Internet Engineering Task Force (IETF)
Request for Comments: 9832
Category: Experimental
ISSN: 2070-1721

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

September 2025

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" or "TCs"), 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 the TEAS Network Slices framework (RFC 9543).

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 the procedures described in RFC 4364 ("BGP/MPLS IP Virtual Private Networks (VPNs)") and follows the NLRI encoding described in RFC 8277 ("Using BGP to Bind MPLS Labels to Address Prefixes") is defined to enable each advertised underlay route to be identified by its class. This new address family is called "BGP Classful Transport" (or "BGP CT").

Status of This Memo

This document is not an Internet Standards Track specification; it is published for examination, experimental implementation, and evaluation.

This document defines an Experimental Protocol for the Internet community. This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Not all documents approved by the IESG are candidates for any level of Internet Standard; see Section 2 of RFC 7841.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at https://www.rfc-editor.org/info/rfc9832.

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 be either in the same domain or across different domains.

[RFC9315] defines "Intent" as:

> 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.

This document prescribes constructs and procedures to realize "Intent" and enable provider networks to forward service traffic based on service-specific Intent from 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, a low-latency path or a best-effort path.

* Expressing a desired Intent. For example, use a low-latency path with a 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 equally to intra-AS and inter-AS (a.k.a. multi-AS) deployments in the style of option A, option B, and option C (Section 10 of [RFC4364]) in provider networks.

Such networks provision intra-domain transport tunnels between a pair of endpoints, typically a service node or a border node that service traffic traverses through. 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 TC 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) is 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 over IPv6 (SRv6) are provided in [BGP-CT-SRv6] and an overview is provided in Section 7.13.

Customers need to be able to express 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 characteristics (e.g., low-latency), into identifiable classes that can pass "Intent-aware" traffic. These constructs enable services to signal their Intent to use 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 (BGP CT)", which 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.

Appendix C provides an outline of the design philosophy behind this specification. In particular, readers who are already familiar with one or more BGP VPN technologies may want to review this appendix before reading the main body of the specification.

2. Terminology

2.1. Abbreviations

ABR: Area Border Router (readvertises BGP CT or BGP LU routes with NH self)

AFI: Address Family Identifier

AS: Autonomous System

ASBR: Autonomous System Border Router

ASN: Autonomous System Number

BGP VPN: VPNs built using RD or 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

CCA: Community Carrying Attribute

CsC: Carriers' Carriers (serving the 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

FRR: Fast Reroute (Preprogrammed NH leg in forwarding)

iSN: Ingress Service Node

L-ISIS: Labeled ISIS (see RFC 8667)

LSP: Label Switched Path

MPLS: Multiprotocol Label Switching

NH: Next Hop

NLRI: Network Layer Reachability Information

PE: Provider Edge

PIC: Prefix Independent Convergence

PNH: Protocol Next Hop (address carried in a BGP UPDATE message)

RD: Route Distinguisher

RD:EP: Route Distinguisher and Endpoint (in a BGP Prefix)

RSVP-TE: Resource Reservation Protocol - Traffic Engineering

RT: Route Target (as used in Route Target extended community)

RTC: Route Target Constraint [RFC4684]

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: Transport Class - Best Effort

TE: Traffic Engineering

TEA: Tunnel Encapsulation Attribute (attribute code 23)

TRDB: Transport Route Database

UHP: Ultimate Hop Popping

VRF: Virtual Routing and Forwarding (used with a table)

2.2. Definitions and Notations

BGP CCA:
A BGP attribute that carries community. Examples of BGP CCAs are COMMUNITIES (attribute code 8), EXTENDED COMMUNITIES (attribute code 16), IPv6 Address Specific Extended Community (attribute code 25), and LARGE_COMMUNITY (attribute code 32).

color:0:100 or col-100:
This notation denotes a Color extended community as defined in [RFC9012] with the "Flags" field set to 0 and the "Color Value" 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 NH resolution.

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.

Provider Namespace:
Internal Infrastructure address space in a provider network managed by the operator.

Resolution Scheme:
A construct comprising of an ordered set of TRDBs to resolve NH 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.

Service Prefix:
A destination in "data traffic". Routes to these prefixes are

carried in a Service Family.

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 Flexible Algorithm (Flex-Algo), or SRTE.

Transport Layer:
A layer in the network that contains Transport Tunnels and Transport Families.

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 SNs and 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 that can make choices of new technology or hardware as needed with fewer constraints than brownfield network.

Transport Class:
A construct to group transport tunnels offering similar SLAs (see Section 4.1).

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 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 TRDB that collects its tunnel routes.

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

2.3. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

3. Architecture Overview

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

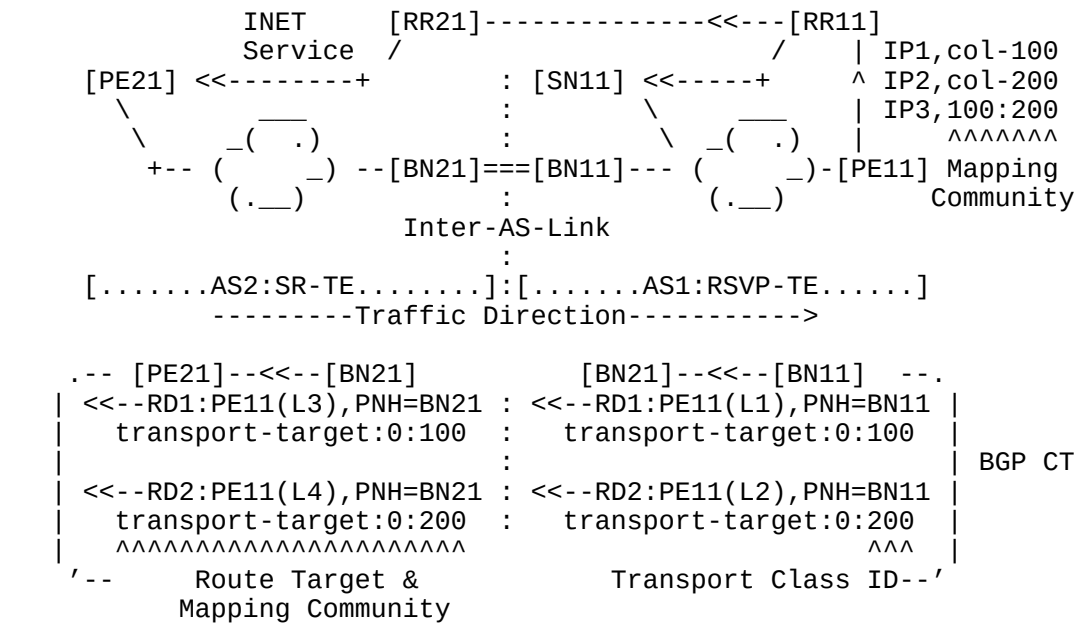


Figure 1: BGP CT Overview with Example Topology

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

* The "Transport Class" construct (see Section 4) to group underlay tunnels.

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

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

Figure 1 depicts the intra-AS and inter-AS application of these constructs. Interactions between SN1 and PE11 describe the intra-AS usage. Interactions between PE21 and PE11 describe the inter-AS usage.

The example topology is an inter-AS option C network (Section 10 of [RFC4364]) with two AS domains; each domain contains tunnels serving two Intents, e.g., 'low-latency' denoted by color 100 and 'high-bandwidth' denoted by color 200. AS1 is an RSVP-TE network; AS2 is an SRTE network. BGP CT and BGP LU are transport families used between the two AS domains. IP1, IP2, and 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 TRDB (see Section 4.2). In Figure 1, RSVP-TE publishes its underlay tunnels into TRDBs created for Transport Classes 100 and 200 at BN11 and SN11 within AS1; Similarly, SR-TE publishes its underlay tunnels into TRDBs created for Transport Classes 100 and 200 at PE21 within AS2.

Resolution Schemes are used to realize Intent. A Resolution Scheme is identified by its "Mapping Community" and contains an ordered list of Transport Classes. Overlay routes carry an indication of the desired Intent using a BGP community, which assumes the role of "Mapping Community".

Egress SN "PE11" advertises service routes with desired Mapping Community, e.g., color:0:100.

For the intra-AS case, SN1 maps this intra-AS route on RSVP-TE tunnels with TC ID 100 by using the Resolution Scheme for color:0:100.

