Network Working Group
Internet-Draft
Intended status: Experimental
Expires: 28 December 2023

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

26 June 2023

BGP Classful Transport Planes
draft-ietf-idr-bgp-ct-09

Abstract

This document specifies ~~a~~an experimental mechanism, referred to as "Intent Driven Service Mapping" to express association of overlay routes with underlay routes satisfying a certain Service Level Agreement (SLA) using BGP. The document describes a framework for classifying underlay routes into transport classes and mapping service routes to a specific transport class.

The "Transport class" construct maps to a desired SLA and can be used to realize the "Topology Slice" in 5G Network slicing architecture. This document specifies BGP ~~protocol~~ procedures that enable dissemination of such service mapping information that may span multiple cooperating administrative domains. These domains may be administered by the same provider or by closely ~~co-ordinating~~ coordinating provider networks.

A new BGP transport layer address family (SAFI 76) is defined for this purpose that uses RFC~~-~~ 4364 technology and follows RFC~~-~~ 8277 NLRI encoding. This new address family is called "BGP Classful Transport", ~~aka~~a.k.a., BGP CT.

BGP CT makes it possible to advertise multiple tunnels to the same destination address, thus avoiding need of multiple loopbacks on the egress node.

~~It carries transport prefixes across tunnel domain boundaries (e.g. in Inter-AS Option-C networks), which is parallel to BGP LU (SAFI 4). It disseminates "Transport class" information for the transport prefixes across the participating domains, which is not possible with BGP LU. This makes the end-to-end network a "Transport Class" aware tunneled network.~~

~~Just like BGP LU (SAFI 4), BGP CT family (SAFI 76) is used in inter-AS option-C networks. The Service Mapping procedures described in this document apply in the same manner to Intra-AS service endpoints as well as Inter-AS option-A, option-B, option-C variations. Examples of these variations are given in Appendix A.~~

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 RFC 2119 [RFC2119] RFC 8174 [RFC8174] when, and only when, they appear in all capitals, as shown here.

a mis en forme : En-tête

Commenté [BMI1]: Please shorten the abstract and focus on the main contributions.

Commenté [BMI2]: Redundant

a mis en forme : Pied de page

a mis en forme : En-tête

Status of This Memo

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

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

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

This Internet-Draft will expire on 28 December 2023.

Copyright Notice

Table of Contents

a mis en forme : Pied de page

a mis en forme : En-tête

a mis en forme : Pied de page

a mis en forme : En-tête

1. Introduction

Commenté [BMI3]: The introduction dives quickly into protocol machinery without explaining first the motivation/rationale. I would suggest some text to be added.

Having a figure with the involved entities would be helpful.

The mechanisms defined in this document enable brownfield networks deployed using existing technologies like RSVP-TE and greenfield networks that use technologies like SPRING to achieve 'Intent Driven Service Mapping'.

Commenté [BMI4]: Should first define what is meant by this concept. Not sure at this stage, we got why this is mentioned and how this is useful.

To facilitate this, the tunnels in a network can be grouped by the purpose they serve into a "Transport Class". These tunnels ~~could~~may be created using any signaling protocol including but not limited to LDP, RSVP-TE, BGP LU, or SPRING. The tunnels may use MPLS, IPv4, or IPv6 forwarding and carry one of the signaled payload types (e.g~~.~~., MPLS). Tunnels may exist between different ~~pair~~pairs of endpoints. Multiple tunnels may exist between the same pair of endpoints.

Commenté [BMI5]: Please add pointers.

Commenté [BMI6]: This is not a protocol

Commenté [BMI7]: Where/by whom?

A Transport Class consists of tunnels created by various protocols that satisfy the properties of the class. For example, a "Gold" transport class may consist of tunnels that traverse the shortest path with fast re-route protection. A "Silver" transport class may hold tunnels that traverse shortest paths without protection. A "To NbrAS Foo" transport class may hold tunnels that exit to neighboring Autonomous System (AS) Foo, and so on.

Commenté [BMI8]: Does this mean that a transport class can't be crated by configuration, including manual configuration?

Commenté [BMI9]: Checken and egg :-)

Commenté [BMI10]: Use consistent form (`Transport Class or transport class). Please pick one form.`

Commenté [BMI11]: I guess you meant the data conveyed in the tunnel. Please clarify in the text. Some tunnels may be defined by termination points at the boundaries.

The extensions specified in this document can be used to create a BGP transport tunnel that potentially spans domains while preserving its

a mis en forme : Pied de page

Transport Class. Examples of domain are ~~Autonomous System (~~an AS~~)~~ or an IGP area. Within each domain, there is a second level underlay tunnel used by the BGP to cross the domain. The second level underlay tunnels could be ~~heterogeneous; each domain may~~ use a different type of tunnel (e.g~~.~~., MPLS, IP encapsulation, GRE, or SRv6) and use a different signaling mechanism. A domain boundary is demarcated by a rewrite of the BGP ~~nexthop~~ next hop to 'self' while readvertising BGP CT transport routes. Examples of domain boundary are inter-AS links and inter-region ABRs. The path uses MPLS label-switching when crossing domain boundaries and uses the native intra-AS tunnel of the desired transport class when traversing within a domain.

Overlay routes carry ~~sufficient~~ an indication of the desired Transport Classes in the form of a BGP community called the "Mapping community". The ~~"~~route ~~resolution"~~selection procedure on the ingress node selects an appropriate tunnel whose destination matches (LPM) the ~~nexthop~~ next hop of the overlay route ~~belonging~~that belongs to the corresponding Transport Class. If the overlay route is carried in BGP, the protocol ~~nexthop~~ next hop (~~or~~ PNH) is carried as an attribute of the route.

The PNH of the overlay route is also referred to as "Service Endpoint" (SEP). The SEP may exist in the same domain as the service ingress node or lie in a different domain, which is adjacent or non-adjacent. In the former case, reachability to ~~the~~ a SEP is provided by an intra-domain tunneling protocol and in the latter case, reachability to ~~the~~ a SEP is via BGP transport families (e.g~~.~~., Subsequent Address Family Identifier (SAFI) 4 or 76).

In the context of this ~~architecture~~document, the intra-domain ~~transport~~ protocols (e.g~~.~~., RSVP-TE, SRTE) are also "Transport Class aware". They publish ingress routes in the Transport Route Database associated with the Transport Class at the tunnel ingress node. These routes are used to resolve BGP routes including BGP CT which may be further readvertised to adjacent domains to extend this tunnel. How exactly ~~the transport protocols area~~ a protocol is made transport class aware is outside the scope of this document.

This document describes mechanisms to:

* Model a "Transport Class" as a "Transport Route Database" on a router and to collect tunnel ingress routes of a certain class.

* Enable service routes to resolve over an intended Transport Class by virtue of carrying the appropriate "Mapping Community", which results in using the corresponding Transport Route Database for finding ~~nexthop~~next hop reachability.

a mis en forme : En-tête

Commenté [BMI12]: Please expand and add pointers.

Commenté [BMI13]: Expand

Commenté [BMI14]: This one is introduced in this document. I would introduce it first before citing it in examples.

Commenté [BMI15]: Let's avoid confusing TSV people

Commenté [BMI16]: Please expand and add a pointer.

Commenté [BMI17]: This refers to which entity?

Commenté [BMI18]: What is an "ingress route"? Why not simply "route"?

Commenté [BMI19]: Please indicate this is a new database.

Commenté [BMI20]: I don't parse this. Please consider rewording.

Commenté [BMI21]: To be defined first.

Commenté [BMI22]: To be consistent with the base BGP RFC. Please change all the occurrences in the document.

a mis en forme : Pied de page

* Publish and maintain tunnel ingress routes in a Transport Route Database via BGP without any path hiding using BGP VPN and Add-path procedures~~,~~. ~~such that~~That is overlay routes in the receiving domains ~~can~~ are also resolve over tunnels of the associated Transport Class.

* Provide an approach ~~way~~ for cooperating domains to reconcile any differences in extended community namespaces and interoperate between different transport signaling protocols in each domain.

~~In this~~ This document ~~we focus~~focuses mainly on MPLS as the intra-domain transport tunnel forwarding technology, but the mechanisms described here would work in similar manner for non-MPLS (~~(e.g. IP, GRE, UDP or SRv6.~~ Section 17) transport tunnel forwarding technologies~~ too.~~

This document assumes MPLS forwarding ~~as the de facto standard~~ when crossing domain boundaries. However, the mechanisms specified in this document can also support different forwarding technologies~~ (e.g. SRv6)~~. For example, Section 17 ~~(SRv6 support) in this document~~ describes the application of BGP CT over SRv6 data plane.

This document realizes "Intent" as defined in ~~Intent-based Networking: Concepts and Definitions~~ [RFC9315] and prescribes procedures that use the transport class as a construct to express intent for specific contexts. The procedures defined in this document provide homogenous building blocks to achieve Intent-based Networking.

~~The document Intent-aware Routing using Color~~ [Intent-Routing-Color] describes various use cases and applications of procedures described in this document.

xx

x. Experiment Goals & Success Criteria

2. Terminology

2.1 Acronyms and Abbreviations

LSP: Label Switched Path.

TE~~-~~: Traffic Engineering.

TC: Transport Class.

SN~~-~~: Service Node~~. A router that sends or receives BGP Service routes (e.g. SAFIs 1, 128) with self as nexthop.~~

eSN~~-~~: Egress Service Node. ~~A router that sends BGP Service routes (e.g. SAFIs 1, 128) with self as nexthop.~~

iSN~~-~~: Ingress Service Node~~. A router that receives BGP Service routes (e.g. SAFIs 1, 128).~~

a mis en forme : En-tête

Commenté [BMI23]: Cite references.

Commenté [BMI24]: Do we really need to have these details at this stage?

Commenté [BMI25]: I guess some setup/agreement is needed to make use of the procedure of the multi domain case. Such prerequisite should be listed “somewhere” in the spec and insert a pointer here.

Commenté [BMI26]: You may add an applicability scope section with this kind of assumptions.

Commenté [BMI27]: I’m not sure this claim is needed/justified. The proposal in the draft covers only one specific aspect of service delivery/provision. I see hardly how we can justify this claim.

You may have a dedication section where you can analyze/position this work vs. intent-based.

Commenté [BMI28]: I was expecting to see a mention of rfc9012 in this section. Is it normal that no mention is included in the introduction?

Commenté [BMI29]: Given the intended status, some experiments goals and success criteria should be documented.

Commenté [BMI30]: The current set is not homogenous. Please split acronyms vs. definitions.

a mis en forme : Pied de page

a mis en forme : En-tête

BN~~-~~: Border Node~~. A router that sends or receives BGP Transport routes (e.g. SAFI 4, 76) with self as nexthop.~~

TN~~-~~: Transport Node, P~~-~~ router.

