Workgroup: Network Working Group
Internet-Draft: draft-ietf-idr-bgp-ct-30
Published: 18 March 2024
Intended Status: Experimental
Expires: 19 September 2024

Authors: K. Vairavakkalai, Ed.   N. Venkataraman, Ed.
Juniper Networks, Inc.   Juniper Networks, Inc.

BGP Classful Transport Planes

Abstract

This document specifies a mechanism referred to as "Intent Driven Service Mapping". The mechanism uses BGP to express intent based association of overlay routes with underlay routes having specific Traffic Engineering (TE) characteristics satisfying a certain Service Level Agreement (SLA). This is achieved by defining new constructs to group underlay routes with sufficiently similar TE characteristics into identifiable classes (called "Transport Classes"), that overlay routes use as an ordered set to resolve reachability (Resolution Schemes) towards service endpoints. These constructs can be used, for example, to realize the "IETF Network Slice" defined in TEAS Network Slices framework.

Additionally, this document specifies protocol procedures for BGP that enable dissemination of service mapping information in a network that may span multiple cooperating administrative domains. These domains may be administered either by the same provider or by closely coordinating providers. A new BGP address family that leverages RFC 4364 procedures and follows RFC 8277 NLRI encoding is defined to advertise underlay routes with its identified class. This new address family is called "BGP Classful Transport", a.k.a., BGP CT.

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 RFC 2119 [RFC2119] RFC 8174 [RFC8174] when, and only when, they appear in all capitals, as shown here.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

Provider networks typically span across multiple domains where each domain can either represent an Autonomous System (AS) or an Interior Gateway Protocol (IGP) region within an AS. In these networks, several services are provisioned between different pairs of service endpoints (e.g., Provider Edge (PE) nodes), that can either be in the same domain or across different domains.

This document realizes "Intent" as defined in [RFC9315] and prescribes constructs and procedures that enable provider networks to be able to forward service traffic based on service specific intent, end-to-end across service endpoints.

The mechanisms described in this document achieve "Intent Driven Service Mapping" between any pair of service endpoints by:

> Provisioning end-to-end "intent-aware" paths using BGP. For example, low latency path, best effort path.

Expressing a desired intent. For example, use low latency path with fallback to the best effort path.

Forwarding service traffic "only" using end-to-end "intent-aware" paths honoring that desired intent.

The constructs and procedures defined in this document apply homogeneously to intra-AS as well as inter-AS (a.k.a. multi-AS) Option A, Option B and Option C (Section 10, [RFC4364]) style deployments in provider networks.

Provider networks that are deployed using such styles provision intra-domain transport tunnels between a pair of endpoints, typically a service node or a border node, that service traffic use to traverse that domain. These tunnels are signaled using various tunneling protocols depending on the forwarding architecture used in the domain, which can be Multiprotocol Label Switching (MPLS), Internet Protocol version 4 (IPv4), or Internet Protocol version 6 (IPv6).

The mechanisms defined in this document allow different tunneling technologies to become Transport Class aware. These can be applied homogeneously to intra-domain tunneling technologies used in existing brownfield networks as well as new greenfield networks. For clarity, only some tunneling technologies are detailed in this document. In some examples only MPLS Traffic Engineering (TE) examples are described. Other tunneling technologies have been described in detail in other documents and only an overview has been included in this document. For example, the details for Segment Routing (SRv6) are provided in [BGP-CT-SRv6], and an overview is provided in Section 7.13.

Customers need to be able to signal desired Intent to the network, and the network needs to have constructs able to enact the customer's intent. The network constructs defined in this document are used to classify and group these intra-domain tunnels based on various characteristics, like TE characeteristics (e.g., low latency), into identifiable classes that can pass "intent-aware" traffic. These constructs enable services to express their desired intent on using one or more identifiable classes, and mechanisms to selectively map traffic onto "intent-aware" tunnels for these classes.

This document introduces a new BGP address family called "BGP Classful Transport", that extends/stitches intent-aware intra-domain tunnels belonging to the same class across domain boundaries, to establish end-to-end intent-aware paths between service endpoints.

[Intent-Routing-Color] describes various use cases and applications of the procedures described in this document.

2. Terminology

ABR: Area Border Router

AFI: Address Family Identifier

AS: Autonomous System

ASBR: Autonomous System Border Router

ASN: Autonomous System Number

BGP VPN: VPNs built using RD, RT; architecture described in RFC4364

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

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

BN: Border Node

CBF: Class Based Forwarding

CsC: Carrier serving Carrier VPN

DSCP: Differentiated Services Code Point

EP: Endpoint of a tunnel, e.g. a loopback address in the network

EPE: Egress Peer Engineering

eSN: Egress Service Node

FEC: Forwarding Equivalence Class

iSN: Ingress Service Node

LPM: Longest Prefix Match

LSP: Label Switched Path

MNH: BGP MultiNexthop attribute

MPLS: Multi Protocol Label Switching

NLRI: Network Layer Reachability Information

PE: Provider Edge

PHP: Penultimate Hop Pop

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

RD: Route Distinguisher

RSVP-TE: Resource Reservation Protocol - Traffic Engineering

RT: Route Target extended community

RTC: Route Target Constrain

SAFI: Subsequent Address Family Identifier

SID: Segment Identifier

SLA: Service Level Agreement

SN: Service Node

SR: Segment Routing

SRTE: Segment Routing Traffic Engineering

TC: Transport Class

TC ID: Transport Class Identifier

TC-BE: Best Effort Transport Class

TE: Traffic Engineering

TRDB: Transport Route Database

UHP: Ultimate Hop Pop

VRF: Virtual Routing and Forwarding table

2.1. Definitions and Notations

BGP Community Carrying Attribute (CCA) : A BGP attribute that carries community. Examples of BGP CCA are: Communities (attr code 8), Extended Communities (attr code 16), IPv6 Address Specific Extended Community (attr code 25), Large community (attr code 32).

color:0:100 : This notation denotes a Color extended community as defined in RFC 9012 with the Flags field set to 0 and the color field set to 100.

End to End Tunnel: A tunnel spanning several adjacent tunnel domains created by "stitching" them together using MPLS labels or an equivalent identifier based on the forwarding architecture.

Import processing: Receive side processing of an overlay route, including things like import policy application, resolution scheme selection and next hop resolution.

Intent: A set of operational goals (that a network should meet) and outcomes (that a network is supposed to deliver) defined in a declarative manner without specifying how to achieve or implement them, as defined in Section 2 of [RFC9315].

Mapping Community: Any BGP CCA (e.g., Community, Extended Community) on an overlay route that maps to a Resolution Scheme. For example, color:0:100, transport-target:0:100.

Resolution Scheme: A construct comprising of an ordered set of TRDBs to resolve next hop reachability, for realizing a desired intent.

Service Family: A BGP address family used for advertising routes for destinations in "data traffic". For example, AFI/SAFIs 1/1 or 1/128.

Transport Family: A BGP address family used for advertising tunnels, which are in turn used by service routes for resolution. For example, AFI/SAFIs 1/4 or 1/76.

Transport Tunnel : A tunnel over which a service may place traffic. Such a tunnel can be provisioned or signaled using a variety of means. For example, Generic Routing Encapsulation (GRE), UDP, LDP, RSVP-TE, IGP FLEX-ALGO or SRTE.

Tunnel Route: A Route to Tunnel Destination/Endpoint that is installed at the headend (ingress) of the tunnel.

Tunnel Domain: A domain of the network containing Service Nodes (SNs) and Border Nodes (BNs) under a single administrative control that has tunnels between them.

Brownfield network: An existing network that is already in service, deploying a chosen set of technologies and hardware. Enhancements and upgrades to such network deployments protect return on investment, and should consider continuity of service.

Greenfield network: A new network deployment which can make choice of new technology or hardware as needed, with fewer constraints than brownfield network.

Transport Class: A construct to group transport tunnels offering similar SLA.

Transport Class RT: A Route Target Extended Community used to identify a specific Transport Class.

transport-target:0:100 : This notation denotes a Transport Class RT extended community as defined in this document with the "Transport Class ID" field set to 100.

Transport Route Database: At the SN and BN, a Transport Class has an associated Transport Route Database that collects its Tunnel Routes.

Transport Plane: An end-to-end plane consisting of transport tunnels belonging to the same Transport Class.

3. Architecture Overview

This section describes the BGP CT architecture with a brief illustration.

```
                INET     [RR21]--------------<<---[RR11]
                Service  /                       /   | IP1, color:0:100
       [PE21] <<--------+       | [SN11] <<-----+    | IP2, color:0:200
          \       ___           |       \    ___     | IP3, 100:200
           \    _(   )          |        \ _(   )    ^<<    ^^^^^^^^^^^
            +--(     _)--[BN21]===[BN11]--(     _)--[PE11]    Mapping
                (___)           |          (___)             Community
                          Inter-AS-Link
                                |
      [.......AS2:SR-TE........]|[.......AS1:RSVP-TE......]
       -------->---------MPLS Forwarding--------->--------

          [PE21]--<<--[BN21]          [BN21]--<<--[BN11]
     { <<-RD1:PE11(L3),PNH=BN21 | <<-RD1:PE11(L1),PNH=BN11 }
     |   transport-target:0:100 |   transport-target:0:100 | BGP
     |                          |                          | Classful
     | <<-RD2:PE11(L4),PNH=BN21 | <<-RD2:PE11(L2),PNH=BN11 | Transport
     {   transport-target:0:200 |   transport-target:0:200 }
         ^^^^^^^^^^^^^^^^^^^^^^                        ^^^
             Route Target &                 Transport Class ID
            Mapping Community

  at SN11 and PE21,

       Scheme1: color:0:100, (TRDB[TC-100], TRDB[TC-BE])
       Scheme2: color:0:200, (TRDB[TC-200], TRDB[TC-BE])
       Scheme3:     100:200, (TRDB[TC-100], TRDB[TC-200])
       ^^^^^^^                ^^^^               ^^^^^^
  Resolution Schemes   Transport Route DB    Transport Class
```

Figure 1: BGP CT Overview with Example Topology

To achieve end-to-end "Intent Driven Service Mapping", this document defines the following constructs and BGP extensions:

The "Transport Class" (Section 4) construct to group underlay tunnels with sufficiently similar TE characteristics.

The "Resolution Scheme" (Section 5) construct for overlay routes with Mapping Community to resolve next hop reachability from either one or an ordered set of Transport Classes.

The "BGP Classful Transport" (Section 6) address family to extend these constructs to adjacent domains.

Figure 1 depicts the intra-AS and inter-AS application of these constructs. It uses an example topology of Inter-AS option C network with two AS domains. AS1 is a RSVP-TE network, AS2 is a SRTE network. BGP CT and BGP LU are transport layer families used between the two AS domains. IP1, IP2, IP3 are service prefixes (AFI/SAFI: 1/1) behind egress PE11.

PE21, SN11 and PE11 are the SNs in this network. SN11 is an ingress PE with intra domain reachability to PE11. PE21 is an ingress PE with inter domain reachability to PE11.

The tunneling mechanisms are made "Transport Class" aware. They publish their underlay tunnels for a Transport Class into an associated "Transport Route Database" (TRDB) (Section 4.2). In Figure 1, RSVP-TE publishes its underlay tunnels into TRDBs created for Transport Class 100 and 200 at BN11 and SN11 within AS1; Similarly, SR-TE publishes its underlay tunnels into TRDBs created for Transport Class 100 and 200 at PE21 within AS2.

The underlay route in a TRDB can be advertised in BGP to extend an underlay tunnel to adjacent domains. A new BGP transport layer address family called "BGP Classful Transport", also known as BGP CT (AFI/SAFIs 1/76, 2/76) is defined for this purpose. BGP CT makes it possible to advertise multiple tunnels to the same destination address, thus avoiding the need for multiple loopbacks on the Egress Service Node (eSN).

The BGP CT address family carries transport prefixes across tunnel domain boundaries, which is parallel to BGP LU (AFI/SAFIs 1/4 or 2/4). It disseminates "Transport Class" information for the transport prefixes across the participating domains while avoiding the need of per-transport class loopback. This is not possible with BGP LU without using per-color loopback. This makes the end-to-end network a "Transport Class" aware tunneled network.

In Figure 1, BGP CT routes are originated at BN11 in AS1 with next hop "self" towards BN21 in AS2 to extend available RSVP-TE tunnels for Transport Class 100 and 200 in AS1. BN21 propagates these routes with next hop "self" onto PE21, which resolves the BGP CT routes over SRTE tunnels belonging to same transport class.

Overlay routes carry sufficient indication of the desired Transport Classes using a BGP community which assumes the role of as a "Mapping Community". A Resolution Scheme is identified by its "Mapping Community", where its configuration can either be auto-generated based on TC ID or done manually.

The following text illustrates CT architecture having the property of providing tiered fallback options at a per-route granularity. In Figure 1, the Resolution Schemes are shown and the following next hop resolutions are done by SN11 and PE21 for the service routes of prefixes IP1, IP2, IP3:

> Resolve IP1 next hop over available tunnels in TRDB for Transport Class 100 with fallback to TRDB for best effort.
>
> Resolve IP2 next hop over available tunnels in TRDB for Transport Class 200 with fallback to TRDB for best effort.
>
> Resolve IP3 next hop over available tunnels in TRDB for Transport Class 100 with fallback to TRDB for Transport Class 200.

In Figure 1, SN11 resolves IP1, IP2 and IP3 directly over RSVP-TE tunnels in AS1. PE21 resolves IP1, IP2 and IP3 over extended BGP CT tunnels that resolve over SR-TE tunnels in AS2.

This document describes procedures using MPLS forwarding architecture. However, these procedures would work in a similar manner for non-MPLS forwarding architectures as well. Section 7.13 describes the application of BGP CT over SRv6 data plane.

4. Transport Class

Transport Class is a construct that groups transport tunnels offering similar SLA within the administrative domain of a provider network or closely coordinated provider networks.

A Transport Class is uniquely identified on a box by a 32-bit "Transport Class ID", that is assigned by the operator. The operator consistently provisions a Transport Class on participating nodes (SNs and BNs) in a domain with its unique Transport Class ID.

A Transport Class is also configured with RD and import/export RT attributes. Creation of a Transport Class instantiates its corresponding TRDB and Resolution Schemes on that node.

All nodes within a domain agree on a common Transport Class ID namespace. However, two co-operating domains may not always agree on the same namespace. Procedures to manage differences in Transport Class ID namespaces between co-operating domains are specified in Section 11.2.2.

Transport Class ID conveys the Color of tunnels in a Transport Class. The terms 'Transport Class ID' and 'Color' are used interchangeably in this document.

4.1. Classifying TE tunnels

TE tunnels can be classified into a Transport Class based on the TE attributes they possess and the TE characteristics that the operator defines for that Transport Class. Due to the fact that multiple TE tunneling protocols exist, their TE attributes and characteristics may not be equal but sufficiently similar. Some examples of such classifications are as follows:

> Tunnels (RSVP-TE, IGP FLEX-ALGO, SR-TE) that support latency sensitive routing.
>
> RSVP-TE Tunnels that only go over admin-group with Green links.
>
> Tunnels (RSVP-TE, SR-TE) that offer Fast Reroute.
>
> Tunnels (RSVP-TE, SR-TE) that share resources in the network based on Shared Risk Link Groups defined by TE policy.
>
> Tunnels (RSVP-TE, SR-TE, BGP CT) that avoid certain nodes in the network based on RSVP-TE ERO, SR-TE policy or BGP policy.

An operator may configure a SN/BN to classify a tunnel into an appropriate Transport Class. How exactly these tunnels are made Transport Class aware is implementation specific and outside the scope of this document.