For the inter-AS case, the underlay route in a TRDB is advertised in BGP to extend an underlay tunnel to adjacent domains. A new BGP transport 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 eSN.

The BGP CT address family carries transport prefixes across tunnel domain boundaries. Its design and operation are analogous 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-TC loopback. This is not possible with BGP LU without using per-color loopback. This dissemination 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 NH "self" towards BN21 in AS2 to extend available RSVP-TE tunnels for Transport Classes 100 and 200 in AS1. BN21 propagates these routes with NH "self" to PE21, which resolves the BGP CT routes over SRTE tunnels belonging to same Transport Class, thus forming a BGP CT tunnel for each TC ID at PE21.

PE21 maps the inter-AS service routes received with color:0:100 from AS1 on BGP CT tunnel with TC ID 100 by using the Resolution Scheme for color:0:100. Note that this procedure is same as that followed by SN1 in the intra-AS case.

The following text illustrates how CT architecture provides tiered fallback options at a per-route granularity. Figure 1 shows the Resolution Schemes in use, which make the following NH resolution happen at SN11 (intra-AS) and PE21 (inter-AS) for the service routes of prefixes IP1, IP2, and IP3:

* Resolve IP1 NH over available tunnels in TRDB for Transport Class 100 with fallback to TRDB for best effort.

* Resolve IP2 NH over available tunnels in TRDB for Transport Class 200 with fallback to TRDB for best effort.

* Resolve IP3 NH 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 the SRv6 data plane.

4. Transport Class

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

A Transport Class is uniquely identified 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 cooperating domains may not always agree on the same namespace. Procedures to manage differences in Transport Class ID namespaces between cooperating 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 FRR.

* 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 Explicit Route Object (ERO), SR-TE policy, or BGP policy.

An operator may configure an 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 ID" is encoded as part of the Transport Class RT, which is a new Route Target extended community defined in

Section 4.3.

An SN/BN receiving the transport routes via BGP with sufficient signaling information to identify a Transport Class can associate those tunnel routes with 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 [RFC9830] with "Color:Endpoint" as the NLRI that encodes the Transport Class as an integer 'Color' in its Policy Color field, 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:EP, 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 the Path Computation Element Communication Protocol (PCEP) to signal attributes of an SR Policy that 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 (TRDB)

A 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 NH reachability for EP using the transport routes within the scope of the TRDBs.

Tunnel EP addresses in a TRDB belong to the provider namespace representing the core transport region.

An implementation may realize the TRDB as a "Routing Table" referred to in Section 9.1.2.1 of [RFC4271], which is used only for resolving NH reachability in the 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 an NH.

