Network Working Group
Internet-Draft
Intended status: Experimental
Expires: December 24, 2017

K. Vairavakkalai
M. Jeyananth
Juniper Networks, Inc.

June 22, 2017

BGP signalled private MPLS-labels
draft-kaliraj-bess-bgp-sig-private-mpls-labels-00

Abstract

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

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

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

Requirements Language

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

Status of This Memo

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

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

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

This Internet-Draft will expire on December 24, 2017.

Copyright Notice

Table of Contents

1. Introduction 3
2. Motivation 3
3. Constructs and building blocks 4
3.1. Context Protocol Nexthop Address 4
3.2. MPLS context FIB 4
3.3. Context Label 5
3.4. Roles of nodes in a MPLS-plane 5
3.4.1. Edge-nodes (PLER) 5
3.4.2. Transit-nodes (PLSR) 5
3.5. Sending traffic into the MPLS plane 5
4. Terminology 6
5. BGP families, routes and encoding 7
5.1. New address-families 7
5.1.1. AFI: MPLS, SAFI: 128 7
5.1.2. AFI: MPLS, SAFI: 1 8
5.2. Routes and Operational procedures 8
5.2.1. "Context-Nexthop" discovery route 8
5.2.2. "Private Label" routes 10
6. Example of Usecases 12
6.1. Mezanine transport layer in a Seamless-MPLS network . . . 12
6.2. Service Forwarding Helper usecase 12
6.3. Standard BGP API to a MPLS network's forwarding-plane . . 13
6.4. Traffic engineering and Security advantages 13
7. IANA Considerations 13
8. Security Considerations 14
9. Acknowledgements 14
10. References 14
11. Normative References 14
Authors' Addresses 14

1. Introduction

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

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

It can be noted that, mechanism described in this document is nothing but a [RFC-4364] style BGP VPN where the FEC is MPLS-Label, instead of IP-prefix. This document defines new address-families (AFI: MPLS, SAFI: VPN-Unicast, Unicast) and associated signaling mechanisms to create and use MPLS forwarding-contexts in a network. The concepts of MPLS-Context-tables and upstream allocation are described in [RFC-5331].

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

2. Motivation

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

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

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

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

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

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

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

3. Constructs and building blocks

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

3.1. Context Protocol Nexthop Address

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

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

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

3.2. MPLS context FIB

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

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

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

3.3. Context Label

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

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

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

3.4. Roles of nodes in a MPLS-plane

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

3.4.1. Edge-nodes (PLER)

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

3.4.2. Transit-nodes (PLSR)

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

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

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

3.5. Sending traffic into the MPLS plane

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

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

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

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

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

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

4. Terminology

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

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

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

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

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

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

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

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

MPLS FIB : MPLS Forwarding table

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

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

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

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

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

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

5. BGP families, routes and encoding

This section describes the new constructs defined by this document.

5.1. New address-families

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

5.1.1. AFI: MPLS, SAFI: 128

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

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

5.1.2. AFI: MPLS, SAFI: 1

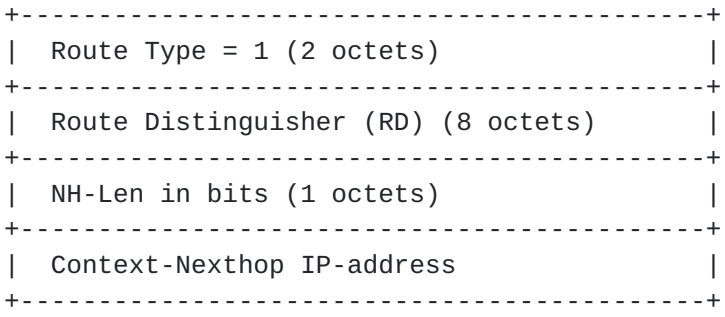
This address-family is used to exchange private label-routes in private MPLS-FIBs to routers that are connected using a private network-interface.

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

5.2. Routes and Operational procedures

5.2.1. "Context-Nexthop" discovery route

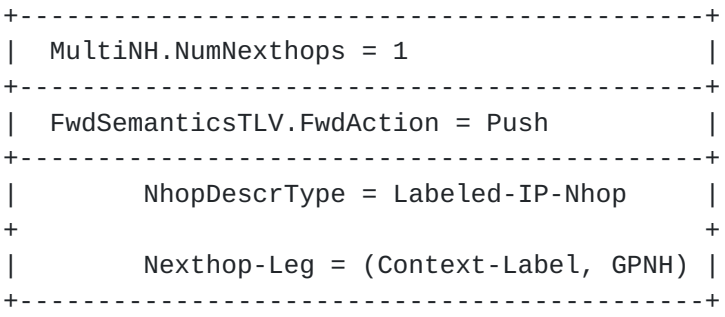
NLRI prefix



The Context-NH discovery route contains the following path-attributes:

- The BGP MultiNexthop-attribute [BGP_MULTI_NH] with forwarding-semantic:
  - Push <Context-Label> to GPNH (for AFI, SAFI: "MPLS, VpnUni"), OR
  - Forward to GPNH (for AFI, SAFI: "MPLS, Uni")
- Route-Target extended community, identifying the private FIB-layer

MultiNexthop BGP-attribute



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

A speaker (re)advertising a Context-Nexthop discovery-route with "next-hop self" MUST allocate a new Context-Label with a forwarding semantic of "Swap Received-Context-Label, Forward to Received-GPNH". This new Context-label along with self-GPNH is advertised in the Multinexthop-attribute [MULTI_NH] attached to the re-advertised Context-nexthop discovery route.