Commenté [BMI31]: Why introducing a new term here, rather than using simply P nodes ?

Commenté [BMI32]: Add a pointer to Section 5.3.1 of [RFC4026]

BGP VPN~~-~~: VPNs built using RFC4364 mechanisms.

BGP LU: BGP Labeled Unicast family (SAFI 4)

BGP CT: BGP Classful Transport family (SAFI 76)

ASN: Autonomous System Number~~.~~

RT~~-~~: Route Target~~Route-Target extended community.~~

RD~~-~~: Route Distinguisher~~Route-Distinguisher.~~

RTC~~-~~: Route Target Constrain~~.~~

VRF: Virtual Router Forwarding~~ Table.~~

CsC: Carrier serving Carrier~~ VPN.~~

PNH~~-~~: Protocol-~~Nexthop~~Next hop address carried in a BGP Update message.

MNH~~-~~: BGP MultiNexthop attribute~~.~~

FEC~~-~~: Forwarding Equivalence Class~~.~~

RSVP-TE~~-~~: Resource Reservation Protocol - Traffic Engineering~~.~~

SR~~-~~: Segment Routing~~.~~

SRTE~~-~~: Segment Routing Traffic Engineering~~.~~

SID~~-~~: SR Segment Identifier~~.~~

EP~~-~~: Endpoint, a loopback address in the network.

SEP~~-~~: Service Endpoint, the PNH of a Service route.

LPM~~-~~: Longest Prefix Match~~.~~

SLA: Service Level Agreement~~.~~

EPE: Egress Peer Engineering~~.~~

UHP~~ Label~~: Ultimate Hop Pop~~ label.~~

PHP~~ Label~~: Penultimate Hop Pop~~ label.~~

2.2 Definitions

Intent: A set of operational goals (that a network should meet) and

a mis en forme : Pied de page

outcomes (that a network is supposed to deliver) defined in a
declarative manner without specifying how to achieve or implement
them.

Service routes: routes for used for forwarding "data traffic".

Transport routes: …

Classful Transport: xxxx

Service Node: A router that sends or receives BGP Service routes (e.g., SAFIs 1 or 128) with self as next hop.

Egress Service Node: A router that sends BGP Service routes (e.g., SAFIs 1 or 128) with self as next hop.

Ingress Service Node: A router that receives BGP Service routes (e.g., SAFIs 1 or 128).

Border Node: A router that sends or receives BGP Transport routes (e.g., SAFI 4 or 76) with self as next hop.

Service Family~~ :~~: A BGP address family that is used for advertising routes for
"data traffic" (i.e., service routes) as opposed to tunnels (e.g~~.~~., SAFI 1 or 128).

Transport Family~~ :~~: A BGP address family that is used for advertising tunnels,
which are in turn used by service routes for resolution (e.g~~.~~., SAFI 4
or 76).

Transport Tunnel~~-~~: A tunnel over which a service may place traffic
(e.g~~.~~., GRE, UDP encapsulation, LDP, or RSVP-TE~~ or SPRING~~).

Tunnel Ingress Route: ~~Route~~A route to Tunnel Destination/Endpoint installed
at the headend (ingress) of the tunnel by ~~the~~a tunneling protocol.

Tunnel Domain~~-~~: A domain of the network containing Service Nodes (SNs) and Border Nodes (BNs) under
a single administrative control that has tunnels between them. An
end-to-end tunnel spanning several adjacent tunnel domains can be
created by "stitching" them together using labels.

Transport Class~~-~~: A group of transport tunnels offering the same SLA.

Transport Class RT~~-~~: A Route~~-~~ Target extended community used to
identify a specific Transport Class.

Transport Route Database (TRDB): At ~~the~~ SNs and BNs, a Transport Class
has an associated Transport Route Database that ~~collects~~maintains its tunnel
ingress routes.

Transport Plane~~-~~: An end~~-~~-to~~-~~-end plane consisting of transport tunnels

a mis en forme : En-tête

Commenté [BMI33]: As this is copied from RFC9315, I would add a pointer to Section 2 of that RFC.

Commenté [BMI34]: Consider adding a definition for service route.

Commenté [BMI35]: Idem as the previous comment.

Commenté [BMI36]: Please consider adding an entry for this one as well. Thanks.

Commenté [BMI37]: Is this typically a PE? If so, I would say that in the text.

Commenté [BMI38]: Why AFIs are not mentioned as well?

This comment applies for all similar uses in the document?.

Commenté [BMI39]: This reads like an AFI, but the example are about SAFI :-)

Commenté [BMI40]: Idem as the previous comment.

Commenté [BMI41]: By "service nodes"? No?

Commenté [BMI42]: Same or similar SLA?

Do you mean all the clauses of an SLA must be identical or equivalent?

See a discussion on similar concept, e.g., in https://www.rfc-editor.org/rfc/rfc5160.html. Meta-classes are used to ease mapping classes in adjacent domains, without making assumption on how classes are implemented in each domain.

Commenté [BMI43]: Refers to what?

a mis en forme : Pied de page

belonging to the same Transport Class. Tunnels of the same Transport Class are stitched together by BGP CT route readvertisements with ~~nexthop~~ next hop "self" to enable Label-Swap forwarding across domain boundaries.

Mapping Community: The BGP Community/Extended-community on a BGP route that maps it to resolve over a Transport Class. ~~E.g.~~Examples of such a mappings are color:0:100~~,~~ and transport-target:0:100.

3. Transport Class

A Transport Class is defined as a set of transport tunnels that share the same SLA. It is encoded as the Transport Class RT, which is a new Route Target ~~Route-Target~~ extended community (XX).

A Transport Class is configured at ~~SN~~SNs and ~~BN~~BNs with RD and ~~Route Target~~ RT attributes. ~~Creation~~The creation of a Transport Class in a node instantiates its corresponding Transport Route Database~~.~~

~~The~~ in a node.

An operator may configure an SN/BN to classify a tunnel into ~~an~~a ~~appropriate~~ given Transport Class, which causes the tunnel's ingress route to be installed in the corresponding Transport Route Database (TRDB). These routes are used to resolve BGP routes including BGP CT which may be further readvertised to adjacent domains to extend ~~this~~ a tunnel.

Alternatively, a router receiving the transport routes ~~in~~via BGP with appropriate signaling information can associate those ingress routes to the ~~appropriate~~ relevant Transport Class. ~~E.g.~~For example, for Classful Transport family (SAFI 76) routes, the Transport Class RT ~~indicates~~ identifies the Transport Class. For BGP LU family (SAFI 4) routes, import processing based on Communities or inter-AS source~~-~~ peer may be used to place the route in the ~~desired~~ relevant Transport Class.

When the ingress route is received via SRTE [SRTE] with "Color:Endpoint" as the Network Layer Reachability Information (NLRI) that encodes the Transport Class as an integer 'Color', the 'Color' is mapped to a Transport Class during the import processing. The SRTE ingress route for 'Endpoint' is installed in the corresponding ~~Transport Route Database~~TRDB. The SRTE tunnel will be extended by a BGP CT advertisement with NLRI 'RD:Endpoint', Transport Class RT and a new label. The MPLS swap route ~~thus~~ installed for the new label will pop the label and can be thus used to deliver decapsulated traffic into

a mis en forme : En-tête

Commenté [BMI44]: See a previous similar comment. I would add some text to explain the intent here.

Commenté [BMI45]: It is the identified which is encoded, not the class itself.

Commenté [BMI46]: Add a pointer to the section where this is defined.

Commenté [BMI47]: May declare that how configuration is done is deployment specific.

Commenté [BMI48]: How the tunnel is identified in the classification instruction (e.g., endpoint @, interface, etc.)?

Commenté [BMI49]: That is? Please be explicit.

Commenté [BMI50]: Isn't this stitching also conditioned by a policy? Or yu expect a default behavior.

Commenté [BMI51]: BN/SN/etc.?

a mis en forme : Anglais (États-Unis)

Commenté [BMI52]: That is?

a mis en forme : Pied de page

a mis en forme : En-tête

the path determined by an SRTE route.

~~RFC8664~~ [RFC8664] extends Path Computation Element Communication Protocol (PCEP) to carry SRTE Color. This color association learnt from PCEP is also mapped to a Transport Class thus associating the ~~PCEP~~ PCEP-signaled SRTE LSP with the desired Transport Class.

Commenté [BMI53]: There is no such a thing in RFC8664.

Similarly, ~~PCEP-RSVP-COLOR~~ [PCEP-RSVP-COLOR] extends PCEP to carry an RSVP Color. This color association learnt from PCEP is also mapped to a Transport Class thus associating the ~~PCEP~~ PCEP-signaled RSVP-TE LSP with the desired Transport Class.

4. "Transport Class" Route Target Extended Community

This ~~document~~ section defines a new type of Route Target, called "Transport Class" Route Target Extended Community.

"Transport Class" Route Target ~~extended~~ Extended ~~community~~ Community is a transitive ~~extended community~~ Community ~~EXT-COMM~~ [RFC4360] of extended type~~extended-type~~, ~~with a new~~ ~~Format (Type high = 0xa) and SubType as 0x2 (Route Target).~~

Commenté [BMI54]: Be consistent with RFC4360. Thanks.

~~This new Route Target Format~~which has the ~~following encoding~~format shown in Figure 1~~:~~.

Commenté [BMI55]: This is provided in the description text right after the figure.

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

Figure 1: "Transport Class" Route Target Extended Community

Type: This 1~~-~~-octet

~~Type~~ field ~~contains value~~MUST be set to 0xa~~.~~

Commenté [BMI56]: Please a pointer to the IANA section

SubType: This 1~~-~~-octet

~~Subtype~~ field ~~contain~~MUST be set to 0x2~~. This indicates~~ to indicate 'Route Target'.

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

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

The least significant 32-bits of the value field contain the "Transport Class" identifier, which is an unsigned non-zero 32-bit integer.

Commenté [BMI57]: Yeah, but the field is encoded in 4 octets.

Commenté [BMI58]: Who manages/assigns the ID?

a mis en forme : Pied de page

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

~~The remaining 2 octets after SubType field are Reserved. They MUST be set to zero on transmission, SHOULD be ignored on reception and left unaltered.~~

The ~~"~~Transport ~~class~~Class~~"~~ Route Target Extended community follows the mechanisms for VPN route import/export as specified in BGP VPN [RFC4364] and follows the Constrained Route Distribution mechanisms as specified in ~~Route Target Constraints~~ [RFC4684].

A BGP speaker that implements ~~RT Constraint Route Target Constraints~~ [RFC4684] MUST apply the RT Constraint procedures to the ~~"~~Transport ~~C~~class" Route Target Extended community as well.

