Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 8, 2020

K. Vairavakkalai
N. Venkataraman
B. Rajagopalan
Juniper Networks, Inc.

March 07, 2020

BGP Transport VPNs
draft-kaliraj-idr-bgp-transport-vpn-00

Abstract

This document specifies a mechanism, referred to as "service mapping", to express association of overlay routes with underlay routes using BGP. The document describes a framework for service mapping, and specifies BGP protocol procedures that enable dissimination of the service mapping information that may span across administrative domains. It makes it possible to advertise multiple tunnels to the same destination.

A new BGP transport address family is defined for this purpose that uses BGP-VPN [RFC4364] technology and follows MPLS-BGP [RFC8277] NLRI encoding. This new address family is called "Transport-VPN".

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 September 8, 2020.

Table of Contents

1. Introduction

To facilitate service mapping, the tunnels in a network can be grouped by the purpose they serve into a "Transport Class". The tunnels could be created using any signaling protocol, such as LDP, RSVP, BGP-LU or SPRING. The tunnels could also use native IP or IPv6, as long as the tunnels can carry MPLS payload. Tunnels may exist between different pair of end points. Multiple tunnels may exist between the same pair of end points.

Thus, a Transport Class consists of tunnels created by many protocols terminating in various nodes, each satisfying the properties of the class. For example, a "Gold" transport class may consist of tunnels

that traverse the shortest path with fast re-route protection, a "Silver" transport class may hold tunnels that traverse shortest paths without protection, a "To NbrAS Foo" transport class may hold tunnels that exit to neighboring AS Foo, and so on.

The extensions specified in this document can be used to create a BGP transport tunnel that potentially spans domains, while preserving its Transport Class. Examples of domain are Autonomous System (AS), or IGP area. Within each domain, there is a second level underlay tunnel used by BGP to cross the domain. The second level underlay tunnels could be hetrogeneous: Each domain may use a different type of tunnel, or use a differnet signaling protocol. A domain boundary is demarcated by a rewrite of BGP nexthop to 'self' while re-advertising tunnel routes in BGP. The path uses MPLS label-switching when crossing inter-AS links and uses the native intra-AS tunnel of the desired transport class when traversing within a domain.

Overlay routes carry sufficient indication of the Transport Class they should be encapsulated over. A "route resolution" procedure on the ingress node selects from the Transport Class an appropriate tunnel whose destination matches the nexthop of the overlay route. If the overlay route is carried in BGP, the protocol nexthop (or, PNH) is generally carried as an attribute of the route. The PNH of the overlay route is also referred to as "service endpoint". The service endpoint may exist in the same domain as the service ingress node or lie in a different domain, adjacent or non-adjacent.

This document describes mechanisms to:

- Model a "Transport Class" as "Transport RIB" on a router, consisting of tunnel ingress routes of a certain class.

- Enable service routes to resolve over an intended Transport Class by using the corresponding Transport RIB for finding nexthop reachability.

- Advertise tunnel ingress routes in a Transport RIB via BGP without any path hiding, using BGP VPN technology and Add-path. Such that overlay routes in the receiving domains can also resolve over tunnels of associated Transport Class.

- Provide a way for co-operating domains to reconcile between independently administered extended community namespaces, and interoperate between different transport signaling protocols in each domain.

In this document we focus mainly on MPLS LSPs as transport tunnels, but the mechanisms would work in similar manner for non-MPLS transport tunnels too, provided the tunnel can carry MPLS payload.

2. Terminology

LSP: Label Switched Path

TE : Traffic Engineering

SN : Service Node

BN : Border Node

TN : Transport Node, P-router

BGP-VPN : VPNs built using RFC4364 mechanisms

RT : Route-Target extended community

RD : Route-Distinguisher

PNH : Protocol-Nexthop

Service Family : BGP address family used for advertising routes for "data traffic", as opposed to tunnels

Transport Family : BGP address family used for advertising tunnels, which are in turn used by service routes for resolution

Transport Tunnel : A tunnel over which a service may place traffic. These tunnels can be GRE, UDP, LDP, RSVP, or SR-TE

Tunnel Domain : A domain of the network containing SN and BN, under a single administrative control that has a tunnel between SN and BN. An end-to-end tunnel spanning several adjacent tunnel domains can be created by "stitching" them together using labels.

Transport Class : A group of transport tunnels offering the same type of service.

