2 Service TLVs within the BGP Prefix-SID Attribute. Signaling of SRv6 Service SID(s) serves two purposes - first, it indicates that the BGP egress device supports SRv6 overlay and the BGP ingress device receiving this route MUST perform IPv6 encapsulation and optionally insert an SRH [RFC8754] when required; second, it indicates the value of the Service SID(s) to be used in the encapsulation.

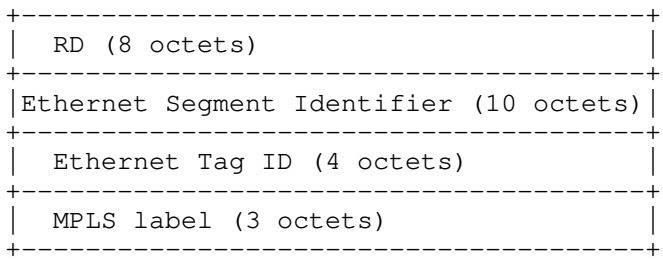
The SRv6 Service SID SHOULD be routable within the AS of the egress PE and serves the dual purpose of providing reachability between ingress PE and egress PE while also encoding the endpoint behavior.

When the egress PE sets the next-hop to a value that is not covered by the SRv6 Locator from which the SRv6 Service SID is allocated, then the ingress PE SHOULD perform reachability check for the SRv6 Service SID in addition to the BGP next-hop reachability procedures.

6.1. Ethernet Auto-discovery route over SRv6 Core

Ethernet Auto-Discovery (A-D) routes are Route Type 1 defined in [RFC7432] and may be used to achieve split horizon filtering, fast convergence and aliasing. EVPN Route Type 1 is also used in EVPN-VPWS as well as in EVPN flexible cross-connect; mainly used to advertise point-to-point services ID.

As a reminder, EVPN Route Type 1 is encoded as follows:



6.1.1. Per-ES A-D route

Per-ES A-D route for SRv6 overlay is advertised as follows:

- BGP next-hop: IPv6 address of an egress PE
- Ethernet Tag ID: set as per [RFC7432]
- MPLS Label: set as per [RFC7432]
- ESI label extended community ESI label field: carries the Argument part of the SRv6 SID when ESI filtering approach is used along with the Transposition Scheme of encoding (Section 4) and otherwise set to Implicit NULL.

A Service SID enclosed in a SRv6 L2 Service TLV within the BGP Prefix-SID attribute is advertised along with the A-D route. The behavior of the Service SID thus signaled is entirely up to the originator of the advertisement. When ESI filtering approach is used, the Service SID is used to signal Arg.FE2 SID argument for applicable End.DT2M SIDs. When local-bias approach is used, the Service SID MAY be of value 0.

6.1.2. Per-EVI A-D route

Per-EVI A-D route for SRv6 overlay is advertised as follows:

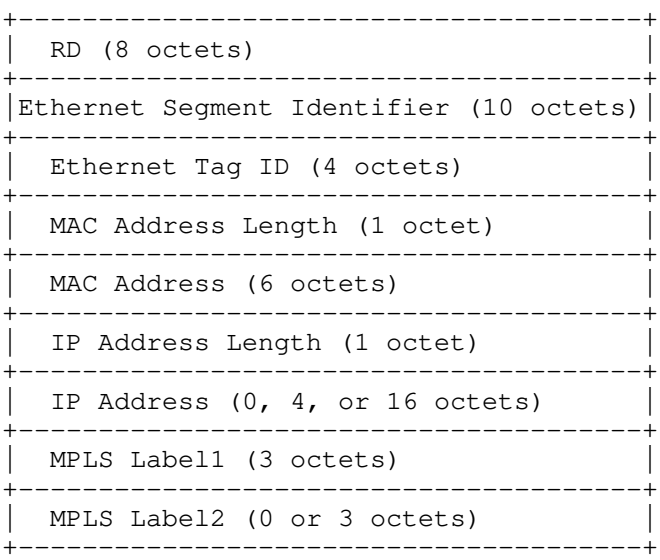
- BGP next-hop: IPv6 address of an egress PE
- Ethernet Tag ID: Set as per [RFC7432], [RFC8214] and [I-D.ietf-bess-evpn-vpws-fxc]
- MPLS Label: carries the Function part of the SRv6 SID when the Transposition Scheme of encoding (Section 4) is used and otherwise set to Implicit NULL.

A Service SID enclosed in a SRv6 L2 Service TLV within the BGP Prefix-SID attribute is advertised along with the A-D route. The behavior of the Service SID thus signaled is entirely up to the originator of the advertisement. In practice, the behavior is END.DX2, END.DX2V or END.DT2U.

6.2. MAC/IP Advertisement route over SRv6 Core

EVPN Route Type 2 is used to advertise unicast traffic MAC+IP address reachability through MP-BGP to all other PEs in a given EVPN instance.

As a reminder, EVPN Route Type 2 is encoded as follows:



- BGP next-hop: IPv6 address of an egress PE

- MPLS Label1: Is associated with the SRv6 L2 Service TLV. It carries the Function part of the SRv6 SID when the Transposition Scheme of encoding (Section 4) is used and otherwise set to Implicit NULL.

- MPLS Label2: Is associated with the SRv6 L3 Service TLV. It carries the Function part of the SRv6 SID when the Transposition Scheme of encoding (Section 4) is used and otherwise set to Implicit NULL.

Service SIDs enclosed in SRv6 L2 Service TLV and optionally in SRv6 L3 Service TLV within the BGP Prefix-SID attribute is advertised along with the MAC/IP Advertisement route.

Described below are different types of Route Type 2 advertisements.

6.2.1. MAC/IP Advertisement route with MAC Only

- MPLS Label1: Is associated with the SRv6 L2 Service TLV. It carries the Function part of the SRv6 SID when the Transposition Scheme of encoding (Section 4) is used and otherwise set to Implicit NULL.

A Service SID enclosed in a SRv6 L2 Service TLV within the BGP Prefix-SID attribute is advertised along with the route. The behavior of the Service SID thus signaled is entirely up to the originator of the advertisement. In practice, the behavior is END.DX2 or END.DT2U.

6.2.2. MAC/IP Advertisement route with MAC+IP

- MPLS Label1: Is associated with the SRv6 L2 Service TLV. It carries the Function part of the SRv6 SID when the Transposition Scheme of encoding (Section 4) is used and otherwise set to Implicit NULL.

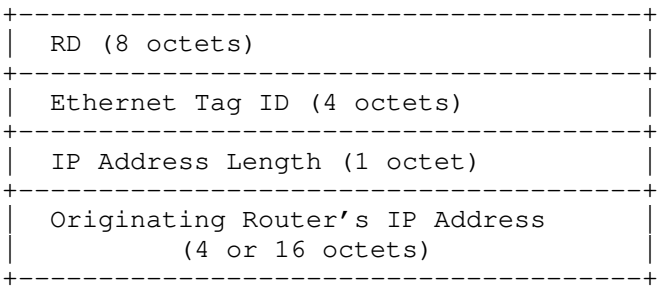
- MPLS Label2: Is associated with the SRv6 L3 Service TLV. It carries the Function part of the SRv6 SID when the Transposition Scheme of encoding (Section 4) is used and otherwise set to Implicit NULL.

An L2 Service SID enclosed in a SRv6 L2 Service TLV within the BGP Prefix-SID attribute is advertised along with the route. In addition, an L3 Service SID enclosed in a SRv6 L3 Service TLV within the BGP Prefix-SID attribute MAY also be advertised along with the route. The behavior of the Service SID(s) thus signaled is entirely up to the originator of the advertisement. In practice, the behavior is END.DX2 or END.DT2U for the L2 Service SID, and END.DT6/4 or END.DX6/4 for the L3 Service SID.

6.3. Inclusive Multicast Ethernet Tag Route over SRv6 Core

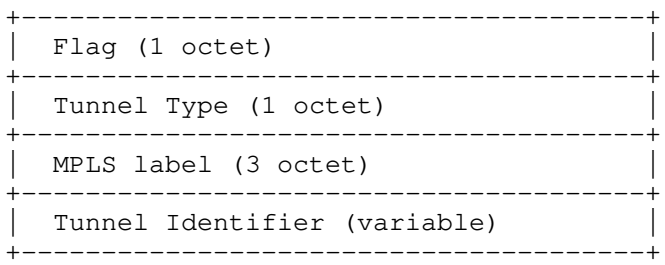
EVPN Route Type 3 is used to advertise multicast traffic reachability information through MP-BGP to all other PEs in a given EVPN instance.

As a reminder, EVPN Route Type 3 is encoded as follows:



- BGP next-hop: IPv6 address of egress PE

PMSI Tunnel Attribute [RFC6514] is used to identify the P-tunnel used for sending BUM traffic. The format of PMSI Tunnel Attribute is encoded as follows for SRv6 Core:



- Flag: zero value defined per [RFC7432]
- Tunnel Type: defined per [RFC6514]
- MPLS label: It carries the Function part of the SRv6 SID when ingress replication is used and the Transposition Scheme of encoding (Section 4) is used and otherwise it is set as defined in [RFC6514]
- Tunnel Identifier: IP address of egress PE

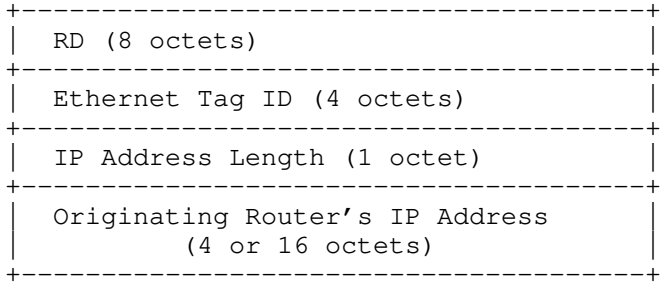
A Service SID enclosed in a SRv6 L2 Service TLV within the BGP Prefix-SID attribute is advertised along with the route. The behavior of the Service SID thus signaled, is entirely up to the originator of the advertisement. In practice, the behavior of the SRv6 SID is as follows:

- END.DT2M behavior.
- When ESI-based filtering is used for Multi-Homing or E-Tree procedures, the ESI Filtering argument (Arg.FE2) of the Service SID carried along with EVPN Route Type 1 route SHOULD be merged together with the applicable End.DT2M SID of Type 3 route advertised by remote PE by doing a bitwise logical-OR operation to create a single SID on the ingress PE. Details of split-horizon ESI-based filtering mechanisms for multihoming are described in [RFC7432]. Details of filtering mechanisms for Leaf-originated BUM traffic in EVPN E-Tree services are provided in [RFC8317].
- When "local-bias" is used as the Multi-Homing split-horizon method, the ESI Filtering argument SHOULD NOT be merged with the corresponding End.DT2M SID on the ingress PE. Details of the "local-bias" procedures are described in [RFC8365].

The setup of multicast trees for use as P-tunnels is outside the scope of this document.

6.4. Ethernet Segment route over SRv6 Core

As a reminder, an Ethernet Segment route i.e. EVPN Route Type 4 is encoded as follows:



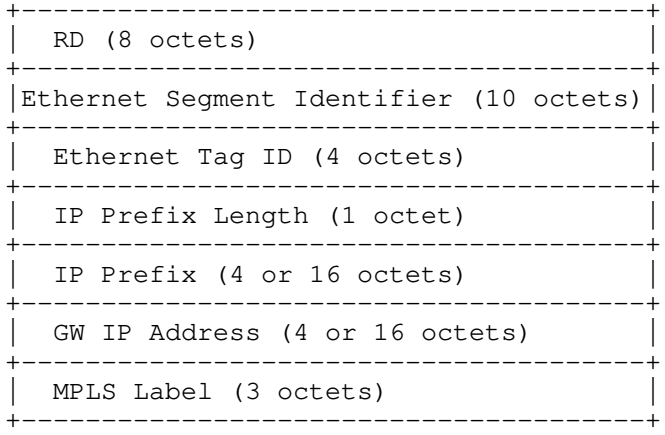
- BGP next-hop: IPv6 address of egress PE

SRv6 Service TLVs within BGP Prefix-SID attribute are not advertised along with this route. The processing of the route has not changed - it remains as described in [RFC7432].

6.5. IP prefix route over SRv6 Core

EVPN Route Type 5 is used to advertise IP address reachability through MP-BGP to all other PEs in a given EVPN instance. IP address may include host IP prefix or any specific subnet.

As a reminder, EVPN Route Type 5 is encoded as follows:



- BGP next-hop: IPv6 address of egress PE

- MPLS Label: It carries the Function part of the SRv6 SID when the Transposition Scheme of encoding (Section 4) is used and otherwise set to Implicit NULL.

SRv6 Service SID is encoded as part of the SRv6 L3 Service TLV. The function of the SRv6 SID is entirely up to the originator of the advertisement. In practice, the behavior is End.DT4/6 or End.DX4/6.

6.6. EVPN multicast routes (Route Types 6, 7, 8) over SRv6 core

These routes do not require the advertisement of SRv6 Service TLVs along with them. Similar to EVPN Route Type 4, the BGP Nexthop is equal to the IPv6 address of egress PE.

7. Implementation Status

The [I-D.matsushima-spring-srv6-deployment-status] describes the current deployment and implementation status of SRv6 which also includes the BGP services over SRv6 as specified in this document.

8. Error Handling

In case of any errors encountered while processing SRv6 Service TLVs, the details of the error SHOULD be logged for further analysis.

If multiple instances of SRv6 L3 Service TLV is encountered, all but the first instance MUST be ignored.

If multiple instances of SRv6 L2 Service TLV is encountered, all but the first instance MUST be ignored.

An SRv6 Service TLV is considered malformed in the following cases:

- the TLV Length is less than 1

- the TLV Length is inconsistent with the length of BGP Prefix-SID attribute

- at least one of the constituent Sub-TLVs is malformed

An SRv6 Service Sub-TLV is considered malformed in the following cases:

- the Sub-TLV Length is inconsistent with the length of the enclosing SRv6 Service TLV

An SRv6 SID Information Sub-TLV is considered malformed in the following cases:

* the Sub-TLV Length is less than 21

* the Sub-TLV Length is inconsistent with the length of the enclosing SRv6 Service TLV

* at least one of the constituent Sub-Sub-TLVs is malformed

An SRv6 Service Data Sub-sub-TLV is considered malformed in the following cases:

o the Sub-Sub-TLV Length is inconsistent with the length of the enclosing SRv6 service Sub-TLV

Any TLV or Sub-TLV or Sub-Sub-TLV is not considered malformed because its Type is unrecognized.

Any TLV or Sub-TLV or Sub-Sub-TLV is not considered malformed because of failing any semantic validation of its Value field.

SRv6 overlay service requires Service SID for forwarding. The treat-as-withdraw action [RFC7606] MUST be performed when at least one malformed SRV6 Service TLV is present in the BGP Prefix-SID attribute.

SRv6 SID value in SRv6 Service Sub-TLV is invalid when SID Structure Sub-Sub-TLV transposition length is greater than 24 or addition of transposition offset and length is greater than 128. Path having such Prefix-SID Attribute should be ineligible during the selection of best path for the corresponding prefix.

9. IANA Considerations

9.1. BGP Prefix-SID TLV Types registry

This document introduces three new TLV Types of the BGP Prefix-SID attribute. IANA has assigned Type values in the registry "BGP Prefix-SID TLV Types" as follows:

Value	Type	Reference
4	Deprecated	<this document>
5	SRv6 L3 Service TLV	<this document>
6	SRv6 L2 Service TLV	<this document>

The value 4 previously corresponded to the SRv6-VPN SID TLV, which was specified in previous versions of this document and used by early implementations of this specification. It was deprecated and replaced by the SRv6 L3 Service and SRv6 L2 Service TLVs.

9.2. SRv6 Service Sub-TLV Types registry

IANA is requested to create and maintain a new registry called "SRv6 Service Sub-TLV Types". The allocation policy for this registry is:

0 : Reserved
1-127 : IETF Review
128-254 : First Come First Served
255 : Reserved

The following Sub-TLV Types are defined in this document:

Value	Type	Reference
1	SRv6 SID Information Sub-TLV	<this document>

9.3. SRv6 Service Data Sub-Sub-TLV Types registry

IANA is requested to create and maintain a new registry called "SRv6 Service Data Sub-Sub-TLV Types". The allocation policy for this registry is:

0 : Reserved
1-127 : IETF Review
128-254 : First Come First Served
255 : Reserved

The following Sub-Sub-TLV Types are defined in this document:

Value	Type	Reference
1	SRv6 SID Structure Sub-Sub-TLV	<this document>

10. Security Considerations

This document specifies extensions to BGP protocol for signalling of services for SRv6. As such, techniques related to authentication of BGP sessions for securing messages between BGP peers as discussed in the BGP specification [RFC4271] and in the security analysis for BGP [RFC4272] apply. The discussion of the use of the TCP Authentication option to protect BGP sessions is found in [RFC5925], while [RFC6952] includes an analysis of BGP keying and authentication issues.

This document does not introduce new services or BGP NLRI types but extends the signaling of existing ones for SRv6. Therefore, the security considerations for the respective BGP services [I-D.ietf-bess-rfc5549revision] [RFC4659] [RFC2545] [RFC7432] [I-D.ietf-bess-evpn-prefix-advertisement] also apply.