The Transport Class Route Target Extended community is carried on Classful Transport family routes and ~~allows~~ is used to ~~associating~~ associate them with appropriate ~~Transport Route Databases~~TRDBs at receiving BGP speakers.

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

5. Transport Route Database

A ~~Transport Route Database~~TRDB is a logical collection of transport routes pertaining to the same Transport Class. Tunnel endpoint addresses in ~~this~~ a TRDB~~database~~ belong to the "Provider Namespace".

Overlay routes that ~~want to~~require the use of a specific Transport Class confine the scope of ~~nexthop~~next hop resolution to the set of routes contained in the corresponding ~~Transport Route Database~~TRDB.

The ~~Transport Route Database~~TRDB can be realized, e.g., as a "Routing Table" referred in Section 9.1.2.1 of [RFC4271] ~~(https://www.rfc-editor.org/rfc/rfc4271#section-9.1.2.1)~~ which is a control plane only database. However, an implementation may choose a different ~~methodology~~ TRDB implementation approach ~~to realize this logical construct in such a way that it~~while still being adhering with ~~supports~~ the procedures defined in this document.

SNs or BNs originate routes for 'Classful Transport' address family from the ~~Transport Route Database~~TRDB. These routes have NLRI "RD:Endpoint", Transport Class RT, and an MPLS label. 'Classful Transport' family routes received with Transport Class RT are imported into its corresponding ~~Transport Route Database~~TRDB.

6. Next ~~hop~~ Hop Resolution Scheme

This section defines the ~~Nexthop~~ next hop ~~Resolution~~ resolution ~~Scheme~~ scheme construct that is used to specify how a service route or a BGP CT route can resolve its

a mis en forme : En-tête

Commenté [BMI59]: Add a pointer to the IANA Section

a mis en forme : Anglais (États-Unis)

Commenté [BMI60]: Not sure to get what is meant by «follows » here.

Do you mean the use of the RT is similar to what is discussed in 4.3.1 of 4364?

Commenté [BMI61]: Idem as the previous comment.

Commenté [BMI62]: The transport network provider. Please be explicit.

Commenté [BMI63]: As some policies are also involved to bind the overlay routes to an underlay.

Commenté [BMI64]: For the specific case of MPLS.

a mis en forme : Surlignage

a mis en forme : Pied de page

next hop~~nexthop~~ using its associated Mapping Community over a specific TRDB
or an ordered set of TRDBs.

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

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

A community ~~known~~ defined as "Mapping Community" is configured for a ~~"~~resolution scheme~~"~~. Mapping community is a "role", ~~and~~ not a new type of community per se~~;~~. Concretely, any BGP community or extended community may play this role. A Mapping Community maps to exactly one ~~Resolution~~resolution ~~Scheme~~scheme. A ~~Resolution~~ resolution ~~Scheme~~ scheme comprises of one primary ~~transport~~ Transport ~~class~~Class and optionally, one or more fallback ~~transport~~ Transport ~~classes~~Classes. The ~~Resolution~~ resolution ~~Scheme~~ scheme is used to ~~realize~~ provide the desired Intent.

~~An e~~Examples of mapping community values ~~is~~ are "color:0:100", described in ~~RFC 9012~~ [RFC9012], or ~~the~~ "transport-target:0:100" described in ~~section~~ Section 4 ~~in this document~~.

A BGP route is associated with a resolution scheme during import processing. The first community on the route that matches a Mapping Community of a locally configured r~~R~~esolution ~~Scheme~~ scheme is considered the effective Mapping Community for the route. The r~~R~~esolution ~~Scheme~~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~~. So, the first community that maps to a ~~Resolution~~ resolution ~~Scheme~~ scheme is chosen as the effective ~~mapping~~ Mapping ~~community~~community.

A transport route received in BGP Classful Transport family SHOULD use a ~~Resolution~~ resolution ~~Scheme~~ scheme that contains the primary Transport Class without any fallback to best effort tunnels. The primary Transport Class is identified by the Transport Class RT carried on the route. Thus, Transport Class RT serves as the Mapping Community for BGP CT routes.

A service route received in a BGP service family MAY map to a ~~Resolution~~ resolution ~~Scheme~~ scheme that contains the primary Transport Class identified by the Mapping Community on the route and a fallback to best effort Transport Class. The primary Transport Class is

a mis en forme : En-tête

Commenté [BMI65]: I guess there is always one primary TC.

Commenté [BMI66]: This may also include the fallback conditions.

Commenté [BMI67]: As this is a new role to be yet known :-)

Commenté [BMI68]: Not sure this is useful.

Commenté [BMI69]: What if a domain in the chain does not support a given class.

Commenté [BMI70]: Under which conditions, the fallback is possible/safe?

If this is an absolute requirement, consider s/SHOULD/MUST. Otherwise, please add a statement when the fallback is OK.

Commenté [BMI71]: Isn't this configuration based ?

a mis en forme : Pied de page

identified by the Mapping Community carried on the route. For ~~e.g~~example,~~.~~ the Color Extended ~~Color c~~Community may serve as the Mapping Community for service routes. "Color:0:<n>" ~~MAY~~ may map to a ~~Resolution~~ resolution ~~s~~Scheme that has primary Transport Class <n> and a fallback to the best-effort Transport Class.

The ~~Resolution~~ resolution ~~Scheme~~ scheme mechanism not only works with ~~SPRING~~ SR-based transport protocols to realize Intent based forwarding, but also with existing tunneling technologies like RSVP TE, GRE, UDP, etc. Not assuming a specific tunneling technology makes the BGP CT architecture backward and forward compatible with existing and newer tunneling protocols, respectively. It is compatible with ~~SPRING~~SR, but there is no specific dependency on ~~SPRING~~SR. It is more generic and has broader applicability.

7. BGP Classful Transport Family NLRI

The Classful Transport (CT) family ~~will~~ uses the existing Address Family Identifier (AFI) of IPv4 or IPv6 and a new SAFI 76 "Classful Transport" that ~~will apply~~ applies to both IPv4 and IPv6 AFIs. These AFI~~, ~~/SAFI pair of values ~~MUST be~~are negotiated as per the ~~in~~ Multiprotocol Extensions capability described in Section 8 of [RFC4760] to be able to send and receive BGP CT routes.

The "Classful Transport" SAFI NLRI itself is encoded similar to what as specified in ~~https://tools.ietf.org/html/rfc8277#section~~Section~~-~~2 of [RFC8277].

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

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

The ~~Procedures~~ procedures described for SAFIs 4 or ~~SAFI~~ 128 in ~~https://tools.ietf.org/html/rfc8277#section~~Section~~-~~ 2 of [RFC8277] apply for SAFI 76 as well. BGP CT routes ~~MAY~~ may carry multiple labels in the NLRI, by negotiating the Multiple Labels Capability as described in ~~https://www.rfc-editor.org/rfc/rfc8277#sections~~Section~~-~~ 2.1 of [RFC8277]. ~~[RFC8277]~~

For ~~easy reference~~convenience, ~~the following f~~Figure 2 illustrates a BGP Classful Transport family NLRI when a single Label is advertised (Multiple Labels Capability is not negotiated):

```
    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
```

a mis en forme : En-tête

Commenté [BMI72]: This is not a protocol !

Commenté [BMI73]: Why the focus on SR?

Commenté [BMI74]: I would simply remove this text. The suggested applicability scope would be better rather than repeating it for every section. Thanks.

Commenté [BMI75]: No need for the normative language here as Section 8 of 4760 has the following:

"To have a bi-directional exchange of routing information for a particular <AFI, SAFI> between a pair of BGP speakers, each such speaker MUST advertise to the other (via the Capability Advertisement mechanism) the capability to support that particular <AFI, SAFI> route."

Commenté [BMI76]: As the SAFI values in that section are #, for example.

Commenté [BMI77]: The normative language is not required here as this is part of the behavior induced by the sentence right before.