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

4.3. Transport Class Route Target

This section defines a new type of Route Target called a "Transport Class Route Target extended community" (also known as a "Transport Class RT"). 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 RT is a transitive extended community [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
```

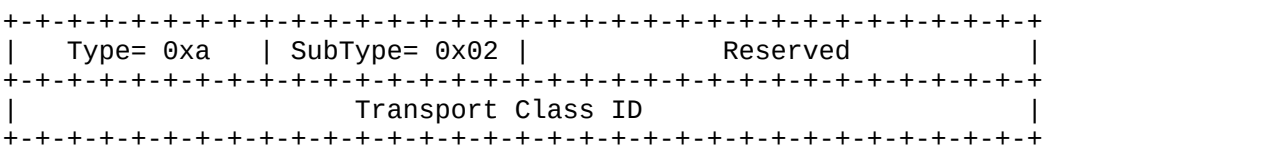


Figure 2: Transport Class RT

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

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

Reserved: A 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 ID", which is an unsigned 32-bit integer.

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

A Transport Class RT with TC ID 100 is denoted as "transport-target:0:100".

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

A BGP speaker that implements procedures described in this document and [RFC4684] MUST also apply the RTC procedures to the Transport Class RT 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 containing a Transport Class RT received from external BGP peers, it is necessary to consider multiple External BGP (EBGP) paths for a given RTC prefix for building the outbound route filter: not just the best path. An implementation MAY provide configuration to control how many EBGP RTC paths are considered.

The Transport Class RT is carried on BGP CT family routes and is used to associate them with appropriate TRDBs at receiving BGP speakers. The Transport Class RT 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 Class RT just like any other BGP community.

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

This document also reserves the non-transitive version of the Transport Class RT (see Section 13.2.1.1.2) for future use. The non-

transitive Transport Class RT is not used. If received, it is considered equivalent in functionality to the transitive Transport Class RT, 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 the Mapping Community in the route.

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

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 in addition to the other roles it may already be playing. For example, the Transport Class RT plays a dual role: as Route Target and 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 a receiving BGP speaker. An implementation SHOULD allow the association of multiple Mapping Communities to a Resolution Scheme. This helps with renumbering and migration scenarios.

An example of a Mapping Community is a Color extended community "color:0:100", described in [RFC9012], or the "transport-target:0:100" described in Section 4.3.

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.

On an overlay route, if more than one Mapping Community exists that map to distinct Resolution Schemes having dissimilar Intents at a receiving node, it is considered a configuration error.

Since a route can carry multiple communities, but only a single Resolution Scheme can be in effect for the route on any given router, it is incumbent on the operator to ensure that communities attached to a route will map to the desired Resolution Scheme at each point in the network.

It should be noted that the Mapping Community role does not require applying Route Target Constraint procedures specified in [RFC4684].

6. BGP Classful Transport Family

The BGP Classful Transport (BGP CT) family uses the existing Address Family Identifier (AFI) of IPv4 or IPv6 and a new SAFI 76 "Classful Transport" that 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 the AFI/SAFI is 1/76, the BGP CT NLRI Prefix consists of an 8-byte RD followed by an IPv4 prefix. When AFI/SAFI is 2/76, the BGP CT 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].

Properties received on a BGP CT route include the Transport Class RT, 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 NH.

6.2. Next Hop Encoding

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

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

When the length of Next hop Address field is 24 (or 48), the next hop address is a 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 of [RFC4659].

When the length of the Next hop Address field is 12, the next hop address is a 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 Types of Encapsulation Information

To ease interoperability between nodes supporting different forwarding technologies, a BGP CT route allows carrying multiple types of 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 MPLS label field (see [RFC8277]). The Implicit NULL label carried in BGP CT route indicates to a receiving node that it should not impose any BGP CT label for this route.

The SID information for SR with respect to the MPLS data plane is carried as specified in the Prefix-SID attribute defined as part of Section 3 of [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 the Tunnel Encapsulation Attribute as specified in [RFC9012].

6.4. Comparison with Other Families Using Encoding from RFC 8277

AFI/SAFI 1/128 (L3VPN) is a family encoded using [RFC8277] 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 a family encoded using [RFC8277] that carries transport prefixes in the NLRI, where the prefixes come from the provider namespace.

AFI/SAFI 1/76 (BGP CT) is a family encoded using [RFC8277] 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 [RFC4364] 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 (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 a "per-prefix" label-allocation scheme for BGP CT prefixes, typically with a number of routes in the hundreds of thousands or less, without affecting SAFI 128 service prefixes, which may represent millions of routes at the time of writing. 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), Virtual Private LAN Service (VPLS) (AFI/SAFI: 25/65), or 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 distribution path of the transport family routes in a network than

the service route distribution path. Service routes (SAFI 128) are exchanged over an EBGP multihop session between ASes with the NH unchanged; whereas BGP CT routes (SAFI 76) are advertised over EBGP single-hop sessions with a "NH self" rewrite over inter-AS links.

The BGP CT SAFI 76 for AFI 1 and 2 is conceptually similar 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 message.

7. Protocol Procedures

This section summarizes the procedures followed by various nodes speaking BGP CT family.

7.1. Preparing the Network to Deploy Classful Transport Planes

It is the 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 RT to identify each Transport Class.

Operators configure the Transport Classes on the SNs and BNs in the network with Transport Class RTs 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 BGP CT 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, a Transport Class RT, and the PNH as the 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 readvertised by the ingress node to BGP peers in neighboring domains.

Alternatively, the ingress node of the tunnel, which is also an ASBR/ ABR in a tunnel’s home domain, may originate the BGP CT route for the tunnel destination with RD:EP in the NLRI, 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.

A unique RD SHOULD be used by the originator of a BGP CT 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 BGP CT 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 ID "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 TRDB is used.

The routes in the TRDBs associated with a 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 the EP in any of the associated TRDBs, the received BGP CT route MUST be considered unresolvable. (See Section 9.1.2.1 of [RFC4271].)

The received BGP CT route MUST be added to the TRDB corresponding to the Transport Class ID "C1" if the TC 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 the EP. Please note that a TRDB is a logical database of tunnel routes belonging to the same Transport Class ID; hence, it only uses the EP as the lookup key (without RD or TC ID).

If no Mapping Community is found on a BGP CT route, the Best-Effort Resolution Scheme is used to resolve the route’s next hop, and the BGP CT route is not added to any TRDB.

7.4. Readvertising BGP CT Route by Border Nodes

This section describes the MPLS label handling when readvertising a BGP CT route with "NH self". When readvertising a BGP CT route with "NH self", a BN allocates an MPLS label to advertise upstream in the BGP CT 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). The MPLS route then pushes received traffic to the transport tunnel or direct interface that the BGP CT 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 "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 a locally configured export policy or via mechanisms such as the ones