SRv6 operates within a trusted SR domain with filtering of traffic at the domain boundaries. These and other security aspects of SRv6 are discussed in the security considerations of [RFC8402] [RFC8754] and apply for deployment of BGP services using SRv6. The SRv6 SIDs used for the BGP Services in this document are defined in [I-D.ietf-spring-srv6-network-programming] and hence the security considerations of that document also apply. The service flows between PE routers using SRv6 SIDs advertised via BGP are expected to be limited within the trusted SR domain (e.g. within a single AS or between multiple ASes within a single provider network). Therefore, precaution is necessary to ensure that the BGP service information (including associated SRv6 SID) advertised via BGP sessions is limited to peers within this trusted SR domain. Security consideration section of [RFC8669] discuss mechanisms to prevent leaking of BGP Prefix-SID attribute, that carries SRv6 SID, outside the SR domain. In the event that these filtering mechanisms, both in the forwarding and control plane, are not implemented properly, it may be possible for nodes outside the SR domain to learn the VPN Service SIDs and use them to direct traffic into VPN networks from outside the SR domain.

11. Acknowledgments

The authors of this document would like to thank Stephane Litkowski, Rishabh Parekh and Xiejingrong for their comments and review of this document.

12. Contributors

Satoru Matsushima
SoftBank

Email: satoru.matsushima@g.softbank.co.jp

Dirk Steinberg
Steinberg Consulting

Email: dws@steinberg.net

Daniel Bernier
Bell Canada

Email: daniel.bernier@bell.ca

Daniel Voyer
Bell Canada

Email: daniel.voyer@bell.ca

Jonn Leddy
Individual

Email: john@leddy.net

Swadesh Agrawal
Cisco

Email: swaagraw@cisco.com

Patrice Brissette
Cisco

Email: pbrisset@cisco.com

Ali Sajassi
Cisco

Email: sajassi@cisco.com

Bart Peirens
Proximus
Belgium

Email: bart.peirens@proximus.com

Darren Dukes
Cisco

Email: ddukes@cisco.com

Pablo Camarilo
Cisco

Email: pcamaril@cisco.com

Shyam Sethuram
Cisco

Email: shyam.ioml@gmail.com

Zafar Ali
Cisco

Email: zali@cisco.com

Ketan Talaulikar
Cisco

Email: ketant@cisco.com

13. References

13.1. Normative References

[I-D.ietf-bess-evpn-igmp-mld-proxy]
Sajassi, A., Thoria, S., Patel, K., Drake, J., and W. Lin, "IGMP and MLD Proxy for EVPN", draft-ietf-bess-evpn-igmp-mld-proxy-05 (work in progress), April 2020.

[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-evpn-vpws-fxc]
Sajassi, A., Brissette, P., Uttaro, J., Drake, J., Lin, W., Boutros, S., and J. Rabadan, "EVPN VPWS Flexible Cross-Connect Service", draft-ietf-bess-evpn-vpws-fxc-01 (work in progress), June 2019.

[I-D.ietf-bess-rfc5549revision]
Litkowski, S., Agrawal, S., ananthamurthy, k., and K. Patel, "Advertising IPv4 Network Layer Reachability Information with an IPv6 Next Hop", draft-ietf-bess-rfc5549revision-04 (work in progress), July 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-16 (work in progress), June 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>.

[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>.

[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>.

[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>.

[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>.

[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>.

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

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

[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>.

[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>.

[RFC8200] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", STD 86, RFC 8200, DOI 10.17487/RFC8200, July 2017, <https://www.rfc-editor.org/info/rfc8200>.

[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>.

[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>.

[RFC8317] Sajassi, A., Ed., Salam, S., Drake, J., Uttaro, J., Boutros, S., and J. Rabadan, "Ethernet-Tree (E-Tree) Support in Ethernet VPN (EVPN) and Provider Backbone Bridging EVPN (PBB-EVPN)", RFC 8317, DOI 10.17487/RFC8317, January 2018, <https://www.rfc-editor.org/info/rfc8317>.

[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>.

[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>.

[RFC8754] Filsfils, C., Ed., Dukes, D., Ed., Previdi, S., Leddy, J., Matsushima, S., and D. Voyer, "IPv6 Segment Routing Header (SRH)", RFC 8754, DOI 10.17487/RFC8754, March 2020, <https://www.rfc-editor.org/info/rfc8754>.

13.2. Informative References

[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-09 (work in progress), May 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-08 (work in progress), July 2020.

[I-D.matsushima-spring-srv6-deployment-status]
Matsushima, S., Filsfils, C., Ali, Z., Li, Z., and K. Rajaraman, "SRv6 Implementation and Deployment Status", draft-matsushima-spring-srv6-deployment-status-07 (work in progress), April 2020.

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

[RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, DOI 10.17487/RFC5925, June 2010, <https://www.rfc-editor.org/info/rfc5925>.

[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>.

Authors' Addresses

Gaurav Dawra (editor)
LinkedIn
USA

Email: gdawra.ietf@gmail.com

Clarence Filsfils
Cisco Systems
Belgium

Email: cfilsfil@cisco.com

Robert Raszuk
Bloomberg LP
USA

Email: robert@raszuk.net

Bruno Decraene
Orange
France

Email: bruno.decraene@orange.com

Shunwan Zhuang
Huawei Technologies
China

Email: zhuangshunwan@huawei.com

Jorge Rabadan
Nokia
USA

Email: jorge.rabadan@nokia.com

MPLS WG
Internet-Draft
Intended status: Standards Track
Expires: September 9, 2020

K. Kompella
W. Lin
Juniper Networks

March 08, 2020

No Further Fast Reroute
draft-kompella-mpls-nffrr-00

Abstract

There are several cases where, once Fast Reroute has taken place (for MPLS protection), a second fast reroute is undesirable, even detrimental. This memo gives several examples of this, and proposes a mechanism to prevent further fast reroutes.

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 September 9, 2020.

Table of Contents

1. Introduction

MPLS Fast Reroute (FRR) [RFC4090] [RFC5286] [RFC7490] is a useful and widely deployed tool for minimizing packet loss in the case of a link or node failure. This has not only proven to be very effective, it is often the reason for using MPLS as a data plane. FRR works for a variety of control plane protocols, including LDP, RSVP-TE, and SPRING. Furthermore, FRR is often used to protect MPLS services such as IP VPN and EVPN.

Having said this, there are case where, once FRR has taken place, if the packet encounters a second failure, a second FRR is not helpful, perhaps even disruptive. For example, the packet may loop until TTL expires. This can lead to link congestion and further packet loss. Thus, the attempt to prevent a packet from being dropped may instead affect many other packets. Note that the "second" failure may simply be another manifestation of the same failure; see Figure 1.

This memo proposes a mechanism for preventing further FRR once in cases where such further protection may be harmful. Several examples where this is the case are demonstrated as motivation. A solution using special-purpose labels (SPLs) is then offered. Some mechanisms

for distributing the capability to avoid further fast reroutes are also discussed, although these may be better placed in other documents in other Working Groups.

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.

2. Motivation

A few cases are given where "further fast reroute" is harmful. Some of the cases are for MPLS services; others for "plain" MPLS forwarding.

2.1. EVPN (VPN/VPLS) Active-active Multihoming

Consider the following topology for multihoming an Ethernet VPN (EVPN [RFC7432]) Customer Edge (CE) device for protection against the failure of a Provider Edge (PE) device or a PE-CE link. To do so, there is a backup MPLS path between PE2 and PE3 (denoted by the starred line).

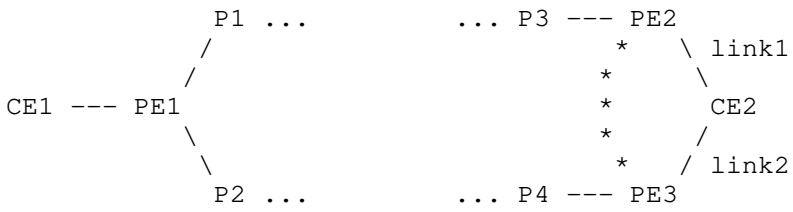


Figure 1: EVPN Multihoming

Suppose (known unicast) traffic goes from CE1 to CE2. With active-active multihoming, this traffic will be load-balanced between PE2 (to CE2 via link link1) and PE3 (to CE2 via link2). If link1 were to fail, PE2 can still get traffic for CE2 by sending it over the backup path to PE3 (and similarly for PE3 if link2 fails).

However, suppose CE2 is down. PE2 will assume link1 is down and send traffic for CE2 to PE3 over the backup path. PE3 (which thinks that link2 is down; note that the single real failure of CE2 being down is manifested as separate failures to PE2 and PE3) will protect this "second" failure by sending traffic for CE2 over the backup path to

PE2. Thus, traffic will ping-pong between PE2 and PE3 until TTL expires.

Thus, the attempt to protect traffic to CE2 may end up doing more harm than good, by congesting the backup path between PE2 and PE3 and by giving PE2 and PE3 useless work to do.

A similar topology can be used in EVPN-Etree [RFC8317], EVPN-VPWS [RFC8214], IP VPN [RFC4364] or VPLS [RFC4761] [RFC4762]. In all these cases, the same looping behavior would occur for unicast traffic if CE2 is down.

2.2. RMR Protection

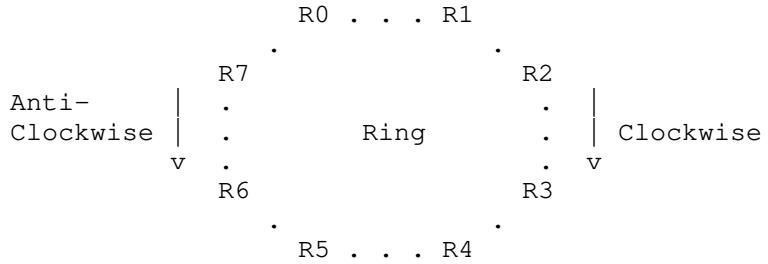


Figure 2: RMR Looping

In Resilient MPLS Rings (RMR), suppose traffic goes from a node, say R0, to a node, say R4, over a clockwise path. Protection consists of switching this traffic onto the anti-clockwise path to R4. This works well if a node or link between R0 or R4 is down. However, if node R4 itself is down, its adjacent neighbor R3, will send the traffic anti-clockwise to R4; when this traffic reaches R4's other neighbor R5, it will return to N3, and so on, until TTL expires. [I-D.ietf-mpls-rmr] provides more details, and offers some means of mitigation. This memo offers a more elegant solution.

2.3. General MPLS forwarding

Consider the following topology:

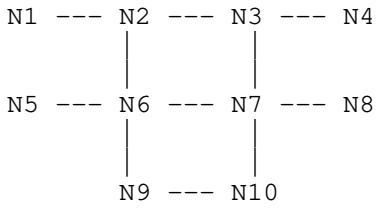


Figure 3: General MPLS Forwarding

Say link protection is configured for links N2-N3 and N6-N7. Link N2-N3 is protected by a bypass tunnel N2-N6-N7-N3, and link N7-N3 is protected by a bypass tunnel N7-N6-N2-N3. (These bypass tunnels may be set up using RSVP-TE [RFC3209] or via SPRING stacks [RFC8660].) Say furthermore that there is an LSP from N1 to N4 with path N1-N2-N3-N4, which asks for link protection. If link N2-N3 fails, traffic will take the path N1-N2-N6-N7-N3-N4.

Suppose, however, links N2-N3 and N7-N3 fail simultaneously. This may happen if they share fate (e.g., go over a common fiber conduit); it may also appear to happen if node N3 fails. Either way, first, the bypass protecting link N2-N3 kicks in, and traffic is sent to N3 via N6 and N7. However, when the traffic hits N7, the bypass for N7-N3 kicks in, and traffic is sent back to N2. Thus the traffic will loop between N2 and N7 until TTL expires, in the process congesting links N2-N6 and N6-N7.

Now consider an LSP: N5-N6-N7-N8. The link N6-N7 may be protected by the bypass N6-N2-N3-N7 or by N6-N9-N10-N7, or by load-balancing between these two bypasses. If both links N2-N3 and N6-N7 fail, then traffic that is protected via bypass N6-N2-N3-N7 will ping-pong between N6 and N2 until TTL expires; traffic protected via bypass N6-N9-N10-N7 will successfully make it to N8. If link N6-N7 is protected by load-balancing across the two bypass paths, then about half the traffic will loop between N6 and N2, and the rest will make it to N8.

While the above description is for protection using a bypass tunnel, the same principle applies to protection using Loop-Free Alternates [RFC5286] [RFC7490] or any of its variants (such as Topology Independent LFA).

3. Solution

To address this issue, we suggest the use of a SPL [RFC7274] called NFFRR (value TBD; suggested: 8). An alternate would be to use an extended SPL, whereby a pair of labels indicates that no further fast route is desired. However, in the case of SPRING MPLS bypass tunnels

(Section 3.2.1) of depth N, this would triple the label stack size. Using regular SPLs instead would only double the stack size.

3.1. NFFRR for MPLS forwarding

To illustrate, we'll first take the example of Figure 3, with MPLS paths signaled using RSVP-TE. This method can be used for paths that use SPRING stacks, but this will be detailed in a later version.

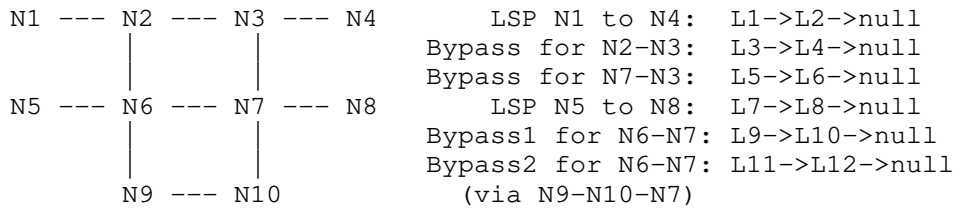


Figure 4: Example Using RSVP-TE LSPs

Node	Action	Next	New Pkt	Comment
N1	push L1	N2	[L1] pkt	ingress
N2	L1 -> L2	N3	[L2] pkt	
N3	pop L2	N4	pkt	PHP
N4	fwd pkt	-	-	continue

Table 1: Forwarding from N1 to N4

Note 1: "[L1 ...]" denotes the label stack on the packet; pkt is the original packet received at ingress. "L1 -> L2" means swap label L1 with L2. "pop L2" means pop the top label L2. "fwd pkt" means forward the packet as usual.

Node	Action	Next	New Pkt	Comment
N2	push L3	N6	[L3] pkt	ingress
N6	L3 -> L4	N7	[L4] pkt	
N7	pop L4	N3	pkt	PHP

Table 2: Forwarding over the bypass for link N2-N3

Node	Action	Next	New Pkt	Comment
N7	push L5	N6	[L5] pkt	ingress
N6	L5 -> L6	N2	[L6] pkt	
N2	pop L6	N3	pkt	PHP

Table 3: Forwarding over Bypass1 for link N7-N3

Node	Action	Next	New Pkt	Comment
N1	push L1	N2	[L1] pkt	ingress
N2	L1 -> L2	N3	[L2] pkt	N3 X
N2	push L3	N6	[L3 L2] pkt	PLR
N6	L3 -> L4	N7	[L4 L2] pkt	
N7	pop L4	N3	[L2] pkt	merge
N3	pop L2	N4	pkt	PHP
N4	fwd pkt	-	-	continue

Table 4: Forwarding from N1 to N4 if link N2-N3 fails

Table 4 is obtained by composing Table 1 and Table 2.

Note 2: "N3 X" means "next hop N3 unavailable (because link N2-N3 failed)".

Node	Action	Next	New Pkt	Comment
N1	push L1	N2	[L1] pkt	ingress
N2	L1 -> L2	N3	[L2] pkt	N3 X
N2	push L3	N6	[L3 L2] pkt	PLR
N6	L3 -> L4	N7	[L4 L2] pkt	
N7	pop L4	N3	[L2] pkt	N3 X’
N7	push L5	N6	[L5 L2] pkt	
N6	L5 -> L6	N2	[L6 L2] pkt	PLR
N2	pop L6	N3	[L2] pkt	N3 X
N2	push L3	N6	[L3 L2]	PLR
etc				loop!

Table 5: Forwarding from N1 to N4 if links N2-N3 and N7-N3 fail

Table 5 is obtained by composing Table 1, Table 2 and Table 3.

Note 3: "N3 X’" means "next hop N3 unavailable because link N7-N3 is down.

Note 4: While the impact of a loop is pretty bad, the impact of an ever-growing label stack (not illustrated here) and possible associated fragmentation on transit nodes may be worse.

3.2. Proposal

An LSR (typically a PLR) that wishes to prevent further FRRs after the first one can push an SPL, namely NFFRR, onto the label stack as follows:

Node	Action	Next	New Pkt	Comment
N1	push L1	N2	[L1] pkt	ingress
N2	L1 -> L2	N3	[L2] pkt	N3 X
N2	push L3, NFFRR	N6	[L3 NFFRR L2] pkt	PLR
N6	L3 -> L4	N7	[L4 NFFRR L2] pkt	
N7	pop L4, NFFRR	N3	[L2] pkt	merge
N3	pop L2	N4	pkt	PHP
N4	fwd pkt	-	-	continue

Table 6: Forwarding from N1 to N4 if link N2-N3 fails with NFFRR

Note 5: N2 can insert an NFFRR label only if it knows that all LSRs in the path can process it correctly. See Section 4 for some details on how this capability is communicated.

Node	Action	Next	New Pkt	Comment
N1	push L1	N2	[L1] pkt	ingress
N2	L1 -> L2	N3	[L2] pkt	N3 X
N2	push L3, NFFRR	N6	[L3 NFFRR L2] pkt	PLR
N6	L3 -> L4	N7	[L4 NFFRR L2] pkt	
N7	pop L4	N3	[NFFRR L2] pkt	N3 X
N7	check NFFRR	-	-	drop pkt

Table 7: Forwarding from N1 to N4 if links N2-N3 and N7-N3 fail with NFFRR

Note 6: "check NFFRR" means that, before N7 applies FRR (because link N7-N3 is down), N7 checks the label below the top label (or in this case, because of PHP, the top label itself). If this is the NFFRR label, N7 drops the packet rather than apply FRR.