Transport Class RT : A BGP-VPN Route-Target used to identify a specific Transport Class

Transport RIB : At the SN and BN, a Transport Class has an associted Transport RIB that holds its tunnel routes.

Transport RTI : A Routing Instance; container of Transport RIB, and associated Transport Class RT and RD.

Transport-VPN : Set of Transport RTIs importing same Transport Class RT. These are inturn stitched together to span across tunnel domain boundaries using a mechanism similar to Inter-AS option-b to swap labels at BN (nexthop-self).

Mapping Community : Community on a service route, that maps it to resolve over a Transport Class

3. Transport Class

A Transport Class is defined as a set of transport tunnels that share certain characteristics useful for underlay selection.

On the wire, a transport class is represented as the Transport Class RT, which is a regular Route-Target extended community.

A Transport Class is configured at SN and BN, along with attributes like RD and Route-Target. Creation of a Transport Class instantiates the associated Transport RIB and a Transport routing instance to contain them all.

The operator may configure a BN to classify a tunnel into an appropriate Transport Class, which causes the tunnel's ingress routes to be installed in the corresponding Transport RIB. These tunnel routes may then be advertised into BGP.

Alternatively, a router receiving the transport routes in BGP with appropriate signaling information can associate those ingress routes to the appropriate Transport Class. E.g. for Transport-VPN family(SAFI TBD) routes, the Transport Class RT indicates the Transport Class. For BGP-LU family(SAFI 4) routes, import policy based on Communities or inter-AS source-peer may be used to place the route in the desired Transport Class.

When the ingress route is received via SRTE [SRTE], which encodes the Transport Class as an integer "Color" in the NLRI as "Color:Endpoint", the Color can be mapped to a Transport Class during import processing. The Color could map to a Community, or Route-Target that installs the ingress route for "Endpoint" in the appropriate Transport RIB. The SRTE route when advertised out to BGP speakers will then be advertised in Transport-VPN family with Transport Class RT and a new label. The MPLS swap route thus installed for the new label will pop the label and deliver decapsulated-traffic into the path determined by SRTE route.

4. Transport RIB

A Transport RIB is a routing-only RIB that is not installed in forwarding path. However, the routes in this RIB are used to resolve reachability of overlay routes' PNH. Transport RIB is created when the Transport Class it represents is configured.

Overlay routes that want to use a specific Transport Class confine the scope of nexthop resolution to the set of routes contained in the corresponding Transport RIB. This Transport RIB is the "Routing Table" referred in Section 9.1.2.1 RFC4271 [1]

Routes in a Transport RIB are exported out in 'Transport-VPN' address family.

5. Transport Routing Instance

A BGP VPN routing instance that is a container for the Transport RIBs. It imports, and exports routes in this RIB with Transport Class RT. Tunnel destination addresses in this routing instance's context come from the "provider namespace". This is different from user VRFs for e.g., which contain prefixes in "customer namespace"

The Transport Routing instance uses the RD and RT configured for the Transport Class.

6. Nexthop Resolution Scheme

An implementation may provide an option for the service route to resolve over less preferred Transport Classes, should the resolution over preferred, or "primary" Transport Class fail.

To accomplish this, the set of service routes may be associated with a user-configured "resolution scheme", which consists of the primary Transport Class, and an ordered list of fallback Transport Classes.

A community called as "Mapping Community" is configured for a "resolution scheme". A Mapping community maps to exactly one resolution scheme.

When a resolution scheme comprises of a primary Transport Class without any fallback, the Transport Class RT associated with the primary Transport Class is used as the Mapping Community.

A BGP service route is associated with a resolution scheme during import processing. The import processing matches against "Mapping Community" on the service route and determines the resolution scheme that should be used when resolving the route's PNH. If the route

contains more than one Mapping Communities, the first one mapping to a resolution scheme is chosen.

A transport route received in BGP Transport-VPN family should use a resolution scheme that contains only the primary Transport Class without any fallbacks. The primary Transport Class is identified by the Transport Class RT carried on the route. Thus Transport Class RT serves as the Mapping Community for Transport-VPN routes.

7. BGP Transport-VPN Family NLRI

The Transport-VPN family will use the existing AFI of IPv4 or IPv6, and a new SAFI TBD "Transport-VPN" that will apply to both IPv4 and IPv6 AFIs.

The "Transport-VPN" SAFI NLRI itself is encoded as specified in https://tools.ietf.org/html/rfc8277#section-2 [RFC8277].