a mis en forme : Pied de page

```
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Length     |                    Label              |Rsrv |S|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                    Route Distinguisher (8 bytes)              |
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    IPv4/IPv6 Prefix                           ~
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2: SAFI 76 "Classful Transport" NLRI Format

Length: 1 octet

Th~~e Length field consists of a single octet. It ~~is field specifies the length in bits of the remainder of the NLRI field.

Note that the length will always be the sum of 20 (number of bits in Label field), plus 3 (number of bits in Rsrv field), plus 1 (number of bits in S field), plus the length in bits of the Prefix (RD:IP prefix).

In an MP_REACH_NLRI attribute whose SAFI is 76, the Prefix (RD + IP prefix) will be 96 bits or less if the AFI is 1 and will be 192 bits or less if the AFI is 2.

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

Label:

The Label field is a 20-bit field containing an MPLS label value (~~see~~ [RFC3032]).

Rsrv:

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

S:

When single label is advertised, this 1-bit field MUST be set to one on transmission and MUST be ignored on reception.

Route Distinguisher:

An 8-octet~~8 byte~~ RD as defined in Section 4.2 of [RFC4364~~ See 4.2~~].

IPv4/IPv6 Prefix:

Includes an IPv4 prefix, if AFI/SAFI 1/76.
Includes IPv6 prefix, if AFI/SAFI 2/76.

Attributes on a Classful Transport route include the Transport Class Route Target ~~Route-Target extended ~~Extended ~~community~~Community, which is used to associate the route

a mis en forme : En-tête

Commenté [BMI78]: I know this is echoing what is 8277, but the use of SHOULD/MUST uses are not consistent with this part of the spec above (which is the correct form, IMO):

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

Commenté [BMI79]: No need to repeat these details, or at least say that the meaning is the same as in RFC8277.

Commenté [BMI80]: BTW, the text uses a mix of octet/byte. I suggest to use "octet", otherwise this may induce some distraction and revive discussions such as byes are not necessary 8 bits and the like.

a mis en forme : Pied de page

with the ~~correct~~ target ~~Transport Route Databases~~TRDBs on SNs and BNs ~~in the network~~.

SAFI 76 routes can be sent with either IPv4 or IPv6 next_hop. The type of next_hop is inferred from the length of the next_hop.

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

When the length of the Nexthop Address field is 12 the nexthop address is of type VPN-IPv4 with 8-octet RD set to zero.

7.1. Carrying multiple ~~encapsulation~~ Encapsulation ~~information~~Information

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

An MPLS Label is carried using the encoding in ~~RFC 8277~~ [RFC8277] ~~encoding~~. A node that does not support MPLS forwarding advertises the special label 3 (Implicit Null) in the ~~RFC 8277~~ MPLS Label field.

SRv6 SID is carried using Prefix SID attribute as specified in ~~RFC 9252~~ [RFC9252], without Transposition Scheme. The Transposition Length is set to 0 and Transposition Offset is set to 0 to indicate nothing is transposed and that the entire SRv6 SID value is encoded in the SID Information Sub-TLV.

UDP tunneling information is carried using the Tunnel Encapsulation Attribute ~~TEA~~ attribute as specified in ~~RFC 9012~~ [RFC9012].

8. Usage of Route Distinguisher and Label Allocation Modes

RDs aid~~s~~ in troubleshooting a ~~BGP CT~~ network that makes use of the BGP CT by uniquely identifying the originator of a route across a ~~multi~~ multi-domain network.

The u~~U~~se of RDs also ~~allows~~ provides an ~~the~~ option for signaling forwarding diversity within the same Transport Class. The same Egress PE can advertise multiple BGP CT routes for an EP belonging to the same Transport Class.

~~E.g.~~For example, multiple “RDx:EP1” prefixes can be advertised for an EP1 to different sets of BGP peers in order to collect traffic statistics for them. In absence of RD, duplicated Transport Class/Color values will be needed in the transport network to achieve such use cases.

In a BGP CT network, the number of routes at an Ingress PE is a

a mis en forme : En-tête

Commenté [BMI81]: What is the purpose of this text ?

Commenté [BMI82]: BTW, is it possible to swap from IPv4 to IPv6 next hops when transporting data of a tunnel that spans multiple domains?

Commenté [BMI83]: Suggest moving this to the SRv6 section.

Commenté [BMI84]: Not usedd as such in 9012

Commenté [BMI85]: Does those domains belong to the same administrative entity?

Commenté [BMI86]: Please map this to the nodes defined above SN/BN, etc.

Commenté [BMI87]: This EP is local or external to the PE ?

Commenté [BMI88]: Not sure to get this example.

a mis en forme : Pied de page

function of unique EPs multiplied by BNs in the ingress domain that do next_hop self. BGP CT provides flexible RD and Label allocation modes to address operational requirements in a multi-domain network.

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

A label is allocated for a BGP CT route when it is advertised with Next_hop self by a SN or a BN. An implementation may use different label allocation modes with BGP CT. The recommended label allocation mode is per-prefix as it provides better traffic convergence properties than per-nexthop 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.

The impacts on the control plane and forwarding behavior for the above modes are detailed with an example in ~~Managing Transport Route Visibility (~~Section 20.3~~)~~.

9. Comparison with ~~other~~ Other ~~families~~ Families using RFC~~-~~ 8277 ~~encoding~~Encoding

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

SAFI 4 (NLRI with MPLS Labels~~BGP LU~~) is an RFC8277 encoded family that carries transport prefixes in the NLRI, where the prefixes come from the provider namespace.

SAFI 76 (Classful-Transport SAFI~~Classful Transport~~) is an RFC8277 encoded family that carries transport prefixes in the NLRI, where the prefixes come from the provider namespace and are contextualized into separate ~~Transport Route Databases~~TRDBs as per ~~using~~ [RFC4364~~ procedures~~].

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

In this document, a new AFI/SAFI is used instead of reusing SAFI 128 to carry these transport routes because it is operationally advantageous to segregate transport and service prefixes into separate address families. ~~E.g.~~For example, such an approach ~~It~~ allows operators to safely enable "per-prefix" label allocation scheme for Classful Transport prefixes without affecting SAFI 128 service prefixes, which may have huge

a mis en forme : En-tête

Commenté [BMI89]: I still do think that having a figure will all involved entities is needed.

Commenté [BMI90]: That is?

Commenté [BMI91]: Use the name as assigned by IANA.

Commenté [BMI92]: As per the current IANA records

Commenté [BMI93]: Add a pointer to /rfc4364#section-10

Commenté [BMI94]: I wonder whether the text starting with « SAFI 76 » is moved to be right after this text. If not, consider s/new/dedicated ...SAFI (76)

a mis en forme : Pied de page

scale. The "per prefix" label allocation scheme keeps the routing churn local during topology changes.

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

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

10. Protocol Procedures

This section summarizes the procedures followed by various CT-aware nodes ~~speaking Classful Transport family~~.

10.1. Preparing the ~~network~~ Network to ~~deploy~~ Deploy Classful Transport ~~planes~~Planes

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

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

Implementations ~~MAY~~ may provide automatic generation and assignment of RD, RT values; they ~~MAY~~ 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.

10.2. ~~Origination~~ Originating ~~of~~ Classful Transport ~~route~~Routes

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

The egress node of the tunnel, i.e., the tunnel endpoint originates the BGP Classful Transport route with NLRI containing RD:TunnelEndpoint, Transport Class RT, and PNH TunnelEndpoint,

a mis en forme : En-tête

Commenté [BMI95]: That is ?

Commenté [BMI96]: Please use the registered names

Commenté [BMI97]: This is more about bootstrapping.

Otherwise, operators should have means to easily discover activated CTs in a network and that consistent mapping rules are configured to boundary nodes.

Should there be policies to control the use of CT per specific prefixes?

Commenté [BMI98]: Shouldn't the resolution instructions be provided as well?

Also, the fallback instructions can be supplied to participating nodes. I would mention those in the text.

Commenté [BMI99]: Does this impact interoperability? I don't think so but I may be mistaken.

Commenté [BMI100]: Please define early in the document what is a network domain.

Commenté [BMI101]: I think that I know what is meant here, but it is preferable to be explicit.

a mis en forme : Pied de page

which will be resolved over the tunnel route~~ in~~ using the TRDB ~~Transport Route Database at the~~of ingress node. When ~~the~~ a tunnel is up, the Classful Transport BGP route will become usable and get re-advertised.

Alternatively, the ingress node may advertise this tunnel destination into BGP as a Classful Transport family route with NLRI RD:TunnelEndpoint, 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 route SHOULD NOT be advertised to the IBGP core that contains the tunnel.

Unique RD SHOULD be used by the originator of a Classful Transport route to disambiguate the multiple BGP advertisements for a transport endpoint.

10.3. Processing ~~Ingress node receiving~~ Classful Transport ~~route~~Routes by Ingress Nodes

~~On~~ Upon ~~receiving~~ receipt of a BGP Classful Transport route with a PNH that is not directly connected (e.g., an IBGP-route), a Mapping Community on the route (the Transport Class RT) is used to ~~indicates~~ decide to which ~~Resolution~~resolution ~~Scheme~~ scheme this route is to be map~~s~~ ~~to~~ped. The resolution scheme for a Transport Class RT with Transport ~~class~~ Class ID "C1" contains ~~Transport Route Database~~TRDB for Transport Class with same ID. In cases where Transport ~~class~~ Class "C1" tunnels are not available in a domain, the administrator ~~MAY~~ may customize the Resolution scheme to map to a different set of transport class available in that domain.

The routes in the associated ~~Transport Route Databases~~TRDBs are used to resolve the received PNH. If the resolution process does not find a matching route in any of the associated ~~Transport Route Databases~~TRDBs, the received BGP CT route MUST be considered unusable for forwarding purpose and be withdrawn.

10.4. ~~Border node r~~Readvertising Classful Transport ~~route~~ Route by Border Nodes with ~~nexthop~~Next Hop ~~self~~Self

~~The~~ BNs allocate~~s~~ ~~an~~ MPLS labels to advertise upstream in Classful Transport NLRI. ~~The~~ A BN also installs an MPLS route for that label that swaps the incoming label with a label received from the downstream BGP speaker or pops the incoming label. It then pushes received traffic to the transport tunnel or direct interface that the Classful Transport route's PNH resolved over.

The label SHOULD be allocated with "per-prefix" label allocation

a mis en forme : En-tête

Commenté [BMI102]: By whom?

Commenté [BMI103]: Shouldn't this be constrained by policy? Not all EBGP peers will be eligible to the CT, no?

Commenté [BMI104]: Under which condition, it is safe to advertise them?

Absent such exception, and assuming you elaborate on the rationale, you may consider s/SHOULD/MUST

Commenté [BMI105]: This is deployment-specific. I would avoid the normative language here.

Commenté [BMI106]: Please consider presenting this as an example.

Commenté [BMI107]: The customization behavior was already introduced earlier. This is not a new one.

Commenté [BMI108]: Including fallback instructions ? if so, I would mention it in the text. Thanks.

a mis en forme : Pied de page

semantics. RD is stripped by the BN from the BGP CT NLRI prefix when a BGP
CT route is added to a ~~Transport Route Database~~TRDB. The IP prefix in
the ~~Transport Route Database~~TRDB context (Transport-Class, IP-prefix)
is used as the key to do per-prefix label allocation. This helps
in avoiding BGP CT route churn throughout the CT network when an instability (e.g., failure)
~~failure happens~~ is experienced in a domain. The failure is not propagated
further than the BN closest to the failure.

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

10.5. Border ~~node~~ Nodes ~~receiving~~ Receiving Classful Transport ~~route~~ Routes on EBGP

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

If the inter-AS links ~~should~~ has to honor Transport Class, then the BN
SHOULD follow procedures of an Ingress node described above and
perform next_hop resolution process. The interface routes SHOULD
be installed in the ~~Transport Route Database~~TRDB belonging to the
associated Transport Class.

10.6. Avoiding ~~path~~Path~~-h~~ Hiding ~~through~~ Through Route Reflectors

When multiple BNs exist such that ~~theu~~ they advertise a "RD:EP" prefix to
Route Reflectors (RRs), the RRs may hide all but one of the BNs, unless ADDPATH
[RFC7911] is used for the Classful Transport family. This is
similar to L3VPN ~~option~~Option~~-~~ B~~ scenarios~~. Hence, ADDPATH SHOULD be
used for Classful Transport family, to avoid path-hiding through
RRs.

10.7. Avoiding ~~loop~~ Loops ~~between~~ Between Route Reflectors~~ in forwarding path~~

A pair of redundant ABRs, each acting as an RR with "next hop~~nexthop~~ self",
may choose each other as best path instead of the upstream ASBR,
causing a traffic forwarding loop. This happens because of
following the path selection rule specified in ~~BGP RR~~ [RFC4456]

a mis en forme : En-tête

Commenté [BMI109]: This is covered by this part :

"The recommended label allocation
mode is per-prefix as it provides better traffic convergence
properties than per-nexthop label allocation mode."

I would keep the behavior in one single place. Thanks.

Commenté [BMI110]: MUST?

Commenté [BMI111]: ?

a mis en forme : Surlignage

Commenté [BMI112]: Please add a pointer to the exact section where this is discussed.

a mis en forme : Pied de page

that tie-breaks on Router-ID~~ROUTER-ID~~ before CLUSTER_LIST. RFC4456 considers ~~pure~~ RRs which ~~is~~ are not in forwarding path. When a RR is in forwarding path and reflects routes with next_hop self, as is the case for ABR BNs in a BGP transport network, this rule may cause loops.

Implementations SHOULD provide a way to alter the tie-breaking rule specified in BGP RR [RFC4456] so as to tie-break on CLUSTER_LIST step before Router-ID~~ROUTER-ID~~ step, when performing path selection for BGP CT routes.

This document suggests the following modification to the BGP Decision Process Tie Breaking rules (Sec~~.~~tion 9.1.2.2~~,~~ of [RFC4271]) that can be applied to path selection of BGP CT family routes:

The following rule SHOULD be inserted between Steps e) and f): a BGP Speaker SHOULD prefer a route with the shorter CLUSTER_LIST length. The CLUSTER_LIST length is zero if a route does not carry the CLUSTER_LIST attribute.

Taking into account ~~Some~~ some other deployment considerations can also help in avoiding this Problem, e.g.,:

- IGP metric should be assigned such that "ABR to redundant ABR" cost is inferior than "ABR to upstream ASBR" cost.

- Tunnels belonging to ~~non~~ non-best effort Transport Classes ~~SHOULD NOT~~should not be provisioned between ABRs. This will ensure that the route received from an ABR with next_hop self will not be usable at a redundant ABR.

This ~~avoids~~ softens the possibility of such loops altogether.

10.8. Ingress ~~node~~ Nodes ~~receiving~~ Receiving ~~service~~ Service ~~route~~ Routes with a Mapping Community

Service routes received with a Mapping Community resolve using ~~Transport Route Databases~~TRDBs determined by the ~~Resolution~~ resolution ~~Scheme~~scheme.

If the resolution process does not find a Tunnel Ingress Route in any of the ~~Transport Route Databases~~TRDBs, the service route MUST be considered unusable for forwarding purpose and be withdrawn.

10.9. ~~Coordinating~~ Coordination ~~between~~ Between ~~domains~~ Domains ~~using~~ Using ~~different~~ Different ~~community~~ Community ~~namespaces~~Namespaces

Cooperating Inter-AS option~~-~~C domains may sometimes not agree on RT, RD, Mapping community or Transport Route Target values because of differences in community namespaces (e.g., during network mergers

a mis en forme : En-tête

Commenté [BMI113]: Deployment specific

Commenté [BMI114]: Not sure this a new behavior. Isn't this already covered in 10.3?

a mis en forme : Pied de page

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 ~~used~~ leveraged with the Classful Transport family also.

The ~~Resolution~~ resolution ~~Schemes~~ schemes SHOULD allow association with multiple Mapping Communities. This helps with renumbering, network mergers, or transitions.

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

This ~~document defines a new format of~~Transport Class Route~~-~~ Target ~~extended~~Extended~~-~~ ~~community~~ Community ~~to carry Transport Class, this~~ is useful to avoid~~s~~ collision with regular Route Target namespace used by service routes.

10.10. Best ~~effort~~ Effort ~~transport~~ Transport ~~class~~Class

It is possible to represent 'Best effort' SLAs also as a Transport ~~class~~Class. Today, BGP LU (SAFI 4) is used to extend the best effort intra domain tunnels to other domains.

Alternatively, BGP CT (SAFI 76) may be used to carry the best effort tunnels also. This document reserves the Transport ~~class~~Class ID value 0 to represent "Best Effort Transport Class ID". However, implementations SHOULD provide configuration to use a different value for this purpose.

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

When a BN or SN receives a BGP CT route with Best Effort Transport ~~class~~ Class ~~route~~ Route ~~target~~ Target as the mapping community, the Best effort Resolution scheme is used for resolving the BGP nex_thop, and the resultant route is installed in the best effort transport route database. If no best effort tunnel was found to resolve the BGP nexthop, 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~~ Mapping ~~community~~Community, and absent local policy, the ~~Best~~ best effort ~~Resolution~~ resolution scheme is used for resolving the BGP nexthop on the route. This behavior is backward

a mis en forme : En-tête

Commenté [BMI115]: It would be valuable to group all deployment considerations in one single section.

Commenté [BMI116]: If another value is used, how this is signaled to peers to ease the mapping?

Commenté [BMI117]: This is a specific case that is already covered in 10.3.

a mis en forme : Pied de page

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.

11. Flowspec Redirect to IP

Flowspec routes using Redirect to IP nexthop is described in BGP
Flow-Spec Redirect to IP Action [FLOWSPEC-REDIR-IP].

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

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

12. BGP CT Egress TE

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

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

13. Interaction with BGP ~~attributes~~ Attributes ~~specifying~~ Specifying
~~nexthop~~ Nexthop ~~address~~ Address and
~~color~~Color

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

The ~~Mechanisms~~ mechanisms described in ~~BGP MultiNexthop Attribute~~
[MULTI-NH-ATTR]
allow a BGP route to carry multiple nexthop addresses. It also
allows specifying 'Transport Class ID' as a qualifier for each
Nexthop address.

It should be noted that in such cases "Transport Class/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 nexthop should resolve over. This document suggests the
following order of precedence, more preferred first:

Transport Class ID SubTLV, in MultiNexthop Attribute.

Color SubTLV, in Tunnel Encapsulation Attribute.

Transport Target Extended community, on BGP CT route.

Color Extended community, on BGP service route.

a mis en forme : En-tête

Commenté [BMI118]: This is a local behavior. Not sure the normative language is justified.

Commenté [BMI119]: I woudl move thsi to be under the suggestion operational consideration section.

Commenté [BMI120]: This is still an individual draft. I would delete this text but include a discussion of the CT applicability there, not here/

Commenté [BMI121]: Again, this is an individual I-D. I suggest to delete this text and move the discussion to that draft.

a mis en forme : Pied de page

The above precedence order follows more specific scoping of Color to less specific scoping.

Transport Class ID specified for Nexthop-Leg subTLV in a MultiNextHop attribute is more specific indication of Color than Color subTLV in a TEA, which in turn is more specific than Mapping Community (Transport Target) on a BGP CT transport route, which is in turn more specific than a Service route scoped Mapping Community (Color Extended community).

14. Signaling Intent ~~across~~ over PE-CE ~~link~~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.

This section describes the mechanisms that enable such a signaling. These procedures use existing AFIs, and service families (SAFI 1) on the PE-CE ~~link~~AC, with a new BGP attribute. It does not require a forklift upgrade of the PE-CE session with a new set of address families.