3.2.1. NFFRR and SPRING

Suppose that, to protect link N2-N3, a bypass tunnel N2-N6-N7-N3 were instantiated using SPRING MPLS [RFC8660], in particular, using adjacency SIDs. If the corresponding labels for links N6-N7 and N7-N3 were L20 and L21, the bypass would consist of pushing the label stack [L20 L21] onto the packet and sending the packet to N6. To indicate that FRR has already occurred and to drop the packet rather than to try to protect the packet again, N2 would have to push [L20 NFFRR L21 NFFRR] onto the packet before sending it to N6. If the packet came from N1 with label L1, N2 would send a packet with label stack [L20 NFFRR L21 NFFRR L2] to N6.

N6 would see L20, pop it, note the NFFRR label and pop it, then attempt to send the packet to N7. If the link N6-N7 is down, N6 drops the packet. Otherwise, N7 gets the packet, sees L21, pops it, sees NFFRR, pops it and tries to send the packet to N3. If link N7-N3 is down, N7 drops the packet. Otherwise, N3 gets the packet with L2, swaps with with L3 and sends it to N4.

Note that with SPRING MPLS, the NFFRR label needs to be repeated for each label in the bypass stack. Hence the request for a "regular" SPL rather than an extended SPL.

3.3. NFFRR for MPLS Services

First, we illustrate known unicast EVPN forwarding:

Node	Action	Next	Packet	Comment
PE1	send to CE2	PE2	[T1 S2] pkt	EVPN
PE2	send to CE2	link1	pkt	done!

Note: T1/T2/T3 are the transport labels for PE1/PE3/PE2 to reach PE2/PE2/PE3 respectively. S2/S3 are the service labels announced by PE2/PE3 for CE2.

Then, we show what happens when CE2 is down without NFFRR:

Node	Action	Next	Packet	Comment
PE1	send to CE2	PE2	[T1 S2] pkt	EVPN
PE2	send to CE2	link1	--	link1 X
PE2	send to CE2	PE3	[T3 S3] pkt	eFRR
PE3	send to CE2	link2	--	link2 X
PE3	send to CE2	PE2	[T2 S2] pkt	eFRR
PE2	send to CE2	link1	--	link1 X
PE2	send to CE2	PE3	[T3 S3] pkt	eFRR
...				loop!

Note: link1/link2 X means link1/link2 is down. eFRR refers to EVPN multihoming FRR.

In the case of MPLS services such as EVPN Figure 1, the NFFRR label is inserted below the service label, as shown below:

Node	Action	Next	Packet	Comment
PE1	send to CE2	PE2	[T1 S2] pkt	EVPN
PE2	send to CE2	link1	--	link1 X
PE2	send to CE2	PE3	[T3 S2 NFFRR] pkt	eFRR
PE3	send to CE2	link2	--	link2 X
PE3	drop pkt	--	--	check NFFRR

Note: "check NFFRR" is as above.

3.4. NFFRR for RMR

As described in Figure 2, packets will loop until TTL expires if the destination node in an RMR ring (here, R4) fails. The solution in this case is that the first node to apply RMR protection (R3) pops the current RMR transport label being used, sees that the next label

is not NFFRR (so protection is allowed), pushes an NFFRR label and then the RMR transport label for the reverse direction.

When R5 receives the packet, it sees that the next link is down, pops the RMR transport label, sees the NFFRR label and drops the packet. Thus, the loop is avoided.

4. Signaling NFFRR Capability

4.1. Signaling NFFRR Capability for MPLS Services with BGP

The ideal choice would be an attribute consisting of a bit vector of node capabilities, one bit of which would be the capability of processing the NFFRR SPL below the BGP service label. This would be used by BGP L2VPN, BGP VPLS, EVPN, E-Tree and E-VPWS. An alternative is to use the BGP Capabilities Optional Parameter [I-D.ietf-idr-next-hop-capability]. Details to be worked out.

4.2. Signaling NFFRR Capability for MPLS Services with Targeted LDP

One approach to signaling NFFRR capability for MPLS services signaled with targeted LDP is to introduce a new LDP TLV called the NFFRR Capability TLV as an Optional Parameter in the Label Mapping Message [RFC5036]. This TLV has Type TBD (suggested: 0x0207) and Length 0.

Another approach is to use LDP Capabilities [RFC5561]; this approach has the advantage that it deals with capabilities on a node basis rather than on a per label mapping basis. However, there don't appear to be other documents using this approach.

4.3. Signaling NFFRR Capability for MPLS Forwarding

The authors suggest signaling a router's ability to process the NFFRR SPL using the Link State Router TE Node Capabilities [RFC5073], which works for both IS-IS and OSPF. A new TE Node Capability bit, the N bit (suggested value 5) indicates that the advertising node is capable of processing the NFFRR SPL.

5. IANA Considerations

If this draft is deemed useful, an SPL for NFFRR will need to be allocated. We suggest the early allocation of label 8 for this.

Furthermore, means of signaling the ability to process the NFFRR SPL should be defined for IS-IS, OSPF, LDP and BGP.

The following update is suggested for the Link State Router TE Node Capabilities registry:

Bit	Name	Reference
5	NFFRR	This docusment

The following update is suggested for the TLV Type Name Space of the Label Distribution Protocol (LDP) Parameters registry:

Type	Name	Reference
0x0207	NFFRR	This docusment

6. Security Considerations

A malicious or compromised LSR can insert NFFRR into a label stack, preventing FRR from occurring. If so, protection will not kick in for failures that could have been protected, and there will be unnecessary packet loss.

7. References

7.1. Normative References

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

[RFC5036] Andersson, L., Ed., Minei, I., Ed., and B. Thomas, Ed., "LDP Specification", RFC 5036, DOI 10.17487/RFC5036, October 2007, <https://www.rfc-editor.org/info/rfc5036>.

[RFC5073] Vasseur, J., Ed. and J. Le Roux, Ed., "IGP Routing Protocol Extensions for Discovery of Traffic Engineering Node Capabilities", RFC 5073, DOI 10.17487/RFC5073, December 2007, <https://www.rfc-editor.org/info/rfc5073>.

[RFC7274] Kompella, K., Andersson, L., and A. Farrel, "Allocating and Retiring Special-Purpose MPLS Labels", RFC 7274, DOI 10.17487/RFC7274, June 2014, <https://www.rfc-editor.org/info/rfc7274>.

[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>.

7.2. Informative References

[I-D.ietf-idr-next-hop-capability]
Decraene, B., Kompella, K., and W. Henderickx, "BGP Next-Hop dependent capabilities", draft-ietf-idr-next-hop-capability-05 (work in progress), June 2019.

[I-D.ietf-mpls-rmr]
Kompella, K. and L. Contreras, "Resilient MPLS Rings", draft-ietf-mpls-rmr-12 (work in progress), October 2019.

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, DOI 10.17487/RFC3209, December 2001, <https://www.rfc-editor.org/info/rfc3209>.

[RFC4090] Pan, P., Ed., Swallow, G., Ed., and A. Atlas, Ed., "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, DOI 10.17487/RFC4090, May 2005, <https://www.rfc-editor.org/info/rfc4090>.

[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>.

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

[RFC4762] Lasserre, M., Ed. and V. Kompella, Ed., "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, DOI 10.17487/RFC4762, January 2007, <https://www.rfc-editor.org/info/rfc4762>.

[RFC5286] Atlas, A., Ed. and A. Zinin, Ed., "Basic Specification for IP Fast Reroute: Loop-Free Alternates", RFC 5286, DOI 10.17487/RFC5286, September 2008, <https://www.rfc-editor.org/info/rfc5286>.

[RFC5561] Thomas, B., Raza, K., Aggarwal, S., Aggarwal, R., and JL. Le Roux, "LDP Capabilities", RFC 5561, DOI 10.17487/RFC5561, July 2009, <https://www.rfc-editor.org/info/rfc5561>.

[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>.

[RFC7490] Bryant, S., Filsfils, C., Previdi, S., Shand, M., and N. So, "Remote Loop-Free Alternate (LFA) Fast Reroute (FRR)", RFC 7490, DOI 10.17487/RFC7490, April 2015, <https://www.rfc-editor.org/info/rfc7490>.

[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>.

[RFC8317] Sajassi, A., Ed., Salam, S., Drake, J., Uttaro, J., Boutros, S., and J. Rabadan, "Ethernet-Tree (E-Tree) Support in Ethernet VPN (EVPN) and Provider Backbone Bridging EVPN (PBB-EVPN)", RFC 8317, DOI 10.17487/RFC8317, January 2018, <https://www.rfc-editor.org/info/rfc8317>.

[RFC8660] Bashandy, A., Ed., Filsfils, C., Ed., Previdi, S., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing with the MPLS Data Plane", RFC 8660, DOI 10.17487/RFC8660, December 2019, <https://www.rfc-editor.org/info/rfc8660>.

Authors' Addresses

Kireeti Kompella
Juniper Networks
1133 Innovation Way
Sunnyvale, CA 94089
United States

Email: kireeti.kompella@gmail.com

Wen Lin
Juniper Networks
1133 Innovation Way
Sunnyvale, CA 94089
United States

Email: wlin@juniper.net

BESS Workgroup
INTERNET-DRAFT
Intended Status: Standards Track

A. Sajassi, Ed.
A. Banerjee
S. Thoria
D. Carrel
Cisco
B. Weis
Individual
J. Drake
Juniper

Expires: January 8, 2020

July 8, 2019

Secure EVPN
draft-sajassi-bess-secure-evpn-02

Abstract

The applications of EVPN-based solutions ([RFC7432] and [RFC8365]) have become pervasive in Data Center, Service Provider, and Enterprise segments. It is being used for fabric overlays and inter-site connectivity in the Data Center market segment, for Layer-2, Layer-3, and IRB VPN services in the Service Provider market segment, and for fabric overlay and WAN connectivity in Enterprise networks. For Data Center and Enterprise applications, there is a need to provide inter-site and WAN connectivity over public Internet in a secured manner with same level of privacy, integrity, and authentication for tenant's traffic as IPsec tunneling using IKEv2. This document presents a solution where BGP point-to-multipoint signaling is leveraged for key and policy exchange among PE devices to create private pair-wise IPsec Security Associations without IKEv2 point-to-point signaling or any other direct peer-to-peer session establishment messages.

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

Copyright and License Notice

Table of Contents

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.

AC: Attachment Circuit.

ARP: Address Resolution Protocol.

BD: Broadcast Domain. As per [RFC7432], an EVI consists of a single or multiple BDs. In case of VLAN-bundle and VLAN-based service models (see [RFC7432]), a BD is equivalent to an EVI. In case of VLAN-aware bundle service model, an EVI contains multiple BDs. Also, in this document, BD and subnet are equivalent terms.

BD Route Target: refers to the Broadcast Domain assigned Route Target [RFC4364]. In case of VLAN-aware bundle service model, all the BD instances in the MAC-VRF share the same Route Target.

BT: Bridge Table. The instantiation of a BD in a MAC-VRF, as per [RFC7432].

DGW: Data Center Gateway.

Ethernet A-D route: Ethernet Auto-Discovery (A-D) route, as per [RFC7432].

Ethernet NVO tunnel: refers to Network Virtualization Overlay tunnels with Ethernet payload. Examples of this type of tunnels are VXLAN or GENEVE.

EVI: EVPN Instance spanning the NVE/PE devices that are participating on that EVPN, as per [RFC7432].

EVPN: Ethernet Virtual Private Networks, as per [RFC7432].

GRE: Generic Routing Encapsulation.

GW IP: Gateway IP Address.

IPL: IP Prefix Length.

IP NVO tunnel: it refers to Network Virtualization Overlay tunnels with IP payload (no MAC header in the payload).

IP-VRF: A VPN Routing and Forwarding table for IP routes on an NVE/PE. The IP routes could be populated by EVPN and IP-VPN address families. An IP-VRF is also an instantiation of a layer 3 VPN in an NVE/PE.

IRB: Integrated Routing and Bridging interface. It connects an IP-VRF to a BD (or subnet).

MAC-VRF: A Virtual Routing and Forwarding table for Media Access Control (MAC) addresses on an NVE/PE, as per [RFC7432]. A MAC-VRF is also an instantiation of an EVI in an NVE/PE.

ML: MAC address length.

ND: Neighbor Discovery Protocol.

NVE: Network Virtualization Edge.

GENEVE: Generic Network Virtualization Encapsulation, [GENEVE].

NVO: Network Virtualization Overlays.

RT-2: EVPN route type 2, i.e., MAC/IP advertisement route, as defined in [RFC7432].

RT-5: EVPN route type 5, i.e., IP Prefix route. As defined in Section 3 of [EVPN-PREFIX].

SBD: Supplementary Broadcast Domain. A BD that does not have any ACs, only IRB interfaces, and it is used to provide connectivity among all the IP-VRFs of the tenant. The SBD is only required in IP-VRF- to-IP-VRF use-cases (see Section 4.4.).

SN: Subnet.

TS: Tenant System.

VA: Virtual Appliance.

VNI: Virtual Network Identifier. As in [RFC8365], the term is used as a representation of a 24-bit NVO instance identifier, with the understanding that VNI will refer to a VXLAN Network Identifier in VXLAN, or Virtual Network Identifier in GENEVE, etc. unless it is stated otherwise.

VTEP: VXLAN Termination End Point, as in [RFC7348].

VXLAN: Virtual Extensible LAN, as in [RFC7348].

This document also assumes familiarity with the terminology of [RFC7432], [RFC8365] and [RFC7365].

1 Introduction

The applications of EVPN-based solutions have become pervasive in Data Center, Service Provider, and Enterprise segments. It is being used for fabric overlays and inter-site connectivity in the Data Center market segment, for Layer-2, Layer-3, and IRB VPN services in the Service Provider market segment, and for fabric overlay and WAN connectivity in the Enterprise networks. For Data Center and Enterprise applications, there is a need to provide inter-site and WAN connectivity over public Internet in a secured manner with the same level of privacy, integrity, and authentication for tenant's traffic as used in IPsec tunneling using IKEv2. This document presents a solution where BGP point-to-multipoint signaling is leveraged for key and policy exchange among PE devices to create private pair-wise IPsec Security Associations without IKEv2 point-to-point signaling or any other direct peer-to-peer session establishment messages.

EVPN uses BGP as control-plane protocol for distribution of information needed for discovery of PEs participating in a VPN, discovery of PEs participating in a redundancy group, customer MAC addresses and IP prefixes/addresses, aliasing information, tunnel encapsulation types, multicast tunnel types, multicast group memberships, and other info. The advantages of using BGP control plane in EVPN are well understood including the following:

1) A full mesh of BGP sessions among PE devices can be avoided by using Route Reflector (RR) where a PE only needs to setup a single BGP session between itself and the RR as opposed to setting up N BGP sessions to N other remote PEs; therefore, reducing number of BGP sessions from O(N^2) to O(N) in the network. Furthermore, RR hierarchy can be leveraged to scale the number of BGP routes on the RR.

2) MP-BGP route filtering and constrained route distribution can be leveraged to ensure that the control-plane traffic for a given VPN is only distributed to the PEs participating in that VPN.

For setting up point-to-point security association (i.e., IPsec tunnel) between a pair of EVPN PEs, it is important to leverage BGP point-to-multipoint singling architecture using the RR along with its route filtering and constrain mechanisms to achieve the performance and the scale needed for large number of security associations (IPsec tunnels) along with their frequent re-keying requirements. Using BGP signaling along with the RR (instead of peer-to-peer protocol such as IKEv2) reduces number of message exchanges needed for SAs establishment and maintenance from O(N^2) to O(N) in the network.

2 Requirements

The requirements for secured EVPN are captured in the following subsections.

2.1 Tenant's Layer-2 and Layer-3 data & control traffic

Tenant's layer-2 and layer-3 data and control traffic must be protected by IPsec cryptographic methods. This implies not only tenant's data traffic must be protected by IPsec but also tenant's control and routing information that are advertised in BGP must also be protected by IPsec. This in turn implies that BGP session must be protected by IPsec.

2.2 Tenant's Unicast & Multicast Data Protection

Tenant's layer-2 and layer-3 unicast traffic must be protected by IPsec. In addition to that, tenant's layer-2 broadcast, unknown unicast, and multicast traffic as well as tenant's layer-3 multicast traffic must be protected by IPsec when ingress replication or assisted replication are used. The use of BGP P2MP signaling for setting up P2MP SAs in P2MP multicast tunnels is for future study.

2.3 P2MP Signaling for SA setup and Maintenance

BGP P2MP signaling must be used for IPsec SAs setup and maintenance. The BGP signaling must follow P2MP signaling framework per [CONTROLLER-IKE] for IPsec SAs setup and maintenance in order to reduce the number of message exchanges from O(N^2) to O(N) among the participant PE devices.

2.4 Granularity of Security Association Tunnels

The solution must support the setup and maintenance of IPsec SAs at the following level of granularities:

1) Per PE: A single IPsec tunnel between a pair of PEs to be used for all tenants' traffic supported by the pair of PEs.

2) Per tenant: A single IPsec tunnel per tenant per pair of PEs. For example, if there are 1000 tenants supported on a pair of PEs, then 1000 IPsec tunnels are required between that pair of PEs.

3) Per subnet: A single IPsec tunnel per subnet (e.g., per VLAN/EVI) of a tenant on a pair of PEs.

