Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 22 April 2024

K. Vairavakkalai, Ed.
M. Jeyananth
Juniper Networks, Inc.
P.R. Ramadenu
AT&T Services, Inc.
I. Means
AT&T

20 October 2023

BGP Signaled MPLS Namespaces
draft-kaliraj-bess-bgp-sig-private-mpls-labels-07

Abstract

The MPLS forwarding layer in a core network is a shared resource. The MPLS FIB at nodes in this layer contains labels that are dynamically allocated and locally significant at that node. These labels are scoped in context of the global loopback address. Let us call this the global MPLS namespace.

For some usecases like upstream label allocation, it is useful to create private MPLS namespaces (virtual MPLS FIB) over this shared MPLS forwarding layer. This allows installing deterministic label values in the private FIBs created at nodes participating in the private MPLS namespace, while preserving the "locally significant" nature of the underlying shared global MPLS FIB.

This document defines new address families (AFI: 16399, SAFI: 128, or 1) and associated signaling mechanisms to create and use MPLS forwarding contexts in a network. Some example use cases are also described.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

Status of This Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 22 April 2024.

Copyright Notice

Copyright (c) 2023 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction

The MPLS forwarding layer in a core network is a shared resource. The MPLS FIB at nodes in this layer contains labels that are dynamically allocated and locally significant at that node. These labels are scoped in context of the global loopback address. Let us call this the global MPLS namespace.

For some usecases like upstream label allocation, it is useful to create private MPLS namespaces (virtual MPLS FIB) over this shared MPLS forwarding layer. This allows installing deterministic label values in the private FIBs created at nodes participating in the private MPLS namespace, while preserving the "locally significant" nature of the underlying shared global MPLS FIB.

This document defines new address families (AFI: 16399, SAFI: 128, or 1) and associated signaling mechanisms to create and use MPLS forwarding contexts in a network.

The mechanism described in this document reuse [RFC4364] and [RFC8277] procedures to implement Upstream label allocation. The MPLS Namespace family uses BGP VPN style NLRI where the FEC is a MPLS Label, instead of IP prefix. The concepts of MPLS Context tables and upstream allocation are described in [RFC5331].

A BGP speakers participating in a private MPLS namespace creates instance of "MPLS forwarding context" FIB, which is identified using a "Context Protocol Nexthop (CPNH)". A Context label MAY be advertised for the Context Protocol Nexthop (CPNH) using a transport layer protocol or BGP family to other nodes.

2. Terminology

LSR : Label Switch Router

PE : Provider Edge

SFH : Service Forwarding Helper

UHP : Ultimate Hop Pop

MPLS FIB : MPLS Forwarding table

NLRI: Network Layer Reachability Information

AFI: Address Family Identifier

SAFI: Subsequent Address Family Identifier

BN : Border Node

TN : Transport Node, P-router

PE : Provider Edge

BGP VPN : VPNs built using RFC4364 mechanisms

BGP LU: BGP Labeled Unicast family (AFI/SAFIs 1/4, 2/4)

BGP CT: BGP Classful Transport family (AFI/SAFIs, 1/76, 2/76)

RT : Route-Target extended community

RD : Route-Distinguisher

VRF: Virtual Router Forwarding Table

PNH : Protocol Next hop address carried in a BGP Update message

CPNH: Context Protocol Nexthop

MNH : BGP MultiNexthop attribute

FEC : Forwarding Equivalence Class

RSVP-TE : Resource Reservation Protocol - Traffic Engineering

SEP : Service Endpoint, the PNH of a Service route

MPLS: Multi Protocol Label Switching

VNF : Virtual Network Function

vCP : VNF Control Plane

vFP : VNF Forwarding Plane

2.1. Definitions

PLSR: a BGP CT or BGP LU transit node in a private MPLS plane, that does label-swap forwarding for Context label.

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

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

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

Private MPLS FIB Layer (Private MPLS plane): The group of Private MPLS FIBs in the network, connected together via Context labels

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

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

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

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

Service Family : BGP address family used for advertising routes for "data traffic" as opposed to tunnels (e.g. AFI/SAFIs 1/1 or 1/128).

Transport Family : BGP address family used for advertising tunnels, which are in turn used by service routes for resolution (e.g. AFI/ SAFIs 1/4 or 1/76).

3. Motivation

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

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

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

It is desirable to leave the global MPLS FIB layer intact, and build private MPLS FIB-layers on top of it to achieve these requirements. The private MPLS FIBs can then be used by the applications as desired. The private MPLS FIBs need to be created only at the nodes in the network where predictable label-values (external label allocation) is desired. E.g. BNs that need to act as a "Detour-nodes" or "Service-Forwarding-Helpers" that need to mirror service-labels.