described related to BGP Prefix-SID as described in BGP (see [RFC8669]).

7.5. Border Nodes Receiving BGP CT Routes on EBGP

If a route is received with a PNH that is known to be directly connected (for example, an 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 the 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 the 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, BGP ADD-PATH (see [RFC7911]) is helpful.

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

Hence, BGP ADD-PATH (see [RFC7911]) SHOULD be used for the BGP CT 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 Paths

A pair of redundant ABRs, each acting as an RR with the next hop set to itself, may choose each other as the 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] lowers 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 the EP that is not directly connected (for example, an IBGP-route), a Mapping Community (for example, a 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 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 the 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 the 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 the EP in any of the Transport Route Databases, the service route MUST be considered unresolvable. (See Section 9.1.2.1 of [RFC4271]).

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

7.9. Best-Effort Transport Class

It is also possible to represent a 'Best-Effort' SLA as a Transport Class. At the time of writing, 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 the "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 the Transport Class RT that is attached to the BGP CT route that advertises a best-effort tunnel endpoint. Thus, the RT formed is called the "Best-Effort Transport Class RT".

When a BN or SN receives a BGP CT route with Best-Effort Transport Class RT 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 CT 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 specifies the following order of precedence with more-specific scoping of Color preferred to less-specific scoping:

* Color sub-TLV in the Tunnel Encapsulation Attribute.

* Transport Class RT on a BGP CT route.

* Color extended community on a BGP service route.

Color specified in the Color sub-TLV in a TEA is a more-specific indication of "Transport Class ID/Color" than Mapping Community (Transport Class RT) on a BGP CT transport route, which, 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 are 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).

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

7.12. Applicability to IPv6

BGP CT procedures apply equally to IPv4- and IPv6-enabled intra-AS or inter-AS option A, B, and C networks. This section describes the applicability of BGP CT to IPv6 at various layers.

A network that is BGP CT enabled supports IPv6 service families (for example, AFI/SAFI 2/1 or 2/128) and IPv6 transport signaling protocols like SRTEv6, LDPv6, or 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. The 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, and 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, which uses MPLS forwarding for intra-domain tunnels and inter-AS links. The 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

The BGP CT family (AFI/SAFI 2/76) may be used to set up inter-domain tunnels of a certain Transport Class when using a 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 and illustrated in a separate document (see [BGP-CT-SRv6]). As noted in that document, a BGP CT route update for SRv6 includes a BGP attribute containing SRv6 SID information (e.g.,

a BGP Prefix-SID [RFC9252]) with the Transposition Scheme disabled.

7.14. Error-Handling Considerations

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

If a BGP speaker considers a received "Transport Class" extended community (the 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, the 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 IP address ranges as described in [RFC6890]. 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.

Though this section illustrates the use of IPv4, as described in Section 7.12, these procedures work equally for IPv6 as well.

8.1. Reference Topology

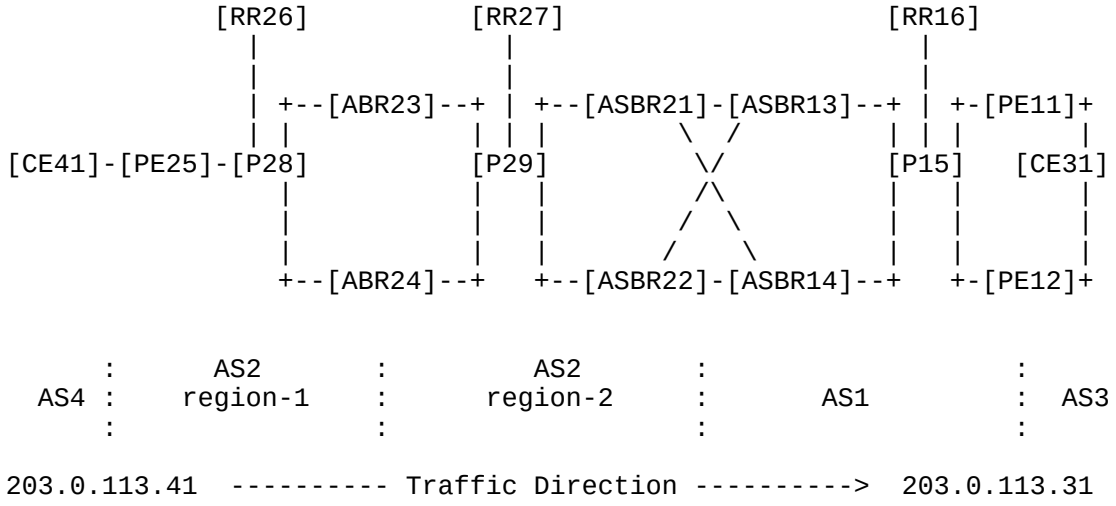


Figure 3: Multi-Domain BGP CT Network

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

AS2 is further divided into two regions. There are three tunnel domains in the provider’s space:

* AS1 uses ISIS Flex-Algo (see[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 and "Bronze" with Transport Class ID 200. These Transport Classes are provisioned at the PEs and the Border nodes (ABRs and ASBRs) in the network.

The following tunnels exist for the 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 Flex-Algo tunnel

* ASBR14_to_PE11_bronze - ISIS Flex-Algo tunnel

These tunnels are either provisioned or autodiscovered to belong to Transport Class IDs 100 or 200.

8.2. Service Layer Route Exchange

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

The PEs see each other as the next hop in the BGP UPDATE message for the service family routes. BGP ADD-PATH send and receive are enabled on both directions on the EBGP multihop session between RR16 and RR26 for AFI/SAFIs: 1/1 and 1/128. BGP ADD-PATH send is negotiated in the RR to PE direction in each AS. This is to avoid the path-hiding service routes at the 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 the same RD.

Forwarding happens using service routes installed at service nodes PE25, PE11, and 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 the next hop as itself to PE11 and PE12. CE31 can attach a Mapping Community color:0:100 on this route to indicate its request for a Gold SLA. Or, PE11 can attach the same using locally configured policies.

Consider CE31 getting VPN service from PE11. The RD1:203.0.113.31 route is readvertised in AFI/SAFI 1/128 by PE11 with the next hop set to itself (192.0.2.11) and label V-L1 to RR16 with the Mapping

Community color:0:100 attached. RR16 advertises this route with the BGP ADD-PATH ID set to RR26, which readvertises to PE25 with the 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/SAFIs: 1/1 and 1/128 reach PE25 via RR16, RR26 with the next hop unchanged, as PE11 or PE12.

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

8.3. Transport Layer Route Propagation

Egress nodes PE11 and PE12 negotiate a BGP CT family with transport ASBRs ASBR13 and ASBR14. These egress nodes originate BGP CT routes for tunnel endpoint addresses that are advertised as a 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; the Bronze SLA transport plane is used to highlight the path-hiding aspects.

For Gold tunnels, PE11 is provisioned with Transport Class having TC ID 100, RD value 192.0.2.11:100, and a transport-target:0:100. For Bronze tunnels, PE11 is provisioned with Transport Class 200, RD value 192.0.2.11:200, and transport-target:0:200. Similarly, for Gold tunnels, PE12 is provisioned with Transport Class having TC ID 100, RD value 192.0.2.12:100, and a transport-target:0:100. For Bronze tunnels, PE12 is provisioned with Transport Class having TC ID 200, RD value 192.0.2.12:200, and transport-target:0:200. Note that, in this example, the BGP CT routes carry only the Transport Class RT 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 Class RT and unique RDs.

ASBR13 and ASBR14 negotiate BGP CT family with transport ASBRs ASBR21 and ASBR22 in neighboring ASes. ASBR21 and ASBR22 negotiate BGP CT family with RR27 in region 2, which reflects BGP CT routes to ABR23 and ABR24. ABR23 and ABR24 negotiate BGP CT family with ingress node PE25 in region 1. The BGP LU family is also negotiated on these sessions alongside the BGP CT family. The BGP LU family carries Best-Effort Transport Class routes; BGP CT carries Gold and Bronze Transport Class routes.

PE11 is provisioned to originate a BGP CT route for endpoint PE11, with a Gold SLA. 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 a 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, the next hop set to itself, and transport-target:0:100. An 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 the BGP CT family to ASBRs ASBR21 and 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 set to itself, and transport-target:0:100. An 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, the BGP CT route with a Bronze SLA to endpoint PE11 is originated by PE11 with an NLRI containing RD prefix 192.0.2.11:200:192.0.2.11 and an appropriate label. The use of distinct RDs for Gold and Bronze allows both Gold and Bronze advertisements to traverse path-selection pinch points 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 the 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 and ASBR14 and can compute ECMP/FRR towards them. ASBR21 readvertises the BGP CT route for 192.0.2.11:100:192.0.2.11 with a next hop set to itself (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 labels B-L1 and B-L2 and forward to ASBR13 and ASBR14 respectively; this is an ECMP route. RR27 readvertises this BGP CT route to ABR23 and ABR24 with the 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 and ASBR14, and it readvertises with the next hop set to itself (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 labels B-L1 and B-L2 and forward to ASBR13 and ASBR14, respectively. RR27 also readvertises this BGP CT route to ABR23 and ABR24 with the label and next hop unchanged.