4) Per IP address: A single IPsec tunnel per pair of IP addresses of a tenant on a pair of PEs.

5) Per MAC address: A single IPsec tunnel per pair of MAC addresses of a tenant on a pair of PEs.

6) Per Attachment Circuit: A single IPsec tunnel per pair of Attachment Circuits between a pair of PEs.

2.5 Support for Policy and DH-Group List

The solution must support a single policy and DH group for all SAs as well as supporting multiple policies and DH groups among the SAs.

3 BGP Component

The architecture that encompasses device-to-controller trust model, has several components among which is the signaling component. Secure EVPN Signaling, as defined in this document, is the BGP signaling component of the overall Architecture. We will briefly describe this Architecture here to further facilitate understanding how Secure EVPN fits into the overall architecture. The Architecture describes the components needed to create BGP based SD-WANs and how these components work together. Our intention is to list these components here along with their brief description and to describe this Architecture in details in a separate document where to specify the details for other parts of this architecture besides the BGP signaling component which is described in this document.

The Architecture consists of four components. These components are Zero Touch Bring-up, Configuration Management, Orchestration, and Signaling. In addition to these components, secure communications must be provided between the edge nodes and all servers/devices providing the architecture components.

3.1 Zero Touch Bring-up (ZTB)

The first component is a zero touch capability that allows an edge device to find and join its SD-WAN with little to no assistance other than power and network connectivity. The goal is to use existing work in this area. The requirements are that an edge device can locate its ZTB server/component of its SD-WAN controller in a secure manner and to proceed to receive its configuration.

3.2 Configuration Management

After an edge device joins its SD-WAN, it needs to be configured.

Configuration covers all device configuration, not just the configuration related to Secure EVPN. The previous Zero Touch Bring-up component will have directed the edge device, either directly or indirectly, to its configuration server/component. One example of a configuration server is the I2NSF Controller. After a device has been configured, it can engage in the next two components. Configuration may include updates over time and is not a one time only component.

3.3 Orchestration

This component is optional. It allows for more dynamic updates of configuration and statistics information. Orchestration can be more dynamic than configuration.

3.4 Signaling

Signaling is the component described in this document. The functionality of a Route Reflector is well understood. Here we describe the signaling component of BGP SD-WAN Architecture and the BGP extension/signaling for IPsec key management and policy.

4 Solution Description

This solution uses BGP P2MP signaling where an originating PE only send a message to the Route Reflector (RR) and then the RR reflects that message to the interested recipient PEs. The framework for such signaling is described in [CONTROLLER-IKE] and it is referred to as device-to-controller trust model. This trust model is significantly different than the traditional peer-to-peer trust model where a P2P signaling protocol such as IKEv2 [RFC7296] is used in which the PE devices directly authenticate each other and agree upon security policy and keying material to protect communications between themselves. The device-to-controller trust model leverages P2MP signaling via the controller (e.g., the RR) to achieve much better scale and performance for establishment and maintenance of large number of pair-wise Security Associations (SAs) among the PEs.

This device-to-controller trust model first secures the control channel between each device and the controller using peer-to-peer protocol such as IKEv2 [RFC7296] to establish P2P SAs between each PE and the RR. It then uses this secured control channel for P2MP signaling in establishment of P2P SAs between each pair of PE devices.

Each PE advertises to other PEs via the RR the information needed in establishment of pair-wise SAs between itself an every other remote PEs. These pieces of information are sent as Sub-TLVs of IPSec tunnel type in BGP Tunnel Encapsulation attribute. These Sub-TLVs are detailed in section 5 and are based on the DIM message components from [CONTROLLER-IKE] and the IKEv2 specification [RFC7296]. The IPsec tunnel TLVs along with its Sub-TLVs are sent along with the BGP route (NLRI) for a given level of granularity.

If only a single SA is required per pair of PE devices to multiplex user traffic for all tenants, then IPsec tunnel TLV is advertised along with IPv4 or IPv6 NLRI representing loopback address of the originating PE. It should be noted that this is not a VPN route but rather an IPv4 or IPv6 route.

If a SA is required per tenant between a pair of PE devices, then IPsec tunnel TLV can be advertised along with EVPN IMET route representing the tenant or can be advertised along with a new EVPN route representing the tenant.

If a SA is required per tenant’s subnet (e.g., per VLAN) between a pair of PE devices, then IPsec tunnel TLV is advertised along with EVPN IMET route.

If a SA is required between a pair of tenant’s devices represented by a pair of IP addresses, then IPsec tunnel TLV is advertised along with EVPN IP Prefix Advertisement Route or EVPN MAC/IP Advertisement route.

If a SA is required between a pair of tenant’s devices represented by a pair of MAC addresses, then IPsec tunnel TLV is advertised along with EVPN MAC/IP Advertisement route.

If a SA is required between a pair of Attachment Circuits (ACs) on two PE devices (where an AC can be represented by <VLAN, port>), then IPsec tunnel TLV is advertised along with EVPN Ethernet AD route.

4.1 Inheritance of Security Policies

Operationally, it is easy to configure a security association between a pair of PEs using BGP signaling. This is the default security association that is used for traffic that flows between peers. However, in the event more finer granularity of security association is desired on the traffic flows, it is possible to set up SAs between a pair of tenants, a pair of subnets within a tenant, a pair of IPs between a subnet, and a pair of MACs between a subnet using the appropriate EVPN routes as described above. In the event, there are no security TLVs associated with an EVPN route, there is a strict

order in the manner security associations are inherited for such a route. This results in an EVPN route inheriting the security associations of the parent in a hierarchical fashion. For example, traffic between an IP pair is protected using security TLVs announced along with the EVPN IP Prefix Advertisement Route or EVPN MAC/IP Advertisement route as a first choice. If such TLVs are missing with the associated route, then one checks to see if the subnets the IPs are associated with has security TLVs with the EVPN IMET route. If they are present, those associations are used in securing the traffic. In the absence of them, the peer security associations are used. The order in which security associations are inherited are from the granular to the coarser, namely, IP/MAC associated TLVs with the EVPN route being the first preference, and the subnet, the tenant, and the peer associations preferred in that fashion.

It should be noted that when a security association is made it is possible for it to be re-used by a large number of traffic flows. For example, a tenant security association may be associated with a number of child subnet routes. Clearly it is mandatory to keep a tenant security association alive, if there are one or more subnet routes that want to use that association. Logically, the security associations between a pair of entities creates a single secure tunnel. It is thus possible to classify the incoming traffic in the most granular sense {IP/MAC, subnet, tenant, peer} to a particular secure tunnel that falls within its route hierarchy. The policy that is applied to such traffic is independent from its use of an existing or a new secure tunnel. It is clear that since any number of classified traffic flows can use a security association, such a security association will not be torn down, if at least there is one policy using such a secure tunnel.

4.2 Distribution of Public Keys and Policies

One of the requirements for this solution is to support a single DH group and a single policy for all SAs as well as to support multiple DH groups and policies among the SAs. The following subsections describe what pieces of information (what Sub-TLVs) are needed to be exchanged to support a single DH group and a single policy versus multiple DH groups and multiple policies.

4.2.1 Minimal DIM

For SA establishment, at the minimum, a PE needs to advertise to other PEs, its DIM values as specified in [CONTROLLER-IKE]. These include:

ID Tunnel ID
N Nonce

RC Rekey Counter
I Indication of initial policy distribution
KE DH public value.

When this minimal set of DIM values is sent, then it is assumed that all peer PEs share the same policy for which DH group to use, as well as which IPSec SA policy to employ. Section 5.1 defines the Minimal DIM sub-TLV as part of IPsec tunnel TLV in BGP Tunnel Encapsulation Attribute.

4.2.2 Multiple Policies

There can be scenarios for which there is a need to have multiple policy options. This can happen when there is a need for policy change and smooth migration among all PE devices to the new policy is required. It can also happen if different PE devices have different capabilities within the network. In these scenarios, PE devices need to be able to choose the correct policy to use for each other. This multi-policy scheme is described in section 6 of [CONTROLLER-IKE]. In order to support this multi-policy feature, a PE device MUST distribute a policy list. This list consists of multiple distinct policies in order of preference, where the first policy is the most preferred one. The receiving PE selects the policy by taking the received list (starting with the first policy) and comparing that against its own list and choosing the first one found in common. If there is no match, this indicates a configuration error and the PEs MUST NOT establish new SAs until a message is received that does produce a match.

4.2.2.1 Multiple DH-groups

It can be the case that not all peers use the same DH group. When multiple DH groups are supported, the peer may include multiple KE Sub-TLVs. The order of the KE Sub-TLVs determines the preference. The preference and selection methods are specified in Section 6 of [CONTROLLER-IKE].

4.2.2.2 Multiple or Single ESP SA policies

In order to specify an ESP SA Policy, a DIM may include one or more SA Sub-TLVs. When all peers are configured by a controller with the same ESP SA policy, they MAY leave the SA out of the DIM. This minimizes messaging when group configuration is static and known. However, it may also be desirable to include the SA. If a single SA is included, the peer is indicating what ESP SA policy it uses, but is not willing to negotiate. If multiple SA Sub-TLVs are included, the peer is indicating that it is willing to negotiate. The order of

the SA Sub-TLVs determines the preference. The preference and selection methods are specified in Section 6 of [CONTROLLER-IKE].

4.3 Initial IPsec SAs Generation

The procedure for generation of initial IPsec SAs is described in section 3 of [CONTROLLER-IKE]. This section gives a summary of it in context of BGP signaling. When a PE device first comes up and wants to setup an IPsec SA between itself and each of the interested remote PEs, it generates a DH pair along for each [what word here? "tennant"?] using an algorithm defined in the IKEv2 Diffie-Hellman Group Transform IDs [IKEv2-IANA]. The originating PE distributes the DH public value along with the other values in the DIM (using IPsec Tunnel TLV in Tunnel Encapsulation Attribute) to other remote PEs via the RR. Each receiving PE uses this DH public number and the corresponding nonce in creation of IPsec SA pair to the originating PE - i.e., an outbound SA and an inbound SA. The detail procedures are described in section 5.2 of [CONTROLLER-IKE].

4.4 Re-Keying

A PE can initiate re-keying at any time due to local time or volume based policy or due to the result of cipher counter nearing its final value. The rekey process is performed individually for each remote PE. If rekeying is performed with multiple PEs simultaneously, then the decision process and rules described in this rekey are performed independently for each PE. Section 4 of [CONTROLLER-IKE] describes this rekeying process in details and gives examples for a single IPsec device (e.g., a single PE) rekey versus multiple PE devices rekey simultaneously.

4.5 IPsec Databases

The Peer Authorization Database (PAD), the Security Policy Database (SPD), and the Security Association Database (SAD) all need to be setup as defined in the IPsec Security Architecture [RFC4301]. Section 5 of [CONTROLLER-IKE] gives a summary description of how these databases are setup for the controller-based model where key is exchanged via P2MP signaling via the controller (i.e., the RR) and the policy can be either signaled via the RR (in case of multiple policies) or configured by the management station (in case of single policy).

5 Encapsulation

Vast majority of Encapsulation for Network Virtualization Overlay (NVO) networks in deployment are based on UDP/IP with UDP destination port ID indicating the type of NVO encapsulation (e.g., VxLAN, GPE, GENEVE, GUE) and UDP source port ID representing flow entropy for load-balancing of the traffic within the fabric based on n-tuple that includes UDP header. When encrypting NVO encapsulated packets using IP Encapsulating Security Payload (ESP), the following two options can be used: a) adding a UDP header before ESP header (e.g., UDP header in clear) and b) no UDP header before ESP header (e.g., standard ESP encapsulation). The following subsection describe these encapsulation in further details.

5.1 Standard ESP Encapsulation

When standard IP Encapsulating Security Payload (ESP) is used (without outer UDP header) for encryption of NVO packets, it is used in transport mode as depicted below. When such encapsulation is used, for BGP signaling, the Tunnel Type of Tunnel Encapsulation TLV is set to ESP-Transport and the Tunnel Type of Encapsulation Extended Community is set to NVO encapsulation type (e.g., VxLAN, GENEVE, GPE, etc.). This implies that the customer packets are first encapsulated using NVO encapsulation type and then it is further encapsulated & encrypted using ESP-Transport mode.

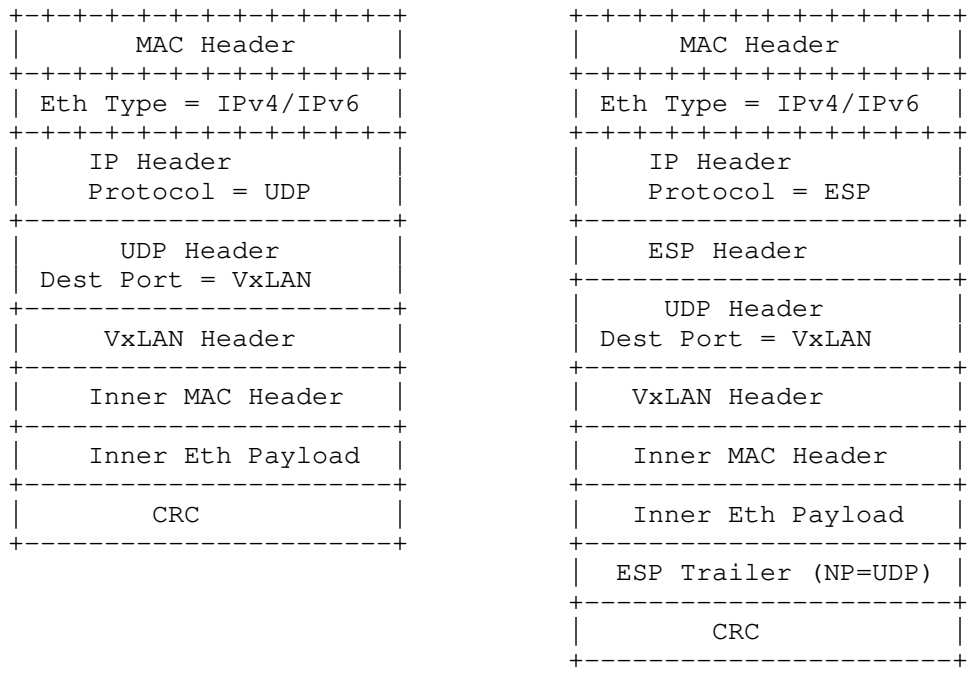


Figure 3: VxLAN Encapsulation within ESP

5.2 ESP Encapsulation within UDP packet

In scenarios where NAT traversal is required ([RFC3948]) or where load balancing using UDP header is required, then ESP encapsulation within UDP packet as depicted in the following figure is used. The ESP for NVO applications is in transport mode. The outer UDP header (before the ESP header) has its source port set to flow entropy and its destination port set to 4500 (indicating ESP header follows). A non-zero SPI value in ESP header implies that this is a data packet (i.e., it is not an IKE packet). The Next Protocol field in the ESP trailer indicates what follows the ESP header, is a UDP header. This inner UDP header has a destination port ID that identifies NVO encapsulation type (e.g., VxLAN). Optimization of this packet format where only a single UDP header is used (only the outer UDP header) is for future study.

When such encapsulation is used, for BGP signaling, the Tunnel Type of Tunnel Encapsulation TLV is set to ESP-in-UDP-Transport and the Tunnel Type of Encapsulation Extended Community is set to NVO

encapsulation type (e.g., VxLAN, GENEVE, GPE, etc.). This implies that the customer packets are first encapsulated using NVO encapsulation type and then it is further encapsulated & encrypted using ESP-in-UDP with Transport mode.

[RFC3948]

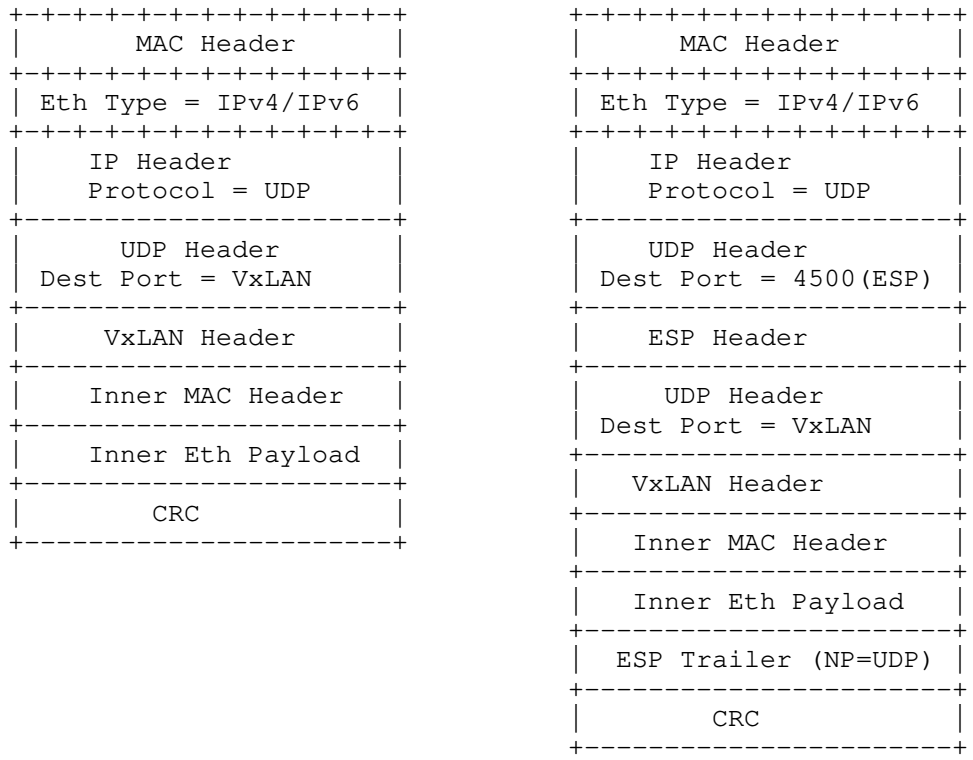


Figure 4: VxLAN Encapsulation within ESP Within UDP

6 BGP Encoding

This document defines two new Tunnel Types along with its associated sub-TLVs for The Tunnel Encapsulation Attribute [TUNNEL-ENCAP]. These tunnel types correspond to ESP-Transport and ESP-in-UDP-Transport as described in section 4. The following sub-TLVs apply to both tunnel types unless stated otherwise.