```
                        ---Gold----->
               [CE1]-----[PE1]---[P]----[PE2]-----[CE2]
                        ---Bronze--->
               11.0.0.0                         22.0.0.0
                   ----  Traffic direction ---->
```

Figure 1: ~~Intent on~~Example of a Topology with PE-CE ~~link~~Links.

14.1. Using DSCP in MultiNexthop ~~attribute~~Attribute

~~One s~~Such an indication can be in form of DSCP code point (~~RFC2474~~ [RFC2474]) in the IP header.

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

Let PE1 be configured to map DSCP1 to Gold Transport class, and DSCP2 to Bronze Transport class. Based on the DSCP code point received on the IP traffic from CE1, PE1 forwards the IP packet over a Gold or Bronze tunnel. Thus, the forwarding is not based on ~~just~~ the destination IP address, but also the DSCP code point. This is known as Class Based Forwarding (CBF). Today CBF is configured at the PE1 device roles and CE1 doesn't receive any indication in BGP signaling regarding what DSCP code points are being offered by the provider network.

With a BGP MultiNexthop Attribute [MULTI-NH-ATTR] attached to a SAFI 1 service route, it is possible to extend the PE-CE BGP ~~signalling~~signaling (if used) to communicate such information to the CE1. In the above example, the MNH contains two Nexthop Legs, described by two Forwarding Instruction TLVs. Each Nexthop Leg contains PE1's peering self address in Endpoint Identifier TLV [MNH-EP] , the color Gold or

a mis en forme : En-tête

Commenté [BMI122]: This is deployment specific. This is similar to QoS marking.

Commenté [BMI123]: Why specifically in data packets? A control channel can be exposed to request specific classification rules. This can also be managed out of managed (via controllers). See for example the classification rules in RFC9182

Commenté [BMI124]: Which session? Do you mean the routing session between CE/PE?

Commenté [BMI125]: As the session can be using other protocols. See RFC9182

Commenté [BMI126]: I'm not sure we need to call for specific solutions here.

a mis en forme : Pied de page

Bronze encoded in the Transport class ID TLV [MNH-TC] , and associated DSCP code point indicating Gold or Bronze transport class encoded in the Payload Encapsulation Info TLV [MNH-ENCAP-DSCP]~~ ~~. This allows the CE to discover what transport classes exist in the provider network, and which DSCP codepoint to encode so that traffic is forwarded using the desired transport class in the provided network.

14.2. ~~MPLS ~~MPLS-enabled CE

If the PE-CE link is MPLS enabled, a distinct MPLS label can also be used to express Intent in data packets from CE. Enabling MPLS forwarding on PE-CE links comes with some security implications. This section gives details on these aspects.

Consider the ingress PE1 receiving a VPN prefix RD:Pfx1 received with VPN label VL1, nexthop as PE2 and a mapping community containing TC1 as 'Transport class ID'. PE1 can allocate a MPLS Label PVL1 for the tuple "VPN Label, PNH Address, Transport class ID" and advertise to CE1.

Label PVL1 may identifies a service function at any node in the network, e.g., a Firewall device or egress node PE2. And, for the same service prefix, a distinct label may be advertised to different CEs, such that incoming traffic from different CEs to the same service prefix can be diverted to a distinct devices in the network for further processing. This provides Ingress Peer Engineering control to the network.

PE1 installs a MPLS FIB route for PVL1 with nexthop as "Swap VL1, Push TL1 towards PE2". TL1 is the BGP CT label received for the tuple 'PE2, TC1'. In forwarding, when MPLS packet with label PVL1 is received from CE1, PVL1 Swaps to label VL1 and pushes the BGP CT label TL1. PE1 advertises the label "PVL1" in the MULTI_NH_ATTR to CE1. PE1 forwards based on MPLS label without performing any IP lookup. This allows for PE1 to be a low IP FIB device and still support CBF by using MPLS Label inferred PHB. The number of MPLS Labels consumed at PE1 for this approach will be proportional to the number of Service functions and Intents that are exposed to CE1.

A BGP MultiNexthop Attribute [MULTI-NH-ATTR] is attached to a SAFI 1 service route to convey the MPLS Label information to CE1. In the above example, the MNH contains two Nexthop Legs, described by two Forwarding Instruction TLVs. Each Nexthop Leg contains PE1's peering self address in Endpoint Identifier TLV [MNH-EP] , the color Gold or Bronze encoded in the Transport class ID TLV [MNH-TC] , and associated MPLS Label "PVL1" or "PVL2" encoded in the Payload Encapsulation Info TLV [MNH-ENCAP-MPLS] . This allows the CE to discover what transport classes exist in the provider network, and which MPLS Label to encode so that traffic is forwarded using the desired transport class.

14.2.1. Secure MPLS F~~f~~orwarding on ~~inter~~Inter-AS ~~link~~Link

The MPLS enabled PE-CE ~~link ~~attachement circuit is considered connecting to an untrusted domain. Such interfaces can be secured against MPLS label spoofing by a walled garden approach using "MPLS context tables".

a mis en forme : En-tête

Commenté [BMI127]: Idem as above

Commenté [BMI128]: I don't think this is specific to this document

Commenté [BMI129]: You mean the part behind the CE?

a mis en forme : Pied de page

a mis en forme : En-tête

The PE1-CE1 interface can be confined to a specific MPLS context table "A" corresponding to the BGP peer. Such that only the routes for labels advertised to CE1 are installed in MPLS context table "A".

This ensures that if CE1 sends MPLS packet with a label that was not advertised to the CE1, the packet will be dropped.

Further, the routes for labels PVL1, PVL2 installed in MPLS context table "A" can match on 'Bottom of stack' bit being 'one', ensuring a MPLS packet is accepted from CE1 only if it has no more than one label in the label stack.

However, the PE itself may not be able to perform any checks based on inner payload in the MPLS packet since it performs label swap forwarding. Such inner payload based checks may be offloaded to a downstream node that forwards and processes inner payload, e.g., a IP FIB router. These security aspects should be considered when using MPLS enabled CE devices.

15. Scaling ~~considerations~~Considerations

15.1. Avoiding ~~unintended~~ Unintended ~~spread~~ Spread of BGP CT R~~r~~outes ~~across~~ Across ~~domains~~Domains

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

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

15.2. Constrained D~~d~~istribution of PNHs to SNs (~~On~~ On-Demand Next H~~h~~op)

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

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

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

a mis en forme : Pied de page

a mis en forme : En-tête

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

The RT carrying <eSN>:<TC> MAY be an IP-address specific regular RT (BGP attribute code 16), IPv6-address specific RT (BGP attribute code 25), or a Wide-communities based RT (BGP attribute code 34) as described in Route Target Constrain Extension [RTC-Ext]. This document recommends using Wide-communities based RT for the same.

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

Commenté [BMI130]: MUST?

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 Nexthop" of BGP CT routes, which help with the scaling of MPLS forwarding state at SN and BN.

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

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

Commenté [BMI131]: What happens if it doesn't? Please document those.

15.3. Limiting ~~scope~~ The ~~of visibility~~ Visibility Scope of PE ~~loopback~~ Loopback as PNHs

Commenté [BMI132]: Advertising these loopbacks may also be problematic for a security stand point. I guess this should be discussed in the sec cons.

It may be even more desirable to limit the number of PNHs that are globally visible in the network. This is possible using mechanism described in MPLS Namespaces [MPLS-NAMESPACES].

~~Such that~~that the 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

Commenté [BMI133]: Where this is defined?

a mis en forme : Pied de page

a mis en forme : En-tête

network, swapping the SN scoped service label with a CPNH scoped private namespace label.

This provides much greater advantage in terms of scaling and convergence. Changes to implement this feature are required only on the local region's BNs and RRs.

16. OAM ~~considerations~~Considerations

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

The 'Target FEC Stack' sub-TLV for IPv4 Classful Transport has a Sub-Type of [TBD1], and a length of 13. The Value field consists of the RD advertised with the Classful Transport prefix, the IPv4 prefix (with trailing 0 bits to make 32 bits in all) and a prefix length encoded as ~~follows~~shown in Figure X.~~:~~

Commenté [BMI134]: Add a note for the RFC editor to update this once the value is assigned.