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

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

4. Constructs and Building Blocks

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

4.1. Context Protocol Nexthop Address

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

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

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

4.2. MPLS Context FIB

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

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

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

4.3. Context Label

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

Context Label -> Pop and Lookup in MPLS-context FIB

Advertising the "context label in conjunction with the GPNH" tells the network how to reach a "RD:CPNH".

4.4. Roles of Nodes in a MPLS Plane

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

4.4.1. Edge Nodes (PLER)

Private Label Edge-routers (PLER) have MPLS context FIB that belong to the MPLS plane. They advertise the presence of this context FIB using transport layer address families like BGP CT (SAFI 76) or BGP LU (SAFI 4), and private label routes from this FIB are advertused using new BGP AFI/SAFI described in this document.

4.4.2. Transit Nodes (PLSR)

These are just Border-nodes that do label-swap forwarding for the context labels they see in the Context-Protocol-Nexthop advertisement routes (BGP CT or BGP LU) going thru them. They basically stitch/ extend the label switched path to a PLER's CPNH when they re-advertise the CPNH routes with next hop as self.

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

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

4.5. Sending Traffic into a MPLS Plane

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

To send data traffic into this private MPLS plane, the sender MUST use as handle either a "Context label" advertised by a node or a "Private interface" owned by the MPLS context FIB at the node. The MPLS context FIB is created for an application that needs a private MPLS plane.

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

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

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

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

5. BGP Families, Routes and Encoding

This section describes the new constructs defined by this document.

5.1. New Address Families for "MPLS Namespace Signaling"

This document defines a new AFI: "MPLS Namespaces" (IANA code 16399). And two new address-families, using SAFIs 128 and 1. These address families are used to signal MPLS namespaces in BGP. To send or receive routes of these address families, these AFI, SAFI pair of values MUST be negotiated in Multiprotocol Extensions capability described in RFC4760 [RFC4760]

5.1.1. AFI: 16399, SAFI: 128

This address-family is used to exchange private label-routes in private MPLS FIBs at routers that are connected using a common network interface. The private label route has NLRI prefix format "RD:PrivateLabel" and contains Route-Target extended-community identifying the private FIB Layer (VPN) the route belongs to. The nexthop of these routes is set to either the GPNH or the CPNH of the BGP-speaker advertising the RFC-8277 label.

Any transport layer protocol is used to advertise the Context label that the receiving router uses to send traffic into the private MPLS FIB. The Context label installed in the global MPLS FIB points to the private MPLS FIB. The Context label is required when the connecting-interface is a shared common interface that terminates into the global MPLS FIB.

Routes of this address-family can be sent with either IPv4 or IPv6 nexthop. The type of nexthop is inferred from the length of the nexthop.

When the length of Next Hop Address field is 24 (or 48) the nexthop address is of type VPN-IPv6 with 8-octet RD set to zero (potentially followed by the link-local VPN-IPv6 address of the next hop with an 8-octet RD).

When the length of Next Hop Address field is 12 the nexthop address is of type VPN-IPv4 with 8-octet RD.

5.1.2. AFI: 16399, SAFI: 1

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

Because the interface is private, and terminates directly into the private MPLS FIB, a Context label is not required to access the private MPLS FIB and NLRI prefix format is just "PrivateLabel/24", without the RD.

Routes of this address-family can be sent with either IPv4 or IPv6 nexthop. The type of nexthop is inferred from the length of the nexthop.

When the length of Next Hop Address field is 16 (or 32) the nexthop address is of type IPv6 (potentially followed by the link-local IPv6 address of the next hop).

When the length of Next Hop Address field is 4 the nexthop address is a 4 octet IPv4 address.

5.2. Routes and Operational Procedures

5.2.1. "Context-Nexthop" Discovery Route

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

This route is advertised with the following path-attributes:

* BGP Nexthop attribute (code 14, MP_REACH) carrying GPNH address.

* Route-Target extended community, identifying the Transport class, if applicable.

The "Context-Nexthop discovery route" is originated by each speaker who acts as a PLER. The "RD:Context-nexthop" uniquely identifies the private MPLS FIB at the speaker. The "Context-nexthop address" uniquely identifies the private MPLS plane in the network. The Context label advertised in this route has a local forwarding semantic of "Pop, Lookup in Private MPLS FIB".

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

5.2.2. MPLS Namespace "Private Label" Routes

The Private Label routes are carried in the new address-family "MPLS VpnUnicast" (AFI:16399, SAFI:128) aka "MPLS namespace signaling", defined in this document.