6.1 The Base (Minimal Set) DIM Sub-TLV

The Base DIM is described in 3.2.1. One and only one Base DIM may be sent in the IPSec Tunnel TLV.

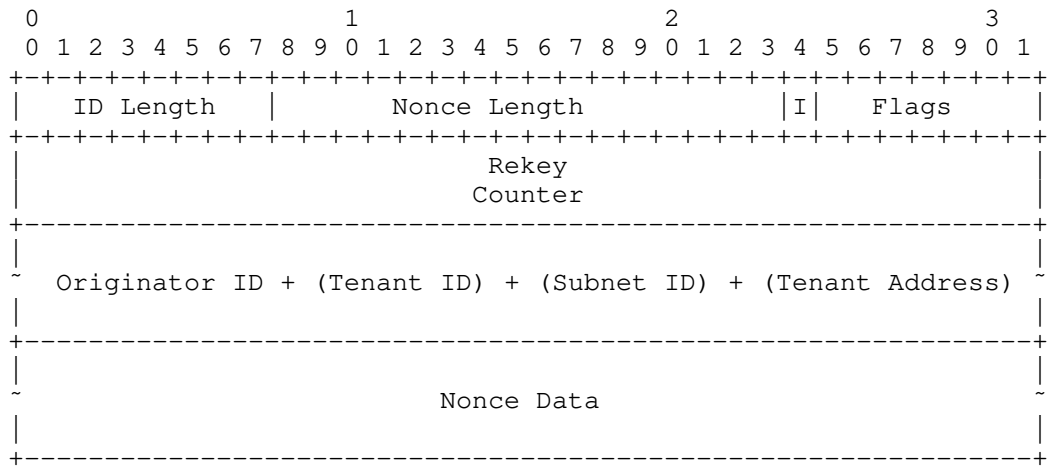


Figure 5: The Base DIM Sub-TLV

ID Length (16 bits) is the length of the Originator ID + (Tenant ID) + (Subnet ID) + (Tenant Address) in bytes.

Nonce Length (8 bits) is the length of the Nonce Data in bytes

I (1 bit) is the initial contact flag from [CONTROLLER-IKE]

Flags (7 bits) are reserved and MUST be set to zero on transmit and ignored on receipt.

The Rekey Counter is a 64 bit rekey counter as specified in [CONTROLLER-IKE]

The Originator ID + (Tenant ID) + (Subnet ID) + (Tenant Address) is the tunnel identifier and uniquely identifies the tunnel. Depending on the granularity of the tunnel, the fields in () may not be used - i.e., for a tunnel at the PE level of granularity, only Originator ID is required.

The Nonce Data is the nonce described in [CONTROLLER-IKE]. Its length is a multiple of 32 bits. Nonce lengths should be chosen to meet minimum requirements described in IKEv2 [RFC7296].

6.2 Key Exchange Sub-TLV

The KE Sub-TLV is described in 3.2.1 and 3.2.2.1. A KE is always required. One or more KE Sub-TLVs may be included in the IPSec Tunnel TLV.

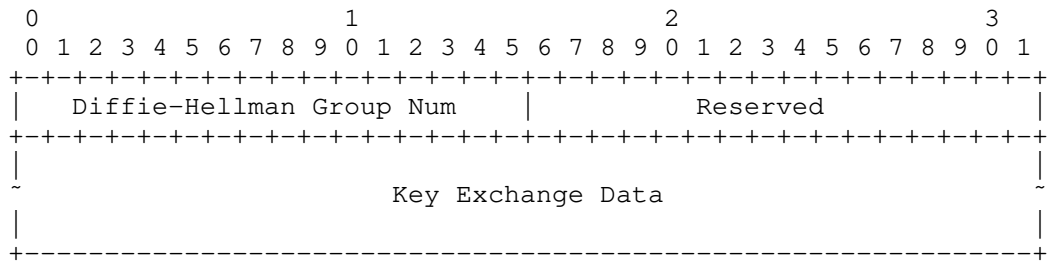


Figure 6: Key Exchange Sub-TLV

Diffie-Hellman Group Num 916 bits) identifies the Diffie-Hellman group in the Key Exchange Data was computed. Diffie-Hellman group numbers are discussed in IKEv2 [RFC7296] Appendix B and [RFC5114].

The Key Exchange payload is constructed by copying one's Diffie-Hellman public value into the "Key Exchange Data" portion of the payload. The length of the Diffie-Hellman public value is described for MOPD groups in [RFC7296] and for ECP groups in [RFC4753].

6.3 ESP SA Proposals Sub-TLV

The SA Sub-TLV is described in 3.2.2.2. Zero or more SA Sub-TLVs may be included in the IPSec Tunnel TLV.

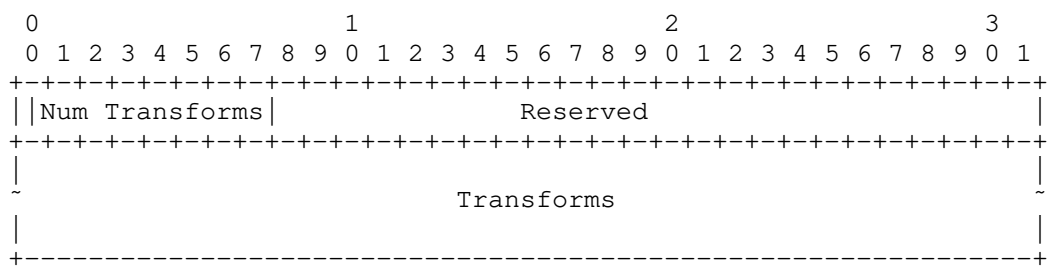


Figure 8: ESP SA Proposals Sub-TLV

Num Transforms is the number of transforms included.

Reserved is not used and MUST be set to zero on transmit and MUST be ignored on receipt.

6.3.1 Transform Substructure

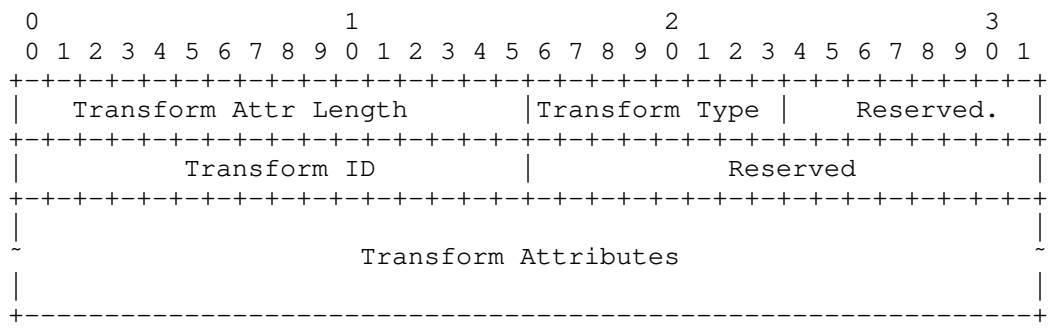


Figure 9: Transform Substructure Sub-TLV

The Transform Attr Length is the length of the Transform Attributes field.

The Transform Type is from Section 3.3.2 of [RFC7296] and [IKEV2IANA]. Only the values ENCR, INTEG, and ESN are allowed.

The Transform ID specifies the transform identification value from [IKEV2IANA].

Reserved is unused and MUST be zero on transmit and MUST be ignored on receipt.

The Transform Attributes are taken directly from 3.3.5 of [RFC7296].

7 Applicability to other VPN types

Although P2MP BGP signaling for establishment and maintenance of SAs among PE devices is described in this document in context of EVPN, there is no reason why it cannot be extended to other VPN technologies such as IP-VPN [RFC4364], VPLS [RFC4761] & [RFC4762], and MVPN [RFC6513] & [RFC6514] with ingress replication. The reason EVPN has been chosen is because of its pervasiveness in DC, SP, and Enterprise applications and because of its ability to support SA establishment at different granularity levels such as: per PE, Per tenant, per subnet, per Ethernet Segment, per IP address, and per MAC. For other VPN technology types, a much smaller granularity levels can be supported. For example for VPLS, only the granularity of per PE and per subnet can be supported. For per-PE granularity level, the mechanism is the same among all the VPN technologies as IPsec tunnel type (and its associated TLV and sub-TLVs) are sent along with the PE's loopback IPv4 (or IPv6) address. For VPLS, if

per-subnet (per bridge domain) granularity level needs to be supported, then the IPsec tunnel type and TLV are sent along with VPLS AD route.

The following table lists what level of granularity can be supported by a given VPN technology and with what BGP route.

Functionality	EVPN	IP-VPN	MVPN	VPLS
per PE	IPv4/v6 route	IPv4/v6 route	IPv4/v6 rte	IPv4/v6
per tenant	IMET (or new)	lpbk (or new)	I-PMSI	N/A
per subnet	IMET	N/A	N/A	VPLS AD
per IP	EVPN RT2/RT5	VPN IP rt	*,G or S,G	N/A
per MAC	EVPN RT2	N/A	N/A	N/A

8 Acknowledgements

9 Security Considerations

10 IANA Considerations

A new transitive extended community Type of 0x06 and Sub-Type of TBD for EVPN Attachment Circuit Extended Community needs to be allocated by IANA.

10 References

11.1 Normative References

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

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC2119

Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017.

[RFC7432] Sajassi et al., "BGP MPLS Based Ethernet VPN", RFC 7432, February, 2015.

[RFC8365] Sajassi et al., "A Network Virtualization Overlay Solution Using Ethernet VPN (EVPN)", RFC 8365, March, 2018.

[TUNNEL-ENCAP] Rosen et al., "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-03, November 2016.

[CONTROLLER-IKE] Carrel et al., "IPsec Key Exchange using a Controller", draft-carrel-ipsecme-controller-ike-00, July, 2018.

[IKEV2IANA] IANA, "Internet Key Exchange Version 2 (IKEv2) Parameters", <http://www.iana.org/assignments/ikev2-parameters/>.

[RFC3948] Huttunen et al., "UDP Encapsulation of IPsec ESP Packets", RFC 3948, January 2005.

[IKEV2-IANA] IANA, "Internet Key Exchange Version 2 (IKEv2) Parameters", February 2016, www.iana.org/assignments/ikev2-parameters/ikev2-parameters.xhtml.

[RFC4301] Kent, S. and K. Seo, "Security Architecture for the Internet Protocol", RFC 4301, DOI 10.17487/RFC4301, December 2005.

11.2 Informative References

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

[RFC4761] Kompella, K., et. al., "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.

[RFC4762] Kompella, K., et. al., "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.

[RFC6513] Rosen, E., et. al., "Multicast in MPLS/BGP IP VPNs", RFC

6513, February 2012.

[RFC6514] Rosen, E., et. al., "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, February 2012.

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

[802.1Q] "IEEE Standard for Local and metropolitan area networks - Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks", IEEE Std 802.1Q(tm), 2014 Edition, November 2014.

[RFC7348] Mahalingam, M., et al., "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.

[GENEVE] Gross, J., et al., "Geneve: Generic Network Virtualization Encapsulation", Work in Progress, draft-ietf-nvo3-geneve-06, March 2018.

Authors' Addresses

Ali Sajassi
Cisco
Email: sajassi@cisco.com

Ayan Banerjee
Cisco
Email: ayabaner@cisco.com

Samir Thoria
Cisco
Email: sthoria@cisco.com

David Carrel
Cisco
Email: carrel@cisco.com

Brian Weis
Individual

Email: bew.stds@gmail.com

John Drake
Juniper
Email: jdrake@juniper.net

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: January 13, 2021

S. Gringeri
J. Whittaker
Verizon
C. Schmutzer, Ed.
L. Della Chiesa
N. Nainar, Ed.
C. Pignataro
Cisco Systems, Inc.

July 12, 2020

Private Line Emulation over Packet Switched Networks
draft-schmutzer-bess-ple-00

Abstract

This document describes a method for encapsulating high-speed bit-streams as virtual private wire services (VPWS) over packet switched networks (PSN) providing complete signal transport transparency.

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 January 13, 2021.

Copyright Notice

Table of Contents

1. Introduction and Motivations

This document describes a method for encapsulating high-speed bit-streams as VPWS over packet switched networks (PSN). This emulation suits applications where complete signal transparency is required and data interpretation of the PE would be counter productive.

One example is two ethernet connected CEs and the need for synchronous ethernet operation between then without the intermediate PEs interfering. Another example is addressing common ethernet control protocol transparency concerns for carrier ethernet services, beyond the behavior definitions of MEF specifications.

The mechanisms described in this document allow the transport of signals from many technologies such as ethernet, fibre channel, SONET/SDH [GR253]/[G.707] and OTN [G.709] by treating them as bit-stream payload defined in Section 3.3.3 of [RFC3985].

2. Requirements Notation

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. Terminology and Reference Model

3.1. Terminology

- ACH – Associated Channel Header
- AIS – Alarm Indication Signal
- CBR – Constant Bit Rate
- CE – Customer Edge
- CSRC – Contributing SouRCe
- ES – Errored Second
- FEC – Forward Error Correction
- IWF – InterWorking Function
- LDP – Label Distribution Protocol
- LF – Local Fault
- MPLS – Multi Protocol Label Switching
- NSP – Native Service Processor
- ODUk – Optical Data Unit k
- OTN – Optical Transport Network
- OTUk – Optical Transport Unit k
- PCS – Physical Coding Sublayer

- PE - Provider Edge
- PLE - Private Line Emulation
- PLOS - Packet Loss Of Signal
- PSN - Packet Switched Network
- P2P - Point-to-Point
- QOS - Quality Of Service
- RSVP-TE - Resource Reservation Protocol Traffic Engineering
- RTCP - RTP Control Protocol
- RTP - Realtime Transport Protocol
- SES - Severely Errored Seconds
- SDH - Synchronous Digital Hierarchy
- SRTP - Secure Realtime Transport Protocol
- SRv6 - Segment Routing over IPv6 Dataplane
- SSRC - Synchronization SouRCe
- SONET - Synchronous Optical Network
- TCP - Transmission Control Protocol
- UAS - Unavailable Seconds
- VPWS - Virtual Private Wire Service

Similarly to [RFC4553] and [RFC5086] the term Interworking Function (IWF) is used to describe the functional block that encapsulates bit streams into PLE packets and in the reverse direction decapsulates PLE packets and reconstructs bit streams.

3.2. Reference Models

The generic models defined in [RFC4664] are applicable to PLE.

PLE embraces the minimum intervention principle outlined in section 3.3.5 of [RFC3985] whereas the data is flowing through the PLE encapsulation layer as received without modifications.

For some applications the NSP function is responsible for performing operations on the native data received from the CE. Examples are terminating FEC in case of 100GE or terminating the OTUk layer for OTN. After the NSP the IWF is generating the payload of the VPWS which carried via a PSN tunnel.

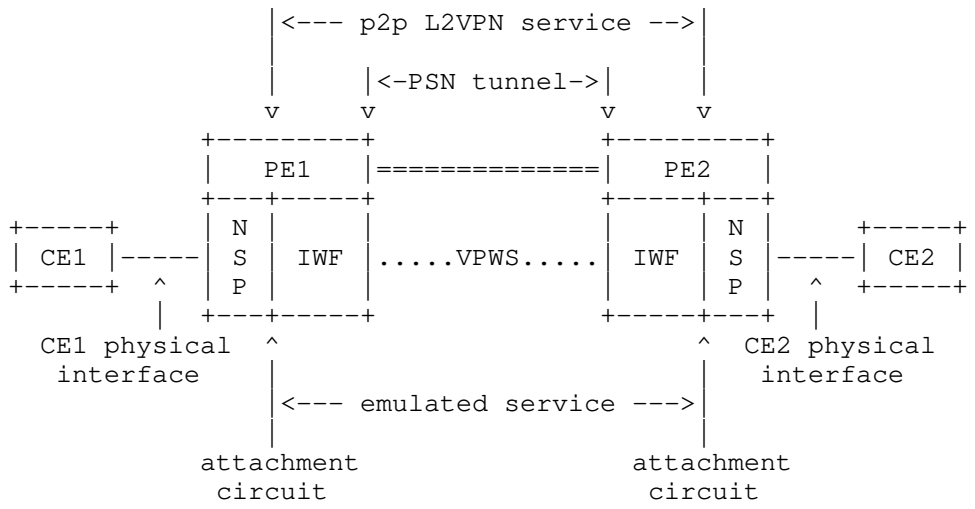


Figure 1: PLE Reference Model

To allow the clock of the transported signal to be carried across the PLE domain in a transparent way the network synchronization reference model and deployment scenario outlined in section 4.3.2 of [RFC4197] is applicable.