BGP ADD-PATH is enabled for the BGP CT family on the sessions between RR27 and the ASBRs and 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 and ABR24 without any path hiding. Thus, ABR23 is given visibility of both available next hops for the Gold SLA.

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

ABR23 also receives the route with next hop 192.0.2.22 and label B-L4 from RR27. The transport-target:0:100 on this route acts as the Mapping Community and instructs ABR23 to strictly resolve the next hop using routes in TC 100 TRDB only. ABR23 successfully resolves the next hop to point to ABR23_to_ASBR22_gold tunnel. ABR23 readvertises this BGP CT route with the next hop set to itself (loopback address 192.0.2.23) and a new label B-L5 to PE25. A 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. 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 set to itself towards PE25.

Thus formed, the BGP CT LSP 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 the 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 the destination 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 labels V-L1 (innermost), B-L5, and top label PE25_to_ABR23_gold tunnel. This MPLS packet is thus transmitted to ABR23 using the 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 the Gold SLA.

ASBR22 similarly performs a lookup for label B-L4 in the global MPLS FIB, finds the route that swaps label B-L4 with label B-L2, and forwards it 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 the 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 the 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 the 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 the ASBR22_to_ASBR13 link goes down, traffic with a Gold SLA going to PE11 will need 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 an 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 the 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 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 a Best-Effort Transport Class as the backup path, this withdrawal of a 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 the 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 (PIC). The repair time will be proportional to time taken for withdrawing the BGP CT route.

These examples demonstrate the various levels of fail-safe 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 resources for MPLS labels and next hops in the network. An ASBR of a domain can be provisioned to allow routes with only the Transport Class RTs 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 (PNHs) advertised to an SN or BN can be constrained using BGP Classful Transport and RTC (see [RFC4684].

An egress SN MAY advertise a BGP CT route for RD:eSN with two Route Targets: transport-target:0:<TC> and an RT carrying <eSN>:<TC>, where TC is the Transport Class identifier and eSN is the IP address used by the 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 the RT of transport-target:0:<TC> is used, the pruning happens in granularity of Transport Class ID (Color), not BGP next hop; a BGP CT route will only be advertised into a domain with at least one PE that imports 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 the BGP extended community attribute (attribute code 16).

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

An ingress SN MAY import BGP CT routes with a Route Target carrying <eSN>:<TC>. The ingress SN may learn the eSN values by configuration or it may discover them from the BGP next hop field in the BGP VPN service routes received from the eSN. A BGP ingress SN receiving a BGP service route with a next hop of eSN generates an 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 RTC is in use, as described here, BGP CT routes will be constrained to follow the same path of propagation as the RTC routes. Therefore, 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.

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 helps with the scaling of MPLS forwarding state at the 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 confines 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 the 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 Operations, Administration, and Maintenance (OAM) procedures specified in [RFC8029] also apply to BGP CT.

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

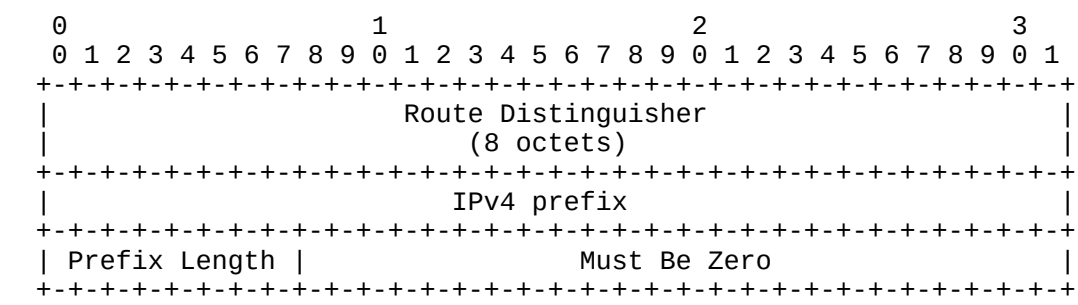


Figure 4: BGP CT IPv4 FEC

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

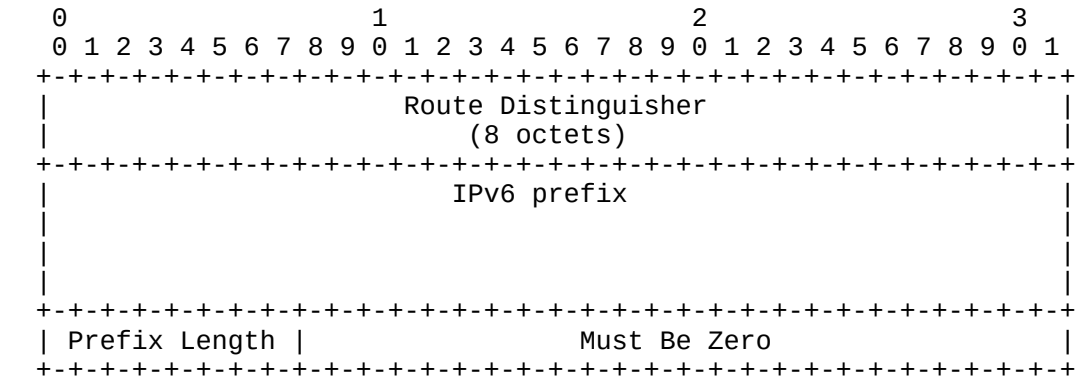


Figure 5: BGP CT IPv6 FEC

These prefix layouts are inherited from Sections 3.2.5 and 3.2.6 of [RFC8029].

10.2. Usage of RD 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. An 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., a Label) in order to collect traffic statistics at the SN for each BN. In the absence of an 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 be either 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 the next hop set to itself by an SN or a BN. An implementation may use different label-allocation modes with BGP CT. Per-prefix is the recommended label-allocation mode as it provides better traffic convergence properties than a per-NH 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 have the next hop set to themselves. 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 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:

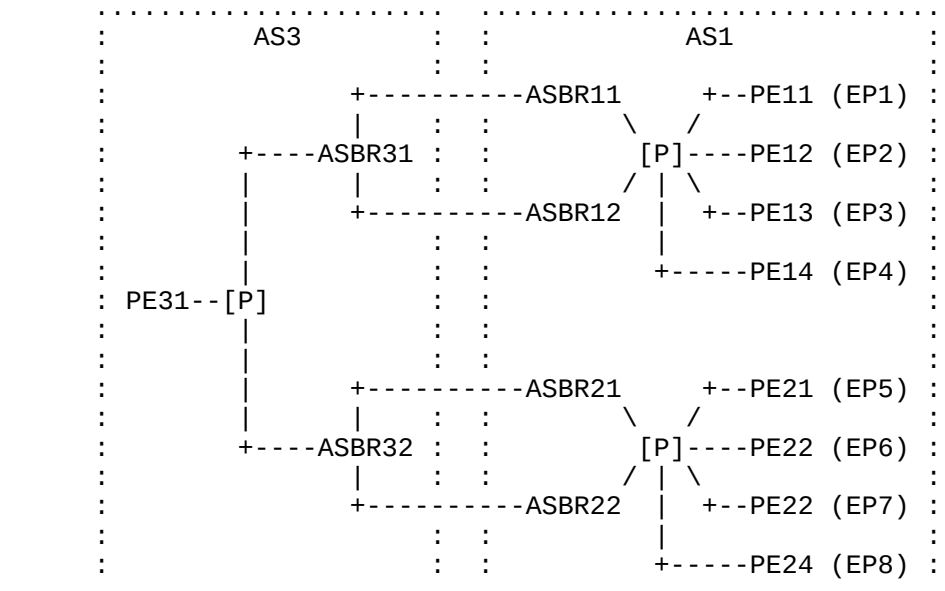


Figure 6: Managing Transport-Route Visibility in Multi-Domain Networks

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-Modes), and different Per-Prefix label-allocation modes (PP-Modes).

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

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 Class RT 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 BGP CT 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 RT is useful to avoid collision with regular Route Target namespace used by service routes.

11.2. Managing Intent at Service and Transport Layers

Section 8 shows multiple domains that agree on a color namespace (Agreeing Color Domains) and contain tunnels with an 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 cooperating 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.

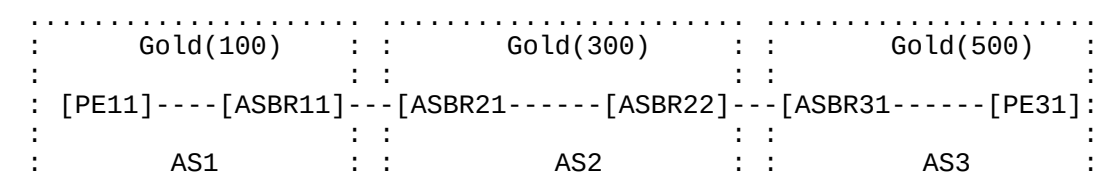
As explained in Section 5, a Mapping Community carried on a service route maps to a Resolution Scheme. The Mapping Community values for the service route can be abstract and are not required to match the transport color namespace. This abstract Mapping Community value representing a global service layer 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.



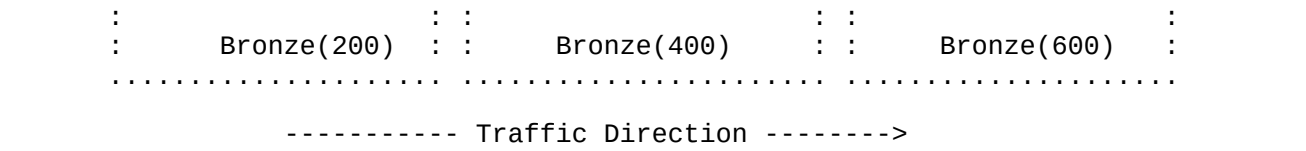


Figure 8: Transport Layer with Non-agreeing Color Domains

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

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

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

* In AS3, the 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; in AS2, it maps to TRDB with color 300; and in AS1, it maps to 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 RT 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 rewritten 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 Class RT rewrite requires coordination of color values between domains in the transport layer. This method avoids the need to rewrite 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-domain 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.

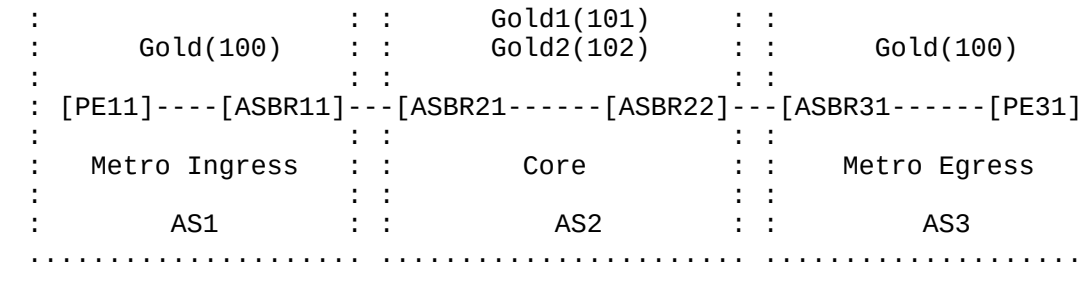


Figure 9: Transport Layer with Heterogeneous Color Domains

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

* In AS1, the Gold SLA is represented by color 100.

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

* In AS3, the Gold SLA is represented by color 100.

This problem can be solved by the two approaches described in Sections 11.2.3.1 and 11.2.3.2.

11.2.3.1. Duplicate Tunnels Approach

In this approach, duplicate tunnels that satisfy the 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 colors 101 and 102.