When AFI is IPv4 the "Prefix" portion of Transport-VPN family NLRI consists of an 8-byte RD followed by an IPv4 prefix. When AFI is IPv6 the "Prefix" consists of an 8-byte RD followed by an IPv6 prefix.

Attributes on a Transport-VPN route include the Route-Target extended community, which is used to leak the route into the right Transport RIBs on SNs and BNs in the network.

8. Comparison with other families using RFC-8277 encoding

SAFI 128 (Inet-VPN) is a RF8277 encoded family that carries service prefixes in the NLRI, where the prefixes come from the customer namespaces, and are contexualized into separate user virtual service RIBs called VRFs, using RFC4364 procedures.

SAFI 4 (BGP-LU) is a RFC8277 encoded family that carries transport prefixes in the NLRI, where the prefixes come from the provider namespace.

SAFI TBD (Transport-VPN) is a RFC8277 encoded family that carries transport prefixes in the NLRI, where the prefixes come from the provider namespace, but are contexualized into separate Transport RIBs, using RFC4364 procedures.

It is worth noting that SAFI 128 has been used to carry transport prefixes in "L3VPN Inter-AS Carrier's carrier" scenario, where BGP-LU/LDP prefixes in CsC VRF are advertised in SAFI 128 to the remote-end baby carrier.

In this document a new AFI/SAFI is used instead of reusing SAFI 128 to carry these transport routes, because it is operationally advantageous to segregate transport and service prefixes into separate address families, RIBs. E.g. It allows to safely enable "per-prefix" label allocation scheme for Transport-VPN prefixes without affecting SAFI 128 service prefixes which may have huge scale. "per prefix" label allocation scheme keeps the routing churn local during topology changes. A new family also facilitates having a different readvertisement path of the transport family routes in a network than the service route readvertisement path. viz. Service routes are exchanged over an EBGP multihop sessions between Autonomous systems with nexthop unchanged; whereas Transport-VPN routes are readvertised over EBGP single hop sessions with "nexthop-self" rewrite over inter-AS links.

The Transport-VPN family is similar in vein to BGP-LU, in that it carries transport prefixes. The only difference is, it also carries in Route Target an indication of which Transport Class the transport prefix belongs to, and uses RD to disambiguate multiple instances of the same transport prefix in a BGP Update.

9. Protocol Procedures

This section summarizes the procedures followed by various nodes speaking Transport-VPN family

Preparing the network for deploying Transport-VPNs

Operator decides on the Transport Classes that exist in the network, and allocates a Route-Target to identify each Transport Class.

Operator configures Transport Classes on the SNs and BNs in the network with unique Route-Distinguishers and Route-Targets.

Implementations may provide automatic generation and assignment of RD, RT values for a transport routing instance; they should also provide a way to manually override the automatic mechanism, in order to deal with any conflicts that may arise with existing RD, RT values in the network.

Origination of Transport-VPN route:

At the ingress node of the tunnel's egress domain, the tunneling protocols install routes in the Transport RIB associated with the Transport Class the tunnel belongs to. The ingress node then advertises this tunnel route into BGP as a Transport-VPN route

with NLRI RD:TunnelEndpoint, attaching a Route-Target that identifies the Transport Class.

Alternatively, the egress node of the tunnel i.e. the tunnel endpoint can originate the BGP Transport-VPN route, with NLRI RD:TunnelEndpoint and PNH TunnelEndpoint, which will resolve over the tunnel route at the ingress node. When the tunnel is up, the Transport-VPN route will become usable and get re-advertised.

Unique RD is used by the originator of a Transport-VPN route to disambiguate the multiple BGP advertisements for a transport end point.

Ingress node receiving Transport-VPN route

On receiving a BGP Transport-VPN route with a PNH that is not directly connected, e.g. an IBGP-route, the Route-Target on the route indicates which Transport Class this route belongs to. The routes in the associated Transport RIB are used to resolve the received PNH. If there does not exist a route in the Transport RIB for the PNH, the Transport-VPN route is considered unusable, and MUST not be re-advertised further.

Border node readvertising Transport-VPN route with nexthop self:

The BN allocates an MPLS label to advertise upstream in Transport-VPN NLRI. The BN also installs an MPLS swap-route for that label that swaps the incoming label with a label received from the downstream BGP speaker, or pops the incoming label. And then pushes received traffic to the transport tunnel or direct interface that the Transport-VPN route's PNH resolved over.