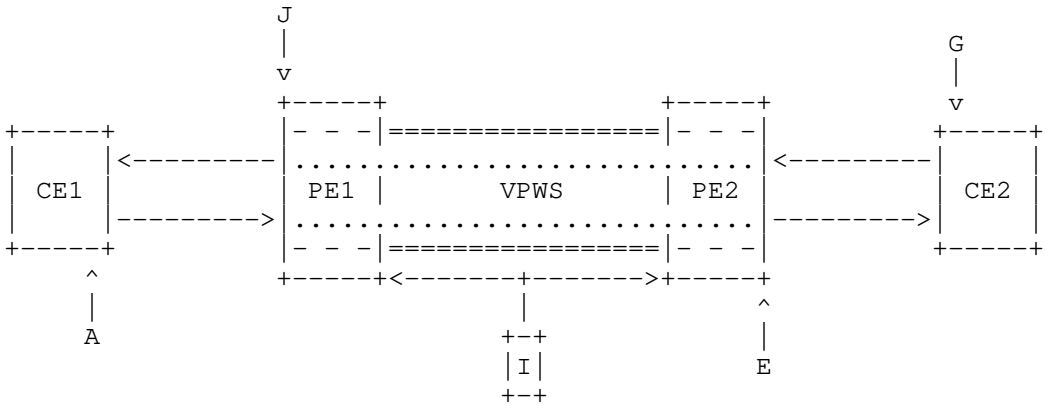


Figure 2: Relative Network Scenario Timing

The attachment circuit clock E is generated by PE2 in reference to a common clock I. For this to work the difference between clock I and A MUST be explicitly transferred between the PE1 and PE2 using the timestamp inside the RTP header.

For the reverse direction PE1 does generate the clock J in reference to clock I and the clock difference between I and G.

4. PLE Encapsulation Layer

The basic packet format used by PLE is shown in the below figure.

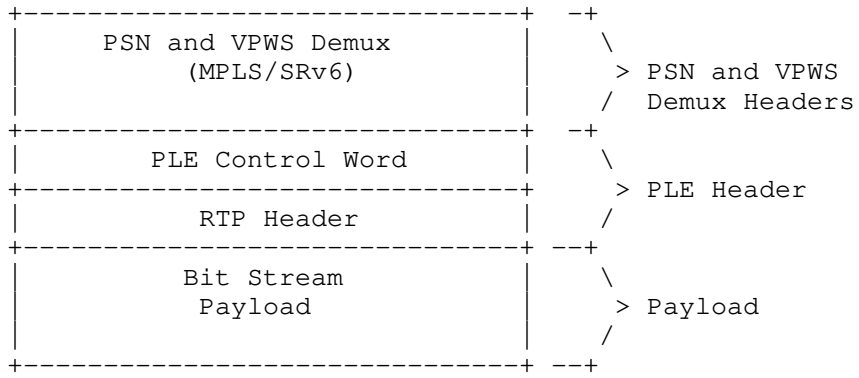


Figure 3: PLE Encapsulation Layer

4.1. PSN and VPWS Demultiplexing Headers

This document does not imply any specific technology to be used for implementing the VPWS demultiplexing and PSN layers.

When a MPLS PSN layer is used. A VPWS label provides the demultiplexing mechanism as described in section 5.4.2 of [RFC3985]. The PSN tunnel can be a simple best path Label Switched Path (LSP) established using LDP [RFC5036] or Segment Routing [RFC8402] or a traffic engineered LSP established using RSVP-TE [RFC3209] or SR-TE [SRPOLICY].

When PLE is applied to a SRv6 based PSN, the mechanisms defined in [RFC8402] and the End.DX2 endpoint behavior defined in [SRV6NETPROG] do apply.

4.2. PLE Header

The PLE header MUST contain the PLE control word (4 bytes) and MUST include a fixed size RTP header [RFC3550]. The RTP header MUST immediately follow the PLE control word.

4.2.1. PLE Control Word

The format of the PLE control word is inline with the guidance in [RFC4385] and as shown in the below figure:

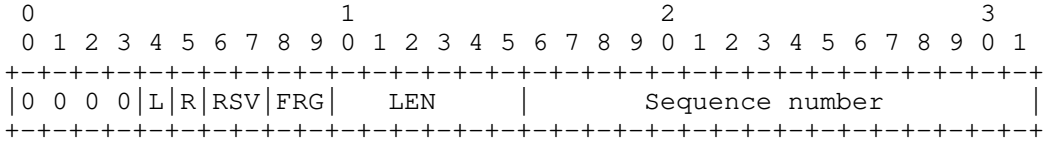


Figure 4: PLE Control Word

The first nibble is used to differentiate if it is a control word or Associated Channel Header (ACH). The first nibble MUST be set to 0000b to indicate that this header is a control word as defined in section 3 of [RFC4385].

The other fields in the control word are used as defined below:

L

Set by the PE to indicate that data carried in the payload is invalid due to an attachment circuit fault (client signal

failure). The downstream PE MUST play out an appropriate replacement data. The NSP MAY inject an appropriate native fault propagation signal.

R

Set by the downstream PE to indicate that the IWF experiences packet loss from the PSN or a server layer backward fault indication is present in the NSP. The R bit MUST be cleared by the PE once the packet loss state or fault indication has cleared.

RSV

These bits are reserved for future use. This field MUST be set to zero by the sender and ignored by the receiver.

FRG

These bits MUST be set to zero by the sender and ignored by the receiver except for frame aligned payloads; see Section 5.2

LEN

In accordance to [RFC4385] section 3 the length field MUST always be set to zero as there is no padding added to the PLE packet. To detect malformed packets the default, preconfigured or signaled payload size MUST be assumed.

Sequence Number

The sequence number field is used to provide a common PW sequencing function as well as detection of lost packets. It MUST be generated in accordance with the rules defined in Section 5.1 of [RFC3550] for the RTP sequence number and MUST be incremented with every PLE packet being sent.

4.2.2. RTP Header

The RTP header MUST be included and is used for explicit transfer of timing information. The RTP header is purely a formal reuse and RTP mechanisms, such as header extensions, contributing source (CSRC) list, padding, RTP Control Protocol (RTCP), RTP header compression, Secure Realtime Transport Protocol (SRTP), etc., are not applicable to PLE VPWS.

The format of the RTP header is as shown in the below figure:

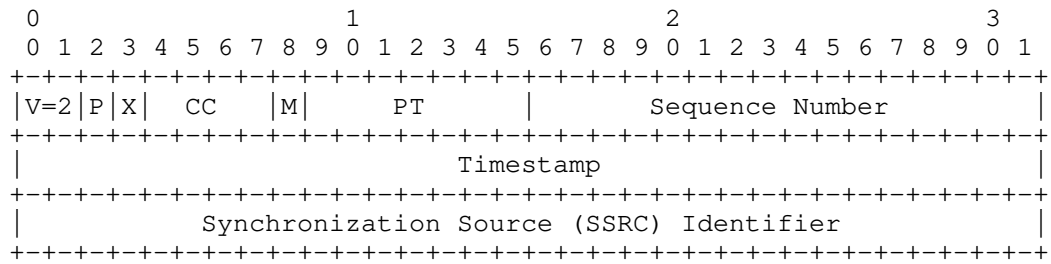


Figure 5: RTP Header

V: Version

The version field MUST be set to 2.

P: Padding

The padding flag MUST be set to zero by the sender and ignored by the receiver.

X: Header Extension

The X bit MUST be set to zero by sender and ignored by receiver.

CC: CSRC Count

The CC field MUST be set to zero by the sender and ignored by the receiver.

M: Marker

The M bit MUST be set to zero by sender and ignored by receiver.

PT: Payload Type

A PT value MUST be allocated from the range of dynamic values define by [RFC3551] for each direction of the VPWS. The same PT value MAY be reused both for direction and between different PLE VPWS.

Sequence Number

The packet sequence number MUST continuously cycle from 0 to 0xFFFF. It is generated and processed in accordance with the rules established in [RFC3550]. The PLE receiver MUST sequence packets according to the Sequence Number field of the PLE control

word and MAY verify correct sequencing using RTP Sequence Number field.

Timestamp

Timestamp values are used in accordance with the rules established in [RFC3550]. Frequency of the clock used for generating timestamps MUST be 25 MHz based on a local reference.

SSRC: Synchronization Source

The SSRC field MAY be used for detection of misconnections.

5. PLE Payload Layer

5.1. Constant Bit Rate Payload

A bit-stream is mapped into a packet with a fixed payload size ignoring any structure being present. The number of bytes MUST be defined during VPWS setup, MUST be the same in both directions of the VPWS and MUST remain unchanged for the lifetime of the VPWS.

All PLE implementations MUST be capable of supporting the default payload size of 480 bytes.

For PCS based CE interface types supporting FEC the NSP function MUST terminate the FEC and pass the PCS encoded signal to the IWF function.

For PCS based CE interface types supporting virtual lanes (i.e. 100GE) a PLE payload MUST carry information from all virtual lanes in a bit interleaved manner after the NSP function has performed PCS layer de-skew and re-ordering.

A PLE implementation MUST support the transport of all service types except ODUk bit-streams using the constant bit rate payload.

5.2. ODUk Frame aligned Payload

In case of OTN PLE does only transport the ODUk layer to be bandwidth efficient. This means the OTUk layer which does include the FEC is terminated by NSP function. As OTN is performing frame alignment at the OTUk layer the bit-stream must be carried frame aligned.

A ODUk frame consists of 3824 columns and 4 rows which results in a frame size of 15296 bytes. As common PSN MTU sizes are in the range of at most 9200 bytes the ODUk frame has to be fragmented during PLE payload encapsulation. The used payload size has to be a integer

fraction of the full 15296 bytes to allow for ODUk frame alignment. All PLE implementations MUST support the payload size of 478 bytes.

The two FRG bits in the PLE control word MUST be used to indicate first, intermediate, and last fragment of the encapsulated ODUk frame as described in section 4.1 of [RFC4623].

All PLE implementations MUST support the transport ODUk bit-streams using the frame aligned payload.

6. PLE Operation

6.1. Common Considerations

A PLE VPWS can be established using manual configuration or leveraging mechanisms of a signalling protocol

Furthermore emulation of bit-stream signals using PLE is only possible when the two attachment circuits of the VPWS are of the same type (OC192, 10GBASE-R, ODU2, etc) and are using the same PLE payload type and payload size. This can be ensured via manual configuration or via a signalling protocol

Extensions to the PWE3 [RFC4447] and EVPN-VPWS [RFC8214] control protocols are described in a separate document [PLESIG].

6.2. PLE IWF Operation

6.2.1. PSN-bound Encapsulation Behavior

After the VPWS is set up, the PSN-bound IWF does perform the following steps:

- Packetise the data received from the CE is into a fixed size PLE payloads
- Add PLE control word and RTP header with sequence numbers, flags and timestamps properly set
- Add the VPWS demultiplexer and PSN headers
- Transmit the resulting packets over the PSN
- Set L bit in the PLE control word whenever attachment circuit detects a fault
- Set R bit in the PLE control word whenever the local CE-bound IWF is in packet loss state

6.2.2. CE-bound Decapsulation Behavior

The CE-bound IWF is responsible for removing the PSN and VPWS demultiplexing headers, PLE control word and RTP header from the received packet stream and play-out of the bit-stream to the local attachment circuit.

A de-jitter buffer MUST be implemented where the PLE packets are stored upon arrival. The size of this buffer SHOULD be locally configurable to allow accommodation of specific PSN packet delay variation expected.

The CE-bound IWF SHOULD use the sequence number in the control word to detect lost packets. It MAY use the sequence number in the RTP header for the same purposes.

The payload of a lost packet MUST be replaced with equivalent amount of replacement data. The contents of the replacement data MAY be locally configurable. All PLE implementations MUST support generation of "0xAA" as replacement data. The alternating sequence of 0s and 1s of the "0xAA" pattern does ensure clock synchronization is maintained.

Whenever the VPWS is not operationally up, the CE-bound NSP function MUST inject the appropriate native downstream fault indication signal (for example ODUk-AIS or ethernet LF).

Whenever a VPWS comes up, the CE-bound IWF will start receiving PLE packets and will store them in the jitter buffer. The CE-bound NSP function will continue to inject the appropriate native downstream fault indication signal until a pre-configured amount of payloads is stored in the jitter buffer.

After the pre-configured amount of payload is present in the jitter buffer the CE-bound IWF transitions to the normal operation state and the content of the jitter buffer is played out to the CE in accordance with the required clock. In this state the CE-bound IWF does perform egress clock recovery.

Whenever the L bit is set in the PLE control word of a received PLE packet the CE-bound NSP function SHOULD inject the appropriate native downstream fault indication signal instead of playing out the payload.

If the CE-bound IWF detects loss of a pre-configured number of consecutive packets, the de-jitter buffer under- or over-runs, it enters packet loss (PLOS) state . While in this state CE-bound NSP function SHOULD inject the appropriate native downstream fault

indication signal. Also the PSN-bound IWF SHOULD set the R bit in the PLE control word of every packet transmitted.

The CE-bound IWF exits the packet loss state after a pre-configured amount of valid PLE packets have been received.

Whenever the R bit is set in the PLE control word of a received PLE packet the PLE performance monitoring statistics SHOULD get updated.

6.3. PLE Performance Monitoring

PLE SHOULD provide the following functions to monitor the network performance to be inline with expectations of transport network operators.

The near-end performance monitors defined for PLE are as follows:

ES-PLE : PLE Errored Seconds

SES-PLE : PLE Severely Errored Seconds

UAS-PLE : PLE Unavailable Seconds

Each second that contains at least one lost packet defect SHALL be counted as ES-PLE. Each second that contains a PLOS defect SHALL be counted as SES-PLE.

UAS-PLE SHALL be counted after configurable number of consecutive SES-PLE have been observed, and no longer counted after a configurable number of consecutive seconds without SES-PLE have been observed. Default value for each is 10 seconds.

Once unavailability is detected, ES and SES counts SHALL be inhibited up to the point where the unavailability was started. Once unavailability is removed, ES and SES that occurred along the clearing period SHALL be added to the ES and SES counts.

A PLE far-end performance monitor is providing insight into the CE-bound IWF at the far end of the PSN. The statistics are based on the PLE-RDI indication carried in the PLE control word via the R bit.

The PLE VPWS performance monitors are derived from the definitions in accordance with [G.826]

6.4. QoS and Congestion Control

The PSN carrying PLE VPWS may be subject to congestion, but PLE VPWS representing constant bit-rate (CBR) flows cannot respond to congestion in a TCP-friendly manner as described in [RFC2913].

Hence the PSN providing connectivity for the PLE VPWS between PE devices MUST be Diffserv [RFC2475] enabled and MUST provide a per domain behavior [RFC3086] that guarantees low jitter and low loss.

To achieve the desired per domain behavior PLE VPWS SHOULD be carried over traffic-engineering paths through the PSN with bandwidth reservation and admission control applied.

7. Security Considerations

As PLE is leveraging VPWS as transport mechanism the security considerations described in [RFC7432] and [RFC3985] are applicable.

8. IANA Considerations

Applicable signalling extensions are out of the scope of this document.

PLE does not introduce additional requirements from IANA.

9. Acknowledgements

To be updated.

10. References

10.1. Normative References

[PLESIG] IETF, "Private Line Emulation VPWS Signalling", <https://tools.ietf.org/html/draft-schmutzer-bess-ple-vpws-signalling>.

[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>.

[RFC2475] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, DOI 10.17487/RFC2475, December 1998, <https://www.rfc-editor.org/info/rfc2475>.

[RFC3086] Nichols, K. and B. Carpenter, "Definition of Differentiated Services Per Domain Behaviors and Rules for their Specification", RFC 3086, DOI 10.17487/RFC3086, April 2001, <https://www.rfc-editor.org/info/rfc3086>.

[RFC3550] Schulzrinne, H., Casner, S., Frederick, R., and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications", STD 64, RFC 3550, DOI 10.17487/RFC3550, July 2003, <https://www.rfc-editor.org/info/rfc3550>.

[RFC3551] Schulzrinne, H. and S. Casner, "RTP Profile for Audio and Video Conferences with Minimal Control", STD 65, RFC 3551, DOI 10.17487/RFC3551, July 2003, <https://www.rfc-editor.org/info/rfc3551>.

[RFC3985] Bryant, S., Ed. and P. Pate, Ed., "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, DOI 10.17487/RFC3985, March 2005, <https://www.rfc-editor.org/info/rfc3985>.

[RFC4197] Riegel, M., Ed., "Requirements for Edge-to-Edge Emulation of Time Division Multiplexed (TDM) Circuits over Packet Switching Networks", RFC 4197, DOI 10.17487/RFC4197, October 2005, <https://www.rfc-editor.org/info/rfc4197>.

[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, <https://www.rfc-editor.org/info/rfc4385>.

[RFC4447] Martini, L., Ed., Rosen, E., El-Aawar, N., Smith, T., and G. Heron, "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", RFC 4447, DOI 10.17487/RFC4447, April 2006, <https://www.rfc-editor.org/info/rfc4447>.

[RFC4623] Malis, A. and M. Townsley, "Pseudowire Emulation Edge-to-Edge (PWE3) Fragmentation and Reassembly", RFC 4623, DOI 10.17487/RFC4623, August 2006, <https://www.rfc-editor.org/info/rfc4623>.

[RFC4664] Andersson, L., Ed. and E. Rosen, Ed., "Framework for Layer 2 Virtual Private Networks (L2VPNs)", RFC 4664, DOI 10.17487/RFC4664, September 2006, <https://www.rfc-editor.org/info/rfc4664>.

[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>.

[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>.

[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>.

10.2. Informative References

[G.707] ITU-T, "Network node interface for the synchronous digital hierarchy (SDH)", <https://www.itu.int/rec/T-REC-G.707>.