Overlay routes received with a Mapping Community (e.g., transport-target or color community) can resolve over these tunnels in the TRDB with matching colors 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, SNs/BNs in all three ASes provision the transport classes 100, 101, and 102. SNs and BNs in AS1 and AS3 are provisioned with customized Resolution Schemes that resolve routes

with transport-target:0:101 or transport-target:0:102 using color 100 TRDB.

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

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

The following description explains the remaining procedures with color 101 as an example.

At ASBR31, the Route Target role of transport-target:0:101 on this BGP CT route gives instruction to add the route to color 101 TRDB. ASBR31 is provisioned with a customized Resolution Scheme that resolves the routes carrying Mapping Community transport-target:0:101 to resolve using color 100 TRDB. This route is then readvertised 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 readvertised 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 readvertised to ASBR11 with transport-target:0:101.

At ASBR11, the Route Target role of transport-target:0:101 on this BGP CT route gives instruction 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 readvertised from color 101 TRDB to PE11 with transport-target:0:101.

At PE11, the Route Target role of transport-target:0:101 on this BGP CT route gives instruction 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.

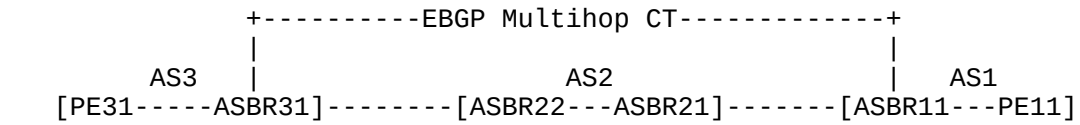
Similar processing is done for color 102 routes also at ASBR31, ASBR22, ASBR21, ASBR11, and PE11.

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 an 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.



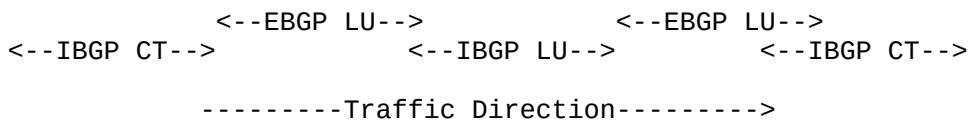


Figure 10: BGP CT in AS1 and AS3 Connected by BGP LU in AS2

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 the 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 the 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 they are advertised between ASBR31 and ASBR11 with the next hop set to themselves.

In AS2, which does not support BGP CT, a separate loopback may be used on ASBR22 and ASBR21 to represent Gold and Bronze SLAs, namely 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

The RD:PE11 BGP CT route is originated from PE11 towards ASBR11 with transport-target 'gold.' ASBR11 readvertises this route with the next hop set to ASBR11_lpbk_gold on the EBGP multihop session towards ASBR31. ASBR11 originates a BGP LU route for endpoint ASBR11_lpbk_gold on an EBGP session to ASBR21 with a 'gold SLA' community and a 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 the 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 the next hop set to itself and a new label.

ASBR31 adds the ASBR11_lpbk_gold route received via EBGP LU from ASBR22 to a '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 the 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 the next hop set to itself, thus stitching with the BGP LU LSP in AS2. Intra-domain traffic forwarding in AS1 and AS3 follows the procedures as explained in 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 an ASBR21 loopback assigned to the available SLA. This eases migration in case of a heterogeneous color domain 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 when using the BGP CT family.

11.3.2.1. Interoperation Between MPLS and SRv6 Nodes

BGP speakers may carry MPLS labels and SRv6 SIDs in BGP CT SAFI 76 for AFI 1 or 2 routes using protocol encoding as described in Section 6.3.

MPLS labels are carried using the encoding described in [RFC8277], and SRv6 SIDs are carried using the Prefix-SID attribute as specified in Section 7.13.

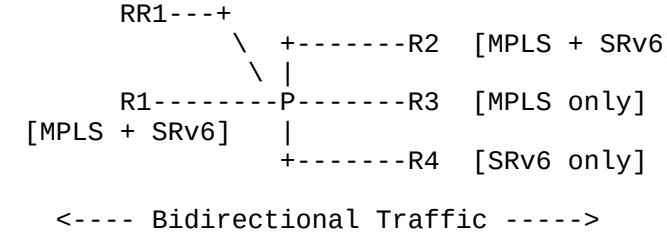


Figure 11: BGP CT Interoperation Between MPLS and SRv6 Nodes

This example shows a provider network with a mix of devices that have 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 a BGP session with Route Reflector RR1, which reflects routes between these nodes with the next hop unchanged. The BGP CT family is negotiated on these sessions.

R1 and R2 send and receive both MPLS labels and SRv6 SIDs in the BGP CT control plane routes. This allows them to be ingress and egress

for both MPLS and SRv6 data planes. The MPLS label is carried using the encoding described in [RFC8277], and an SRv6 SID is carried using the Prefix-SID attribute as specified in Section 7.13 without the Transposition Scheme. In this way, either MPLS or SRv6 forwarding can be used between R1 and R2.

R1 and R3 send and receive an MPLS label in the BGP CT control plane routes using the encoding described in [RFC8277]. This allows them to be ingress and egress for MPLS data plane. R1 will carry an SRv6 SID in the Prefix-SID attribute, which will not be used by R3. In order to interoperate with MPLS-only device R3, R1 MUST NOT use the 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 SIDs in the BGP CT control plane routes using the BGP Prefix-SID attribute, without a Transposition Scheme. This allows them to be ingress and egress for the SRv6 data plane. R4 will carry the special MPLS label with a value of 3 (Implicit NULL) in the encoding described in [RFC8277], which tells R1 not to push any MPLS label for this BGP CT route towards R4. The MPLS label advertised by R1 in an NLRI as described in [RFC8277] will not be used by R4. SRv6 forwarding will be used between R1 and R4.

Note that, in this example, 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 and 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 the encoding described in [RFC8277], and UDP (or IP) tunneling information is carried using the TEA attribute or the Encapsulation extended community as specified in [RFC9012].

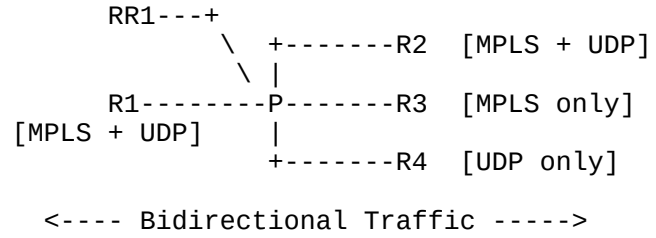


Figure 12: BGP CT Interop Between MPLS and UDP Tunneling Nodes

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 the next hop unchanged. The BGP CT family is negotiated on these sessions.

R1 and R2 send and receive both MPLS labels 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. The MPLS label is carried using the encoding described in [RFC8277]. As specified in [RFC9012], UDP tunneling information is carried using the Tunnel Encapsulation Attribute (attribute code 23) or the "barebones" Tunnel TLV carried in Encapsulation extended community. Either MPLS or UDP tunnel forwarding can be used between R1 and R2.

R1 and R3 send and receive MPLS labels in the BGP CT control plane routes using the encoding described in [RFC8277]. This allows them to be ingress and egress for MPLS data plane. R1 will carry UDP tunneling info in the TEA, 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 the BGP TEA. This allows them to be ingress and egress for UDP tunneled data plane. R4 will carry MPLS special label 3 (Implicit NULL) in the encoding described in [RFC8277], 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 that, in this example, 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 and 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.

[INTAREA-TUNNELS] discusses these considerations in more detail.

11.5. Use of DSCP

BGP CT specifies procedures for Intent-Driven Service Mapping in a service provider network and defines the '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 the form of a DSCP (see [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 the transport layer.

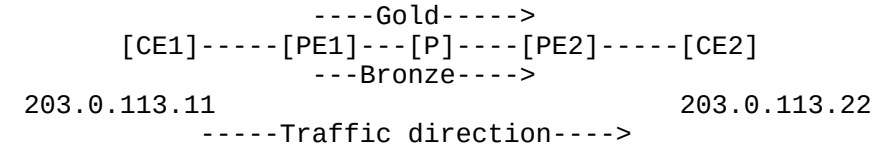


Figure 13: Example Topology with DSCP on PE-CE Links

Let PE1 be configured to map DSCP1 to the Gold TC and DSCP2 to the Bronze TC. Based on the DSCP received on the IP traffic from the 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. 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 devices by out-of-band mechanisms. This allows the administrator of CE1 to discover what Transport Classes exist in

the provider network and which DSCP 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 based on the DSCP peering agreement with CE2.

12. Applicability to Network Slicing

In Network Slicing, the IETF Network Slice Controller (NSC) is responsible for customizing and setting up the underlying transport (e.g., RSVP-TE, SRTE tunnels with desired characteristics) and resources (e.g., policies/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 [RFC9543].

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

13.1. New BGP SAFI

IANA has assigned BGP SAFI code 76 for the "Classful Transport (CT)" SAFI.

Registry Group: Subsequent Address Family Identifiers (SAFI) Parameters

Registry Name: SAFI Values

Value	Description	Reference
76	Classful Transport (CT)	RFC 9832

Table 1

This will be used to create new AFI/SAFI pairs for IPv4 and IPv6 BGP CT families, namely:

* IPv4 BGP CT: AFI/SAFI = 1/76, for carrying IPv4 prefixes.

* IPv6 BGP CT: AFI/SAFI = 2/76, for carrying IPv6 prefixes.

13.2. New Format for BGP Extended Community