```
     0                   1                   2                   3
     0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

    |                      Route Distinguisher                      |
    |                          (8 octets)                           |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

    |                         IPv4 prefix                           |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

    | Prefix Length |                Must Be Zero                   |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Commenté [BMI135]: One would expect the prefix length to be encoded before the prefix, but this is inherited from RFC8029. You may say so.

Figure ~~2~~3: Classful Transport IPv4 FEC

Commenté [BMI136]: Please check the numbering of the figures + call them explicitly in the text.

The 'Target FEC Stack' sub-TLV for IPv6 Classful Transport has a Sub-Type of [TBD2], and a length of 25. The Value field consists of the RD advertised with the Classful Transport prefix, the IPv6 prefix (with trailing 0 bits to make 128 bits in all) and a prefix length encoded as ~~follows~~shown in Figure X.~~:~~

```
     0                   1                   2                   3
     0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

    |                      Route Distinguisher                      |
    |                          (8 octets)                           |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

    |                         IPv6 prefix                           |
    |                                                               |

    |                                                               |

    |                                                               |

    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

a mis en forme : Pied de page

a mis en forme : En-tête

```
      | Prefix Length |                 Must Be Zero                  |
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Commenté [BMI137]: Idem as for IPv4 comment

Figure ~~3~~4: Classful Transport IPv6 FEC

Commenté [BMI138]: Idem as for IPv4 comment

17. Applicability to Network Slicing

In Network Slicing, the Transport Slice Controller (TSC) is responsible for ~~sets~~ customizing and setting up~~up~~ the ~~Topology~~ underlying transport (e.g., RSVP-TE, SR-TE tunnels with desired characteristics) and resources (e.g., polices/shapers) in a ~~transport~~ network to create ~~a Transport~~an IETF Network Slice. The Transport Class construct described in this document represents the "Topology Slice" portion of this equation.

Commenté [BMI139]: Add a pointer to the TEAS slicing framework.

Commenté [BMI140]: Please map this to the terms used in the teas framework as there is no such a thing in the slice framework.

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

Further, the ~~Network Slice Controller~~TSC can use the Mapping Community on the service route to map traffic to the desired ~~Transport Slice~~IETF Network Slice.

18. SRv6 ~~support~~Support

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

[RFC8986] specifies the SRv6 Endpoint behaviors (End USD, End.BM, End.B6.Encaps). [SRV6-INTER-DOMAIN] specifies the SRv6 Endpoint behaviors (END.REPLACE, END.REPLACEB6 and END.DB6). These are leveraged for BGP CT routes with SRv6 data plane.

The BGP Classful Transport route update for SRv6 MUST include an attribute containing SRv6 SID information. This may be either the BGP Prefix-SID attribute as specified in [RFC9252] or the BGP MultiNexthop attribute as specified in BGP MultiNexthop Attribute [MULTI-NH-ATTR] section 5.5.3.3. If the Prefix-SID attribute is used, it MUST NOT include SRv6 SID structure for Transposition described in [RFC9252].

It should be noted that prefixes carried in BGP CT family are transport layer end-points, e.g., PE loopback addresses. Thus, the SRv6 SID carried in a BGP CT route is also a transport layer identifier. For an illustration of BGP CT deployment in SRv6 newtorks, ~~please~~ refer to Appendix D~~ Appendix D~~.

19. Illustration of BGP CT ~~procedures~~ Procedures in Inter AS ~~option~~Option~~-~~ C

19.1. Reference Topology

Commenté [BMI141]: Would be good to have some IPv6 examples as well.

Commenté [BMI142]: The mapping between SN/BN should be provided.

```
                     [RR26]      [RR27]                       [RR16]
                        |           |                            |
```

a mis en forme : Pied de page

```
                   |            |                             |
                   |+-[ABR23]--+|+--[ASBR21]---[ASBR13]-+|+--[PE11]--+
                   ||           |||          `   /        |||          |
   [CE41]--[PE25]--[P28]       [P29]          `/         [P15]     [CE31]
                   |           |  |           /`         |  |          |
                   |           |  |          /  `        |  |          |
                   |           |  |         /     `      |  |          |
                   +--[ABR24]--+ +--[ASBR22]---[ASBR14]-+ +--[PE12]--+
         |                |                  |                   |
         +                +                  +                   +
      CE |     region-1   |   region-2       |                   |CE
     AS4               ...AS2...                    AS1            AS3

   41.41.41.41  ------------ Traffic Direction ---------->   31.31.31.31
                           Figure 4: Multi-Domain BGP CT Network
```

This example shows a provider network that consists of two ~~Autonomous systems~~ASes, AS1~~,~~ and AS2. They are serving customers AS3, AS4 respectively.

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

AS2 is further divided into two regions. So, there are three tunnel domains in provider's space. AS1 uses ISIS Flex-Algo intra-domain tunnels, whereas AS2 uses RSVP-TE intra-domain tunnels.

The network ~~has~~ exposes two Transport ~~classes~~Classes: Gold with ~~transport~~ Transport ~~class~~ Class id 100, Bronze with ~~transport~~ Transport ~~class~~ Class id 200. These ~~transport~~ Transport ~~classes~~ Classes are provisioned at the PEs and the ~~Border~~ BNs~~nodes~~ (ABRs, ASBRs) in the network.

Following tunnels exist for Gold ~~transport~~ Transport ~~class~~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

Following tunnels exist for Bronze Transport Class:~~transport class.~~

a mis en forme : En-tête

Commenté [BMI143]: Please use IP addresses that are reserved for documentation

Commenté [BMI144]: Please check the numbering of the figures

Commenté [BMI145]: Based on which criteria the mapping is done between adjacent domains?

Commenté [BMI146]: I don't think the SLA points to a class, but that the network selects a class that satisfies an SLA.

Otherwise, there is a dependency on the engineering vs. service offering, which is undesirable.

Commenté [BMI147]: Cite a ref

a mis en forme : Pied de page

PE25_to_ABR23_bronze - RSVP-TE tunnel

ABR23_to_ASBR21_bronze - RSVP-TE tunnel

ABR23_to_ASBR22_bronze - RSVP-TE tunnel

ABR24_to_ASBR21_bronze - RSVP-TE tunnel

ASBR13_to_PE12_bronze - ISIS FlexAlgo tunnel

ASBR14_to_PE11_bronze - ISIS FlexAlgo tunnel

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

19.2. Service Layer ~~R~~route ~~exchange~~Exchange

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

The PEs see each other as nexthop in the BGP Update for service family (SAFIs 1, 128) routes. Addpath send, receive is enabled on both directions on the EBGP multihop session between RR16 and RR26 for SAFIs 1, 128. Addpath send is negotiated in the RR to PE direction in each AS. This is to avoid path hiding of service routes at RR. E.g. SAFI 1 routes advertised by both PE11 and PE12. Or, SAFI 128 routes originated by both PE11 and PE12 using same RD.

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

As an example, CE31 advertises a route for prefix 31.31.31.31 with nexthop as self to PE11, PE12. CE31 can attach a Mapping Community Color:0:100 on this route, to indicate its request for Gold SLA. Or, PE11 can attach the same using locally configured policies.