Border node receiving Transport-VPN route on EBGP :

If the route is received with PNH that is known to be directly connected, e.g. EBGP single-hop peering address, the directly connected interface is checked for MPLS forwarding capability. No other nexthop resolution process is performed, as the inter-AS link can be used for any Transport Class.

If the inter-AS links should honor Transport Class, then the BN should follow procedures of an Ingress node described above, and perform nexthop resolution process. The interface routes should be installed in the Transport RIB belonging to the associated Transport Class.

Avoiding path-hiding through Route Reflectors

When multiple BNs exist that advertise a RDn:PEn prefix to RRs, the RRs may hide all but one of the BNs, unless ADDPATH [RFC7911] is used for the Transport-VPN family. This is similar to L3VPN option-B scenarios. Hence ADDPATH should be used for Transport-VPN family, to avoid path-hiding through RRs.

Ingress node receiving service route with mapping community

Service routes received with mapping community resolve using Transport RIBs determined by the resolution scheme. If the resolution process does not find an usable Transport-VPN route or tunnel route in any of the Transport RIBs, the service route MUST be considered unusable for forwarding purpose.

Coordinating between domains using different community namespaces.

Domains not agreeing on RT, RD, Mapping-community values because of independently administered community namespaces may deploy mechanisms to map and rewrite the Route-target values on domain boundaries, using per ASBR import policies. This is no different than any other BGP VPN family. Mechanisms employed in inter-AS VPN deployments may be used with the Transport-VPN family also.

Though RD can also be rewritten on domain boundaries, deploying unique RDs is strongly recommended, because it helps in trouble shooting by uniquely identifying originator of a route, and avoids path-hiding.

Future versions of this document may define a new format of Route-Target extended-community to carry Transport Class, to avoid collision with regular Route Target namespace used by service routes.

10. OAM considerations

TBD

11. IANA Considerations

This document makes following requests of IANA.

New BGP SAFI code for "Transport-VPN". Value TBD.

This will be used to create new AFI,SAFI pairs for IPv4, IPv6 Transport-VPN families. viz:

- "Inet, Transport-VPN". AFI/SAFI = "1/TBD" for carrying IPv4 Transport-VPN prefixes.

- "Inet6, Transport-VPN". AFI/SAFI = "2/TBD" for carrying IPv6 Transport-VPN prefixes.

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

12. Security Considerations

Mechanisms described in this document carry Transport routes in a new BGP address family. That minimizes possibility of these routes leaking outside the expected domain or mixing with service routes.

When redistributing between SAFI 4 and SAFI TBD Transport-VPN routes, there is a possibility of SAFI 4 routes mixing with SAFI 1 service routes. To avoid such scenarios, it is recommended that implementations support keeping SAFI 4 routes in a separate transport RIB, distinct from service RIB that contain SAFI 1 service routes.

13. Acknowledgements

The authors thank Jeff Haas, John Scudder, Navaneetha Krishnan, Ravi M R, Chandrasekar Ramachandran, Shradha Hegde, Richard Roberts, Krzysztof Szarkowicz, John E Drake, Srihari Sangli, Vijay Kestur, Santosh Kolenchery for the valuable discussions.

The decision to not reuse SAFI 128 and create a new address-family to carry these transport-routes was based on suggestion made by Richard Roberts and Krzysztof Szarkowicz.

14. References

14.1. 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>.

[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>.

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

[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>.

[SRTE] Previdi, S., Ed., "Advertising Segment Routing Policies in BGP", 11 2019, <https://tools.ietf.org/html/draft-ietf-idr-segment-routing-te-policy-08>.

14.2. URIs

[1] https://www.rfc-editor.org/rfc/rfc4271#section-9.1.2.1

Authors' Addresses

Kaliraj Vairavakkalai
Juniper Networks, Inc.
1133 Innovation Way,
Sunnyvale, CA 94089
US

Email: kaliraj@juniper.net

Natarajan Venkataraman
Juniper Networks, Inc.
1133 Innovation Way,
Sunnyvale, CA 94089
US

Email: natv@juniper.net

Balaji Rajagopalan
Juniper Networks, Inc.
Electra, Exora Business Park~Marathahalli - Sarjapur Outer Ring Road,
Bangalore, KA 560103
India

Email: balajir@juniper.net