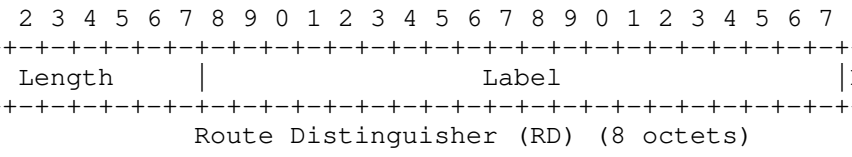
The NLRI format follows the specifications in [RFC8277], with the "Prefix" portion of the NLRI comprising of the RD and "Private MPLS Label" encoded as shown below.

In a MP_REACH_NLRI attribute whose AFI/SAFI is MPLS/128, the "Length" field will be 112 bits or less, comprising of the Label, RD and "Private MPLS Label".

In a MP_REACH_NLRI attribute whose AFI/SAFI is MPLS/1, the "Length" field will be 48 bits or less, comprising of the Label, and "Private MPLS Label".

NLRI Prefix (Private Label route, AFI:16399, SAFI:128)

This picture shows NLRI format when the RFC-8277 Multiple Labels Capability is not used:



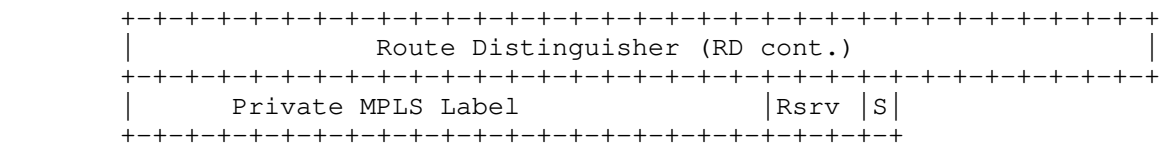


Fig 1: RFC-8277 NLRI with one Label.

- Length:
  The Length field consists of a single octet. It specifies the length in bits of the remainder of the NLRI field.

  In a MP_REACH_NLRI attribute whose AFI/SAFI is MPLS/128, the "Length" field will be 112 bits or less, comprising of the Label, RD and "Private MPLS Label".

  As specified in [RFC4760], the actual length of the NLRI field will be the number of bits specified in the Length field, rounded up to the nearest integral number of octets.

- Label:
  The Label field is a 20-bit field containing an MPLS label value (see [RFC3032]). This label is locally significant, downstream allocated at the speaker identified in the BGP Nexthop field in MP_REACH_NLRI (code 14). This label is pushed in nexthop of the route installed in MPLS context FIB at receiving router.

- Route Distinguisher (RD):
  The 8 byte Route Distinguisher as specified in [RFC4760].

- Private MPLS Label:
  The "Private MPLS Label" field is a 20-bit field containing an MPLS label value (see [RFC3032]). This is an upstream assigned MPLS label, used as destination of route installed in MPLS context FIB at the receiving router.

- Rsrv:
  This 3-bit field SHOULD be set to zero on transmission and MUST be ignored on reception.

- S:
  This 1-bit field MUST be set to one on transmission and MUST be ignored on reception.

Attributes on this route:

* BGP Nexthop attribute (code 14, MP_REACH) carrying a GPNH address. (OR)

* The MultiNextHop attribute [MNH] with forwarding-semantic:

  - "Forward to RD:CPNH"

* Route-Target extended-community, identifying the private FIB-layer

MultiNexthop BGP-attribute (Private Label route)

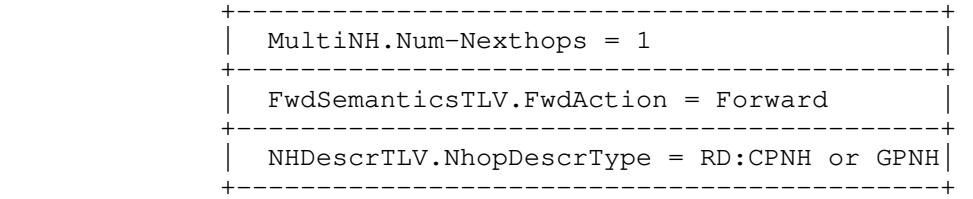


Fig 2: MultiNexthop attr of Private Label route

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

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

Even if the speaker sets nexthop-address to Self because of regular BGP readvertisement-rules, Label Prefix MUST NOT be altered, and the received NLRI "RD:Private-Label1" MUST be re-advertised as-is. Such that value of label "Private-Label1" doesn't change while the packet traverses multiple nodes in the private MPLS FIB layer.