[G.709] International Telecommunication Union (ITU), "G.709: Interfaces for the optical transport network", <https://www.itu.int/rec/T-REC-G.709>.

[G.826] ITU-T, "End-to-end error performance parameters and objectives for international, constant bit-rate digital paths and connections", <https://www.itu.int/rec/T-REC-G.826>.

[GR253] Telcordia, "SONET Transport Systems : Common Generic Criteria", <https://telecom-info.telcordia.com>.

[RFC2913] Klyne, G., "MIME Content Types in Media Feature Expressions", RFC 2913, DOI 10.17487/RFC2913, September 2000, <https://www.rfc-editor.org/info/rfc2913>.

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, DOI 10.17487/RFC3209, December 2001, <https://www.rfc-editor.org/info/rfc3209>.

[RFC4553] Vainshtein, A., Ed. and YJ. Stein, Ed., "Structure-Agnostic Time Division Multiplexing (TDM) over Packet (SAToP)", RFC 4553, DOI 10.17487/RFC4553, June 2006, <https://www.rfc-editor.org/info/rfc4553>.

[RFC5036] Andersson, L., Ed., Minei, I., Ed., and B. Thomas, Ed., "LDP Specification", RFC 5036, DOI 10.17487/RFC5036, October 2007, <https://www.rfc-editor.org/info/rfc5036>.

[RFC5086] Vainshtein, A., Ed., Sasson, I., Metz, E., Frost, T., and P. Pate, "Structure-Aware Time Division Multiplexed (TDM) Circuit Emulation Service over Packet Switched Network (CESoPSN)", RFC 5086, DOI 10.17487/RFC5086, December 2007, <https://www.rfc-editor.org/info/rfc5086>.

[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>.

[SRPOLICY] IETF, "Segment Routing Policy Architecture", <https://tools.ietf.org/html/draft-ietf-spring-segment-routing-policy>.

[SRV6NETPROG] IETF, "SRv6 Network Programming", <https://tools.ietf.org/html/draft-ietf-spring-srv6-network-programming>.

Authors' Addresses

Steven Gringeri
Verizon

Email: steven.gringeri@verizon.com

Jeremy Whittaker
Verizon

Email: jeremy.whittaker@verizon.com

Christian Schmutzer (editor)
Cisco Systems, Inc.

Email: cschmutz@cisco.com

Luca Della Chiesa
Cisco Systems, Inc.

Email: ldellach@cisco.com

Nagendra Kumar Nainar (editor)
Cisco Systems, Inc.

Email: naikumar@cisco.com

Carlos Pignataro
Cisco Systems, Inc.

Email: cpignata@cisco.com

BESS Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 13, 2021

S. Gringeri
J. Whittaker
Verizon
C. Schmutzer, Ed.
P. Brissette
Cisco Systems, Inc.

July 12, 2020

Private Line Emulation VPWS Signalling
draft-schmutzer-bess-ple-vpws-signalling-00

Abstract

This document specifies the mechanisms to signal Virtual Private Wire Services (VPWS) carrying bit-stream signals over Packet Switched Networks (PSN).

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 January 13, 2021.

Copyright Notice

Table of Contents

1. Introduction

Virtual Private Wire Service (VPWS) is a widely deployed technology for providing point-to-point (P2P) services for various layer 2 and also layer 1 technologies. Initially VPWS were define in the Pseudowire Emulation Edge-to-Edge (PWE3) architecture [RFC3985] for Frame Relay, ATM, HDLC, PPP, Ethernet, TDM and SONET/SDH.

Later on the adoption of BGP [RFC4271] as a protocol for delivering ethernet layer 2 services led to Ethernet VPNs [RFC7432] and EVPN-VPWS [RFC8214] respectively.

This document focuses on bit stream VPWS instance types which already got introduced in [RFC3985] and describes mechanisms that can be used to discover and establish such VPWS instances by a means of a dynamic protocol rather than manual configuration on both endpoints.

Possible bit stream VPWS instance types are:

- TDM services using SAToP [RFC4553]
- TDM services using CESoP [RFC5086]
- SONET/SDH services using CEP [RFC4842]
- High-speed private line services using PLE [PLE]

A generic VPWS reference model similar to the one defined in [RFC3985] and [PLE] is shown in Figure 1. Data received from a CEs is encapsulated by PEs into the respective VPWS established between the attachment circuits of the local and remote PE and transmitted across the Packet Switched Network (PSN) using a PSN tunnel.

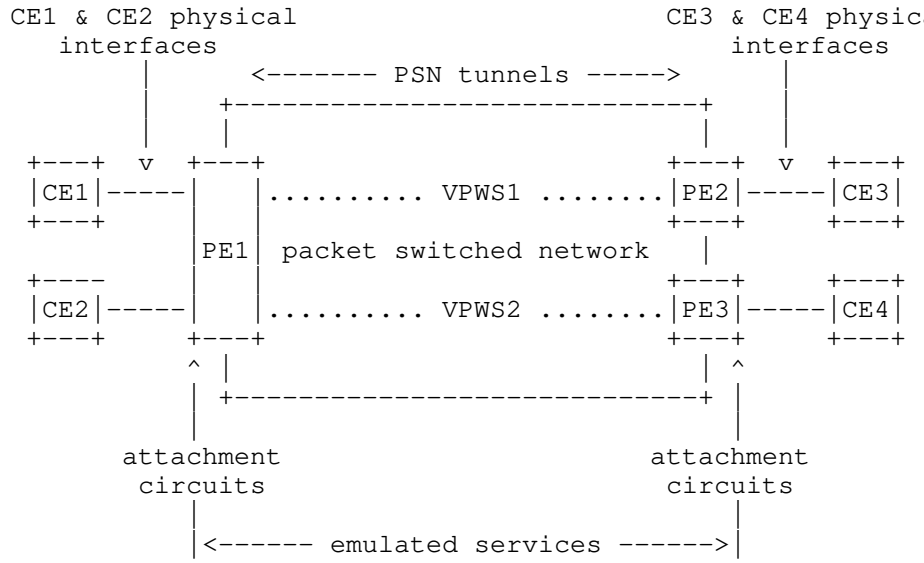


Figure 1: VPWS Reference Model

In the example shown in Figure 1 there are two CE nodes (CE1 and CE2) connected to the same PE node (PE1). CE3 is connected to PE2 and CE4 is connected to PE3. There are two VPWS instances established. VPWS1 between CE1 and CE3 and VPWS2 between CE2 and CE4. For traffic to be carried across the network PSN tunnels between PE1 and PE2 and between PE1 and PE3 are needed.

In order for a bit stream VPWS instance to come up, the attachment circuit parameters must be identical on both endpoints. The control plane mechanisms described in this document are leveraged to meet this requirement. Mechanisms for misconnection detection and protection switch coordination are also described.

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.

1.2. Terminology

- AIS - Alarm Indication Signal
- AFI - Address Family Identifier
- ATM - Asynchronous Transfer Mode
- BGP - Border Gateway Protocol
- CBR - Constant Bit Rate
- CE - Customer Edge
- CEP - SONET/SDH Circuit Emulation over Packet
- CESoP - Structure-aware TDM Circuit Emulation Service over Packet Switched Network
- DF - Designated Forwarder
- EAD - Ethernet Auto Discovery
- FC - Fibre Channel
- EBM - Equipped Bit Mask
- EVI - EVPN Instance
- EVPN - Ethernet Virtual Private Network
- HDLC - High-level Data Link Control
- LDP - Label Distribution Protocol
- MPLS - Multi Protocol Label Switching
- MTU - Maximum Transmission Unit
- NDF - Non-Designated Forwarder
- NLRI - Network Layer Reachability Information
- OC - Optical Carrier
- ODUk - Optical Data Unit k
- PDH - Plesynchronous Digital Hierarchy

- PE - Provider Edge
- PLE - Private Line Emulation
- PPP - Point-to-Point Protocol
- PSN - Packet Switched Network
- PW - Pseudo Wire
- PWE3 - Pseudowire Emulation Edge-to-Edge
- P2P - Point-to-Point
- RTP - Realtime Transport Protocol
- SAFI - Subsequent Address Family Identifier
- SAToP - Structure Agnostic TDM over Packet
- SDH - Synchronous Digital Hierarchy
- SONET - Synchronous Optical Network
- SRv6 - Segment Routing over IPv6 Dataplane
- STM - Synchronous Transport Module
- STS - Synchronous Transport Signal
- TDM - Time Division Multiplexing
- TLV - Type Length Value
- UNE - Unequipped
- VC - Virtual Circuit
- VPWS - Virtual Private Wire Service
- VT - Virtual Tributary
-

2. Solution Requirements

The scope of this document is to specify signalling mechanisms for PLE bit-stream services as defined in [PLE].

For legacy bit-stream services that have been defined for TDM [RFC4553] and [RFC5086] as well as SONET/SDH [RFC4842], additional signalling mechanisms are described to complement the mechanisms defined in [RFC5287].

To avoid redefining PW types for [RFC8214] the notion of "PW type" from [RFC8077] is maintained and only a new PW type for [PLE] has been assigned by IANA.

- TBD1 - Private Line Emulation (PLE) over Packet

The concept of "CEP type" from [RFC5287] to distinguish different connection types that use the same PW type is adopted. In this document it is referred to as "PLE/CEP type". Two new connection types are defined (see Section 4.2.3).

To unambiguously identify the rate of an attachment circuit, also the concept of "CEP/TDM bit-rate" from [RFC5287] is adopted and called "PLE bit-rate" herein.

The VPWS signalling requirements are as follows:

- EVPN-VPWS [RFC8214] as signalling protocol MUST be supported
- LDP [RFC8077] MAY be supported as VPWS signalling protocol
- Implementations MUST support MPLS as underlay PSN
- The VPWS instance MAY be signalled as SRv6 overlay service per [srv6_overlay] leveraging on [srv6_netprog] using the End.DX2 function. In such case, the implementation MUST support SRv6 as underlay PSN.
- The use of control word as defined in [RFC4553], [RFC5086], [RFC4842] and [PLE] MUST be signalled.
- The PW type MUST be signalled and the PE nodes MUST validate that the PW type is identical on both endpoint.
- For CEP [RFC4842] and PLE [PLE] the PLE/CEP type MUST be signalled and the PE nodes MUST validate that the PLE/CEP type is identical on both endpoints.

- The PLE/CEP/TDM bit-rate MUST be signalled if the attachment circuit rate can not be unambiguously identified from the PW type alone and the PE nodes MUST validate that the attachment circuit rate is identical on both endpoints.

- A non-default payload size MAY be signalled. Both PE nodes MUST validate that the payload size is identical on both endpoints.

- A locally configured connection identifier as defined in Section 4.2.6 SHOULD be exchanged and MAY be used to identify a misconnection by comparing the locally configured identifier with the received identifier from the remote PE node.

- Multi-homed PE scenarios per [RFC7432] and [RFC8214] SHOULD be supported where the load-balancing mode single-active MUST be supported. Port-active load-balancing mode MAY also be supported.

- For multi-homed PE scenarios non-revertive mode MUST and revertive mode SHOULD be supported in compliance to [pref_df]

3. Service Types

The following sections list all possible service types that are supported by the proposed signalling mechanisms.

3.1. Ethernet Service Types

Service Type	Encapsulation Standard	PW Type	PLE/CEP Type	PLE/CEP/TDM Bit-rate
1000BASE-X	[PLE]	TBD1	0x3	1,250,000
10GBASE-R	[PLE]	TBD1	0x3	10,312,500
25GBASE-R	[PLE]	TBD1	0x3	25,791,300
40GBASE-R	[PLE]	TBD1	0x3	41,250,000
100GBASE-R	[PLE]	TBD1	0x3	103,125,000

3.2. Fibre Channel Service Types

Service Type	Encapsulation Standard	PW Type	PLE/CEP Type	PLE/CEP/TDM Bit-rate
1GFC	[PLE]	TBD1	0x3	1,062,500
2GFC	[PLE]	TBD1	0x3	2,125,000
4GFC	[PLE]	TBD1	0x3	4,250,000
8GFC	[PLE]	TBD1	0x3	8,500,000
10GFC	[PLE]	TBD1	0x3	19,518,750
16GFC	[PLE]	TBD1	0x3	14,025,000
32GFC	[PLE]	TBD1	0x3	28,050,000
128GFC	[PLE]	TBD1	0x3	112,200,000

3.3. OTN Service Types

Service Type	Encapsulation Standard	PW Type	PLE/CEP Type	PLE/CEP/TDM Bit-rate
ODU0	[PLE]	TBD1	0x4	1,244,160
ODU1	[PLE]	TBD1	0x4	2,498,775
ODU2	[PLE]	TBD1	0x4	10,037,273
ODU2e	[PLE]	TBD1	0x4	10,399,525
ODU3	[PLE]	TBD1	0x4	40,319,218
ODU4	[PLE]	TBD1	0x4	104,794,445

3.4. TDM Service Types

Service Type	Encapsulation Standard	PW Type	PLE/CEP Type	PLE/CEP/TDM Bit-rate
CESoPSN basic mode	[RFC5086]	0x0015	N/A	N
CESoPSN with CAS	[RFC5086]	0x0017	N/A	N
E1	[RFC4553]	0x0011	N/A	32
DS1	[RFC4553]	0x0012	N/A	24
DS1 octet-aligned	[RFC4553]	0x0012	N/A	25
E3	[RFC4553]	0x0013	N/A	535
T3	[RFC4553]	0x0014	N/A	699

N is the number of DS0 channels in the attachment circuit

3.5. SONET/SDH Service Types

Service Type	Encapsulation Standard	PW Type	PLE/CEP Type	PLE/CEP/TDM Bit-rate
VT1.5/VC-11	[RFC4842]	0x0010	0x1	26
VT2/VC-12	[RFC4842]	0x0010	0x1	35
VT3	[RFC4842]	0x0010	0x1	53
VT6/VC-2	[RFC4842]	0x0010	0x1	107
STS-Nc	[RFC4842]	0x0010	0x0	783*N
VC-4-Mc	[RFC4842]	0x0010	0x0	783*3*M
Fract. STS1/VC-3	[RFC4842]	0x0010	0x2	783
Fract. VC-4	[RFC4842]	0x0010	0x2	783*4
Async STS1/VC-3	[RFC4842]	0x0010	0x2	783
OC3/STM1	[PLE]	TBD1	0x3	155,520
OC12/STM4	[PLE]	TBD1	0x3	622,080
OC48/STM16	[PLE]	TBD1	0x3	2,488,320
OC192/STM64	[PLE]	TBD1	0x3	9,953,280
OC768/STM256	[PLE]	TBD1	0x3	39,813,120

N=1,3,12,48,192,768 and M=1,4,16,64,256

4. EVPN-VPWS signalling

4.1. Reuse of existing BGP EVPN-VPWS capabilities

A PLE VPWS instance is identified by a pair of per-EVI ethernet A-D routes advertised by two PE nodes establishing the VPWS in accordance to [RFC8214].

The EVPN layer 2 attribute extended community defined in [RFC8214] MUST be supported and added to the per-EVI ethernet A-D route.

- C bit set to 1 to indicate Control Word MUST be present.
- P and B bits are set by dual-homing PEs as per [RFC8214] and [pref_df]
- L2 MTU MUST be set to zero and ignored by the receiver

4.2. BGP PLE Attribute

To exchange and validate bit-stream specific attachment circuit parameters during the VPWS setup, a new BGP path attribute called "BGP PLE attribute" is defined.

The BGP PLE attribute defined in this document can be attached to EVPN VPWS routes [RFC8214]. The usage for other Address Family Identifier (AFI) / Subsequent Address Family Identifier (SAFI) combinations is not defined herein but may be specified in future specifications.

The BGP PLE attribute is an optional and transitive BGP path attribute. The attribute type code TBD2 has been assigned by IANA (see section Section 6)

The format is defined as a set of Type/Length/Value (TLV) triplets, described in the following sections and listed in Table 1. This attribute SHOULD only be included with EVPN Network Layer Reachability Information (NLRI).

TLV Type	Name	Length	Mandatory
1	PW Type TLV	3	Y
2	PLE/CEP/TDM Bit-rate TLV	5	Y
3	PLE/CEP Options TLV	3	Y 1*
4	TDM Options TLV	13	Y 2*
5	PLE/CEP/TDM Payload Bytes TLV	3	N
6	Endpoint-ID TLV	0..80	N

1* PLE/CEP only, 2* TDM only

Table 1: BGP PLE attribute TLVs

For a particular PSN it is expected that the network operator will choose a common set of parameters per VPWS type, hence efficient BGP update packing as discussed in section 12 of [RFC4277] is expected to happen.

4.2.1. PW Type TLV

The PW Type TLV MUST be present in the BGP PLE attribute to signal what type of VPWS instance has to be established. Valid PW types for the mechanisms described in this document can be found in Section 3.

The PW Type TLV format is shown in Figure 2.

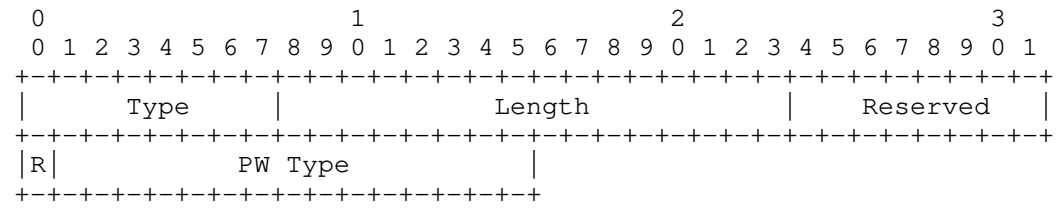


Figure 2: PW type TLV

Type : 1

Length : 3

The total length in octets of the value portion of the TLV.