When a tunnel is made Transport Class aware, it causes the Tunnel Route to be installed in the corresponding TRDB of that Transport Class. These routes are used to resolve overlay routes, including BGP CT. The BGP CT routes may be further readvertised to adjacent domains to extend these tunnels. While readvertising BGP CT routes, the "Transport Class" identifier is encoded as part of the Transport Class RT, which is a new Route Target extended community defined in Section 4.3.

A SN/BN receiving the transport routes via BGP with sufficient signaling information to identify a Transport Class can associate those tunnel routes to the corresponding Transport Class. For example, in BGP CT family routes, the Transport Class RT indicates the Transport Class. For BGP LU family routes, import processing based on Communities or Inter-AS source-peer may be used to place the route in the desired Transport Class.

When the tunnel route is received via [SRTE] with "Color:Endpoint" as the NLRI that encodes the Transport Class as an integer 'Color', the 'Color' is mapped to a Transport Class during the import processing. The SRTE tunnel route for this 'Endpoint' is installed in the corresponding TRDB. The SRTE tunnel will be extended by a BGP CT advertisement with NLRI 'RD:Endpoint', Transport Class RT and a new label. The MPLS swap route thus installed for the new label will pop the label and forward the decapsulated traffic into the path determined by the SRTE route for further encapsulation.

[PCEP-SRPOLICY] extends Path Computation Element Communication Protocol (PCEP) to signal attributes of an SR Policy which include Color. This Color is mapped to a Transport Class thus associating the SR Policy with the desired Transport Class.

Similarly, [PCEP-RSVP-COLOR] extends PCEP to carry the Color attribute for its use with RSVP-TE LSPs . This Color is mapped to a Transport Class thus associating the RSVP-TE LSP with the desired Transport Class.

4.2. Transport Route Database

A Transport Route Database (TRDB) is a logical collection of transport routes pertaining to the same Transport Class. In any node, every Transport Class has an associated TRDB. Resolution Schemes resolve next hop reachability for EP using the transport routes within the scope of the TRDBs.

Tunnel endpoint addresses (EP) in a TRDB belong to the "Provider Namespace" representing the core transport region.

An implementation may realize the TRDB as a "Routing Table" referred in Section 9.1.2.1 of RFC4271 which is used only for resolving next hop reachability in control plane. An implementation may choose a different datastructure to realize this logical construct while still adhering to the procedures defined in this document. The tunnel routes in a TRDB require no footprint in the forwarding plane unless they are used to resolve a next hop.

SNs or BNs originate routes for the "Classful Transport" address family from the TRDB. These routes have "RD:Endpoint" in the NLRI, carry a Transport Class RT, and an MPLS label or equivalent identifier in different forwarding architecture. "Classful Transport" family routes received with Transport Class RT are installed into their respective TRDB.

4.3. "Transport Class" Route Target Extended Community

This section defines a new type of Route Target, called a "Transport Class" Route Target Extended Community; also known as a Transport Target. The procedures for use of this extended community with BGP CT routes (AFI/SAFI: 1/76 or 2/76) are described below.

The "Transport Class" Route Target Extended Community is a transitive extended community EXT-COMM [RFC4360] of extended type, which has the format as shown in Figure 2.

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   Type= 0xa   | SubType= 0x02 |            Reserved           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       Transport Class ID                      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Type: 1-octet field MUST be set to 0xa to indicate 'Transport Class'.

SubType: 1-octet field MUST be set to 0x2 to indicate 'Route Target'.

Reserved: 2-octet reserved bits field.
This field MUST be set to zero on transmission.
This field SHOULD be ignored on reception, and
MUST be left unaltered.

Transport Class ID: This field is encoded in 4 octets.

This field contains the "Transport Class" identifier, which is an unsigned 32-bit integer.

This document reserves the Transport class ID value 0 to represent "Best Effort Transport Class ID".

Figure 2: "Transport Class" Route Target Extended Community

The VPN route import/export mechanisms specified in BGP/MPLS IP VPNs [RFC4364] and the Constrained Route Distribution mechanisms specified in Route Target Constrain [RFC4684] are applied using Route Target extendend community. These mechanisms are applied to BGP CT routes (AFI/SAFI: 1/76 or 2/76) using "Transport Class Route Target Extended community".

A BGP speaker that implements Route Target Constrain [RFC4684] MUST also apply the RTC procedures to the Transport Class Route Target Extended communities carried on BGP CT routes (AFI/SAFI: 1/76 or

2/76). An RTC route is generated for each Route Target imported by locally provisioned Transport Classes.

Further, when processing RT membership NLRIs received from external BGP peers, it is necessary to consider multiple EBGP paths for a given RTC prefix for building the outbound route filter, and not just the best path. An implementation MAY provide configuration to control how many EBGP RTC paths are considered.

The Transport Class Route Target Extended community is carried on BGP CT family routes and is used to associate them with appropriate TRDBs at receiving BGP speakers. The Transport Target is carried unaltered on the BGP CT route across BGP CT negotiated sessions except for scenarios described in Section 11.2.2. Implementations should provide policy mechanisms to perform match, strip, or rewrite operations on a Transport Target just like any other BGP community.

Defining a new type code for the Transport Class Route Target Extended community avoids conflicting with any VPN Route Target assignments already in use for service families.

This document also reserves the Non-Transitive version of Transport Class extended community (Section 13.2.1.1.2) for future use. The "Non-Transitive Transport Class" Route Target Extended Community is not used. If received, it is considered equivalent in functionality to the Transitive Transport Class Route Target Extended Community, except for the difference in Transitive bit flag.

5. Resolution Scheme

A Resolution Scheme is a construct that consists of a specific TRDB or an ordered set of TRDBs. An overlay route is associated with a resolution scheme during import processing, based on Mapping Community on the route.

Resolution Schemes enable a BGP speaker to resolve next hop reachability for overlay routes over the appropriate underlay tunnels within the scope of the TRDBs. Longest Prefix Match (LPM) of the next hop is performed within the identified TRDB.

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

To accomplish this, the "Resolution Scheme" is configured with the primary Transport Class, and an ordered list of fallback Transport Classes. Two Resolution Schemes are considered equivalent in Intent if they consist of the same ordered set of TRDBs.

Operators must ensure that Resolution Schemes for a mapping community are provisioned consistently on various nodes participating in a BGP CT network, based on desired Intent and transport classes available in that domain.

5.1. Mapping Community

A "Mapping Community" is used to signal the desired Intent on an overlay route. At an ingress node receiving the route, it maps the overlay route to a "Resolution Scheme" used to resolve the route's next hop.

A Mapping Community is a "role" and not a new type of community; any BGP Community Carrying Attribute (e.g. Community or Extended Community) may play this role, besides the other roles it may already be playing. For example, the Transport Class Route Target Extended Community plays both roles of being a Route Target as well as a Mapping Community.

Operator provisioning ensures that the ingress and egress SNs agree on the BGP CCA and community namespace to use for the Mapping Community.

A Mapping Community maps to exactly one Resolution Scheme at receiving BGP speaker. An implementation SHOULD allow associating multiple Mapping Communities to a Resolution Scheme. This helps with renumbering and migration scenarios.

An example of mapping community is "color:0:100", described in [RFC9012], or the "transport-target:0:100" described in Section 4.3 in this document.

The order of communities on an overlay route does not affect the determining of Mapping community in effect.

The first community on the overlay route that matches a Mapping Community of a locally configured Resolution Scheme is considered the effective Mapping Community for the route. The Resolution Scheme thus found is used when resolving the route's PNH. If a route contains more than one Mapping Community, it indicates that the route considers these distinct Mapping Communities as equivalent in Intent.

If more than one distinct Mapping Communities on an overlay route map to distinct Resolution Schemes with dissimilar Intents at a receiving node, it is considered a configuration error. Operators should avoid such configuration errors when attaching mapping communities on overlay routes.

It should be noted that the Mapping Community role does not require applying Route Target Constrain procedures specified in RFC 4684.

6. BGP Classful Transport Family

The BGP Classful Transport (BGP CT) family will use the existing Address Family Identifier (AFI) of IPv4 or IPv6 and a new SAFI 76 "Classful Transport" that will applies to both IPv4 and IPv6 AFIs.

The AFI/SAFI 1/76 MUST be negotiated as per the Multiprotocol Extensions capability described in Section 8 of [RFC4760] to be able to send and receive BGP CT routes for IPv4 endpoint prefixes.

The AFI/SAFI 2/76 MUST be negotiated as per the Multiprotocol Extensions capability described in Section 8 of [RFC4760] to be able to send and receive BGP CT routes for IPv6 endpoint prefixes.

6.1. NLRI Encoding

The "Classful Transport" SAFI NLRI has the same encoding as specified in Section 2 of [RFC8277].

When AFI/SAFI is 1/76, the Classful Transport NLRI Prefix consists of an 8-byte RD followed by an IPv4 prefix. When AFI/SAFI is 2/76, the Classful Transport NLRI Prefix consists of an 8-byte RD followed by an IPv6 prefix.

The procedures described for AFI/SAFIs 1/4 or 1/128 in Section 2 of [RFC8277] apply for AFI/SAFI 1/76 also. The procedures described for AFI/SAFIs 2/4 or 2/128 in Section 2 of [RFC8277] apply for AFI/SAFI 2/76 also.

BGP CT routes MAY carry multiple labels in the NLRI, by negotiating the Multiple Labels Capability as described in Section 2.1 of [RFC8277]

Attributes on a Classful Transport route include the Transport Class Route Target extended community, which is used to associate the route with the correct TRDBs on SNs and BNs in the network and either an IPv4 or an IPv6 next hop.

6.2. Next Hop Encoding

When the length of the Next hop Address field is 4, the next hop address is of type IPv4 address.

When the length of Next hop Address field is 16 (or 32), the next hop address is of type IPv6 address (potentially followed by the link-local IPv6 address of the next hop). This follows Section 3 in [RFC2545]

When the length of Next hop Address field is 24 (or 48), the next hop address is of type VPN-IPv6 with an 8-octet RD set to zero (potentially followed by the link-local VPN-IPv6 address of the next hop with an 8-octet RD set to zero). This follows Section 3.2.1.1 in [RFC4659]

When the length of the Next hop Address field is 12, the next hop address is of type VPN-IPv4 with 8-octet RD set to zero.

If the length of the Next hop Address field contains any other values, it is considered an error and is handled via BGP session reset as per Section 7.11 of [RFC7606].

6.3. Carrying multiple Encapsulation Information

To ease interoperability between nodes supporting different forwarding technologies, a BGP CT route allows carrying multiple encapsulation information.

An MPLS Label is carried using the encoding in [RFC8277]. A node that does not support MPLS forwarding advertises the special label 3 (Implicit NULL) in the RFC 8277 MPLS Label field. The Implicit NULL label carried in BGP CT route indicates to receiving node that it should not impose any BGP CT label for this route.

The SID information for SR with respect to MPLS Data Plane is carried as specified in Prefix SID attribute defined as part of Section 3 in [RFC8669].

The SID information for SR with respect to SRv6 Data Plane is carried as specified in Section 7.13.

UDP tunneling information is carried using Tunnel Encapsulation Attribute as specified in [RFC9012].

6.4. Comparison with Other Families using RFC-8277 Encoding

AFI/SAFI 1/128 (MPLS-labeled VPN address) is an RFC8277 encoded family that carries service prefixes in the NLRI, where the prefixes come from the customer namespaces and are contextualized into separate user virtual service RIBs called VRFs as per [RFC4364].

AFI/SAFI 1/4 (BGP LU) is an RFC8277 encoded family that carries transport prefixes in the NLRI, where the prefixes come from the provider namespace.

AFI/SAFI 1/76 (Classful Transport SAFI) is an RFC8277 encoded family that carries transport prefixes in the NLRI, where the prefixes come from the provider namespace and are contextualized into separate TRDB, following mechanisms similar to RFC 4364 procedures.

It is worth noting that AFI/SAFI 1/128 has been used to carry transport prefixes in "L3VPN Inter-AS Carrier's carrier" scenario as defined in Section 10 of [RFC4364], where BGP LU/LDP prefixes in CsC VRF are advertised in AFI/SAFI 1/128 towards the remote-end client carrier.

In this document, SAFI 76 (BGP CT) is used instead of reusing SAFI 128 (L3VPN) for AFIs 1 or 2 to carry these transport routes because it is operationally advantageous to segregate transport and service prefixes into separate address families. For example, such an approach allows operators to safely enable "per-prefix" label allocation scheme for Classful Transport prefixes, typically with a space complexity of O(1K) to O(100K), without affecting SAFI 128 service prefixes with a space complexity of O(1M). The "per prefix" label allocation scheme localizes routing churn during topology changes.

Service routes continue to be carried in their existing AFI/SAFIs without any change. For example, L3VPN (AFI/SAFI: 1/128 and 2/128), EVPN (AFI/SAFI: 25/70), VPLS (AFI/SAFI: 25/65), Internet (AFI/SAFI: 1/1 or 2/1). These service routes can resolve over BGP CT (AFI/SAFI: 1/76 or 2/76) transport routes.

A new SAFI 76 for AFI 1 and AFI 2 also facilitates having a different readvertisement path of the transport family routes in a network than the service route readvertisement path. Service routes (Inet-VPN SAFI 128) are exchanged over an EBGP multihop session between ASes with next hop unchanged; whereas Classful Transport routes (SAFI 76) are readvertised over EBGP single hop sessions with "next hop self" rewrite over inter-AS links.

The BGP CT SAFI 76 for AFI 1 and 2 is similar in vein to BGP LU SAFI 4, in that it carries transport prefixes. The only difference is that it also carries in a Route Target an indication of which Transport Class the transport prefix belongs to, and uses the RD to disambiguate multiple instances of the same transport prefix in a BGP Update.

7. Protocol Procedures

This section summarizes the procedures followed by various nodes speaking Classful Transport family.

7.1. Preparing the network to deploy Classful Transport planes

It is responsibility of the operators to decide the Transport Classes to enable and use in their network. They are also expected to allocate a Transport Class Route Target to identify each Transport Class.

Operators configure the Transport Classes on the SNs and BNs in the network with Transport Class Route Targets and appropriate Route-Distinguishers.

Implementations MAY provide automatic generation and assignment of RD, RT values. They MAY also provide a way to manually override the automatic mechanism in order to deal with any conflicts that may arise with existing RD, RT values in different network domains participating in the deployment.

7.2. Originating Classful Transport Routes

BGP CT routes are sent only to BGP peers that have negotiated the Multiprotocol Extensions capability described in Section 8 of [RFC4760] to be able to send and receive BGP CT routes.

At the ingress node of the tunnel's home domain, the tunneling protocols install tunnel routes in the TRDB associated with the Transport Class to which the tunnel belongs.

The egress node of the tunnel, i.e. the tunnel endpoint (EP), originates the BGP CT route with RD:EP in the NLRI, Transport Class RT and PNH as EP. This BGP CT route will be resolved over the tunnel route in TRDB at the ingress node. When the tunnel is up, the Classful Transport BGP route will become usable and get re-advertised by the ingress node to BGP peers in neighboring domains.

Alternatively, the ingress node of the tunnel, which is also an ASBR/ABR in tunnel's home domain, may originate the BGP CT route for the tunnel destination with NLRI RD:EP, attaching a Transport Class Route Target that identifies the Transport Class. This BGP CT route is advertised to EBGP peers and IBGP peers in neighboring domains.

This originated route SHOULD NOT be advertised to the IBGP core that contains the tunnel. This may be implemented by mechanisms such as policy configuration. The impact of not prohibiting such advertisements is outside the scope of this document.