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

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

The label route MUST be installed with forwarding-semantic as specified in the received MultiNextHop attribute. As an example, a Detour node MAY receive the private label route with a forwarding-semantic of "Forward to RD:CPNH" operation. And an Egress node MAY

receive a private label route with a forwarding-semantic pointing to a resource it houses. Note that such a Private label BGP route MAY be received from external-application also.

5.2.2.1. Resolving Received Private Label Routes

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

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

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

6. Example of Usecases

6.1. Label Spoof Protection in Inter-AS Option C Network

In certain deployments, some domains of an Inter AS Option C network may be located in an untrusted geography. Even though such domains are administered by the same operator, employing security mechanisms may be desirable on interfaces connecting such domains.

This section describes how an Inter domain Option C MPLS network can be protected against Label spoofing, using MPLS Namespaces technology.

The inter-AS labeled traffic will be protected against spoofing, such that the transport ASBRs will accept labeled traffic on inter-AS links only if the MPLS label stack matches the transport and service MPLS labels that have been advertised in BGP (LU and L3VPN) families to the peers in untrusted zone.

In order to achieve this security, new functionality is required on only the BNs, PEs or RRs in the trusted zone.

This section illustrates the mechanisms using BGP LU as transport family and L3VPN as service family. But the mechanisms described will work in similar manner for other labeled transport families (e.g., BGP CT) and service families (e.g., L3VPNv6, EVPN, VPLS) as-well.

6.1.1. Reference Topology

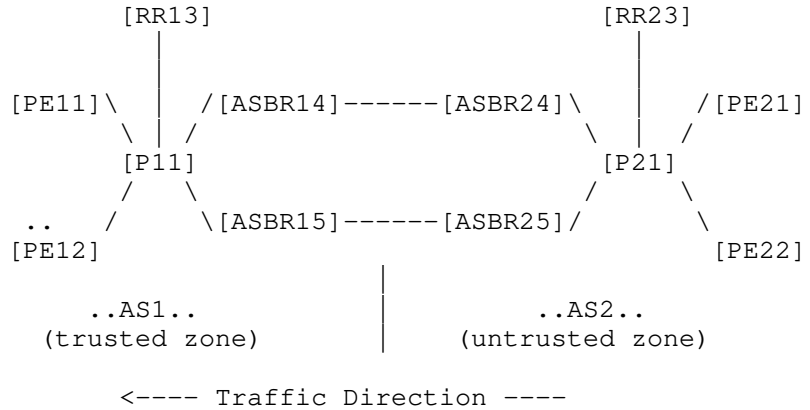


Figure 1: Inter-AS Option C Network with a domain in untrusted zone

Figure 1 shows an Inter-AS Option C network with two domains. AS1 is in a trusted geography, and AS2 is in an untrusted geography.

BGP LU (AFI/SAFI: 1/4) is negotiated on EBGP sessions between ASBR14 - ASBR24 and ASBR15 - ASBR25. BGP LU is also negotiated on IBGP sessions in AS1 between RR13 and the nodes PE11, PE12, ASBR13, and ASBR14; also in AS2 between RR23 and the nodes PE21, PE22, ASBR24, and ASBR25. The ASBRs readvertise the BGP LU routes rewriting next hop to self. The RRs readvertise the BGP LU routes with the next hop unchanged.

L3VPN Service routes are present only at PEs and RRs in the two ASes. L3VPN family (AFI/SAFI: 1/128) is negotiated between PE11, PE12 and RR13. RR13 has multihop EBGP peering with RR23 and negotiates AFI/SAFI: 1/128. RR23 further peers with PE21, PE22 in AS2. The RRs readvertise the L3VPN service routes with next hop unchanged.

In this example loopback addresses of all PEs in one AS are reachable via BGP LU to the other AS.

Following sections describe the control plane and forwarding plane mechanics to deploy label spoofing protection using MPLS Namespaces in this network.

Traffic direction being described is AS2 to AS1, since focus is on traffic entering a trusted zone from an untrusted zone.

6.1.2. Spoof protection for Transport Labels

6.1.2.1. MPLS Namespace to Confine Untrusted Interfaces

At ASBR14 and ASBR15, the interfaces connecting to the BGP peers in untrusted zone are provisioned to terminate in a separate MPLS Namespace, lets call it "From-AS2" namespace. It identifies traffic that is allowed from AS2. This namespace contains a distinct MPLS FIB, which is different from the global MPLS FIB. MPLS packets received on these interfaces are forwarded based on lookup in this MPLS FIB.