IANA has assigned a Format type (Type high = 0xa) of Extended Community [RFC4360] for the Transport Class from the following registries in the "Border Gateway Protocol (BGP) Extended Communities" registry group:

* the "BGP Transitive Extended Community Types" registry and

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

The same low-order six bits have been assigned 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".

Following [RFC7153], assignments in the following subsections have been made.

13.2.1. Existing Registries

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

Table 2

(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

Table 3

(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 has added the following subregistry under the "Border Gateway Protocol (BGP) Extended Communities" registry group:

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 Procedure
0x00-0xbf	First Come First Served
0xc0-0xff	IETF Review

Table 4

Sub-Type Value	Name
0x02	Route Target

Table 5

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

IANA has added the following subregistry under the "Border Gateway Protocol (BGP) Extended Communities" registry group:

Registry Group: Border Gateway Protocol (BGP) Extended Communities

Registry Name: Non-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 0x4a.

Range	Registration Procedure
0x00-0xbf	First Come First Served
0xc0-0xff	IETF Review

Table 6

Sub-Type Value	Name
0x02	Route Target

Table 7

13.3. MPLS OAM Code Points

The following two code points have been assigned 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

Table 8

14. Transport Class ID Registry

This RFC documents the "Transport Class ID" registry and its assigned values. The value ranges in this registry are either assigned by this document or reserved for Private Use. Because the registry is complete, it is being published in this RFC rather than as an IANA-maintained registry. However, note that IANA-related terminology [BCP26] is used here.

Registry Name: Transport Class ID

The Registration Procedures are as follows:

Value	Registration Procedure
0	IETF Review
1-4294967295	Private Use

Table 9

As shown in the table below, the Transport Class ID value 0 is Reserved to represent the "Best-Effort Transport Class ID". This is used in the ’Transport Class ID’ field of a Transport Class RT that represents the Best-Effort Transport Class.

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

Table 10

As noted in Sections 4 and 7.10, ’Transport Class ID’ is interchangeable with ’Color’. For purposes of backward compatibility with usage of a ’Color’ field in a Color extended community as specified in [RFC9012] and [RFC9256], the range 1-4294967295 uses ’Private Use’ as the Registration Procedure.

15. Security Considerations

This document uses the mechanisms from [RFC4760] 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. These routes stay in the part of network where the new address family is negotiated and don’t leak out into the Internet.

Furthermore, procedures defined in Section 9.1 mitigate the risk of unintended propagation of BGP CT routes across inter-AS boundaries even when a BGP CT family is negotiated. BGP import and export policies are used to control the BGP CT reachability information exchanged across AS boundaries. This mitigates the risk of advertising internal loopback addresses outside the administrative control of the provider network.

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 the TCP Authentication Option [RFC5925] and the Generalized TTL Security Mechanism [RFC5082].

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 the MPLS data plane; hence, its security considerations apply.

BGP CT routes distribute SRv6 SIDs for SRv6 data planes; hence, the security considerations of Section 9.3 of [RFC9252] apply. Moreover, the 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 route 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, BGPsec solutions ([RFC8205] and Origin Validation [RFC8210][RFC6811]) may be extended in future to work for non-Internet SAFIs (SAFIs other than 1).

The restriction of the applicability of the BGP CT SAFI 76 to its intended well-defined scope and utilizing [RFC8212] limits the likelihood of traffic diversions.

16. References

16.1. 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>.

[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>.

[RFC9830] Previdi, S., Filsfils, C., Talaulikar, K., Ed., Mattes, P., and D. Jain, "Advertising Segment Routing Policies in BGP", RFC 9830, DOI 10.17487/RFC9830, September 2025, <https://www.rfc-editor.org/info/rfc9830>.

16.2. Informative References

[BCP26] Best Current Practice 26, <https://www.rfc-editor.org/info/bcp26>. At the time of writing, this BCP comprises the following:

Cotton, M., Leiba, B., and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 8126, DOI 10.17487/RFC8126, June 2017, <https://www.rfc-editor.org/info/rfc8126>.

[BGP-CT-SRv6] Vairavakkalai, K., Ed. and N. Venkataraman, Ed., "BGP CT - Adaptation to SRv6 dataplane", Work in Progress, Internet-Draft, draft-ietf-idr-bgp-ct-srv6-07, 22 August 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-bgp-ct-srv6-07>.

[BGP-FWD-RR] Vairavakkalai, K., Ed. and N. Venkataraman, Ed., "BGP Route Reflector with Next Hop Self", Work in Progress, Internet-Draft, draft-ietf-idr-bgp-fwd-rr-04, 22 August

2025, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-bgp-fwd-rr-04>.

[BGP-LU-EPE]
Gredler, H., Ed., Vairavakkalai, K., Ed., R, C., Rajagopalan, B., Aries, E., and L. Fang, "Egress Peer Engineering using BGP-LU", Work in Progress, Internet-Draft, draft-gredler-idr-bgplu-epe-16, 14 October 2024, <https://datatracker.ietf.org/doc/html/draft-gredler-idr-bgplu-epe-16>.

[FLOWSPEC-REDIR-IP]
Haas, J., Henderickx, W., and A. Simpson, "BGP Flow-Spec Redirect-to-IP Action", Work in Progress, Internet-Draft, draft-ietf-idr-flowspec-redirect-ip-04, 2 September 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-flowspec-redirect-ip-04>.

[INTAREA-TUNNELS]
Touch, J. D. and M. Townsley, "IP Tunnels in the Internet Architecture", Work in Progress, Internet-Draft, draft-ietf-intarea-tunnels-15, 9 May 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-intarea-tunnels-15>.

[Intent-Routing-Color]
Hegde, S., Rao, D., Uttaro, J., Bogdanov, A., and L. Jalil, "Problem statement for Inter-domain Intent-aware Routing using Color", Work in Progress, Internet-Draft, draft-hr-spring-intentaware-routing-using-color-04, 31 January 2025, <https://datatracker.ietf.org/doc/html/draft-hr-spring-intentaware-routing-using-color-04>.

[MNH]
Vairavakkalai, K., Ed., Jeganathan, J. M., Nanduri, M., and A. R. Lingala, "BGP MultiNexthop Attribute", Work in Progress, Internet-Draft, draft-ietf-idr-multinexthop-attribute-04, 25 March 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-multinexthop-attribute-04>.

[MPLS-NS]
Vairavakkalai, K., Ed., Jeganathan, J. M., Ramadenu, P., and I. Means, "BGP Signaled MPLS Namespaces", Work in Progress, Internet-Draft, draft-kaliraj-bess-bgp-mpls-namespaces-01, 21 August 2025, <https://datatracker.ietf.org/doc/html/draft-kaliraj-bess-bgp-mpls-namespaces-01>.

[PACKING-TEST]
"update-packing-test-results.txt", 1a75d4d, 25 June 2023, <https://github.com/ietf-wg-idr/draft-ietf-idr-bgp-ct/blob/main/update-packing-test-results.txt>.

[PCEP-RSVP-COLOR]
Rajagopalan, B., Beeram, V. P., Peng, S., Koldychev, M., and G. S. Mishra, "Path Computation Element Protocol (PCEP) Extension for Color", Work in Progress, Internet-Draft, draft-ietf-pce-pcep-color-12, 26 February 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-pce-pcep-color-12>.

[PCEP-SRPOLICY]
Koldychev, M., Sivabalan, S., Sidor, S., Barth, C., Peng, S., and H. Bidgoli, "Path Computation Element Communication Protocol (PCEP) Extensions for Segment Routing (SR) Policy Candidate Paths", Work in Progress, Internet-Draft, draft-ietf-pce-segment-routing-policy-cp-

27, 4 April 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-pce-segment-routing-policy-cp-27>.

[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>.

[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>.

[RFC8210] Bush, R. and R. Austein, "The Resource Public Key Infrastructure (RPKI) to Router Protocol, Version 1", RFC 8210, DOI 10.17487/RFC8210, September 2017, <https://www.rfc-editor.org/info/rfc8210>.

[RFC9256] Filsfils, C., Talaulikar, K., Ed., Voyer, D., Bogdanov, A., and P. Mattes, "Segment Routing Policy Architecture", RFC 9256, DOI 10.17487/RFC9256, July 2022, <https://www.rfc-editor.org/info/rfc9256>.

[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>.

[RFC9543] Farrel, A., Ed., Drake, J., Ed., Rokui, R., Homma, S., Makhijani, K., Contreras, L., and J. Tantsura, "A Framework for Network Slices in Networks Built from IETF Technologies", RFC 9543, DOI 10.17487/RFC9543, March 2024, <https://www.rfc-editor.org/info/rfc9543>.