Unique RD SHOULD be used by the originator of a Classful Transport route to disambiguate the multiple BGP advertisements for a transport endpoint. An administrator may use duplicate RDs based on local choice, understanding the impact on path diversity and troubleshooting, as described in Section 10.2.

7.3. Processing Classful Transport Routes by Ingress Nodes

Upon receipt of a BGP CT route with a PNH EP that is not directly connected (e.g. an IBGP-route), a Mapping Community (the

Transport Class RT) on the route is used to decide to which resolution scheme this route is to be mapped.

The resolution scheme for a Transport Class RT with Transport Class ID "C1" contains the TRDB of a Transport Class with same ID. The administrator MAY customize the resolution scheme for Transport Class "C1" to map to a different ordered list of TRDBs. If the resolution scheme for TC ID "C1" is not found, the resolution scheme containing the "Best Effort" transport class TRDB is used.

The routes in the TRDBs associated with selected resolution scheme are used to resolve the received PNH EP. The order of TRDBs in the resolution scheme is followed when resolving the received PNH, such that a route in a backup TRDB is used only when a matching route was not found for EP in the primary TRDBs preceding it. This achieves the fallback desired by the resolution scheme.

If the resolution process does not find a matching route for EP in any of the associated TRDBs, the received BGP CT route MUST be considered unresolvable. (See RFC 4271, Section 9.1.2.1).

The received BGP CT route MUST be added to the TRDB corresponding to the Transport Class "C1", if the transport class is provisioned locally. This step applies only if the Transport Class RT is received on a BGP CT family route. The RD in the BGP CT NLRI prefix RD:EP is ignored when the BGP CT route for EP is added to the TRDB, so that overlay routes can resolve over this BGP CT tunnel route by performing a lookup for EP. Please note that a TRDB is a logical database of tunnel routes belonging to the same Transport Class ID, hence it uses only the EP as the lookup key without RD or TC-ID.

If no Mapping Community was found on a BGP CT route, the best effort resolution scheme is used for resolving the route's next hop, and the BGP CT route is not added to any TRDB.

7.4. Readvertising Classful Transport Route by Border Nodes

This section describes the MPLS label handling when readvertising a BGP CT route with Next Hop set to Self. When readvertising a BGP CT route with Next Hop set to Self, a BN allocates an MPLS label to advertise upstream in Classful Transport NLRI. The BN also installs an MPLS route for that label that swaps the incoming label with the label received from the downstream BGP speaker (or pops the incoming label if the label received from the downstream BGP speaker was Implicit-NULL). It then pushes received traffic to the transport tunnel or direct interface that the Classful Transport route's PNH resolved over.

The label SHOULD be allocated with "per-prefix" label allocation semantics. The IP prefix in the TRDB context (Transport-Class, IP-prefix) is used as the key to do per-prefix label allocation. This helps in avoiding BGP CT route churn throughout the CT network when an instability (e.g., link failure) is experienced in a domain. The failure is not propagated further than the BN closest to the failure. If a different label allocation mode is used, the impact on end to end convergence should be considered.

The value of the advertised MPLS label is locally significant, and is dynamic by default. A BN may provide an option to allocate a value from a statically provisioned range. This can be achieved using locally configured export policy, or via mechanisms such as the ones described in BGP Prefix-SID [RFC8669].

7.5. Border Nodes Receiving Classful Transport Routes on EBGP

If a route is received with a PNH that is known to be directly connected (for example, EBGP single-hop neighbor address), the directly connected interface is checked for MPLS forwarding capability. No other next hop resolution process is performed since the inter-AS link can be used for any Transport Class.

If the inter-AS links need to honor Transport Class, then the BN MUST follow procedures of an Ingress node (Section 7.3) and perform the next hop resolution process. In order to make the link Transport Class aware, the route to directly connected PNH is installed in the TRDB belonging to the associated Transport Class.

7.6. Avoiding Path Hiding Through Route Reflectors

When multiple instances of a given RD:EP exist with different forwarding characteristics, then BGP ADD-PATH [RFC7911] is helpful.

When multiple BNs exist such that they advertise a "RD:EP" prefix to Route Reflectors (RRs), the RRs may hide all but one of the BNs, unless BGP ADD-PATH [RFC7911] is used for the Classful Transport family. This is similar to L3VPN Option B scenarios.

Hence, BGP ADD-PATH [RFC7911] SHOULD be used for Classful Transport family, to avoid path-hiding through RRs so that the RR sends multiple CT routes for RD:EP to its clients. This improves the convergence time when the path via one of the multiple BNs fails.

7.7. Avoiding Loops Between Route Reflectors in Forwarding Path

A pair of redundant ABRs, each acting as an RR with next hop self, may choose each other as best path instead of the upstream ASBR, causing a traffic forwarding loop.

This problem can happen for routes of any BGP address family, including BGP CT and BGP LU.

Using one or more of the approaches described in [BGP-FWD-RR] softens the possibility of such loops in a network with redundant ABRs.

7.8. Ingress Nodes Receiving Service Routes with a Mapping Community

Upon receipt of a BGP service route (for example, AFI/SAFI: 1/1, 2/1) with a PNH as EP that is not directly connected (for example, an IBGP-route), a Mapping Community (for example, Color Extended Community) on the route is used to decide to which resolution scheme this route is to be mapped.

The resolution scheme for a Color Extended Community with Color "C1" contains a TRDB for a Transport Class with same ID, followed by the Best Effort TRDB. The administrator MAY customize the resolution scheme to map to a different ordered list of TRDBs. If the resolution scheme for TC ID "C1" is not found, the resolution scheme containing the "Best Effort" transport class TRDB is used.

If no Mapping Community was found on the overlay route, the "Best Effort" resolution scheme is used for resolving the route's next hop. This behavior is backward compatible to behavior of an implementation that does not follow procedures described in this document.

The routes in the TRDBs associated with selected resolution scheme are used to resolve the received PNH EP. The order of TRDBs in a resolution scheme is followed when resolving the received PNH, such that a route in a backup TRDB is used only when a matching route was not found for EP in the primary TRDBs preceding it. This achieves the fallback desired by the resolution scheme.

If the resolution process does not find a Tunnel Route for EP in any of the Transport Route Databases, the service route MUST be considered unresolvable (See RFC 4271, Section 9.1.2.1).

Note: For an illustration of above procedures in a MPLS network, refer to Section 8.

7.9. Best Effort Transport Class

It is possible to represent 'Best effort' SLA also as a Transport Class. Today, BGP LU is used to extend the best effort intra domain tunnels to other domains.

Alternatively, BGP CT may also be used to carry the best effort tunnels. This document reserves the Transport Class ID value 0 to represent "Best Effort Transport Class ID". However, implementations SHOULD provide configuration to use a different value for this purpose. Procedures to manage differences in Transport Class ID namespaces between domains are provided in Section 11.2.2.

The "Best Effort Transport Class ID" value is used in the "Transport Class ID" field of Transport Route Target Extended Community that is attached to the BGP CT route that advertises a best effort tunnel endpoint. The RT thus formed is called the "Best Effort Transport Class Route Target".

When a BN or SN receives a BGP CT route with Best Effort Transport Class Route Target as the mapping community, the Best effort resolution scheme is used for resolving the BGP next hop, and the resultant route is installed in the best effort transport route database. If no best effort tunnel was found to resolve the BGP next hop, the BGP CT route MUST be considered unusable, and not be propagated further.

When a BGP speaker receives an overlay route without any explicit Mapping Community, and absent local policy, the best effort resolution scheme is used for resolving the BGP next hop on the route. This behavior is backward compatible to behavior of an implementation that does not follow procedures described in this document.

Implementations MAY provide configuration to selectively install BGP CT routes to the Forwarding Information Base (FIB), to provide reachability for control plane peering towards endpoints in other domains.

7.10. Interaction with BGP Attributes Specifying Next Hop Address and Color

The Tunnel Encapsulation Attribute, described in [RFC9012] can be used to request a specific type of tunnel encapsulation. This attribute may apply to BGP service routes or transport routes, including BGP Classful Transport family routes.

It should be noted that in such cases "Transport Class ID/Color" can exist in multiple places on the same route, and a precedence order needs to be established to determine which Transport Class the route's next hop should resolve over. This document suggests the following order of precedence, more specific scoping of Color preferred to less specific scoping:

Color SubTLV, in Tunnel Encapsulation Attribute.

Transport Target Extended community, on BGP CT route.

Color Extended community, on BGP service route.

Color specified in the Color subTLV in a TEA is a more specific indication of "Transport Class ID/Color" than Mapping Community (Transport Target) on a BGP CT transport route, which is in turn is more specific than a Service route scoped Mapping Community (Color Extended community).

Any BGP attributes or mechanisms defined in future that carry Transport Class ID/Color on the route are expected to specify the order of precedence relative to the above.

7.11. Applicability to Flowspec Redirect to IP

Flowspec routes using Redirect to IP next hop is described in [FLOWSPEC-REDIR-IP]

Such Flowspec BGP routes with Redirect to IP next hop MAY be attached with a Mapping Community (e.g. Color:0:100), which allows redirecting the flow traffic over a tunnel to the IP next hop satisfying the desired SLA (e.g. Transport Class color 100).

Flowspec BGP family acts as just another service that can make use of BGP CT architecture to achieve Flow based forwarding with SLAs.

7.12. Applicability to IPv6

This section describes applicability of BGP CT to IPv6 at various layers. BGP CT procedures apply equally to an IPv6 enabled Intra-AS or Inter-AS Option A, B, C network.

A BGP CT enabled network supports IPv6 service families (for example, AFI/SAFI 2/1 or 2/128) and IPv6 transport signaling protocols like SRTEv6, LDPv6, RSVP-TEv6.

Procedures in this document also apply to a network with Pure IPv6 core, that uses MPLS forwarding for intra-domain tunnels and inter-AS links. BGP CTv6 family (AFI/SAFI: 2/76) is used to carry global IPv6 address tunnel endpoints in the NLRI. Service family routes (for example, AFI/SAFI: 1/1, 2/1, 1/128, 2/128) are also advertised with those Global IPv6 addresses as next hop.

Procedures in this document also apply to a 6PE network with an IPv4 core, that uses MPLS forwarding for intra-domain tunnels and Inter-AS links. BGP CTv6 family (AFI/SAFI: 2/76) is used to carry IPv4 Mapped IPv6 address tunnel endpoints in the NLRI. IPv6 Service family routes (for example, AFI/SAFI: 2/1, 2/128) are also advertised with those IPv4 Mapped IPv6 addresses as next hop.

The PE-CE attachment circuits may use IPv4 addresses only, IPv6 addresses only, or both IPv4 and IPv6 addresses.

7.13. SRv6 Support

This section describes how BGP CT family (AFI/SAFI 2/76) may be used to set up inter-domain tunnels of a certain Transport Class, when using Segment Routing over IPv6 (SRv6) data plane on the inter-AS links or as an intra-AS tunneling mechanism.

Details of SRv6 Endpoint behaviors used by BGP CT and the procedures are specified in a separate document [BGP-CT-SRv6], along with illustration. As noted in this document, BGP CT route update for SRv6 includes a BGP attribute containing SRv6 SID information (e.g. Prefix SID [RFC9252]) with Transposition scheme disabled.

7.14. Error Handling Considerations

If a BGP speaker receives both Transitive (Section 13.2.1.1.1) and Non-Transitive (Section 13.2.1.1.2) versions of Transport Class extended community on a route, only the Transitive one is used.

If a BGP speaker considers a received "Transport Class" extended community (Transitive or Non-Transitive version), or any other part of a BGP CT route invalid for some reason, but is able to successfully parse the NLRI and attributes, Treat-as-withdraw approach from [RFC7606] is used. The route is kept as Unusable, with appropriate diagnostic information, to aid troubleshooting.

8. Illustration of BGP CT Procedures

This section illustrates BGP CT procedures in an Inter AS Option C MPLS network.

All Illustrations in this document make use of [RFC6890] IP address ranges. The range 192.0.2.0/24 is used to represent transport endpoints like loopback addresses. The range 203.0.113.0/24 is used to represent service route prefixes advertised in AFI/SAFIs: 1/1 or 1/128.

8.1. Reference Topology

```
                [RR26]      [RR27]                       [RR16]
                 |            |                             |
                 |            |                             |
                 |+-[ABR23]--+|+--[ASBR21]---[ASBR13]-+|+--[PE11]--+
                 ||          |||          `  /        |||          |
[CE41]--[PE25]--[P28]       [P29]          `/        [P15]     [CE31]
                 |           | |           /`         | |          |
                 |           | |          /  `        | |          |
                 |           | |         /    `       | |          |
                 +--[ABR24]--+ +--[ASBR22]---[ASBR14]-+ +--[PE12]--+


       |      AS2         |         AS2       |                   |
   AS4 +    region-1    +      region-2    +       AS1         + AS3
       |                |                  |                   |


203.0.113.41  ------------ Traffic Direction ---------->  203.0.113.31
```

Figure 3: Multi-Domain BGP CT Network

This example shows a provider MPLS network that consists of two ASes, AS1 and AS2. They are serving customers AS3, AS4 respectively. Traffic direction being described is CE41 to CE31. CE31 may request a specific SLA (for example, mapped to Gold for this example), when traversing these provider networks.

AS2 is further divided into two regions. There are three tunnel domains in provider's space: AS1 uses ISIS Flex-Algo [RFC9350] intra-domain tunnels. AS2 uses RSVP-TE intra-domain tunnels. MPLS forwarding is used within these domains and on inter-domain links.

The network exposes two Transport Classes: "Gold" with Transport Class ID 100, "Bronze" with Transport Class ID 200. These Transport Classes are provisioned at the PEs and the Border nodes (ABRs, ASBRs) in the network.

The following tunnels exist for Gold Transport Class.

PE25_to_ABR23_gold - RSVP-TE tunnel

PE25_to_ABR24_gold - RSVP-TE tunnel

ABR23_to_ASBR22_gold - RSVP-TE tunnel

ASBR13_to_PE11_gold - SRTE tunnel

ASBR14_to_PE11_gold - SRTE tunnel

The following tunnels exist for Bronze Transport Class.

PE25_to_ABR23_bronze - RSVP-TE tunnel

ABR23_to_ASBR21_bronze - RSVP-TE tunnel

ABR23_to_ASBR22_bronze - RSVP-TE tunnel

ABR24_to_ASBR21_bronze - RSVP-TE tunnel

ASBR13_to_PE12_bronze - ISIS FlexAlgo tunnel

ASBR14_to_PE11_bronze - ISIS FlexAlgo tunnel

These tunnels are either provisioned or auto-discovered to belong to Transport Classes 100 or 200.

8.2. Service Layer Route Exchange

Service nodes PE11, PE12 negotiate service families (AFI: 1 and SAFIs 1, 128) on the BGP session with RR16. Service helpers RR16 and RR26 exchange these service routes with next hop unchanged over a multihop EBGP session between the two AS. PE25 negotiates service families (AFI: 1 and SAFIs 1, 128) with RR26.

The PEs see each other as next hop in the BGP Update for the service family routes. BGP ADD-PATH send and receive is enabled on both directions on the EBGP multihop session between RR16 and RR26 for AFI:1 and SAFIs 1, 128. BGP ADD-PATH send is negotiated in the RR to PE direction in each AS. This is to avoid path hiding of service routes at RR; i.e., AFI/SAFI 1/1 routes advertised by both PE11 and PE12. Or, AFI/SAFI 1/128 routes originated by both PE11 and PE12 using same RD.