5.2.1.1. Crossing Tunneled domain boundary

"Nexthop-attributes" include BGP Nexthop attribute (code 3), Nexthop-field inside MP_Reach attribute (code 14) or the Multi-Nexthop BGP attribute (code TBD). Two nodes are deemed to be in same tunneled-domain-boundary if they have some sort of transport-tunnel reachability between them (LDP, RSVP, BGP-LU).

A node receiving a "Context-nexthop discovery route" MAY re-advertise it to other BGP speakers who have negotiated the address-family carrying the route. While doing so, the node SHOULD NOT reset the RD:GPNH next-hop address carried in the "Nexthop-attributes" if the re-advertisement does not cross tunneled-domain boundaries.

If a Context-nexthop discovery route is re-advertised across tunneled-domain-boundaries, the re-advertising node MUST set nexthop-address carried in the "Nexthop-attributes" to Self's GPNH, and allocate a new non-reserved label. The route advertised further MUST carry a Multi-nexthop attribute with a forwarding semantic of:

o "SWAP <Received Context-Label> and Forward to Received-GPNH".

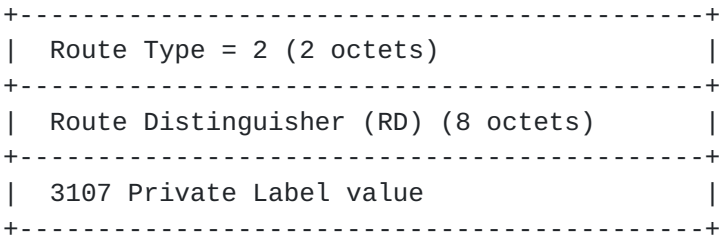
This new-context-label is installed in the global MPLS FIB at the advertising node. And is used as the Context-Label in the re-advertised RD:CPNH route's Multi-Nexthop attribute, with a forwarding-semantic of:

- "Push <New-Context-Label> and Forward to Advertising-GPNH"

.

5.2.2. "Private Label" routes

NLRI prefix (Private Label route)

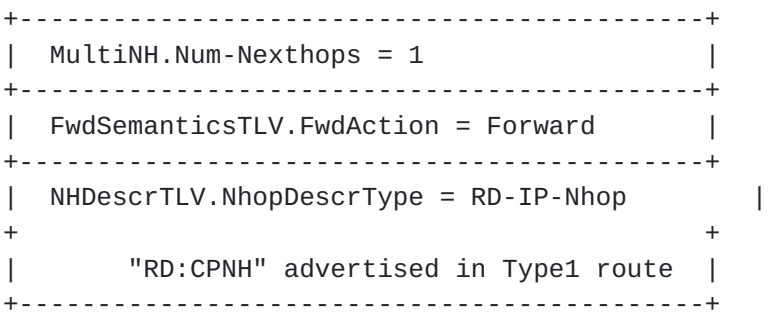


Private-Label-Value: The (upstream assigned) label value

Attributes on this route:

- The Multi-nexthop attribute with forwarding-semantic:
  - "Forward to RD:CPNH"
- Route-Target extended-community, identifying the private FIB-layer

MultiNexthop BGP-attribute (Private Label route)



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

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

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

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

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

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

5.2.2.1. Resolving received Private Label-routes

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

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

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

6. Example of Usecases

6.1. Mezanine transport layer in a Seamless-MPLS network

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

- Protocol Nexthop - PE loopback address (GPNH)
- Service Label - PE advertised locally signifcant label that identifies the service

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

An alternate model could be: to advertise the Service-routes with a protocol-nexthop of CPNH (without RD), with a forwarding-semantic of:

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

This model fully decouples the service-layer from the transport-layer identifiers, by making the Service-routes refer to the CPNH and Private-Labels. Thus the underlying transport-layer can change (nodes representing a Private-label can be added or removed) without any changes to the service-routes. Which present good scaling properties for the network.

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

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

6.2. Service Forwarding Helper usecase

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

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

6.3. Standard BGP API to a MPLS network's forwarding-plane

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

This also avoids vendor specific-API dependencies for external-allocators (controller softwares), and vice-versa.

This mechanism also increases the overal MPLS label-space available in the network, because it creates per-app label-forwarding-contexts (namespaces), instead of reserving/splitting the global MPLS FIB among various applications.

6.4. Traffic engineering and Security advantages

- Ability of ingress to steer mpls-traffic thru specific detour loose-hop nodes using predictable-labels' stack.
- Provide label-spoofing protection at edge-nodes - by virtue of using separate mpls-forwarding-contexts
- Allow private-MPLS label usage to spread across multiple-domains/AS and work seamlessly with existing technologies like Inter-AS VPN option C.

7. IANA Considerations

This document makes following requests of IANA.

New BGP AFI code:

- <TBD> for "MPLS"

Which will be used to create new BGP AFI-SAFI pairs:

- MPLS Uni(SAFI:1),
- MPLS VpnUni(SAFI:128)

.

New NLRI Route-types for these AFI SAFIs:

- Type 1: Context-Nexthop-Discovery-route.

- Type 2: Private-Label route

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

8. Security Considerations

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

9. Acknowledgements

The authors thank Jeffrey (Zhaohui) Zhang, Ron Bonica, Jeff Haas and John Scudder for the valuable discussions.

10. References

[MULTI_NH] https://www.ietf.org/id/draft-kaliraj-idr-multinexthop-attribute-00.txt

[RFC-4364] BGP/MPLS IP Virtual Private Networks (VPNs)

[RFC-5331] MPLS Upstream Label Assignment and Context-Specific Label Space

11. Normative References

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

Authors' Addresses

Kaliraj Vairavakkalai
Juniper Networks, Inc.
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: kaliraj@juniper.net

Minto Jeyananth
Juniper Networks, Inc.
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: kaliraj@juniper.net