ASBR14 and ASBR15 advertise BGP LU routes for PE11, PE12 loopbacks to peers in AS2 with next hop self. Routes for the labels advertised in these routes are installed in the "From-AS2" MPLS namespace. Thus, MPLS packets received on these interfaces will be accepted only if the outermost label is installed in this MPLS namespace FIB. Packets with unknown labels will be discarded.

This provides spoof protection for the transport labels advertised in BGP LU.

6.1.2.2. UHP Labels for PE Loopbacks

The border nodes ASBR14 and ASBR15 use UHP labels in BGP LU routes when advertising a AS1 PE loopback to neighbors in AS2. This label serves as Context Label that identifies traffic sent by AS2 towards that PE in AS1.

The route for Context Label advertised to AS2 neighbors is installed in the "From-AS2" MPLS namespace FIB. This route is installed with a nexthop which has the forwarding semantic as "Pop, Lookup in MPLS-namespace for the PE".

In this manner, the incoming MPLS traffic is validated against the outermost label to match an advertised PE label, and then sent for futher processing in context of the corresponding PE MPLS namespace.

6.1.3. Spoof protection for Service Labels

6.1.3.1. MPLS Namespace for Traffic Destined to a PE

At ASBR14 and ASBR15, a separate MPLS Namespace is created for PE11 and PE12. Lets call them "To-PE1" and "To-PE2" namespaces.

The namespace "To-PE11" identifies traffic direction towards PE11. MPLS packets destined towards PE11 are forwarded based on lookup in this MPLS namespace FIB.

The namespace "To-PE12" identifies traffic direction towards PE12. MPLS packets destined towards PE12 are forwarded based on lookup in this MPLS namespace FIB.

Packets are directed to these namespaces after being processed in the "From-AS2" MPLS namespace FIB.

6.1.3.2. BGP MPLS Namespaces Family Routes

Correspondingly, MPLS Namespaces "To-PE11" and "To-PE12" are created at RR13 which acts as an external label allocator for these namespaces at these ASBRs. The namespace To-PE11 has an associated Route Target RT-PE11. The namespace To-PE12 has an associated Route Target RT-PE12. These Route Targets are exported by the RR and imported by the ASBRs.

In AS1, the route reflector RR13 negotiates MPLS Namespace Signaling family (AFI/SAFI: 16399/128) with the border nodes ASBR14 and ASBR15.

Using the MPLS namespace signaling family, the RR13 insalls the VPN service labels advertised by PE11 and PE12 into their corresponding namespaces at the ASBRs.

Consider PE11 advertising to RR13 a VPN prefix RD:Pfx1 with VPN label VL1, next hop as PE11. RR13 advertises this route with next hop and label unchanged to RR23. When doing so, RR13 originates a MPLS namespace signaling family (AFI/SAFI: 16399/128) route with NLRI RDx:VL1, next hop as PE11, label field containing VL1, and the Route Target RT-PE11.

ASBR14 receives this route and installs in the "To-PE11" MPLS namespace FIB, based on matching import route target RT-PE11. The received next hop PE11 is resolved to map to available tunnel from ASBR14 to PE11. The MPLS route for label VL1 is installed to the "To-PE11" MPLS namespace FIB. This ensures that packets sent by AS2 with VPN label as VL1 will be forwarded properly to PE11. But if an unknown inner label was sent by AS2, such a packet will be dropped after lookup in "To-PE11" MPLS FIB.

Similar mechanism works for labels advertised by PE12, using "To-PE12" MPLS namespace RIB and FIB at RR and ASBRs.

In this manner, protection is provided against nodes in AS2 spoofing service label also.

6.1.4. Applicability to Inter-AS Option B

These mechanisms can be used in Inter-AS Option B scenarios as-well. In such cases, the procedures specified in Section 6.1.2.1 are applied to L3VPN family routes instead of BGP LU routes. MPLS namespace signaling family (AFI/SAFI: 16399/128) is not used in this case.

In Inter-AS Option B scenarios, ASBR14 and ASBR15 re-advertise BGP L3VPN (AFI/SAFI: 1/128) routes from PE11, PE12 to peers in AS2 with next hop self. Routes for the labels advertised in these routes are installed in the "From-AS2" MPLS namespace. Thus, MPLS packets received on these interfaces will be accepted only if the outermost label is installed in this MPLS namespace FIB. Packets with unknown labels will be discarded.

This provides spoof protection for the L3VPN service labels advertised in BGP L3VPN (AFI/SAFI: 1/128) family.