Reserved / R :

For future use. MUST be set to ZERO and ignored by receiver.

PW Type :

A 15-bit quantity containing a value that represents the type of VPWS. Assigned Values are specified in "IANA Allocations for Pseudowire Edge to Edge Emulation (PWE3)" [RFC4446].

4.2.2. PLE/CEP/TDM Bit-rate TLV

The PLE/CEP/TDM Bit-rate TLV is MANDATORY but MAY be omitted if the attachment circuit type can be unambiguously derived from the PW Type carried in the PW Type TLV. The PLE/CEP/TDM Bit-rate TLV format is shown in Figure 3.

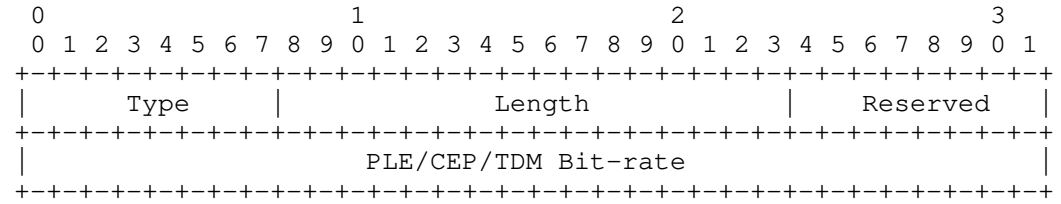


Figure 3: PLE/CEP/TDM Bit-rate TLV

Type : 2

Length : 5

The total length in octets of the value portion of the TLV.

Reserved :

8-bit field for future use. MUST be set to ZERO and ignored by receiver.

PLE/CEP/TDM Bit-rate :

A four byte field denoting the desired payload size to be used. Rules defined in [RFC5287] do apply for signalling TDM VPWS. Rules for CEP VPWS are defined in [RFC4842].

* For PLE [PLE] the bit rate MUST be set to the data rate in units of 1-kbps of the PLE payload.

* Guidelines for setting the bit rate for SAToP VPWS and CESoP VPWS can be found in [RFC5287]. And for CEP VPWS in [RFC4842].

4.2.3. PLE/CEP Options TLV

The PLE/CEP Options TLV MUST be present when signalling CEP and PLE VPWS instances. The PLE/CPE Options TLV format is shown in Figure 4.

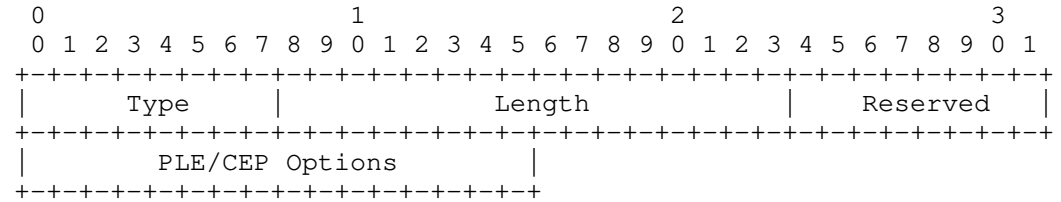


Figure 4: PLE/CEP Options TLV

Type : 3

Length : 3

The total length in octets of the value portion of the TLV.

Reserved :

8-bit field for future use. MUST be set to ZERO and ignored by receiver.

PLE/CEP Options :

A two byte field with the format as shown in Figure 5

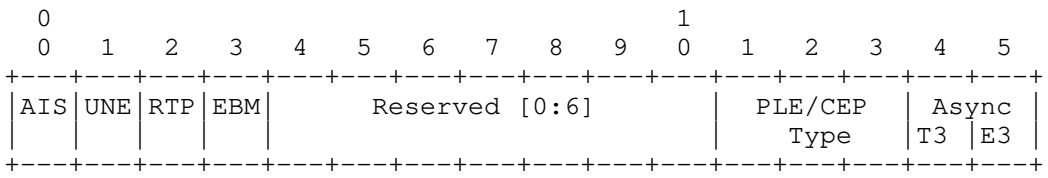


Figure 5: PLE/CEP Options

AIS, UNE, RTP, EBM :

These bits MUST be set to zero and ignored by the receiver except for CEP VPWS. Guidelines for CEP are defined in [RFC4842]

Reserved :

7-bit field for future use. MUST be set to ZERO and ignored by receiver.

CEP/PLE Type :

Indicates the connection type for CEP and PLE. CEP connection types are defined in [RFC4842]. Two new values for PLE are defined in this document:

0x3 - Constand Bit Rate (CBR) PLE payload

0x4 - ODUk frame aligned PLE payload

Async :

These bits MUST be set to zero and ignored by the receiver except for CEP VPWS. Guidelines for CEP are defined in [RFC4842]

4.2.4. TDM options TLV

Whether when signalling TDM VPWS the TDM Options TLV MUST be present or MAY be omitted when signalling TDM VPWS instances is defined in [RFC5287]. The TDM Options TLV format is shown in Figure 6.

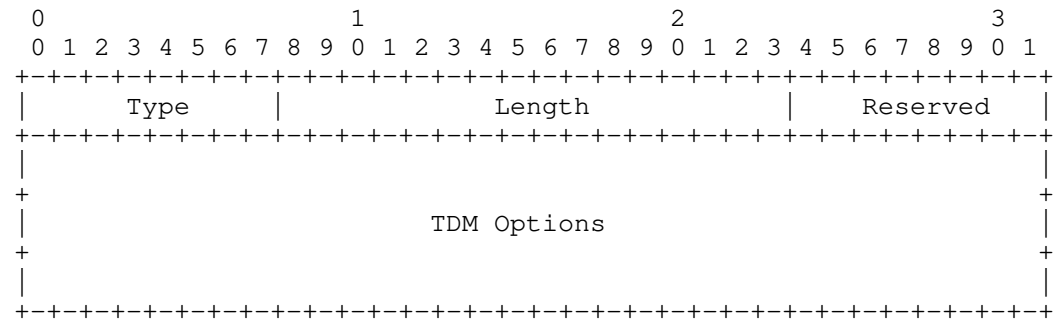


Figure 6: TDM Options TLV

Type : 4

Length : 13

The total length in octets of the value portion of the TLV.

Reserved :

8-bit field for future use. MUST be set to ZERO and ignored by receiver.

TDM Options :

A twelve byte field with the format as defined in section 3.8 of [RFC5287]

4.2.5. PLE/CEP/TDM Payload Bytes TLV

The PLE/CEP/TDM Payload Bytes TLV MAY be included if a non-default payload size is to be used. If this TLV is omitted then the default payload sizes defined in [RFC4553], [RFC5086], [RFC4842] and [PLE] MUST be assumed. The format of the PLE/CEP/TDM Payload Bytes TLV is shown in Figure 7.

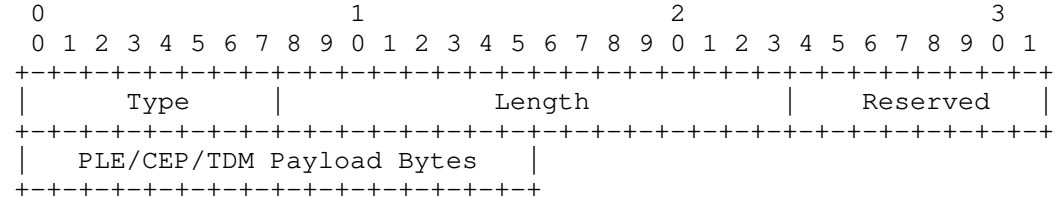


Figure 7: PLE/CEP/TDM Payload Bytes TLV

Type : 5

Length : 3

The total length in octets of the value portion of the TLV.

Reserved :

8-bit field for future use. MUST be set to ZERO and ignored by receiver.

PLE/CEP/TDM Payload Bytes :

A two byte field denoting the desired payload size to be used. Rules defined in [RFC5287] do apply for signalling TDM VPWS. Rules for CEP VPWS are defined in [RFC4842].

4.2.6. Endpoint-ID TLV

The Endpoint-ID TLV MAY be included to allow for misconnection detection. The Endpoint-ID TLV format is shown in Figure 8.

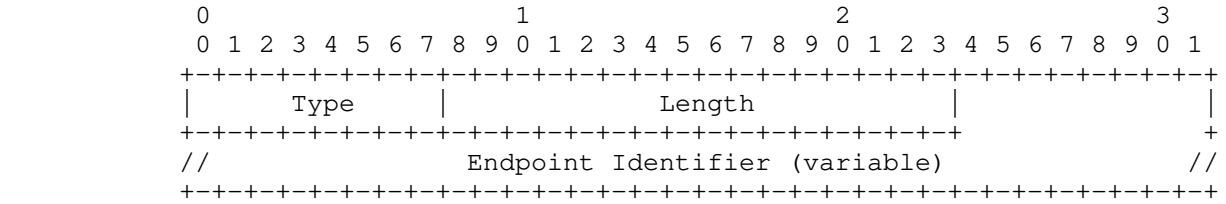


Figure 8: Endpoint-ID TLV

Type : 6

Length : 0-80

The total length in octets of the value portion of the TLV.

Endpoint Identifier :

Arbitrary string of variable length from 0 to 80 octets used to describe the attachment circuit to the remote PE node.

4.3. Control Plane Operations

The deployment model shown in figure 3 of [RFC8214] does equally apply to the operations defined in this document.

4.3.1. VPWS Setup and Teardown

After an attachment circuit has been configured to be part of a VPWS instance and has not declared any local defect, the PE node announces his endpoint using a per-EVI ethernet A-D route to other PEs in the PSN via BGP. The Ethernet Tag ID is set to the VPWS instance identifier and the BGP PLE attribute is included to carry mandatory and optional bit-stream specific attachment circuit parameters.

Both endpoints receiving the EVPN per-EVI A-D route, validate the end to end connectivity by comparing BGP PLE attributes. Upon successful validation, the VPWS instance comes up and traffic can flow through the PSN. In the scenario where the validation phase fails, the remote PE reachability information is simply ignored and totally dismiss as a destination candidate. The VPWS instance validation is performed as follow:

- The mandatory PW type parameter MUST be identical
- The mandatory PLE/CEP/TDM Bit-rate parameter MUST be identical. This MAY be skipped if this parameter was not signaled because the attachment circuit rate can be unambiguously derived from the PW type [RFC5287].
- For CEP and PLE, the mandatory CEP/PLE Type parameter signalled via the CEP/PLE Options TLV MUST be identical
- If the payload size was signalled via the optional PLE/CEP/TDM Payload Bytes TLV it MUST be identical and supported by the PE node. Else the default payload size MUST be assumed.
- If any of the previous statements is no true or any of the signal CEP/PLE or TDM options is not supported by the PE node, the VPWS instance must stay down and a appropriate defect MUST be declared.

PLE is structure agnostic for SONET/SDH service types and hence can not validate whether a mix of SONET and SDH attachment circuits are connected (by incident) via VPWS. The detection of such misconfiguration is the responsibility of the operator managing the CE nodes.

In case of multi-homed CEs the mechanisms defined in [RFC8214] apply but are limited to the single-active and port-active scenarios.

Whenever the VPWS instance configuration is removed, the PE node MUST widthdraw its associated per-EVI ethernet A-D route.

4.3.2. Failure Scenarios

4.3.2.1. Single-homed CEs

Whenever a attachment circuit does declare a local fault the following operations MUST happen:

- Operations defined in [RFC4553], [RFC5086], [RFC4842] and [PLE] MUST happen.
- The per-EVI ethernet A-D route MAY be withdrawn

Whenever the CE-bound IWF does enter packet loss state the operations defined in [RFC4553], [RFC5086], [RFC4842] and [PLE] MUST happen.

4.3.2.2. Multi-homed CEs

Figure 9 demonstrates a multi-homing scenario. CE1 is connected to PE1 and PE2 where PE1 is the designated forwarder while PE2 is the non designated forwarder.

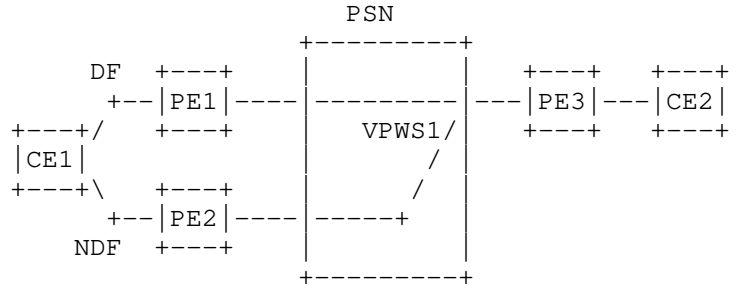


Figure 9: EVPN-VPWS multi-homing redundancy

In Figure 9 PE1 and PE2 are configured for single-active load-balancing mode. Both PEs are advertising a per-ES ethernet A-D route with the same non-zero Ethernet Segment (ES) value and the single-active bit set. This non-zero ES value is called Ethernet Segment Identifier (ESI).

In this example PE1 is elected as Designated Forwarder (DF) for the shared ESI where as PE2 is the Non-Designated Forwarder (NDF) for that segment. The signalling of primary / backup follows exactly the procedure defined in [RFC8214] where P and B bits of the layer 2 attribute extended community are used to settle proper connectivity.

Upon link failure between CE1 and PE1, PE1 and PE2 follows EVPN Ethernet Segment DF Election procedures described in [RFC8214] and [pref_df] for EVPN-VPWS. PE1 leverage mass-withdraw mechanism to tell PE3 to steer traffic over backup connectivity. The per-EVI ethernet A-D route advertisement remains intact. The main purpose is to keep reachability information available for fast convergence purpose. Therefore, the per-EVI ethernet A-D route MAY be withdrawn only under local fault and MUST be withdraw when the circuit is un-configured.

Port-active operation happens in the same way as single-active load-balancing mode described before but at the port level instead of being at the sub-interface level.

5. VPWS signalling using LDP

This section is already under construction and will be soon be publicly announced

6. IANA Considerations

This document defines a new BGP path attribute known as the BGP PLE attribute. IANA is requested to assign attribute code type TBD2 to the BGP PLE attribute from the "BGP Path Attributes" registry.

This document defines a new PW Type for PLE VPWS. IANA is requested to assign a PW type value TBD1 from the "MPLS Pseudowire Types" registry.

7. Security Considerations

The same Security Considerations described in [RFC8214] and [RFC5287] are valid for this document.

8. Acknowledgements

9. References

9.1. Normative References

[PLE] Schmutzer, C., "draft-cisco-bess-mpls-ple-00", 2020.

[pref_df] IETF, "Preference-based EVPN DF Election", <https://tools.ietf.org/html/draft-ietf-bess-evpn-pref-df>.

[RFC4446] Martini, L., "IANA Allocations for Pseudowire Edge to Edge Emulation (PWE3)", BCP 116, RFC 4446, DOI 10.17487/RFC4446, April 2006, <https://www.rfc-editor.org/info/rfc4446>.

[RFC4553] Vainshtein, A., Ed. and YJ. Stein, Ed., "Structure-Agnostic Time Division Multiplexing (TDM) over Packet (SAToP)", RFC 4553, DOI 10.17487/RFC4553, June 2006, <https://www.rfc-editor.org/info/rfc4553>.

[RFC4842] Malis, A., Pate, P., Cohen, R., Ed., and D. Zelig, "Synchronous Optical Network/Synchronous Digital Hierarchy (SONET/SDH) Circuit Emulation over Packet (CEP)", RFC 4842, DOI 10.17487/RFC4842, April 2007, <https://www.rfc-editor.org/info/rfc4842>.

[RFC5086] Vainshtein, A., Ed., Sasson, I., Metz, E., Frost, T., and P. Pate, "Structure-Aware Time Division Multiplexed (TDM) Circuit Emulation Service over Packet Switched Network (CESoPSN)", RFC 5086, DOI 10.17487/RFC5086, December 2007, <https://www.rfc-editor.org/info/rfc5086>.

[RFC5287] Vainshtein, A. and Y(J). Stein, "Control Protocol Extensions for the Setup of Time-Division Multiplexing (TDM) Pseudowires in MPLS Networks", RFC 5287, DOI 10.17487/RFC5287, August 2008, <https://www.rfc-editor.org/info/rfc5287>.

[RFC8077] Martini, L., Ed. and G. Heron, Ed., "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", STD 84, RFC 8077, DOI 10.17487/RFC8077, February 2017, <https://www.rfc-editor.org/info/rfc8077>.

[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>.

[srv6_netprog] IETF, "SRv6 Network Programming", <https://tools.ietf.org/html/draft-ietf-spring-srv6-network-programming>.

[srv6_overlay] IETF, "SRv6 BGP based Overlay services", <https://tools.ietf.org/html/draft-ietf-bess-srv6-services>.

9.2. Informative References

[RFC3985] Bryant, S., Ed. and P. Pate, Ed., "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, DOI 10.17487/RFC3985, March 2005, <https://www.rfc-editor.org/info/rfc3985>.

[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>.

[RFC4277] McPherson, D. and K. Patel, "Experience with the BGP-4 Protocol", RFC 4277, DOI 10.17487/RFC4277, January 2006, <https://www.rfc-editor.org/info/rfc4277>.

[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

Steven Gringeri
Verizon

Email: steven.gringeri@verizon.com

Jeremy Whittaker
Verizon

Email: jeremy.whittaker@verizon.com

Christian Schmutzer (editor)
Cisco Systems, Inc.
Vienna
Austria

Email: cschmutz@cisco.com

Patrice Brissette
Cisco Systems, Inc.
Ottawa, ON
Canada

Email: pbrisset@cisco.com