Consider, CE31 is getting VPN service from PE11. The RD1:31.31.31.31 route is readvertised in SAFI 128 by PE11 with nexthop self (1.1.1.1) and label V-L1, to RR16 with the Mapping Community Color:0:100 attached. RR16 advertises this route with Addpath-ID to RR26 which readvertises to PE25 with nexthop unchanged. Now, PE25 can resolve the PNH 1.1.1.1 using transport routes received in BGP CT or BGP LU.

Using Addpath, service routes advertised by PE11 and PE12 for SAFIs 1, 128 reach PE25 via RR16, RR26 with the nexthop unchanged, as PE11 or PE12.

The IP FIB at PE25 VRF will have a route for 31.31.31.31 with a nexthop when resolved, that points to a Gold tunnel in ingress domain.

19.3. Transport Layer ~~R~~route ~~propagation~~Propagation

a mis en forme : En-tête

Commenté [BMI148]: What about AFIs ?

a mis en forme : Pied de page

a mis en forme : En-tête

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

Commenté [BMI149]: That is ?

PE11 is provisioned with transport class 100, RD value 1.1.1.1:10 and a transport-target:0:100 for Gold tunnels. And a Transport class 200 with RD value 1.1.1.1:20, and transport route target 0:200 for Bronze tunnels. Similarly, PE12 is provisioned with transport class 100, RD value 1.1.1.2:10 and a transport-target:0:100 for Gold tunnels. And transport class 200, RD value 1.1.1.2:20 with transport-target:0:200 for Bronze tunnels. Note that in this example, the BGP CT routes carry only the transport class route target, and no IP address format route target.

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

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

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

PE11 is provisioned to originate BGP CT route with Gold SLA to endpoint PE11. This route is sent with NLRI RD prefix 1.1.1.1:10:1.1.1.1, Label B-L0, nexthop 1.1.1.1 and a route target extended community transport-target:0:100. Label B-L0 can either be Implicit Null (Label 3) or an Ultimate Hop Pop (UHP) label.

This route is received by ASBR13 and it resolves over the tunnel ASBR13_to_PE11_gold. The route is then readvertised by ASBR13 in BGP CT family to ASBRs ASBR21, ASBR22 according to export policy. This route is sent with same NLRI RD prefix 1.1.1.1:10:1.1.1.1, Label B-L1, nexthop self, and transport-target:0:100. MPLS swap route is installed at ASBR13 for B-L1 with a nexthop pointing to ASBR13_to_PE11_gold tunnel.

Similarly, ASBR14 also receives BGP CT route for 1.1.1.1:10:1.1.1.1 from PE11 and it resolves over the tunnel ASBR14_to_PE11_gold. The route is then readvertised by ASBR14 in BGP CT family to ASBRs ASBR21, ASBR22 according to export policy. This route is sent with same NLRI RD prefix 1.1.1.1:10:1.1.1.1, Label B-L2, nexthop self, and transport-target:0:100. MPLS swap route is installed at ASBR14 for B-L1 with a nexthop pointing to ASBR14_to_PE11_gold tunnel.

In the Bronze plane, BGP CT route with Bronze SLA to endpoint PE11 is originated by PE11 with a NLRI containing RD prefix

a mis en forme : Pied de page

1.1.1.1:20:1.1.1.1, and appropriate label. The RD allows both Gold and Bronze advertisements traverse path selection pinchpoints without any path hiding at RRs or ASBRs. And route target extended community transport-target:0:200 lets the route resolve over Bronze tunnels in the network, similar to the process being described for Gold SLA path.

Moving back to the Gold plane, ASBR21 receives the Gold SLA BGP CT routes for NLRI RD prefix 1.1.1.1:10:1.1.1.1 over the single hop EBGP sessions from ASBR13, ASBR14, and can compute ECMP/FRR towards them. ASBR21 readvertises BGP CT route for 1.1.1.1:10:1.1.1.1 with nexthop

self (loopback address 2.2.2.1) to RR27, advertising a new label B-L3. MPLS swap route is installed for label B-L3 at ASBR21 to swap to received label B-L1, B-L2 and forward to ASBR13, ASBR14 respectively. RR27 readvertises this BGP CT route to ABR23, ABR24 with label and nexthop unchanged.

Similarly, ASBR22 receives BGP CT route 1.1.1.1:10:1.1.1.1 over the single hop EBGP sessions from ASBR13, ASBR14, and readvertises with nexthop self (loopback address 2.2.2.2) to RR27, advertising a new label B-L4. MPLS swap route is installed for label B-L4 at ASBR22 to swap to received label B-L1, B-L2 and forward to ASBR13, ASBR14 respectively. RR27 readvertises this BGP CT route also to ABR23, ABR24 with label and nexthop unchanged.

Addpath is enabled for BGP CT family on the sessions between RR27 and ASBRs, ABRs such that routes for 1.1.1.1:10:1.1.1.1 with the nexthops

ASBR21 and ASBR22 are reflected to ABR23, ABR24 without any path hiding. Thus giving ABR23 visibility of both available nexthops for Gold SLA.

ABR23 receives the route with nexthop 2.2.2.1, label B-L3 from RR27. The route target "transport-target:0:100" on this route acts as Mapping Community, and instructs ABR23 to strictly resolve the nexthop using transport class 100 routes only. ABR23 is unable to find a route for 2.2.2.1 with transport class 100. Thus, it considers this route unusable and does not propagate it further. This prunes ASBR21 from Gold SLA tunneled path.

ABR23 also receives the route with nexthop 2.2.2.2, label B-L4 from RR27. The route target "transport-target:0:100" on this route acts as Mapping Community, and instructs ABR23 to strictly resolve the nexthop using transport class 100 routes only. ABR23 successfully resolves the nexthop to point to ABR23_to_ASBR22_gold tunnel. ABR23 readvertises this BGP CT route with nexthop self (loopback address 2.2.2.3) and a new label B-L5 to PE25. Swap route for B-L5 is installed by ABR23 to swap to label B-L4, and forward into ABR23_to_ASBR22_gold tunnel.

PE25 receives the BGP CT route for prefix 1.1.1.1:10:1.1.1.1 with label B-L5, nexthop 2.2.2.3 and transport-target:0:100 from RR26. And it similarly resolves the nexthop 2.2.2.3 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 egress-domain is extended by BGP CT until the ingress-node PE25 in

a mis en forme : En-tête

a mis en forme : Pied de page

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 1.1.1.1:10:1.1.1.1 with nexthop self towards PE25.

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

When PE25 receives service routes from RR26 with nexthop 1.1.1.1 and mapping community Color:0:100, it resolves over this BGP CT route 1.1.1.1:10:1.1.1.1. Thus pushing label B-L5, and pushing as top label the labels associated with PE25_to_ABR23_gold tunnel.

19.4. Data ~~plane~~ Plane ~~view~~View

19.4.1. Steady ~~state~~State

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

CE41 transmits an IP packet with destination as 31.31.31.31. 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 MPLS packet with label V-L1(innermost), B-L5, and top label as PE25_to_ABR23_gold tunnel. This MPLS packet is thus transmitted to ABR23 using Gold SLA.

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

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

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

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

19.4.2. Local ~~repair~~ Repair of ~~primary~~ Primary ~~path~~Path

This section describes how the data plane at ASBR22 reacts when the

a mis en forme : En-tête

a mis en forme : Pied de page

a mis en forme : En-tête

link between ASBR22 and ASBR13 experiences a failure, and an alternate path exists.

Assuming ASBR22_to_ASBR13 link goes down, such that traffic with Gold SLA going to PE11 needs repair. ASBR22 has an alternate BGP CT route for 1.1.1.1:10:1.1.1.1 from ASBR14. This has been preprogrammed in forwarding by ASBR22 as FRR backup nexthop 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).

19.4.3. Absorbing ~~failure~~ Failure of ~~primary~~ Primary ~~path~~Path~~.~~ : Fallback to ~~best~~Best~~-~~ E~~e~~ffort ~~tunnels~~Tunnels~~.~~

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

Assuming tunnel ABR23_to_ASBR22_gold goes down, such that now end-to-end Gold path does not exist in the network. This makes the BGP CT route for RD prefix 1.1.1.1:10:1.1.1.1 unusable at ABR23. This makes ABR23 send a BGP withdrawal for 1.1.1.1:10:1.1.1.1 to PE25.

Withdrawal for 1.1.1.1:10:1.1.1.1 allows PE25 to react to the loss of gold path to 1.1.1.1. Assuming PE25 is provisioned to use best-effort transport class as the backup path, this withdrawal of BGP CT route allows PE25 to adjust the nexthop of the VPN Service-route to push the labels provided by the BGP LU route. That repairs the traffic to go via best effort path. PE25 can also be provisioned to use Bronze transport class as the backup path. The repair will happen in similar manner in that case as-well.

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

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

20. Deployment ~~considerations~~Considerations~~.~~

20.1. Managing Intent at Service and Transport layers~~.~~

Illustration of BGP CT Procedures (Section 19) shows multiple domains that agree on a color name space (Agreeing Color Domains) and contain tunnels with equivalent set of colors (Homogenous Color Domains).

However, in the real world, this may not always be guaranteed. Two domains may independently manage their color namespaces, these are known as Non-Agreeing Color Domains. Two domains may have tunnels with unequal set 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. Example described in this section use Inter AS option C domains. But similar mechanisms will work for

a mis en forme : Pied de page

Inter AS ~~option~~ Option A and Inter AS option B scenarios as-well.

20.1.1. Service ~~layer~~ Layer Color Management

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

As explained in Nexthop Resolution Scheme (Section 6), mapping community carried on service route maps to a resolution scheme. The mapping community values for the service route can be abstract and does not require 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.

20.1.2. 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 namespace to another.

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

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

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

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

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

Figure 5: Transport Layer with Non-agreeing Color Domains

In the above topology, we have three Autonomous Systems. All the nodes in the topology supports BGP CT.

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

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

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

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

The service route carries an abstract mapping community that maps to

a mis en forme : En-tête

a mis en forme : Pied de page

a mis en forme : En-tête

the required SLA. For example, Service routes that need to resolve over gold transport tunnels, carries a mapping community color:0:100500. In AS3 it maps to a resolution scheme containing TRDB with color 500 whereas in AS2 it maps to a TRDB with color 300 and in AS1 it maps to a TRDB with color 100. Co-ordination is needed to provision the resolution schemes in each domain as explained above.

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

Inter-domain traffic forwarding in the above topology works as explained in Section 19.

Transport-target re-write requires co-ordination of color values between domains in the transport layer. This method avoids the need to re-write service route mapping community, keeping the service layer homogenous and simple to manage. Co-ordinating transport-class route-target between adjacent domains is easier than ~~co-ordinating~~coordinating service layer colors deployed in various non-adjacent domains.