Forwarding happens using service routes installed at service nodes PE25, PE11, PE12 only. Service routes received from CEs are not present in any other nodes' FIB in the network.

As an example, CE31 advertises a route for prefix 203.0.113.31 with next hop as self to PE11, PE12. CE31 can attach a Mapping Community Color:0:100 on this route, to indicate its request for Gold SLA. Or, PE11 can attach the same using locally configured policies.

Consider, CE31 is getting VPN service from PE11. The RD1:203.0.113.31 route is readvertised in AFI/SAFI 1/128 by PE11 with next hop self (192.0.2.11) and label V-L1, to RR16 with the Mapping Community Color:0:100 attached. RR16 advertises this route with BGP ADD-PATH ID to RR26 which readvertises to PE25 with next hop unchanged. Now, PE25 can resolve the PNH 192.0.2.11 using transport routes received in BGP CT or BGP LU.

Using BGP ADD-PATH, service routes advertised by PE11 and PE12 for AFI:1 SAFIs 1, 128 reach PE25 via RR16, RR26 with the next hop unchanged, as PE11 or PE12.

The IP FIB at PE25 VRF will have a route for 203.0.113.31 with a next hop when resolved, that points to a Gold tunnel in ingress domain.

8.3. Transport Layer Route Propagation

Egress nodes PE11, PE12 negotiate BGP CT family with transport ASBRs ASBR13, ASBR14. These egress nodes originate BGP CT routes for tunnel endpoint addresses, that are advertised as next hop in BGP service routes. In this example, both PEs participate in transport classes Gold and Bronze. The protocol procedures are explained using the Gold SLA transport plane and the Bronze SLA transport plane is used to highlight the path hiding aspects.

PE11 is provisioned with transport class 100, RD value 192.0.2.11:100 and a transport-target:0:100 for Gold tunnels. And a Transport class 200 with RD value 192.0.2.11:200, and transport route target 0:200 for Bronze tunnels. Similarly, PE12 is provisioned with transport class 100, RD value 192.0.2.12:100 and a transport-target:0:100 for Gold tunnels. And transport class 200, RD value 192.0.2.12:200 with transport-target:0:200 for Bronze tunnels. Note that in this example, the BGP CT routes carry only the transport class route target, and no IP address format route target.

The RD value originated by an egress node is not modified by any BGP speakers when the route is readvertised to the ingress node. Thus, the RD can be used to identify the originator (unique RD provisioned) or set of originators (RD reused on multiple nodes).

Similarly, these transport classes are also configured on ASBRs, ABRs and PEs with same Transport Route Target and unique RDs.

ASBR13 and ASBR14 negotiate BGP CT family with transport ASBRs ASBR21, ASBR22 in neighboring AS. They negotiate BGP CT family with RR27 in region 2, which reflects BGP CT routes to ABR23, ABR24. ABR23, ABR24 negotiate BGP CT family with Ingress node PE25 in region 1. BGP LU family is also negotiated on these sessions alongside BGP CT family. BGP LU carries "best effort" transport class routes, BGP CT carries Gold, Bronze transport class routes.

PE11 is provisioned to originate BGP CT route with Gold SLA to endpoint PE11. This route is sent with NLRI RD prefix 192.0.2.11:100:192.0.2.11, Label B-L0, next hop 192.0.2.11 and a route target extended community transport-target:0:100. Label B-L0 can either be Implicit Null (Label 3) or an Ultimate Hop Pop (UHP) label.

This route is received by ASBR13 and it resolves over the tunnel ASBR13_to_PE11_gold. The route is then readvertised by ASBR13 in BGP CT family to ASBRs ASBR21, ASBR22 according to export policy. This route is sent with same NLRI RD prefix 192.0.2.11:100:192.0.2.11, Label B-L1, next hop self, and transport-target:0:100. MPLS swap route is installed at ASBR13 for B-L1 with a next hop pointing to ASBR13_to_PE11_gold tunnel.

Similarly, ASBR14 also receives a BGP CT route for 192.0.2.11:100:192.0.2.11 from PE11 and it resolves over the tunnel ASBR14_to_PE11_gold. The route is then readvertised by ASBR14 in BGP CT family to ASBRs ASBR21, ASBR22 according to export policy. This route is sent with the same NLRI RD prefix 192.0.2.11:100:192.0.2.11, Label B-L2, next hop self, and transport-target:0:100. MPLS swap route is installed at ASBR14 for B-L1 with a next hop pointing to ASBR14_to_PE11_gold tunnel.

In the Bronze plane, BGP CT route with Bronze SLA to endpoint PE11 is originated by PE11 with a NLRI containing RD prefix 192.0.2.11:200:192.0.2.11, and appropriate label. The RD allows both Gold and Bronze advertisements to traverse path selection pinchpoints without any path hiding at RRs or ASBRs. And route target extended community transport-target:0:200 lets the route resolve over Bronze tunnels in the network, similar to the process being described for Gold SLA path.

Moving back to the Gold plane, ASBR21 receives the Gold SLA BGP CT routes for NLRI RD prefix 192.0.2.11:100:192.0.2.11 over the single hop EBGP sessions from ASBR13, ASBR14, and can compute ECMP/FRR towards them. ASBR21 readvertises BGP CT route for 192.0.2.11:100:192.0.2.11 with next hop self (loopback address 192.0.2.21) to RR27, advertising a new label B-L3. An MPLS swap route is installed for label B-L3 at ASBR21 to swap to received label B-L1, B-L2 and forward to ASBR13, ASBR14 respectively. RR27 readvertises this BGP CT route to ABR23, ABR24 with label and next hop unchanged.

Similarly, ASBR22 receives BGP CT route 192.0.2.11:100:192.0.2.11 over the single hop EBGP sessions from ASBR13, ASBR14, and readvertises with next hop self (loopback address 192.0.2.22) to RR27, advertising a new label B-L4. An MPLS swap route is installed for label B-L4 at ASBR22 to swap to received label B-L1, B-L2 and forward to ASBR13, ASBR14 respectively. RR27 readvertises this BGP CT route also to ABR23, ABR24 with label and next hop unchanged.

BGP ADD-PATH is enabled for BGP CT family on the sessions between RR27 and ASBRs, ABRs such that routes for 192.0.2.11:100:192.0.2.11 with the next hops ASBR21 and ASBR22 are reflected to ABR23, ABR24 without any path hiding. Thus, ABR23 is given visibility of both available next hops for Gold SLA.

ABR23 receives the route with next hop 192.0.2.21, label B-L3 from RR27. The route target "transport-target:0:100" on this route acts as Mapping Community, and instructs ABR23 to strictly resolve the next hop using transport class 100 routes only. ABR23 is unable to find a route for 192.0.2.21 with transport class 100. Thus, it considers this route unusable and does not propagate it further. This prunes ASBR21 from Gold SLA tunneled path.

ABR23 also receives the route with next hop 192.0.2.22, label B-L4 from RR27. The route target "transport-target:0:100" on this route acts as Mapping Community, and instructs ABR23 to strictly resolve the next hop using transport class 100 routes only. ABR23 successfully resolves the next hop to point to ABR23_to_ASBR22_gold tunnel. ABR23 readvertises this BGP CT route with next hop self (loopback address 192.0.2.23) and a new label B-L5 to PE25. Swap route for B-L5 is installed by ABR23 to swap to label B-L4, and forward into ABR23_to_ASBR22_gold tunnel.

PE25 receives the BGP CT route for prefix 192.0.2.11:100:192.0.2.11 with label B-L5, next hop 192.0.2.23 and transport-target:0:100 from RR26. And it similarly resolves the next hop 192.0.2.23 over transport class 100, pushing labels associated with PE25_to_ABR23_gold tunnel.

In this manner, the Gold transport LSP "ASBR13_to_PE11_gold" in the egress domain is extended by BGP CT until the ingress node PE25 in the ingress domain, to create an end-to-end Gold SLA path. MPLS swap routes are installed at ASBR13, ASBR22 and ABR23, when propagating the PE11 BGP CT Gold transport class route 192.0.2.11:100:192.0.2.11 with next hop self towards PE25.

The BGP CT LSP thus formed, originates in PE25, and terminates in ASBR13 (assuming PE11 advertised Implicit Null), traversing over the Gold underlay LSPs in each domain. ASBR13 uses UHP to stitch the BGP CT LSP into the "ASBR13_to_PE11_gold" LSP to traverse the last domain, thus satisfying Gold SLA end-to-end.

When PE25 receives service routes from RR26 with next hop 192.0.2.11 and mapping community Color:0:100, it resolves over this BGP CT route 192.0.2.11:100:192.0.2.11. Thus, pushing label B-L5, and pushing as top label the labels associated with PE25_to_ABR23_gold tunnel.

8.4. Data Plane View

8.4.1. Steady State

This section describes how the data plane looks in steady state.

CE41 transmits an IP packet with destination as 203.0.113.31. On receiving this packet, PE25 performs a lookup in the IP FIB associated with the CE41 interface. This lookup yields the service route that pushes the VPN service label V-L1, BGP CT label B-L5, and labels for PE25_to_ABR23_gold tunnel. Thus, PE25 encapsulates the IP packet in an MPLS packet with label V-L1 (innermost), B-L5, and top label as PE25_to_ABR23_gold tunnel. This MPLS packet is thus transmitted to ABR23 using Gold SLA.

ABR23 decapsulates the packet received on PE25_to_ABR23_gold tunnel as required, and finds the MPLS packet with label B-L5. It performs a lookup for label B-L5 in the global MPLS FIB. This yields the route that swaps label B-L5 with label B-L4, and pushes the top label provided by ABR23_to_ASBR22_gold tunnel. Thus, ABR23 transmits the MPLS packet with label B-L4 to ASBR22, on a tunnel that satisfies Gold SLA.

ASBR22 similarly performs a lookup for label B-L4 in global MPLS FIB, finds the route that swaps label B-L4 with label B-L2, and forwards to ASBR13 over the directly connected MPLS-enabled interface. This interface is a common resource not dedicated to any specific transport class, in this example.

ASBR13 receives the MPLS packet with label B-L2, and performs a lookup in MPLS FIB, finds the route that pops label B-L2, and pushes labels associated with ASBR13_to_PE11_gold tunnel. This transmits the MPLS packet with VPN label V-L1 to PE11 using a tunnel that preserves Gold SLA in AS 1.

PE11 receives the MPLS packet with V-L1, and performs VPN forwarding. Thus transmitting the original IP payload from CE41 to CE31. The payload has traversed path satisfying Gold SLA end-to-end.

8.4.2. Local Repair of Primary Path

This section describes how the data plane at ASBR22 reacts when the link between ASBR22 and ASBR13 experiences a failure, and an alternate path exists.

Assuming ASBR22_to_ASBR13 link goes down, such that traffic with Gold SLA going to PE11 needs repair. ASBR22 has an alternate BGP CT route for 192.0.2.11:100:192.0.2.11 from ASBR14. This has been preprogrammed in forwarding by ASBR22 as FRR backup next hop for label B-L4. This allows the Gold SLA traffic to be locally repaired at ASBR22 without the failure event propagated in the BGP CT network. In this case, ingress node PE25 will not know there was a failure, and traffic restoration will be independent of prefix scale (PIC).

8.4.3. Absorbing Failure of Primary Path: Fallback to Best Effort Tunnels

This section describes how the data plane reacts when a Gold path experiences a failure, but no alternate path exists.

Assume tunnel ABR23_to_ASBR22_gold goes down, such that now no end-to-end Gold path exists in the network. This makes the BGP CT route for RD prefix 192.0.2.11:100:192.0.2.11 is unusable at ABR23. This makes ABR23 send a BGP withdrawal for 192.0.2.11:100:192.0.2.11 to PE25.

The withdrawal for 192.0.2.11:100:192.0.2.11 allows PE25 to react to the loss of the Gold path to 192.0.2.11. Assuming PE25 is provisioned to use best effort transport class as the backup path, this withdrawal of BGP CT route allows PE25 to adjust the next hop of the VPN Service-route to push the labels provided by the BGP LU route. That repairs the traffic to go via the best effort path. PE25 can also be provisioned to use Bronze transport class as the backup path. The repair will happen in similar manner in that case as-well.

Traffic repair to absorb the failure happens at ingress node PE25, in a service prefix scale independent manner. This is called PIC (Prefix scale Independent Convergence). The repair time will be proportional to time taken for withdrawing the BGP CT route.

These examples demonstrate the various levels of failsafe mechanisms available to protect traffic in a BGP CT network.

9. Scaling Considerations

9.1. Avoiding Unintended Spread of BGP CT Routes Across Domains

[RFC8212] suggests BGP speakers require explicit configuration of both BGP Import and Export Policies in order to receive or send routes over EBGP sessions.

It is recommended to follow this for BGP CT routes. It will prohibit unintended advertisement of transport routes throughout the BGP CT transport domain, which may span across multiple AS domains. This will conserve usage of MPLS label and next hop resources in the network. An ASBR of a domain can be provisioned to allow routes with only the Transport Route Targets that are required by SNs in the domain.

9.2. Constrained Distribution of PNHs to SNs (On-Demand Next Hop)

This section describes how the number of Protocol Next hops advertised to a SN or BN can be constrained using BGP Classful Transport and Route Target Constrain (RTC) [RFC4684].

An egress SN MAY advertise a BGP CT route for RD:eSN with two Route Targets: transport-target:0:<TC> and a RT carrying <eSN>:<TC>, where TC is the Transport Class identifier, and eSN is the IP address used by SN as BGP next hop in its service route advertisements.

Note that such use of the IP address specific route target <eSN>:<TC> is optional in a BGP CT network. It is required only if there is a requirement to prune the propagation of the transport route for an egress node eSN to only the set of ingress nodes that need it. When only RT of transport-target:0:<TC> is used, the pruning happens in granularity of Transport Class ID (Color), and not BGP next hop; BGP CT routes will not be advertised into domains with PEs that don't import its transport class.

The transport-target:0:<TC> is the new type of route target (Transport Class RT) defined in this document. It is carried in BGP extended community attribute (BGP attribute code 16).

The RT carrying <eSN>:<TC> MAY be an IP-address specific regular RT (BGP attribute code 16), or IPv6-address specific RT (BGP attribute code 25). It should be noted that the Local Administrator field of these RTs can only carry two octets of information, and thus the <TC> field in this approach is limited to a 2 octets value. Future protocol extensions work is needed to define a BGP CCA that can accomodate an IPv4/IPv6 address along with a 4 octet Local Administrator field.

An ingress SN MAY import BGP CT routes with Route Target carrying <eSN>:<TC>. The ingress SN may learn the eSN values either by configuration, or it may discover them from the BGP next hop field in the BGP VPN service routes received from eSN. A BGP ingress SN receiving a BGP service route with next hop of eSN generates a RTC/Extended-RTC route for Route Target prefix <Origin ASN>:<eSN>/[80|176] in order to learn BGP CT transport routes to reach eSN. This allows constrained distribution of the transport routes to the PNHs actually required by iSN.

When the path of route propagation of BGP CT routes is the same as the RTC routes, a BN would learn the RTC routes advertised by ingress SNs and propagate further. This will allow constraining distribution of BGP CT routes for a PNH to only the necessary BNs in the network, closer to the egress SN.

This mechanism provides "On Demand Next hop" of BGP CT routes, which help with the scaling of MPLS forwarding state at SN and BN.

However, the amount of state carried in RTC family may become proportional to the number of PNHs in the network. To strike a balance, the RTC route advertisements for <Origin ASN>:<eSN>/[80|176] MAY be confined to the BNs in the home region of an ingress SN, or the BNs of a super core.