6.2. Improve Scaling and Convergence of a Seamless MPLS Network

MPLS Namespaces can be used to improve scaling and convergence properties of a scaled BGP MPLS network. It acts like a Mezanine transport layer that decouples the service layer from the actual transport layer.

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

* Protocol Nexthop - PE loopback address (GPNH)

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

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

An alternate model could be: to advertise the service routes with a protocol-nexthop of CPNH identifying a namespace, with a forwarding-semantic of:

* "Push <Private-Label>, and Forward to CPNH"

This model fully decouples the service-layer from the transport-layer identifiers, by making the Service routes refer to the CPNH and Private Labels. Thus the underlying transport layer can change

(nodes representing a Private label can be added or removed) without any changes to the service routes. This presents good convergence scaling properties for the network.

This model also allows anycast traffic forwarding to any resource in the network. Multiple PEs can advertise the same Private label to identify a specific service (e.g. peering with an AS) they are offering.

Once the service route traffic enters the private FIB layer, at the closest entry-point determined by path-selection of CPNH auto-discovery routes; then the Private Labels (with pre determined values) pushed will determine the loose hop path taken by the traffic and also the destination-resource.

This section describes how scaling is achieved in an inter-domain MPLS network, where a domain is an AS or IGP area. Domain boundary is demarcated by a BN performing BGP next hop self action on the transport route.

It considers the scenario suggested in Section 6.3.2.1 of [Intent-Routing-Color] where 300K nodes exist in the network with 5 transport classes.

This may result in 1.5M transport layer routes and MPLS transit routes in all Border Nodes in the network, which may overwhelm the nodes' MPLS forwarding resources.

This section explains how "MPLS Namespaces" is used to scale such a network. This approach reduces the number of PNHs that are globally visible in the network, thus reducing forwarding resource usage network wide. Service route state is kept confined closer to network edge, and any churn is confined within the region containing the point of failure, which improves convergence.

In order to achieve these scaling benefits, new functionality is required only at a Region's Border Nodes and the Regional RRs. All other nodes can remain legacy nodes, and still get the scaling and convergence benefits of this mechanism. This is mainly advantageous to ingress and egress PE devices which may be low end devices not capable of pushing deep label stacks or supporting large number of ECMP next hops. They can enjoy the scaling benefits without needing software upgrades.

6.2.1. Illustration

Let us consider the decomposition of this example network with 300K nodes to be such that there are 300 domains containing 1000 nodes each. The mechanism described here will reduce the forwarding resource usage in all Border Nodes to become a function of number of domains (300) instead of number of nodes (300K). Thus, drastically reducing MPLS transit routes from 1.5M to 1500. The Border Nodes and Regional RRs in a Region do the job of abstracting the 1000 PE loopbacks from the rest of the network. The rest of the network sees this region as 1 BGP next hop, and not as 1000 BGP next hops.

6.2.2. Topology

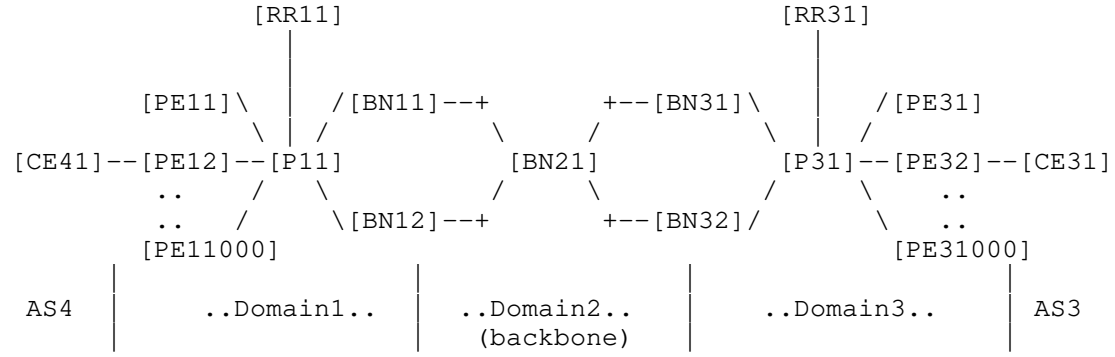


Figure 2: BGP MPLS Namespaces.

This topology in Figure 2 shows a cross section of the network with focus on two domains Domain1 and Domain3 connected via a backbone domain Domain2. Rest of the domains are not shown for brevity. The border nodes have forwarding state pertaining to all domains in the network. The control plane and forwarding plane state in node BN21 can be examined to determine the MPLS scaling characteristics of the network.