Appendix A. Extensibility Considerations

A.1. Signaling Intent over a 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.

Appendix A.10 of [MNH] describes some mechanisms that enable such signaling.

A.2. BGP CT Egress TE

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

The Peer/32 or Peer/128 EPE route MAY be originated in the BGP CT family with the 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 Section 10 of [RFC4364], 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.

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 and 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 and option B deployment scenarios.

B.1. Intra-AS Use Case

B.1.1. Topology

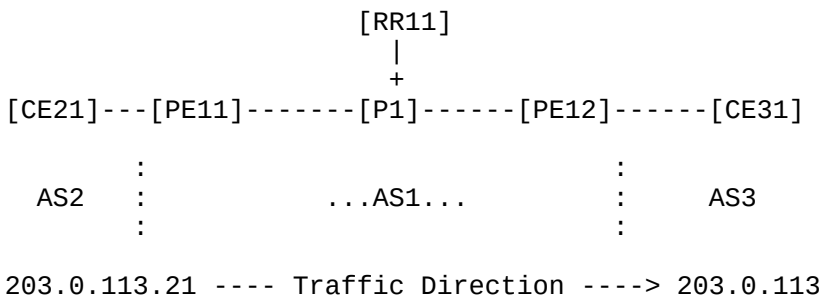


Figure 14: BGP CT Intra-AS

Figure 14 shows a provider network Autonomous System, AS1. It serves customers AS2 and 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 it uses LDP tunnels for best-effort traffic.

The network has two TCs: Gold with TC ID 100 and Bronze with TC ID 200. These TCs are provisioned at the PEs. This creates the Resolution Schemes for these TCs at these PEs.

The following tunnels exist for the Gold TC:

* PE11_to_PE12_gold - RSVP-TE tunnel

* PE12_to_PE11_gold - RSVP-TE tunnel

The following tunnels exist for Bronze TC:

* 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 and 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 the 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 and 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 the next hop set to itself to PE12. CE31 can attach a Mapping Community color:0:100 on this route to indicate its request for a Gold SLA. Or, PE12 can attach the same using locally configured policies.

Consider CE31 getting VPN service from PE12. The RD:203.0.113.31 route is readvertised in AFI/SAFI 1/128 by PE12 with the next hop set to itself (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 the 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 the 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 Use Case

B.2.1. Topology

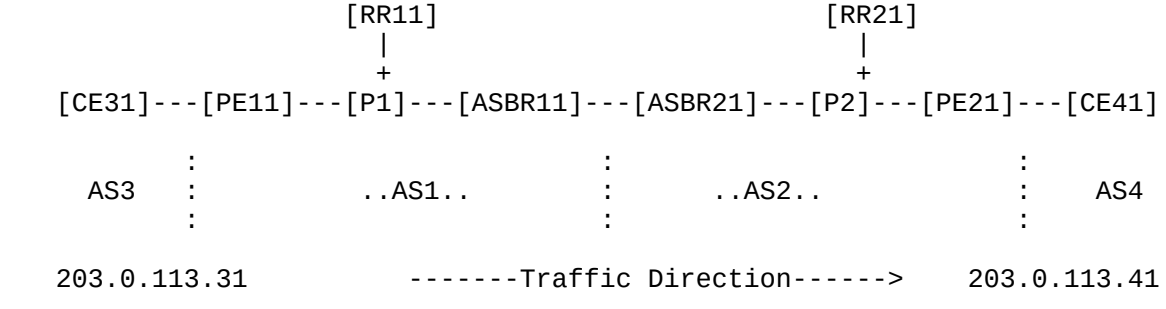


Figure 15: BGP CT Inter-AS option A

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 TCs: Gold with TC ID 100, Bronze with TC ID 200. These TCs are provisioned at the PEs and ASBRs. This creates the Resolution Schemes for these TCs at these PEs and ASBRs.

Following tunnels exist for Gold TC.

* 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 TC.

* 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 TC 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.

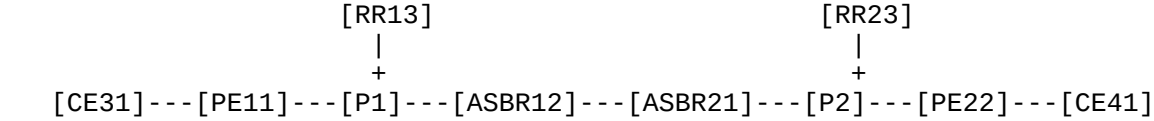
This route is readvertised with the next hop set to itself by ASBR21 to ASBR11 on a BGP session in the VRF. The single-hop EBGP session endpoints are interface addresses. ASBR21 and ASBR11 act like a CE to each other. The previously mentioned process repeats in AS1 until the route reaches PE11 and resolves over the PE11_to_ASBR11_gold RSVP TE tunnel.

Traffic traverses as an unlabeled IP packet on the following legs: CE31-PE11, ASBR11-ASBR21, PE21-CE41. And it uses MPLS forwarding inside the AS1 and 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 an end-to-end SLA.

B.3. Inter-AS Option B Use Case

B.3.1. Topology



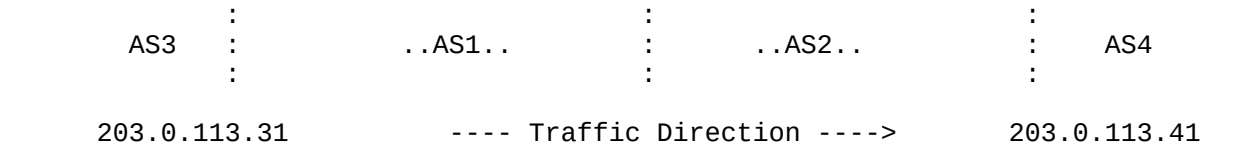


Figure 16: BGP CT Inter-AS option B

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 along with L-ISIS for best-effort tunnels.

The networks have two TCs: Gold with TC ID 100 and Bronze with TC ID 200. These TCs are provisioned at the PEs and ASBRs. This creates the Resolution Schemes for these Transport Classes at these PEs and ASBRs.

The following tunnels exist for Gold TC:

* 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

The following tunnels exist for Bronze TC:

* 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 TC 100 or 200.

B.3.3. Service Layer Route Exchange

Service nodes PE11 and 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 the next hop unchanged.

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

ASBR21 and ASBR12 negotiate AFI/SAFI 1/128 between them and readvertise L3VPN routes with the next hop set to themselves, allocating new labels. The single-hop EBGP session endpoints are interface addresses.

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

Consider CE41 getting VPN service from PE22. The RD:203.0.113.41 route is readvertised in AFI/SAFI 1/128 by PE22 with the next hop set to itself (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 the next hop set to itself 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 the next hop set to itself, 192.0.2.12. PE11 resolves the next hop 192.0.2.12 over PE11_to_ASBR12_gold RSVP TE tunnel.

Traffic traverses as the IP packet on the following legs: CE31-PE11 and PE21-CE41. And it uses MPLS forwarding on the ASBR11-ASBR21 link and inside the 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 an end-to-end SLA.

Appendix C. Why reuse RFCs 8277 and 4364?

[RFC4364] is one of the key design patterns produced by the networking industry. It introduced virtualization and allowed sharing of resources in the service provider space with multiple tenant networks, providing isolated and secure Layer 3 VPN services. This design pattern has been reused since to provide other service layer virtualizations like Layer 2 virtualization (VPLS, L2VPN, EVPN), ISO virtualization, ATM virtualization, and Flowspec VPN.

It is to be noted that these services have different NLRI encodings. The L3VPN service family that binds the MPLS label to an IP prefix uses the encoding described in [RFC8277] and others define different NLRI encodings.

BGP CT reuses the procedures described in [RFC4364] to slice a transport network into multiple transport planes that different service routes can bind to using color.

BGP CT reuses [RFC8277] because it precisely fits the purpose. That is, in an MPLS network, BGP CT needs to bind the 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 the RD:IP_Prefix and carrying a Label for it as specified in [RFC8277] works well for this purpose.

Another advantage of using the precise encoding as defined in [RFC4364] and [RFC8277] is that it allows interoperation 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 the future, if [RFC8277] 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 ([RFC4364] and [RFC8277]) and will adapt to any changes mechanisms from [RFC8277] undergo in future.

This approach leverages the benefits of time-tested design patterns proposed in [RFC4364] and [RFC8277]. Moreover, this approach greatly reduces operational training costs and protocol compatibility considerations as it complements and works well with existing protocol machinery: this problem does not need a brand new NLRI and procedures.

BGP CT design also avoids overloading the NLRI MPLS label field from [RFC8277] with information related to the non-MPLS data plane because it 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 the same BGP UPDATE message. Update packing in BGP CT will be similar to family routes from [RFC8277] 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.

Section 6.3.2.1 of [Intent-Routing-Color] suggests scaling numbers for a 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 an MPLS label were tested. Tests with BGP CT IPv6 endpoints and SRv6 SID are planned.

Tests were conducted with a 1.9 million BGP CT route scale (387K endpoints in 5 TCs). Initial convergence time for all cases was less than 2 minutes, which compares favorably with user expectation for such a scale. This experiment proves that carrying Transport Class information as an attribute keeps BGP convergence within an acceptable range. Details of the experiment and test results are available in [PACKING-TEST].

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

Appendix D. Scaling Using BGP MPLS Namespaces

This document considers the scaling scenario suggested in Section 6.3.2.1 of [Intent-Routing-Color] where 300K nodes exist in the network with 5 TCs.

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.

Acknowledgements

The authors thank Jeff Haas, John Scudder, Susan Hares, Dongjie (Jimmy), Moses Nagarajah, Jeffrey (Zhaohui) Zhang, Joel Halpern, 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, and 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.

Thanks to John Scudder for showing us with example how the Figures can be enhanced using SVG format.

Contributors

The following people contributed substantially to the content of this document and should be considered coauthors:

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

We also acknowledge the contribution of the following individuals:

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

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