Such a BN in the core of the network imports BGP CT routes with Transport-Target:0:<TC> and generates an RTC route for <Origin ASN>:0:<TC>/96, while not propagating the more specific RTC requests for specific PNHs. This lets the BN learn transport routes to all eSN nodes but confine their propagation to ingress SNs.

9.3. Limiting The Visibility Scope of PE Loopback as PNHs

It may be even more desirable to limit the number of PNHs that are globally visible in the network. This is possible using mechanism described in Appendix D.

Such that advertisement of PE loopback addresses as next-hop in BGP service routes is confined to the region they belong to. An anycast IP-address called "Context Protocol Nexthop Address" (CPNH) abstracts the SNs in a region from other regions in the network.

This provides much greater advantage in terms of scaling, convergence and security. Changes to implement this feature are required only on the local region's BNs and RRs, so legacy PE devices can also benefit from this approach.

10. Operations and Manageability Considerations

10.1. MPLS OAM

MPLS OAM procedures specified in [RFC8029] also apply to BGP Classful Transport.

The 'Target FEC Stack' sub-TLV for IPv4 Classful Transport has a Sub-Type of 31744, and a length of 13. The Value field consists of the RD advertised with the Classful Transport prefix, the IPv4 prefix (with trailing 0 bits to make 32 bits in all) and a prefix length encoded as shown in Figure 4.

```
  0                   1                   2                   3
  0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |                      Route Distinguisher                      |
 |                          (8 octets)                           |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |                         IPv4 prefix                           |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 | Prefix Length |                 Must Be Zero                  |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 4: Classful Transport IPv4 FEC

The 'Target FEC Stack' sub-TLV for IPv6 Classful Transport has a Sub-Type of 31745, and a length of 25. The Value field consists of the RD advertised with the Classful Transport prefix, the IPv6 prefix (with trailing 0 bits to make 128 bits in all) and a prefix length encoded as shown in Figure 5.

```
  0                   1                   2                   3
  0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |                      Route Distinguisher                      |
 |                          (8 octets)                           |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |                         IPv6 prefix                           |
 |                                                               |
 |                                                               |
 |                                                               |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 | Prefix Length |                 Must Be Zero                  |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 5: Classful Transport IPv6 FEC

These prefix layouts are inherited from Sections 3.2.5, 3.2.6 in [RFC8029]

10.2. Usage of Route Distinguisher and Label Allocation Modes

RDs aid in troubleshooting provider networks that deploy BGP CT, by uniquely identifying the originator of a route across an administrative domain that may either span multiple domains within a provider network or span closely coordinated provider networks.

The use of RDs also provides an option for signaling forwarding diversity within the same Transport Class. A SN can advertise an EP with the same Transport Class in multiple BGP CT routes with unique RDs.

For example, unique "RDx:EP1" prefixes can be advertised by an SN for an EP1 to different upstream BNs with unique forwarding specific encapsulation (e.g., Label), in order to collect traffic statistics at the SN for each BN. In absence of RD, duplicated Transport Class/ Color values will be needed in the transport network to achieve such use cases.

The allocation of RDs is done at the point of origin of the BGP CT route. This can either be an Egress SN or a BN. The default RD allocation mode is to use a unique RD per originating node for an EP. This mode allows for the ingress to uniquely identify each originated path. Alternatively, the same RD may be provisioned for multiple originators of the same EP. This mode can be used when the ingress does not require full visibility of all nodes originating an EP.

A label is allocated for a BGP CT route when it is advertised with next hop self by a SN or a BN. An implementation may use different label allocation modes with BGP CT. The recommended label allocation mode is per-prefix as it provides better traffic convergence properties than per-next hop label allocation mode. Furthermore, BGP CT offers two flavors for per-prefix label allocation. The first flavor assigns a label for each unique "RD, EP". The second flavor assigns a label for each unique "Transport Class, EP" while ignoring the RD.

In a BGP CT network, the number of routes at an Ingress PE is a function of unique EPs multiplied by BNs in the ingress domain that do next hop self. BGP CT provides flexible RD and Label allocation modes to address operational requirements in a multi-domain network. The impacts on the control plane and forwarding behavior for these modes are detailed with an example in Managing Transport Route Visibility (Section 10.3)

10.3. Managing Transport Route Visibility

This section details the usage of BGP CT RD and label allocation modes to calibrate the level of path visibility and the amount of route and label scale in a multi-domain network.

Consider a multi-domain BGP CT network as illustrated in the following Figure 6:

```
     |-----AS3-----|  |-------AS1------|

              +--------ASBR11     +--PE11 (EP1)
              |             \   /
       +----ASBR31            (P)----PE12 (EP2)
       |      |             / | \
       |      +--------ASBR12  |  +--PE13 (EP3)
       |                       |
       |                       +-----PE14 (EP4)
PE31--(P)
       |
       |
       |      +--------ASBR21     +--PE21 (EP5)
       |      |             \   /
       +----ASBR32            (P)----PE22 (EP6)
              |             / | \
              +--------ASBR22  |  +--PE22 (EP7)
                               |
                               +-----PE24 (EP8)

     |-----AS3-----|  |-------AS2------|
```

Figure 6: Managing Transport Route Visibility in Multi Domain Network

The following table provides a comparison of the BGP CT route and label scale, for varying endpoint path visibility at ingress node PE31 for each TC. It analyzes scenarios where Unicast or Anycast EPs (EP-type) may be originated by different node roles (Origin), using different RD allocation modes (RD-Mode), and different Per-Prefix Label allocation modes (PP-Mode).

EP-type	Origin	RD-Mode	PP-Mode	CT Routes	CT Labels
Unicast	SN	Unique	TC,EP	8	8
Unicast	SN	Unique	RD,EP	8	8
Unicast	BN	Unique	TC,EP	16	8
Unicast	BN	Unique	RD,EP	16	16
Anycast	SN	Unique	TC,EP	8	2
Anycast	SN	Unique	RD,EP	8	8
Anycast	SN	Same	TC,EP	2	2
Anycast	SN	Same	RD,EP	2	2
Anycast	BN	Unique	TC,EP	4	2
Anycast	BN	Unique	RD,EP	4	4
Anycast	BN	Same	TC,EP	2	2
Anycast	BN	Same	RD,EP	2	2

Figure 7: Route and Path Visibility at Ingress Node

In the table shown in Figure 7, route scale at ingress node PE31 is proportional to path diversity in ingress domain (number of ASBRs) and point of origination of BGP CT route. TE granularity at ingress node PE31 is proportional to the number of unique CT labels received, which depends on PP-mode and the path diversity in ingress domain.

Deploying unique RDs is strongly RECOMMENDED because it helps in troubleshooting by uniquely identifying the originator of a route and avoids path-hiding.

In typical deployments originating BGP CT routes at the egress node (SN) is recommended. In this model, using either "RD, EP" or "TC, EP" Per-Prefix label allocation mode repairs traffic locally at the nearest BN for any failures in the network, because the label value does not change.

Originating at BNs with unique RDs induces more routes than when originating at egress SNs. In this model, use of "TC, EP" Per-Prefix label allocation mode repairs traffic locally at the nearest BN for any failures in the network, because the label value does not change.

The previous table in Figure 7 demonstrates that BGP CT allows an operator to control how much path visibility and forwarding diversity is desired in the network, for both Unicast and Anycast endpoints.

11. Deployment Considerations.

11.1. Coordination Between Domains Using Different Community Namespaces

Cooperating Inter-AS Option C domains may sometimes not agree on RT, RD, Mapping community or Transport Route Target values because of differences in community namespaces (e.g. during network mergers or renumbering for expansion). Such deployments may deploy mechanisms to map and rewrite the Route Target values on domain boundaries, using per ASBR import policies. This is no different than any other BGP VPN family. Mechanisms used in inter-AS VPN deployments may be leveraged with the Classful Transport family also.

A resolution scheme allows association with multiple Mapping Communities. This minimizes service disruption during renumbering, network merger or transition scenarios.

The Transport Class Route Target Extended Community is useful to avoid collision with regular Route Target namespace used by service routes.

11.2. Managing Intent at Service and Transport layers.

Illustration of BGP CT Procedures (Section 8) shows multiple domains that agree on a color name space (Agreeing Color Domains) and contain tunnels with equivalent set of colors (Homogenous Color Domains).

However, in the real world, this may not always be guaranteed. Two domains may independently manage their color namespaces; these are known as Non-Agreeing Color Domains. Two domains may have tunnels with unequal sets of colors; these are known as Heterogeneous Color Domains.

This section describes how BGP CT is deployed in such scenarios to preserve end-to-end Intent. Examples described in this section use Inter-AS Option C domains. Similar mechanisms will work for Inter-AS Option A and Inter-AS Option B scenarios as well.

11.2.1. Service Layer Color Management

At the service layer, it is recommended that a global color namespace be maintained across multiple co-operating domains. BGP CT allows indirection using resolution schemes to be able to maintain a global namespace in the service layer. This is possible even if each domain independently maintains its own local transport color namespace.

As explained in Next Hop Resolution Scheme (Section 5) , a mapping community carried on a service route maps to a resolution scheme. The mapping community values for the service route can be abstract and are not required to match the transport color namespace. This abstract mapping community value representing a global service layer intent is mapped to a local transport layer intent available in each domain.

In this manner, it is recommended to keep color namespace management at the service layer and the transport layer decoupled from each other. In the following sections the service layer agrees on a single global namespace.

11.2.2. Non-Agreeing Color Transport Domains

Non-agreeing color domains require a mapping community rewrite on each domain boundary. This rewrite helps to map one domain's color namespace to another domain's color namespace.

The following example illustrates how traffic is stitched and SLA is preserved when domains don't use the same namespace at the transport layer. Each domain specifies the same SLA using different color values.

```
       Gold(100)              Gold(300)               Gold(500)


  [PE11]----[ASBR11]---[ASBR21------[ASBR22]---[ASBR31-------[PE31]
         AS1                     AS2                    AS3


       Bronze(200)          Bronze(400)             Bronze(600)

         ----------- Packet Forwarding Direction -------->
```

Figure 8: Transport Layer with Non-agreeing Color Domains

In the topology shown in Figure 8, we have three Autonomous Systems. All the nodes in the topology support BGP CT.

In AS1 Gold SLA is represented by color 100 and Bronze by 200.

In AS2 Gold SLA is represented by color 300 and Bronze by 400.

In AS3 Gold SLA is represented by color 500 and Bronze by 600.

Though the color values are different, they map to tunnels with sufficiently similar TE characteristics in each domain.

The service route carries an abstract mapping community that maps to the required SLA. For example, Service routes that need to resolve over Gold transport tunnels, carry a mapping community color: 0:100500. In AS3 it maps to a resolution scheme containing TRDB with color 500 whereas in AS2 it maps to a TRDB with color 300 and in AS1 it maps to a TRDB with color 100. Coordination is needed to provision the resolution schemes in each domain as explained previously.

At the AS boundary, the transport-class route-target is rewritten for the BGP CT routes. In the previous topology, at ASBR31, the transport-target:0:500 for Gold tunnels is rewritten to transport-target:0:300 and then advertised to ASBR22. Similarly, the transport-target:0:300 for Gold tunnels are re-written to transport-target:0:100 at ASBR21 before advertising to ASBR11. At PE11, the transport route received with transport-target:0:100 will be added to the color 100 TRDB. The service route received with mapping community color:0:100500 at PE1 maps to the Gold TRDB and resolves over this transport route.

Inter-domain traffic forwarding in the previous topology works as explained in Section 8.

Transport-target re-write requires co-ordination of color values between domains in the transport layer. This method avoids the need to re-write service route mapping communities, keeping the service layer homogenous and simple to manage. Coordinating Transport Class RT between two adjacent color domains at a time is easier than coordinating service layer colors deployed in a global mesh of non-adjacent color domains. This basically allows localizing the problem to a pair of adjacent color domains and solving it.

11.2.3. Heterogeneous Agreeing Color Transport Domains

In a heterogeneous domains scenario, it might not be possible to map a service layer intent to the matching transport color, as the color might not be locally available in a domain.

The following example illustrates how traffic is stitched, when a transit AS contains more shades for an SLA path compared to Ingress and Egress domains. This example shows how service routes can traverse through finer shades when available and take coarse shades otherwise.

```
        <---------- Service Routes AFI/SAFI 1/128 ---------------

                                Gold1(101)
                                Gold2(102)
      Gold(100)                                     Gold(100)

  [PE11]------[ASBR11]----[ASBR21------[ASBR22]----[ASBR31------[PE31]
     Metro-Ingress                 Core            Metro-Egress
         AS1                        AS2                AS3

         ----------- Packet Forwarding Direction -------->
```

Figure 9: Transport Layer with Heterogenous Color Domains

In the preceding topology shown in Figure 9, we have three Autonomous Systems. All the nodes in the topology support BGP CT.

In AS1 Gold SLA is represented by color 100.

In AS2 Gold has finer shades: Gold1 by color 101 and Gold2 by color 102.

In AS3 Gold SLA is represented by color 100.

This problem can be solved by the two following approaches:

11.2.3.1. Duplicate Tunnels Approach

In this approach, duplicate tunnels that satisfy Gold SLA are configured in domains AS1 and AS3, but they are given fine grained colors 101 and 102.

These tunnels will be installed in TRDBs corresponding to transport classes of color 101 and 102.

Overlay routes received with mapping community (e.g.: transport-target or color community) can resolve over these tunnels in the TRDB with matching color by using resolution schemes.

This approach consumes more resources in the transport and forwarding layer, because of the duplicate tunnels.

11.2.3.2. Customized Resolution Schemes Approach

In this approach, resolution schemes in domains AS1 and AS3 are customized to map the received mapping community (e.g., transport-target or color community) over available Gold SLA tunnels. This conserves resource usage with no additional state in the transport or forwarding planes.

Service routes advertised by PE31 that need to resolve over Gold1 transport tunnels carry a mapping community color:0:101. In AS3 and AS1, where Gold1 is not available, it is mapped to color 100 TRDB using a customized resolution scheme. In AS2, Gold1 is available and it maps to color 101 TRDB.

Similarly, service routes advertised by PE31 that need to resolve over Gold2 transport tunnels carry a mapping community color:0:102. In AS3 and AS1, where Gold2 is not available, it is mapped to color 100 TRDB using a customized resolution scheme. In AS2, Gold2 is available and it maps to color 102 TRDB.

To facilitate this mapping, every SN/BN in all AS provisioning required transport classes, viz. 100, 101 and 102. SN and BN in AS1 and AS3 are provisioned with customized resolution schemes that resolve routes with transport-target:0:101 or transport-target:0:102 strictly over color 100 TRDB.

PE31 is provisioned to originate BGP CT route with color 101 for endpoint PE31. This route is sent with NLRI RD prefix RD1:PE31 and route target extended community transport-target:0:101.

Similarly, PE31 is provisioned to originate BGP CT route with color 102 for endpoint PE31. This route is sent with NLRI RD prefix RD2:PE31 and route target extended community transport-target:0:102.

Following description explains the remaining procedures with color 101 as example.

At ASBR31, the route target "transport-target:0:101" on this BGP CT route instructs to add the route to color 101 TRDB. ASBR31 is provisioned with customized resolution scheme that resolves the routes carrying mapping community transport-target:0:101 to resolve using color 100 TRDB. This route is then re-advertised from color 101 TRDB to ASBR22 with route-target:0:101.

At ASBR22, the BGP CT routes received with transport-target:0:101 will be added to color 101 TRDB and strictly resolve over tunnel routes in the same TRDB. This route is re-advertised to ASBR21 with transport-target:0:101.