L3VPN Service routes are present only at ingress and egress PEs. L3VPN family (AFI/SAFI 1/128) is negotiated between PE11..PE11000 and regional route reflector RR11. RR11 has multihop EBGP peering with RR31 and negotiates AFI/SAFI 1/128. RR31 further peers with all PEs PE31..PE31000 in Domain3.

At the Transport layer - in Domain1, PE11..PE11000 negotiate BGP families (AFI/SAFI 1/4, AFI/SAFI 1/76) with BN11, BN12. In Domain2, BN11 and BN12 similarly negotiate the transport families with BN21,

which in turn peers with BN31 and BN32. In Domain3, BN31 and BN32 peer with PEs PE31..PE31000. Each of these BNs change BGP next hop to self, when re advertising the AFI/SAFI 1/4, AFI/SAFI 1/76 transport routes.

When all nodes loopback addresses are visible throughout the network, it will result in 1.5M transport layer routes and MPLS transit routes in BN21.

Following sections describe the control plane and forwarding plane mechanics to reduce this to 1500 routes, when MPLS Namespaces is deployed in this network.

Traffic direction being described is CE41 to CE31. Reverse direction would work in similar way.

Traffic direction being described is CE41 to CE31. Reverse direction would work in similar way.

6.2.3. Context Protocol Nexthop Address (CPNH)

A MPLS Namespace is identified by a Context PNH address. In MPLS forwarding, labels are locally significant to the node advertising it. E.g. labels in default/global MPLS Namespace are scoped by the node's loopback address. The labels belonging to a MPLS Namespace are locally significant in scope of the Context PNH address.

A UHP label called as "Context Label" is advertised for the CPNH in a transport protocol, which points to the MPLS Namespace forwarding context. When Context label is received as outer label in a MPLS packet, it is Popped, and lookup is performed for the MPLS label that appears in the MPLS Namespace identified by the CPNH.

In this example, CPNH is an anycast IP address that represents set of PEs in a domain. E.g. CPNH1 represent all PEs in Domain1. And CPNH3 represents all PEs in Domain3.

6.2.4. Service Forwarding Helper, and Changes to Transport Layer

The border nodes BN11, BN12 maintain the forwarding context for MPLS Namespace identified by CPNH1. They advertise CPNH1 in transport layer routes like AFI/SAFI 1/4 or AFI/SAFI 1/76 with a UHP Context Label CL1. Any transport layer protocol may be used to advertise the UHP Context Label for the CPNH.

In this way, BN11 and BN12 serve as Service Forwarding Helpers for CPNH1 MPLS Namespace. They attract traffic that remote devices send towards the BGP next hop CPNH1, and forward the MPLS packets received with the MPLS labels belonging to the MPLS Namespace identified by CPNH1.

The individual loopback addresses of the PEs need not be advertised outside the local region. E.g. PE11..PE11000 are not advertised beyond BN11, BN12. Only CPNH1 and RR11 addresses are advertised out. RR1 is used for the control plane peering and CPNH1 is used as a forwarding anchor point.

Similarly, Domain3 advertises only RR31 and CPNH3 to Domain2. This significantly reduces the transport route scale and MPLS forwarding resource usage at the border nodes throughout the network.

6.2.5. BGP MPLS Namespace Address family (AFI:16399, SAFI:128)

In Domain1, the regional route reflector RR11 negotiates MPLS Namespace Signaling address family with the border nodes BN11, BN12. RR11 is an external label allocator for the MPLS Namespace identified by CPNH1. RR1 advertises in the MPLS Namespace address family, the labels it allocated in scope of CPNH1. These routes are advertised with a route target that identifies CPNH1. BN11 and BN12 use this route target to import the label route into the forwarding context associated with CPNH1.

Similarly, in Domain3, RR31 negotiates MPLS Namespace Signaling address family with the border nodes BN31, BN32.

6.2.6. Changes to Service Layer Route Exchange

When RR11 re-advertises to RR31 a VPN route RD:Pfx1 received with label VL1 from egress PE11 in Domain1, it sets BGP next hop to CPNH1, and advertises a new label PL1. This label PL1 is allocated within the scope of CPNH1 namespace.

The label PL1 is advertised to BN1, BN2 in MPLS Namespace address family with a route target identifying CPNH1, and BGP next hop PE11 and label VL1 that were received from the egress PE. BN1 and BN2 resolve the path to that BGP next hop PE11 and use as next hop for the PL1 route installed in CPNH1 forwarding context.

The remote PEs in Domain3 consume the BGP updates from Domain1 following regular procedures for AFI/SAFI 1/128. When resolving the BGP next hop CPNH1, they will push the context label that lands the traffic into the correct forwarding context in one of the border nodes.