Similarly, at ASBR21, the BGP CT routes received with transport-target:0:101 will be added to color 101 TRDB and strictly resolve over tunnel routes in the same TRDB. This route is re-advertised to ASBR11 with transport-target:0:101.

At ASBR11, the route target "transport-target:0:101" on this BGP CT route instructs to add the route to color 101 TRDB. ASBR11 is provisioned with a customized resolution scheme that resolves the routes carrying transport-target:0:101 to use color 100 TRDB. This route is then re-advertised from color 101 TRDB to PE11 with route-target:0:101.

At PE11, the route target "transport-target:0:101" on this BGP CT route instructs to add the route to color 101 TRDB. PE11 is provisioned with a customized resolution scheme that resolves the routes carrying transport-target:0:101 to use color 100 TRDB.

When PE11 receives the service route with the mapping community color:0:101 it directly resolves over the BGP CT route in color 101 TRDB, which in turn resolves over tunnel routes in color 100 TRDB.

Similar processing is done for color 102 routes also at ASBR31, ASBR22, ASBR21, ASBR11 and PE11.

In doing so, PE11 can forward traffic via tunnels with color 101, color 102 in the core domain, and color 100 in the metro domains.

11.3. Migration Scenarios.

11.3.1. BGP CT Islands Connected via BGP LU Domain

This section explains how end-to-end SLA can be achieved while transiting a domain that does not support BGP CT. BGP LU is used in such domains to connect the BGP CT islands.

```
              +----------EBGP Multihop CT-------------+
              |                                       |
        AS3   |                   AS2                 |   AS1
  [PE31-----ASBR31]--------[ASBR22---ASBR21]-------[ASBR11---PE11]

                 <-EBGP LU->              <-EBGP LU->
   <---IBGP CT--->          <---IBGP LU--->        <---IBGP CT--->

            ---------Packet Forwarding Direction--------->
```

Figure 10: BGP CT in AS1 and AS3 connected by BGP LU in AS2

In the preceding topology shown in Figure 10, there are three AS domains. AS1 and AS3 support BGP CT, while AS2 does not support BGP CT.

Nodes in AS1, AS2, and AS3 negotiate BGP LU family on IBGP sessions within the domain. Nodes in AS1 and AS3 negotiate BGP CT family on IBGP sessions within the domain. ASBR11 and ASBR21 as well as ASBR22 and ASBR31 negotiate BGP LU family on the EBGP session over directly connected inter-domain links. ASBR11 and ASBR31 have reachability to each other’s loopbacks through BGP LU. ASBR11 and ASBR31 negotiate BGP CT family over a multihop EBGP session formed using BGP LU reachability.

The following tunnels exist for Gold Transport Class

PE11_to_ASBR11_gold - RSVP-TE tunnel

ASBR11_to_PE11_gold - RSVP-TE tunnel

PE31_to_ASBR31_gold - SRTE tunnel

ASBR31_to_PE31_gold - SRTE tunnel

The following tunnels exist for Bronze Transport Class

PE11_to_ASBR11_bronze - RSVP-TE tunnel

ASBR11_to_PE11_bronze - RSVP-TE tunnel

PE31_to_ASBR31_bronze - SRTE tunnel

ASBR31_to_PE31_bronze - SRTE tunnel

These tunnels are provisioned to belong to Transport Classes Gold and Bronze, and are advertised between ASBR31 and ASBR11 with Next hop self.

In AS2, that does not support BGP CT, a separate loopback may be used on ASBR22 and ASBR21 to represent Gold and Bronze SLAs, viz. ASBR22_lpbk_gold, ASBR22_lpbk_bronze, ASBR21_lpbk_gold and ASBR21_lpbk_bronze.

Furthermore, the following tunnels exist in AS2 to satisfy the different SLAs, using per SLA loopback endpoints:

ASBR21_to_ASBR22_lpbk_gold - RSVP-TE tunnel

ASBR22_to_ASBR21_lpbk_gold - RSVP-TE tunnel

ASBR21_to_ASBR22_lpbk_bronze - RSVP-TE tunnel

ASBR22_to_ASBR21_lpbk_bronze - RSVP-TE tunnel

RD:PE11 BGP CT route is originated from PE11 towards ASBR11 with transport-target 'gold.' ASBR11 readvertises this route with next hop set to ASBR11_lpbk_gold on the EBGP multihop session towards ASBR31. ASBR11 originates BGP LU route for endpoint ASBR11_lpbk_gold on EBGP session to ASBR21 with a 'gold SLA' community, and BGP LU route for ASBR11_lpbk_bronze with a 'bronze SLA' community. The SLA community is used by ASBR31 to publish the BGP LU routes in the corresponding BGP CT TRDBs.

ASBR21 readvertises the BGP LU route for endpoint ASBR11_lpbk_gold to ASBR22 with next hop set by local policy config to the unique loopback ASBR21_lpbk_gold by matching the 'gold SLA' community received as part of BGP LU advertisement from ASBR11. ASBR22 receives this route and resolves the next hop over the ASBR22_to_ASBR21_lpbk_gold RSVP-TE tunnel. On successful resolution, ASBR22 readvertises this BGP LU route to ASBR31 with next hop self and a new label.

ASBR31 adds the ASBR11_lpbk_gold route received via EBGP LU from ASBR22 to 'gold' TRDB based on the received 'gold SLA' community. ASBR31 uses this 'gold' TRDB route to resolve the next hop ASBR11_lpbk_gold received on BGP CT route with transport-target 'gold,' for the prefix RD:PE11 received over the EBGP multihop CT session, thus preserving the end-to-end SLA. Now ASBR31 readvertises the BGP CT route for RD:PE11 with next hop as self thus stitching with the BGP LU LSP in AS2. Intra-domain traffic forwarding in AS1 and AS3 follows the procedures as explained in Illustration of CT Procedures (Section 8)

In cases where an SLA cannot be preserved in AS2 because SLA specific tunnels and loopbacks don't exist in AS2, traffic can be carried over available SLAs (e.g.: best effort SLA) by rewriting the next hop to ASBR21 loopback assigned to the available SLA. This eases migration in case of heterogeneous color domains as-well.

11.3.2. BGP CT - Interoperability between MPLS and Other Forwarding Technologies

This section describes how nodes supporting dissimilar encapsulation technologies can interoperate with each other when using BGP CT family.

11.3.2.1. Interop Between MPLS and SRv6 Nodes.

BGP speakers may carry MPLS label and SRv6 SID in BGP CT SAFI 76 for AFIs 1 or 2 routes using protocol encoding as described in Carrying Multiple Encapsulation information (Section 6.3)

MPLS Labels are carried using RFC 8277 encoding, and SRv6 SID is carried using Prefix SID attribute as specified in Section 7.13.

```
                   RR1--+
                          \  +-------R2  [MPLS + SRv6]
                           \ |
                   R1--------P-------R3  [MPLS only]
             [MPLS + SRv6]   |
                             +-------R4  [SRv6 only]

               <---- Bidirectional Traffic ---->
```

Figure 11: BGP CT Interop between MPLS and SRv6 nodes

This example shows a provider network with a mix of devices with different forwarding capabilities. R1 and R2 support forwarding both MPLS and SRv6 packets. R3 supports forwarding MPLS packets only. R4 supports forwarding SRv6 packets only. All these nodes have BGP session with Route Reflector RR1 which reflects routes between these nodes with next hop unchanged. BGP CT family is negotiated on these sessions.

R1 and R2 send and receive both MPLS label and SRv6 SID in the BGP CT control plane routes. This allows them to be ingress and egress for both MPLS and SRv6 data planes. MPLS label is carried using RFC 8277 encoding, and SRv6 SID is carried using Prefix SID attribute as specified in Section 7.13, without Transposition Scheme. In this way, either MPLS or SRv6 forwarding can be used between R1 and R2.

R1 and R3 send and receive MPLS label in the BGP CT control plane routes using RFC 8277 encoding. This allows them to be ingress and egress for MPLS data plane. R1 will carry SRv6 SID in Prefix-SID attribute, which will not be used by R3. In order to interoperate with MPLS only device R3, R1 MUST NOT use SRv6 Transposition scheme described in [RFC9252]. The encoding suggested in Section 7.13 is used by R1. MPLS forwarding will be used between R1 and R3.

R1 and R4 send and receive SRv6 SID in the BGP CT control plane routes using BGP Prefix-SID attribute, without Transposition Scheme. This allows them to be ingress and egress for SRv6 data plane. R4 will carry the special MPLS Label with value 3 (Implicit-NULL) in RFC 8277 encoding, which tells R1 not to push any MPLS label for this BGP CT route towards R4. The MPLS Label advertised by R1 in RFC 8277 NLRI will not be used by R4. SRv6 forwarding will be used between R1 and R4.

Note in this example that R3 and R4 cannot communicate directly with each other, because they don't support a common forwarding technology. The BGP CT routes received at R3, R4 from each other will remain unusable, due to incompatible forwarding technology.

11.3.2.2. Interop Between Nodes Supporting MPLS and UDP Tunneling

This section describes how nodes supporting MPLS forwarding can interoperate with other nodes supporting UDP (or IP) tunneling, when using BGP CT family.

MPLS Labels are carried using RFC 8277 encoding, and UDP (or IP) tunneling information is carried using TEA attribute or the Encapsulation Extended Community as specified in [RFC9012].

```
                      RR1--+
                             \  +-------R2  [MPLS + UDP]
                              \ |
                      R1--------P-------R3  [MPLS only]
                [MPLS + UDP]    |
                                +-------R4  [UDP only]

                  <---- Bidirectional Traffic ---->
```

Figure 12: BGP CT Interop between MPLS and UDP tunneling nodes.

In this example, R1 and R2 support forwarding both MPLS and UDP tunneled packets. R3 supports forwarding MPLS packets only. R4 supports forwarding UDP tunneled packets only. All these nodes have BGP session with Route Reflector RR1 which reflects routes between these nodes with next hop unchanged. BGP CT family is negotiated on these sessions.

R1 and R2 send and receive both MPLS label and UDP tunneling info in the BGP CT control plane routes. This allows them to be ingress and egress for both MPLS and UDP tunneling data planes. MPLS label is carried using RFC 8277 encoding. As specified in [RFC9012], UDP tunneling information is carried using TEA attribute (code 23) or the "barebones" Tunnel TLV carried in Encapsulation Extended

Community. Either MPLS or UDP tunneled forwarding can be used between R1 and R2.

R1 and R3 send and receive MPLS label in the BGP CT control plane routes using RFC 8277 encoding. This allows them to be ingress and egress for MPLS data plane. R1 will carry UDP tunneling info in TEA attribute, which will not be used by R3. MPLS forwarding will be used between R1 and R3.

R1 and R4 send and receive UDP tunneling info in the BGP CT control plane routes using BGP TEA attribute. This allows them to be ingress and egress for UDP tunneled data plane. R4 will carry special MPLS Label with value 3 (Implicit-NULL) in RFC 8277 encoding, which tells R1 not to push any MPLS label for this BGP CT route towards R4. The MPLS Label advertised by R1 will not be used by R4. UDP tunneled forwarding will be used between R1 and R4.

Note in this example that R3 and R4 cannot communicate directly with each other, because they don't support a common forwarding technology. The BGP CT routes received at R3, R4 from each other will remain unusable, due to incompatible forwarding technology.

11.4. MTU Considerations

Operators should coordinate the MTU of the intra-domain tunnels used to prevent Path MTU discovery problems that could appear in deployments. The encapsulation overhead due to the MPLS label stack or equivalent tunnel header in different forwarding architecture should also be considered when determining the Path MTU of the end-to-end BGP CT tunnel.

11.5. Use of DSCP

BGP CT specifies procedures for Intent Driven Service Mapping in a service provider network, and defines 'Transport Class' construct to represent an Intent.

It may be desirable to allow a CE device to indicate in the data packet it sends what treatment it desires (the Intent) when the packet is forwarded within the provider network.

Such an indication can be in form of DSCP code point [RFC2474] in the IP header.

In RFC2474, a Forwarding Class Selector maps to a PHB (Per-hop Behavior). The Transport Class construct is a PHB at transport layer.

```
                        ---Gold----->
              [CE1]-----[PE1]---[P]----[PE2]-----[CE2]
                        ---Bronze--->
        203.0.113.11                             203.0.113.22
                   ----  Traffic direction ---->
```

Figure 13: Example Topology with DSCP on PE-CE Links

Let PE1 be configured to map DSCP1 to Gold Transport class, and DSCP2 to Bronze Transport class. Based on the DSCP code point received on the IP traffic from CE device, PE1 forwards the IP packet over a Gold or Bronze TC tunnel. Thus, the forwarding is not based on just the destination IP address, but also the DSCP code point. This is known as Class Based Forwarding (CBF).

CBF is configured at the PE1 device, mapping the DSCP values to respective Transport Classes. This mapping (DSCP peering agreement) is communicated to CE device by out of band mechanisms. This allows the administrator of CE1 to discover what transport classes exist in the provider network, and which DSCP codepoint to encode so that traffic is forwarded using the desired Transport Class in the provided network. When the IP packet exits the provider network to CE2, PE2 resets the DSCP code point based on DSCP peering agreement with CE2.

12. Applicability to Network Slicing

In Network Slicing, the Network Slice Controller (IETF NSC) is responsible for customizing and setting up the underlying transport (e.g. RSVP-TE, SRTE tunnels with desired characteristics) and resources (e.g., polices/shapers) in a transport network to create an IETF Network Slice.

The Transport Class construct described in this document can be used to realize the "IETF Network Slice" described in Section 4 of [TEAS-NS]

The NSC can use the Transport Class Identifier (Color value) to provision a transport tunnel in a specific IETF Network Slice.

Furthermore, the NSC can use the Mapping Community on the service route to map traffic to the desired IETF Network Slice.

13. IANA Considerations

This document makes the following requests of IANA.

13.1. New BGP SAFI

IANA is requested to assign a new BGP SAFI code for "Classful Transport". Value 76.

Registry Group: Subsequent Address Family Identifiers (SAFI) Parameters

Registry Name: SAFI Values

Value	Description
76	Classful Transport SAFI

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

- "IPv4, Classful Transport". AFI/SAFI = "1/76" for carrying IPv4 Classful Transport prefixes.

- "IPv6, Classful Transport". AFI/SAFI = "2/76" for carrying IPv6 Classful Transport prefixes.

13.2. New Format for BGP Extended Community

IANA is requested to assign a new Format type (Type high = 0xa) of Extended Community EXT-COMM [RFC4360] for Transport Class from the following registries:

the "BGP Transitive Extended Community Types" registry, and

the "BGP Non-Transitive Extended Community Types" registry.

Please assign the same low-order six bits for both allocations.

This document uses this new Format with subtype 0x2 (route target), as a transitive extended community. The Route Target thus formed is called "Transport Class" route target extended community.

The Non-Transitive Transport Class Extended community with subtype 0x2 (route target) is called the "Non-Transitive Transport Class route target extended community".

Taking reference of [RFC7153] , following requests are made:

13.2.1. Existing Registries to be Modified

13.2.1.1. Registries for the "Type" Field

13.2.1.1.1. Transitive Types

This registry contains values of the high-order octet (the "Type" field) of a Transitive Extended Community.

Registry Group: Border Gateway Protocol (BGP) Extended Communities

Registry Name: BGP Transitive Extended Community Types

Type Value	Name
0x0a	Transport Class

(Sub-Types are defined in the "Transitive Transport Class Extended Community Sub-Types" registry)

13.2.1.1.2. Non-Transitive Types

This registry contains values of the high-order octet (the "Type" field) of a Non-transitive Extended Community.

Registry Group: Border Gateway Protocol (BGP) Extended Communities

Registry Name: BGP Non-Transitive Extended Community Types

Type Value	Name
0x4a	Non-Transitive Transport Class

(Sub-Types are defined in the "Non-Transitive Transport Class Extended Community Sub-Types" registry)

13.2.2. New Registries

13.2.2.1. Transitive Transport Class Extended Community Sub-Types Registry

IANA is requested to add the following subregistry under the “Border Gateway Protocol (BGP) Extended Communities”:

Registry Group: Border Gateway Protocol (BGP) Extended Communities

Registry Name: Transitive Transport Class Extended Community Sub-Types

Note:
This registry contains values of the second octet (the "Sub-Type" field) of an extended community when the value of the first octet (the "Type" field) is 0x0a.

Range	Registration Procedures
0x00-0xBF	First Come First Served
0xC0-0xFF	IETF Review

Sub-Type Value	Name
0x02	Route Target

13.2.2.2. Non-Transitive Transport Class Extended Community Sub-Types Registry

IANA is requested to add the following subregistry under the “Border Gateway Protocol (BGP) Extended Communities”:

Registry Group: Border Gateway Protocol (BGP) Extended Communities

Registry Name: Non-Transitive Transport Class Extended Community Sub-Ty

Note:
This registry contains values of the second octet (the "Sub-Type" field) of an extended community when the value of the first octet (the "Type" field) is 0x4a.

Range	Registration Procedures
0x00-0xBF	First Come First Served
0xC0-0xFF	IETF Review

Sub-Type Value	Name
0x02	Route Target

13.3. MPLS OAM Code Points

The following two code points are sought for Target FEC Stack sub-TLVs:

*IPv4 BGP Classful Transport

*IPv6 BGP Classful Transport

Registry Group: Multiprotocol Label Switching (MPLS)
Label Switched Paths (LSPs) Ping Parameters

Registry Name: Sub-TLVs for TLV Types 1, 16, and 21

Sub-Type	Name
31744	IPv4 BGP Classful Transport
31745	IPv6 BGP Classful Transport

13.4. Best Effort Transport Class ID

This document reserves the Transport class ID value 0 to represent "Best Effort Transport Class ID". This is used in the 'Transport Class ID' field of Transport Route Target extended community that represents best effort transport class. Please create a new registry for this.

Registry Group: BGP Classful Transport (BGP CT)

Registry Name: Transport Class ID

Value	Name
0	Best Effort Transport Class ID
1-4294967295	Private Use

Reference: This document.

Registration Procedure(s)

Value	Registration Procedure
0	IETF Review
1-4294967295	Private Use

As noted in Sec 4 and Sec 7.10, 'Transport Class ID' is interchangeable with 'Color'. For purposes of backward compatibility with usage of 'Color' field in Color extended community, the range 1-4294967295 uses 'Private Use' as Registration Procedure.

14. Security Considerations

This document uses [RFC4760] mechanisms to define new BGP address families (AFI/SAFI : 1/76 and 2/76) that carry transport layer endpoints. These address families are explicitly configured and negotiated between BGP speakers, which confines the propagation scope of this reachability information, thus following a 'walled garden' approach.

This mitigates the risk of advertising internal loopback addresses outside the administrative control of the provider network. Furthermore, procedures defined in Section 9.1 mitigate the risk of unintended propagation of BGP CT routes across Inter-AS boundaries even when BGP CT family is negotiated on the EBGP session.

This document does not change the underlying security issues inherent in the existing BGP protocol, such as those described in [RFC4271] and [RFC4272].

Additionally, BGP sessions SHOULD be protected using TCP Authentication Option [RFC5925] and the Generalized TTL Security Mechanism [RFC5082]. To mitigate any risk of manipulating the routing information carried within a new SAFI, BGP origin validation [RFC6811] and BGPsec [RFC8205] MAY be used as means to increase assurance that the information has not been falsified.

Using a separate BGP family and new RT (Transport Class RT) minimizes the possibility of these routes mixing with service routes.

If redistributing between SAFI 76 and SAFI 4 routes for AFIs 1 or 2, 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 76 and SAFI 4 transport routes in separate transport RIBs, distinct from service RIB that contain SAFI 1 service routes.

BGP CT routes distribute label binding using [RFC8277] for MPLS dataplane and hence its security considerations apply.

BGP CT routes distribute SRv6 SIDs for SRv6 dataplanes and hence security considerations of Section 9.3 of [RFC9252] apply. Moreover, SRv6 SID transposition scheme is disabled in BGP CT, as described in Section 7.13, to mitigate the risk of misinterpreting transposed SRv6 SID information as an MPLS label.

As [RFC4272] discusses, BGP is vulnerable to traffic-diversion attacks. This SAFI routes adds a new means by which an attacker could cause the traffic to be diverted from its normal path. Potential consequences include "hijacking" of traffic (insertion of an undesired node in the path, which allows for inspection or modification of traffic, or avoidance of security controls) or denial of service (directing traffic to a node that doesn't desire to receive it).

In order to mitigate the risk of the diversion of traffic from its intended destination, existing BGPsec solution could be extended and supported for this SAFI. The restriction of the applicability of this SAFI to its intended well-defined scope limits the likelihood of traffic diversions. Furthermore, as long as the filtering and appropriate configuration mechanisms discussed previously are applied diligently, risk of the diversion of the traffic is significantly mitigated.

15. References

15.1. 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>.

[RFC2474] Nichols, K., Blake, S., Baker, F., and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474, DOI 10.17487/RFC2474, December 1998, <https://www.rfc-editor.org/info/rfc2474>.

[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>.

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

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

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

[RFC4659] De Clercq, J., Ooms, D., Carugi, M., and F. Le Faucheur, "BGP-MPLS IP Virtual Private Network (VPN) Extension for IPv6 VPN", RFC 4659, DOI 10.17487/RFC4659, September 2006, <https://www.rfc-editor.org/info/rfc4659>.

[RFC4684] Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk, R., Patel, K., and J. Guichard, "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, DOI 10.17487/RFC4684, November 2006, <https://www.rfc-editor.org/info/rfc4684>.

[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>.

[RFC5082] Gill, V., Heasley, J., Meyer, D., Savola, P., Ed., and C. Pignataro, "The Generalized TTL Security Mechanism (GTSM)", RFC 5082, DOI 10.17487/RFC5082, October 2007, <https://www.rfc-editor.org/info/rfc5082>.

[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>.

[RFC6811] Mohapatra, P., Scudder, J., Ward, D., Bush, R., and R. Austein, "BGP Prefix Origin Validation", RFC 6811, DOI 10.17487/RFC6811, January 2013, <https://www.rfc-editor.org/info/rfc6811>.

[RFC7153] Rosen, E. and Y. Rekhter, "IANA Registries for BGP Extended Communities", RFC 7153, DOI 10.17487/RFC7153, March 2014, <https://www.rfc-editor.org/info/rfc7153>.

[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>.

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

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

[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>.

[RFC8205] Lepinski, M., Ed. and K. Sriram, Ed., "BGPsec Protocol Specification", RFC 8205, DOI 10.17487/RFC8205, September 2017, <https://www.rfc-editor.org/info/rfc8205>.

[RFC8212] Mauch, J., Snijders, J., and G. Hankins, "Default External BGP (EBGP) Route Propagation Behavior without Policies", RFC 8212, DOI 10.17487/RFC8212, July 2017, <https://www.rfc-editor.org/info/rfc8212>.

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

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

[RFC9012] Patel, K., Van de Velde, G., Sangli, S., and J. Scudder, "The BGP Tunnel Encapsulation Attribute", RFC 9012, DOI 10.17487/RFC9012, April 2021, <https://www.rfc-editor.org/info/rfc9012>.

[RFC9252] Dawra, G., Ed., Talaulikar, K., Ed., Raszuk, R., Decraene, B., Zhuang, S., and J. Rabadan, "BGP Overlay Services Based on Segment Routing over IPv6 (SRv6)", RFC 9252, DOI 10.17487/RFC9252, July 2022, <https://www.rfc-editor.org/info/rfc9252>.

15.2. Informative References

[BGP-CT-SRv6] Vairavakkalai, Ed. and Venkataraman, Ed., "BGP CT - Adaptation to SRv6 dataplane", 4 March 2024, <https://tools.ietf.org/html/draft-ietf-idr-bgp-ct-srv6-03>.

[BGP-CT-UPDATE-PACKING-TEST] Vairavakkalai, Ed., "BGP CT Update packing Test Results", 25 June 2023, <https://raw.githubusercontent.com/ietf-wg-idr/draft-ietf-idr-bgp-ct/1a75d4d10d4df0f1fd7dcc041c2c868704b092c7/update-packing-test-results.txt>.

[BGP-FWD-RR] Vairavakkalai, Ed. and Venkataraman, Ed., "BGP Route Reflector in Forwarding Path", 17 March 2024, <https://tools.ietf.org/html/draft-ietf-idr-bgp-fwd-rr-02>.

[BGP-LU-EPE] Gredler, Ed., "Egress Peer Engineering using BGP-LU", 16 June 2023, <https://datatracker.ietf.org/doc/html/draft-gredler-idr-bgplu-epe-15>.

[FLOWSPEC-REDIR-IP]
Simpson, Ed., "BGP Flow-Spec Redirect to IP Action", 2 February 2015, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-flowspec-redirect-ip-02>.

[Intent-Routing-Color] Hegde, Ed., "Intent-aware Routing using Color", 23 October 2023, <https://datatracker.ietf.org/doc/html/draft-hr-spring-intentaware-routing-using-color-03#section-6.3.2>.

[MNH] Vairavakkalai, Ed., "BGP MultiNexthop Attribute", 17 March 2024, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-multinexthop-attribute-00>.

[MPLS-NS] Vairavakkalai, Ed., "BGP signalled MPLS namespaces", 20 October 2023, <https://datatracker.ietf.org/doc/html/draft-kaliraj-bess-bgp-sig-private-mpls-labels-07>.

[PCEP-RSVP-COLOR] Rajagopalan, Ed. and Pavan. Beeram, Ed., "Path Computation Element Protocol(PCEP) Extension for RSVP Color", 1 September 2023, <https://datatracker.ietf.org/doc/html/draft-ietf-pce-pcep-color-02>.

[PCEP-SRPOLICY] Koldychev, Ed., Sivabalan, Ed., and Barth, Ed., "PCEP Extensions for SR Policy Candidate Paths", 9 February 2024, <https://www.ietf.org/archive/id/draft-ietf-pce-segment-routing-policy-cp-14.html#name-sr-policy-identifier>.

[RFC6890] Cotton, M., Vegoda, L., Bonica, R., Ed., and B. Haberman, "Special-Purpose IP Address Registries", BCP 153, RFC 6890, DOI 10.17487/RFC6890, April 2013, <https://www.rfc-editor.org/info/rfc6890>.

[RFC9315] Clemm, A., Ciavaglia, L., Granville, L. Z., and J. Tantsura, "Intent-Based Networking - Concepts and Definitions", RFC 9315, DOI 10.17487/RFC9315, October 2022, <https://www.rfc-editor.org/info/rfc9315>.

[RFC9350] Psenak, P., Ed., Hegde, S., Filsfils, C., Talaulikar, K., and A. Gulko, "IGP Flexible Algorithm", RFC 9350, DOI 10.17487/RFC9350, February 2023, <https://www.rfc-editor.org/info/rfc9350>.

[SRTE] Talaulikar, Ed. and S. Previdi, "Advertising Segment Routing Policies in BGP", 23 October 2023, <https://

tools.ietf.org/html/draft-ietf-idr-segment-routing-te-policy-26>.

[TEAS-NS] Farrel, Ed. and Drake, Ed., "A Framework for IETF Network Slices", 14 September 2023, <https://www.ietf.org/archive/id/draft-ietf-teas-ietf-network-slices-25.html#section-4>.

Appendix A. Extensibility considerations

A.1. Signaling Intent over PE-CE Attachment Circuit

It may be desirable to allow a CE device to indicate in the data packet it sends what treatment it desires (the Intent) when the packet is forwarded within the provider network.

Section A.10 in BGP MultiNexthop Attribute [MNH] describes some mechanisms that enable such signaling.

A.2. BGP CT Egress TE

Mechanisms described in [BGP-LU-EPE] also applies to BGP CT family.

The Peer/32 or Peer/128 EPE route MAY be originated in BGP CT family with appropriate Mapping Community (e.g. transport-target:0:100), thus allowing an EPE path to the peer that satisfies the desired SLA.

Appendix B. Applicability to Intra-AS and different Inter-AS deployments.

As described in BGP VPN [RFC4364] Section 10, in an Option C network, service routes (VPN-IPv4) are neither maintained nor distributed by the ASBRs. Transport routes are maintained in the ASBRs and propagated in BGP LU or BGP CT.

Illustration of CT Procedures (Section 8) illustrates how constructs of BGP CT work in an inter-AS Option C deployment. The BGP CT constructs: AFI/SAFI 1/76, Transport Class and Resolution Scheme are used in an inter-AS Option C deployment.

In Intra-AS and Inter-AS option A, option B scenarios, AFI/SAFI 1/76 may not be used, but the Transport Class and Resolution Scheme mechanisms are used to provide service mapping.

This section illustrates how BGP CT constructs work in Intra-AS and Inter-AS Option A, B deployment scenarios.

B.1. Intra-AS usecase

B.1.1. Topology

```
                                    [RR11]
                                     |
                                     +
              [CE21]---[PE11]-------[P1]------[PE12]------[CE31]
                     |                             |
                     +                             +
                     |                             |
               AS2               ...AS1...               AS3

             203.0.113.21 ---- Traffic Direction ----> 203.0.113.31
```

Figure 14: BGP CT Intra-AS.

This example in Figure 14 shows a provider network Autonomous system AS1. It serves customers AS2, AS3. Traffic direction being described is CE21 to CE31. CE31 may request a specific SLA (e.g. Gold for this traffic), when traversing this provider network.

B.1.2. Transport Layer

AS1 uses RSVP-TE intra-domain tunnels between PE11 and PE12. And LDP tunnels for best effort traffic.

The network has two Transport classes: Gold with Transport Class ID 100, Bronze with Transport Class ID 200. These transport classes are provisioned at the PEs. This creates the Resolution Schemes for these transport classes at these PEs.

Following tunnels exist for Gold transport class.

PE11_to_PE12_gold - RSVP-TE tunnel

PE12_to_PE11_gold - RSVP-TE tunnel

Following tunnels exist for Bronze transport class.

PE11_to_PE12_bronze - RSVP-TE tunnel

PE11_to_PE12_bronze - RSVP-TE tunnel

These tunnels are provisioned to belong to transport class 100 or 200.

B.1.3. Service Layer route exchange

Service nodes PE11, PE12 negotiate service families (AFI/SAFI 1/128) on the BGP session with RR11. Service helper RR11 reflects service routes between the two PEs with next hop unchanged. There are no tunnels for transport-class 100 or 200 from RR11 to the PEs.

Forwarding happens using service routes at service nodes PE11, PE12. Routes received from CEs are not present in any other nodes' FIB in the provider network.

CE31 advertises a route for example prefix 203.0.113.31 with next hop self to PE12. CE31 can attach a Mapping Community Color:0:100 on this route, to indicate its request for Gold SLA. Or, PE11 can attach the same using locally configured policies.