6.2.7. Analysis of Forwarding Behavior

The forwarding behavior thus achieved is similar to Inter-AS Option B, without carrying any service routes at the border nodes. Furthermore, the MPLS namespace labels are installed in all the border nodes, which allows for quicker traffic convergence in case of border node failure. The number of border nodes can be increased in a scale out manner, which gives a cookie cutter template to scale a network region.

In conclusion, this mechanism provides both scaling and convergence benefits for the MPLS network, and allows to support huge scale networks.

6.3. VNF Service Forwarding Helper usecase

In a virtualized environment a Service PE node (that comprises of a vCP and multiple vFPs) can mirror MPLS labels (GL1) in its global MPLS FIB to a private forwarding context at an upstream node (SFH) with information on which vFPs are optimal exit-points for that label. Such that the SFH can optimally forward traffic to GL1 to the right vFPs, thus avoiding intra fabric traffic hops.

To do this, the service PE advertises a private label route with RD:GL1 to the SFH node. The route is advertised with a MultiNextHop attribute with one or more legs that have a "Forward to SEPx" semantics. Where SEPx is one of many exit-points at the Service-PE node.

6.4. BGP Based Standard API to Network’s MPLS Forwarding Plane

This mechanism facilitates predictable (external allocator) label values, using a standard BGP family as the API. This gives the external applications a separate MPLS FIB to play with, totally separate from other applications.

This also avoids vendor specific API dependencies between external label allocators (e.g., Controller software), and network routers.

This mechanism also increases the overal MPLS label space available in the network. Because it creates per application label forwarding contexts (namespaces), instead of reserving ranges and splitting the global MPLS FIB among various applications.

6.5. Traffic Engineering and Service Chaining

MPLS namespaces provide an ingress PE the ability to steer MPLS traffic thru specific detour loose hop nodes using predictable label stack.

Labels in a MPLS namespace may be used to identify service chain hops, thus allowing to create a Service Chain consisting of multiple service functions.

Allows private MPLS label usage to spread across multiple domains(e.g., ASes) and works seamlessly with existing technologies like Inter-AS VPN option C.

7. IANA Considerations

This document makes following requests of IANA.

New BGP AFI code ("Address Family Numbers" registry):

* 16399 for "MPLS Namespaces"

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

8. Security Considerations

Using separate mpls forwarding contexts for separate applications and stitching them into separate MPLS planes increases the security attributes of the MPLS network.

9. Acknowledgements

The authors thank Jeffrey (Zhaohui) Zhang, Ron Bonica, Jeff Haas, John Scudder, Jim Uttaro, Israel Means, Torunn Narvestad, Christian Graf, Natarajan Venkataraman, Reshma Das and Aravind Srinivas Srinivasa Prabhakar for the valuable discussions and feedback.

10. References

10.1. Normative References

[BGP-CT] Vairavakkalai, K. and N. Venkataraman, "BGP Classful Transport Planes", 10 July 2023, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-bgp-ct-12>.

[MNH] Vairavakkalai, Ed., "BGP MultiNexthop Attribute", 23 July 2023, <https://datatracker.ietf.org/doc/html/draft-kaliraj-idr-multinexthop-attribute-09>.

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

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

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

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

10.2. Informative References

[Intent-Routing-Color] Hegde, Ed., "Intent-aware Routing using Color", 13 March 2022, <https://datatracker.ietf.org/doc/html/draft-hr-spring-intentaware-routing-using-color-01#section-6.3.2>.

[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>.

[RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, DOI 10.17487/RFC3032, January 2001, <https://www.rfc-editor.org/info/rfc3032>.

Contributors

Moshiko Nayman
Juniper Networks, Inc.
18 Buckingham Dr
Manalapan, New Jersey 07726
United States of America
Email: mnayman@juniper.net

Authors' Addresses

Kaliraj Vairavakkalai (editor)
Juniper Networks, Inc.
1133 Innovation Way,
Sunnyvale, CA 94089
United States of America
Email: kaliraj@juniper.net

Minto Jeyananth
Juniper Networks, Inc.
1133 Innovation Way,
Sunnyvale, CA 94089
United States of America
Email: minto@juniper.net

Praveen Ramadenu
AT&T Services, Inc.
3538 Torrance Blvd, Unit 124
Torrance, CA 90503
United States of America
Email: pr9637@att.com

Israel Means
AT&T
2212 Avenida Mara,
Chula Vista, California 91914
United States of America
Email: israel.means@att.com