Consider, CE31 is getting VPN service from PE12. The RD:203.0.113.31 route is readvertised in AFI/SAFI 1/128 by PE12 with next hop self (192.0.2.12) and label V-L1, to RR11 with the Mapping Community Color:0:100 attached. This AFI/SAFI 1/128 route reaches PE11 via RR11 with the next hop unchanged as PE12 and label V-L1. Now PE11 can resolve the PNH 192.0.2.12 using PE11_to_PE12_gold RSVP TE LSP.

The IP FIB at PE11 VRF will have a route for 203.0.113.31 with a next hop when resolved using Resolution Scheme belonging to the mapping community Color:0:100, points to a PE11_to_PE12_gold tunnel.

BGP CT AFI/SAFI 1/76 is not used in this Intra-AS deployment. But the Transport class and Resolution Scheme constructs are used to preserve end-to-end SLA.

B.2. Inter-AS option A usecase

B.2.1. Topology

```
                  [RR11]                        [RR21]
                    |                             |
                    +                             +
[CE31]---[PE11]----[P1]----[ASBR11]---[ASBR21]---[P2]---[PE21]----[CE41]
        |                           |                           |
        +                           +                           +
        |                           |                           |
  AS3               ..AS1..               ..AS2..                  AS4

203.0.113.31          ---- Traffic Direction ---->         203.0.113.41
```

Figure 15: BGP CT Inter-AS option A.

This example in Figure 15 shows two provider network Autonomous systems AS1, AS2. They serve L3VPN customers AS3, AS4 respectively. The ASBRs ASBR11 and ASBR21 have IP VRFs connected directly. The inter-AS link is IP enabled with no MPLS forwarding.

Traffic direction being described is CE31 to CE41. CE41 may request a specific SLA (e.g. Gold for this traffic), when traversing these provider core networks.

B.2.2. Transport Layer

AS1 uses RSVP-TE intra-domain tunnels between PE11 and ASBR11. And LDP tunnels for best effort traffic. AS2 uses SRTE intra-domain tunnels between ASBR21 and PE21, and L-ISIS for best effort tunnels.

The networks have two Transport classes: Gold with Transport Class ID 100, Bronze with Transport Class ID 200. These transport classes are provisioned at the PEs and ASBRs. This creates the Resolution Schemes for these transport classes at these PEs and ASBRs.

Following tunnels exist for Gold transport class.

PE11_to_ASBR11_gold - RSVP-TE tunnel

ASBR11_to_PE11_gold - RSVP-TE tunnel

PE21_to_ASBR21_gold - SRTE tunnel

ASBR21_to_PE21_gold - SRTE tunnel

Following tunnels exist for Bronze transport class.

PE11_to_ASBR11_bronze - RSVP-TE tunnel

ASBR11_to_PE11_bronze - RSVP-TE tunnel

PE21_to_ASBR21_bronze - SRTE tunnel

ASBR21_to_PE21_bronze - SRTE tunnel

These tunnels are provisioned to belong to transport class 100 or 200.

B.2.3. Service Layer route exchange

Service nodes PE11, ASBR11 negotiate service family (AFI/SAFI 1/128) on the BGP session with RR11. Service helper RR11 reflects service routes between the PE11 and ASBR11 with next hop unchanged.

Similarly, in AS2 PE21, ASBR21 negotiate service family (AFI/SAFI 1/128) on the BGP session with RR21, which reflects service routes between the PE21 and ASBR21 with next hop unchanged.

CE41 advertises a route for example prefix 203.0.113.41 with next hop self to PE21 VRF. CE41 can attach a Mapping Community Color: 0:100 on this route, to indicate its request for Gold SLA. Or, PE21 can attach the same using locally configured policies.

Consider, CE41 is getting VPN service from PE21. The RD:203.0.113.41 route is readvertised in AFI/SAFI 1/128 by PE21 with next hop self (203.0.113.21) and label V-L1 to RR21 with the Mapping Community Color:0:100 attached. This AFI/SAFI 1/128 route reaches ASBR21 via RR21 with the next hop unchanged as PE21 and label V-L1. Now ASBR21 can resolve the PNH 203.0.113.21 using ASBR21_to_PE21_gold SRTE LSP.

The IP FIB at ASBR21 VRF will have a route for 203.0.113.41 with a next hop resolved using Resolution Scheme associated with mapping community Color:0:100, pointing to ASBR21_to_PE21_gold tunnel.

This route is readvertised by ASBR21 on BGP session inside VRF with next hop self. EBGP session peering on interface address. ASBR21 acts like a CE to ASBR11, and the previously mentioned process repeats in AS1, until the route reaches PE11 and resolves over PE11_to_ASBR11_gold RSVP TE tunnel.

Traffic traverses as IP packet on the following legs: CE31-PE11, ASBR11-ASBR21, PE21-CE41. And uses MPLS forwarding inside AS1, AS2 core.

BGP CT AFI/SAFI 1/76 is not used in this Inter-AS Option B deployment. But the Transport class and Resolution Scheme constructs are used to preserve end-to-end SLA.

B.3. Inter-AS option B usecase

B.3.1. Topology

```
                  [RR13]                        [RR23]
                    |                             |
                    +                             +
[CE31]---[PE11]----[P1]----[ASBR12]---[ASBR21]---[P2]---[PE22]----[CE41]
        |                           |                           |
        +                           +                           +
        |                           |                           |
  AS3               ..AS1..               ..AS2..                  AS4

203.0.113.31          ---- Traffic Direction ---->         203.0.113.41
```

Figure 16: BGP CT Inter-AS option B.

This example in Figure 16 shows two provider network Autonomous systems AS1 and AS2. They serve L3VPN customers AS3 and AS4 respectively. The ASBRs ASBR12 and ASBR21 don't have any IP VRFs. The inter-AS link is MPLS forwarding enabled.

Traffic direction being described is CE31 to CE41. CE41 may request a specific SLA (e.g. Gold for this traffic), when traversing these provider core networks.

B.3.2. Transport Layer

AS1 uses RSVP-TE intra-domain tunnels between PE11 and ASBR21. And LDP tunnels for best effort traffic. AS2 uses SRTE intra-domain tunnels between ASBR21 and PE22, and L-ISIS for best effort tunnels.

The networks have two Transport classes: Gold with Transport Class ID 100, Bronze with Transport Class ID 200. These transport classes are provisioned at the PEs and ASBRs. This creates the Resolution Schemes for these transport classes at these PEs and ASBRs.

Following tunnels exist for Gold transport class.

PE11_to_ASBR12_gold - RSVP-TE tunnel

ASBR12_to_PE11_gold - RSVP-TE tunnel

PE22_to_ASBR21_gold - SRTE tunnel

ASBR21_to_PE22_gold - SRTE tunnel

Following tunnels exist for Bronze transport class.

PE11_to_ASBR12_bronze - RSVP-TE tunnel

ASBR12_to_PE11_bronze - RSVP-TE tunnel

PE22_to_ASBR21_bronze - SRTE tunnel

ASBR21_to_PE22_bronze - SRTE tunnel

These tunnels are provisioned to belong to transport class 100 or 200.

B.3.3. Service Layer route exchange

Service nodes PE11, ASBR12 negotiate service family (AFI/SAFI 1/128) on the BGP session with RR13. Service helper RR13 reflects service routes between the PE11 and ASBR12 with next hop unchanged.

Similarly, in AS2 PE22, ASBR21 negotiate service family (AFI/SAFI 1/128) on the BGP session with RR23, which reflects service routes between the PE22 and ASBR21 with next hop unchanged.

ASBR21 and ASBR12 negotiate AFI/SAFI 1/128 between them, and readvertise L3VPN routes with next hop self, allocating new labels. EBGP session peering on interface address.

CE41 advertises a route for example prefix 203.0.113.41 with next hop self to PE22 VRF. CE41 can attach a Mapping Community Color: 0:100 on this route, to indicate its request for Gold SLA. Or, PE22 can attach the same using locally configured policies.

Consider, CE41 is getting VPN service from PE22. The RD:203.0.113.41 route is readvertised in AFI/SAFI 1/128 by PE22 with next hop self (192.0.2.22) and label V-L1 to RR23 with the Mapping Community Color:0:100 attached. This AFI/SAFI 1/128 route reaches ASBR21 via RR23 with the next hop unchanged as PE22 and label V-L1. Now ASBR21 can resolve the PNH 192.0.2.22 using ASBR21_to_PE22_gold SRTE LSP.

Next, ASBR21 readvertises the RD:203.0.113.41 route with next hop self to ASBR12 with a newly allocated MPLS label V-L2. Forwarding for this label is installed to Swap V-L1, and Push labels for ASBR21_to_PE22_gold tunnel.

ASBR12 further readvertises the RD:203.0.113.41 route via RR13 to PE11 with next hop self 192.0.2.12. PE11 resolves the next hop 192.0.2.12 over PE11_to_ASBR12_gold RSVP TE tunnel.

Traffic traverses as IP packet on the following legs: CE31-PE11 and PE21-CE41. And uses MPLS forwarding on ASBR11-ASBR21 link, and inside AS1-AS2 core.

BGP CT AFI/SAFI 1/76 is not used in this Inter-AS Option B deployment. But the Transport class and Resolution Scheme constructs are used to preserve end-to-end SLA.

Appendix C. Why reuse RFC 8277 and RFC 4364?

RFC 4364 is one of the key design patterns produced by networking industry. It introduced virtualization and allowed sharing of resources in service provider space with multiple tenant networks, providing isolated and secure Layer3 VPN services. This design pattern has been reused since to provide other service layer virtualizations like Layer2 virtualization (VPLS, L2VPN, EVPN), ISO virtualization, ATM virtualization, Flowspec VPN.

It is to be noted that these services have different NLRI encoding. L3VPN Service family that binds MPLS label to an IP prefix use RFC 8277 encoding, and others define different NLRI encodings.

BGP CT reuses RFC 4364 procedures to slice a transport network into multiple transport planes that different service routes can bind to, using color.

BGP CT reuses RFC 8277 because it precisely fits the purpose. viz. In a MPLS network, BGP CT needs to bind MPLS label for transport endpoints which are IPv4 or IPv6 endpoints, and disambiguate between multiple instances of those endpoints in multiple transport planes. Hence, use of RD:IP_Prefix and carrying a Label for it as specified in RFC 8277 works well for this purpose.

Another advantage of using the precise encoding as defined in RFC 4364 and RFC 8277 is that it allows to interoperate with BGP speakers that support SAFI 128 for AFIs 1 or 2. This can be useful during transition, until all BGP speakers in the network support BGP CT.

In future, if RFC 8277 evolves into a typed NLRI, that does not carry Label in the NLRI, BGP CT will be compatible with that as-well. In essence, BGP CT encoding is compatible with existing deployed technologies (RFC 4364, RFC 8277) and will adapt to any changes RFC 8277 mechanisms undergo in future.

This approach leverages the benefits of time tested design patterns proposed in RFC 4364 and RFC 8277. Moreover, this approach greatly reduces operational training costs and protocol compatibility considerations, as it complements and works well with existing protocol machineries. This problem does not need reinventing the wheel with brand new NLRI and procedures.

This is a more pragmatic approach. Rather than abandoning time tested design pattern like RFC 4364 and RFC 8277, just to invent something completely new that is not backward compatible with existing deployments. Overloading RFC 8277 NLRI MPLS Label field with information related to non MPLS data plane leads to backward compatibility issues.

C.1. Update packing considerations

BGP CT carries transport class as an attribute. This means routes that don't share the same transport class cannot be packed into same Update message. Update packing in BGP CT will be similar to RFC 8277 family routes carrying attributes like communities or extended communities. Service families like AFI/SAFI 1/128 have considerably more scale than transport families like AFI/SAFI 1/4 or AFI/SAFI 1/76, which carry only loopbacks. Update packing mechanisms that scale for AFI/SAFI 1/128 routes will scale similarly for AFI/SAFI 1/76 routes also.

Section 6.3.2.1 of [Intent-Routing-Color] suggests scaling numbers for transport network where BGP CT can be deployed. Experiments were conducted with this scale to find the convergence time with BGP CT for those scaling numbers. Scenarios involving BGP CT carrying IPv4 and IPv6 endpoints with MPLS label were tested. Tests with BGP CT IPv6 endpoints and SRv6 SID are planned.

Tests were conducted with 1.9 million BGP CT route scale (387K endpoints in 5 transport classes). Initial convergence time for all cases was less than 2 minutes, This experiment proves that carrying transport class information as an attribute keeps BGP convergence within acceptable range. Details of the experiment and test results are available in BGP CT Update packing Test Results [BGP-CT-UPDATE-PACKING-TEST].

Furthermore, even in today's BGP LU deployments each egress node originates BGP LU route for it's loopback, with some attributes like community identifying the originating node or region, and AIGP attribute. These attributes may be unique per egress node, thus do not help with update packing in transport layer family routes.

Appendix D. Scaling using BGP MPLS Namespaces

This document considers scaling 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.

Section 6.2 of [MPLS-NS] describes how MPLS Namespaces mechanism 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 also.

Contributors

Co-Authors

Reshma Das
Juniper Networks, Inc.
1133 Innovation Way,
Sunnyvale, CA 94089
United States of America

Email: dreshma@juniper.net

Israel Means
AT&T
2212 Avenida Mara,
Chula Vista, California 91914
United States of America

Email: israel.means@att.com

Csaba Mate
KIFU, Hungarian NREN
Budapest
35 Vaci street,
1134
Hungary

Email: ietf@nop.hu

Deepak J Gowda
Extreme Networks
55 Commerce Valley Drive West, Suite 300,
Thornhill, Toronto, Ontario L3T 7V9
Canada

Email: dgowda@extremenetworks.com

Other Contributors

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

Email: balajir@juniper.net

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

Email: mrajesh@juniper.net

Chaitanya Yadlapalli
AT&T
200 S Laurel Ave,

Middletown,, NJ 07748
United States of America

Email: cy098d@att.com

Mazen Khaddam
Cox Communications Inc.
Atlanta, GA
United States of America

Email: mazen.khaddam@cox.com

Rafal Jan Szarecki
Google.
1160 N Mathilda Ave, Bldg 5,
Sunnyvale,, CA 94089
United States of America

Email: szarecki@google.com

Xiaohu Xu
China Mobile
Beijing
China

Email: xuxiaohu@cmss.chinamobile.com

Acknowledgements

The authors thank Jeff Haas, John Scudder, Susan Hares, Dongjie (Jimmy), Moses Nagarajah, Jeffrey (Zhaohui) Zhang, Joel Harpern, Jingrong Xie, Mohamed Boucadair, Greg Skinner, Simon Leinen, Navaneetha Krishnan, Ravi M R, Chandrasekar Ramachandran, Shradha Hegde, Colby Barth, Vishnu Pavan Beeram, Sunil Malali, William J Britto, R Shilpa, Ashish Kumar (FE), Sunil Kumar Rawat, Abhishek Chakraborty, Richard Roberts, Krzysztof Szarkowicz, John E Drake, Srihari Sangli, Jim Uttaro, Luay Jalil, Keyur Patel, Ketan Talaulikar, Dhananjaya Rao, Swadesh Agarwal, Robert Raszuk, Ahmed Darwish, Aravind Srinivas Srinivasa Prabhakar, Moshiko Nayman, Chris Tripp, Gyan Mishra, Vijay Kestur, Santosh Kolenchery for all the valuable discussions, constructive criticisms, and review comments.

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

Authors' Addresses

Kaliraj Vairavakkalai (editor)
Juniper Networks, Inc.

1133 Innovation Way,
Sunnyvale, CA 94089
United States of America

Email: kaliraj@juniper.net

Natrajan Venkataraman (editor)
Juniper Networks, Inc.
1133 Innovation Way,
Sunnyvale, CA 94089
United States of America

Email: natv@juniper.net