20.1.3. Heterogeneous Agreeing Color Transport Domains

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

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

```
          <---------- Service Routes SAFI-128 ------------------------

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

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

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

Figure 6: ~~Tranport~~Transport Layer with Heterogenous Color Domains

In the ~~above~~ topology depicted in Figure X, ~~we have~~there are three ~~Autonomous Systems~~ASes. All the nodes in the topology support BGP CT.

a mis en forme : Pied de page

a mis en forme : En-tête

In AS1 Gold SLA is represented by color 100.

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

In AS3 Gold SLA is represented by color 100.

This problem can be solved by two approaches, described below.

20.1.3.1. Duplicate ~~tunnels~~ Tunnels ~~approach~~Approach

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

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

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

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

20.1.3.2. Customized Resolution ~~schemes~~ Schemes ~~approach~~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 transport and forwarding plane.

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.

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

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

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

a mis en forme : Pied de page

a mis en forme : En-tête

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

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

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

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

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

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

20.2. Migration ~~scenarios~~Scenarios~~.~~

20.2.1. BGP CT ~~islands~~ Islands ~~connected~~ Connected via BGP LU ~~domain~~Domain~~.~~

This section explains how an end-to-end SLA can be achieved while transiting a domain that does not support BGP CT (SAFI 76). BGP LU (SAFI 4) is used in such domains to connect the BGP CT islands.

```
                +----------EBGP Multihop CT-------------+
                |                                       |

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

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

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

Figure 7: BGP CT in AS1 and AS3 connected by BGP LU in AS2

a mis en forme : Surlignage

In the above topology, there are three ASes~~ domains~~. AS1 and AS3 supports BGP CT~~. ~~, while AS2 ~~is a domain that~~ 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

a mis en forme : Pied de page

a mis en forme : En-tête

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 interdomain links. ASBR11 and ASBR31 have reachability to each other's loopbacks through BGP LU. ASBR11 and ASBR31 negotiate BGP CT family over a multihop EBGP session formed using BGP LU reachability.

The following tunnels exist for Gold ~~transport~~ Transport ~~class~~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

Following tunnels exist for Bronze Transport class:~~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 class:~~transport class~~ gold and bronze, and are advertised between ASBR31 and ASBR11 with Nexthop

self.

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

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

Commenté [BMI150]: That is ?

ASBR21_to_ASBR22_lpbk_gold - RSVP-TE tunnel

ASBR22_to_ASBR21_lpbk_gold - RSVP-TE tunnel

ASBR21_to_ASBR22_lpbk_bronze - RSVP-TE tunnel

ASBR22_to_ASBR21_lpbk_bronze - RSVP-TE tunnel

RD:PE11 BGP CT route is originated from PE11 towards ASBR11 with transport-target gold. ASBR11 readvertises this route with nexthop

set to ASBR11_lpbk_gold on the EBGP multihop session towards ASBR31. ASBR11 originates BGP LU route for endpoint ASBR11_lpbk_gold on EBGP session to ASBR21 with a 'gold SLA' community, and BGP LU route for ASBR11_lpbk_bronze with a 'bronze SLA' community. The SLA community is used by ASBR31 to publish the BGP LU routes in the corresponding BGP CT TRDBs.

a mis en forme : Pied de page

a mis en forme : En-tête

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

ASBR31 adds the ASBR11_lpbk_gold route received via EBGP LU from ASBR22 to gold TRDB based on the received 'gold SLA' community. ASBR31 uses this gold TRDB route to resolve the nexthop.

.

ASBR11_lpbk_gold received on BGP CT route with transport-target gold, for the prefix RD:PE11 received over the EBGP multihop CT session, thus preserving the end-to-end SLA. Now ASBR31 readvertises the BGP CT route for RD:PE11 with nexthop as self thus stitching with the BGP LU LSP in AS2. Intradomain traffic forwarding in AS1 and AS3 follows the procedures as explained in Illustration of CT Procedures (Section 19).

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

20.2.2. BGP CT - Interop between MPLS and O~~o~~ther ~~forwarding~~Forwarding ~~technologies~~Technologies~~.~~

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

20.2.2.1. Interop ~~between~~ Between MPLS and SRv6 N~~n~~odes.

BGP speakers may carry MPLS label and SRv6 SID in BGP CT SAFI 76 routes using protocol encoding as described in Carrying Multiple Encapsulation information (Section 7.1).

MPLS Labels are carried using RFC 8277 encoding, and SRv6 SID is carried using Prefix SID attribute as specified in RFC 9252 [RFC9252].

```
                      RR1--+
                             \  +-------R2  [MPLS + SRv6]
                              \ |

                      R1--------P-------R3  [MPLS only]
                [MPLS + SRv6]   |
                                +-------R4  [SRv6 only]

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

Figure 8: BGP CT Interop between MPLS and SRv6 nodes

This example shows a provider network with a mix of devices with different forwarding capabilities. R1 and R2 support forwarding both MPLS and SRv6 packets. R3 supports forwarding MPLS packets only. R4 supports forwarding SRv6 packets only. All these nodes have BGP

a mis en forme : Pied de page

a mis en forme : En-tête

session with Route Reflector RR1 which reflects routes between these nodes with nexthop unchanged. BGP CT family is negotiated on these sessions.

R1 and R2 send and receive both MPLS label and SRv6 SID in the BGP CT control plane routes. This allows them to be ingress and egress for both MPLS and SRv6 data planes. MPLS label is carried using RFC 8277 encoding, and SRv6 SID is carried using Prefix SID attribute as specified in RFC 9252 [RFC9252], without Transposition Scheme. The Transposition Length is set to 0 and Transposition Offset is set to 0 to indicate nothing is transposed and that the entire SRv6 SID value is encoded in the SID Information Sub-TLV. In this way, either MPLS or SRv6 forwarding can be used between R1 and R2.

R1 and R3 send and receive MPLS label in the BGP CT control plane routes using RFC 8277 encoding. This allows them to be ingress and egress for MPLS data plane. R1 will carry SRv6 SID in Prefix-SID attribute, which will not be used by R3. In order to interoperate with MPLS only device R3, R1 MUST NOT use SRv6 Transposition scheme described in RFC 9252 [RFC9252]. The Transposition Length is set to 0 and Transposition Offset is set to 0 to indicate nothing is transposed and that the entire SRv6 SID value is encoded in the SID Information Sub-TLV. MPLS forwarding will be used between R1 and R3.

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

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

20.2.2.2. Interop ~~between~~ Between ~~nodes~~ Nodes ~~supporting~~ Supporting MPLS and UDP ~~tunneling~~Tunneling~~.~~

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

MPLS Labels are carried using RFC 8277 encoding, and UDP (or IP) tunneling information is carried using TEA attribute or the Encapsulation Extended Community as specified in ~~RFC 9012~~ [RFC9012].

```
                      RR1--+
                             \  +-------R2  [MPLS + UDP]
                              \ |

                      R1--------P-------R3  [MPLS only]
                [MPLS + UDP]    |
                                +-------R4  [UDP only]

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

a mis en forme : Pied de page

Figure 9: 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 nexthop unchanged. BGP CT family is negotiated on these sessions.

R1 and R2 send and receive both MPLS label and UDP tunneling info in the BGP CT control plane routes. This allows them to be ingress and egress for both MPLS and UDP tunneling data planes. MPLS label is carried using RFC 8277 encoding. As specified in ~~RFC 9012~~ [RFC9012], UDP tunneling information is carried using TEA attribute (code 23) or the "barebones" Tunnel TLV carried in Encapsulation Extended Community. Either MPLS or UDP tunneled forwarding can be used between R1 and R2.

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

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

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

20.3. 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 churn in a multi-domain network.

Consider a multi-domain BGP CT network as illustrated in the figure below.

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

         +--------ASBR11     +--PE11 (EP1)
         |              \   /

  +----ASBR31            (P)----PE12 (EP2)
  |      |              / | \

  |      +--------ASBR12  |  +--PE13 (EP3)
  |                       |

  |                       +-----PE14 (EP4)
```

a mis en forme : En-tête

Commenté [BMI151]: Please call out explicitly the section you are referring to.

a mis en forme : Pied de page

a mis en forme : En-tête

```
PE31--(P)
       |

       |

       |      +--------ASBR21     +--PE21 (EP5)
       |      |              \   /

       +----ASBR32            (P)----PE22 (EP6)
              |              / | \

              +--------ASBR22  |  +--PE22 (EP7)
                               |

                               +-----PE24 (EP8)

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

Figure 10: Multi-Domain Network

The following table details the BGP CT route and path visibility at PE31-- for each TC.

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

Figure 11: Route and Path Visibility at Ingress Node

In the above example, both route churn and TE granularity are directly proportional to the number of CT labels received.

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

21. IANA Considerations

This document makes the following requests ~~of~~ to IANA.

a mis en forme : Pied de page

21.1. New BGP SAFI

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

Registry Group: Subsequent Address Family Identifiers (SAFI) Parameters

Registry Name: SAFI Values

Value	Description
76	Classful-Transport SAFI

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

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

* "Inet6, Classful Transport". AFI/SAFI = "2/76" for carrying IPv6 Classful Transport prefixes.

21.2. New Format for BGP Extended Community

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

the "BGP Transitive Extended Community Types" registry, and

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

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

This document uses this new Format with subtype 0x2 (route target), as a transitive extended community.

The Route Target thus formed is called "Transport Class" route target extended community.

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

21.2.1. Existing ~~registries~~ Registries to be ~~modified~~Modified

21.2.1.1. Registries for the "Type" Field

21.2.1.1.1. 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

a mis en forme : En-tête

Commenté [BMI152]: Please add a pointer to the IANA registry

a mis en forme : Surlignage

Commenté [BMI153]: You may add URLs tohese registries. This wil make IANA operator's life easier. Thanks.

a mis en forme : Pied de page

Type Value	Name
0x0a	Transport Class

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

21.2.1.1.2. 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	Transport Class

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

21.2.2. New ~~registries~~ Registries ~~to be created~~

21.2.2.1. Transitive Transport Class Extended Community Sub-Types Registry

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

~~Registry Group: Border Gateway Protocol (BGP) Extended Communities~~

Registry Name: Transitive Transport Class Extended Community Sub-Types

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

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

Sub-Type Value	Name
0x02	Route Target

21.2.2.2. Non-Transitive Transport Class Extended Community Sub-Types

a mis en forme : En-tête

a mis en forme : Pied de page

Registry

IANA is requested to the following subregistry under the the "Border Gateway Protocol (BGP) Extended Communities":
~~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 Procedures
0x00-0xBF	First Come First Served
0xC0-0xFF	IETF Review

Sub-Type Value	Name
0x02	Route Target

21.3. MPLS OAM ~~C~~ode ~~points~~Points

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

* IPv4 BGP Classful Transport

* IPv6 BGP Classful Transport

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

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

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

21.4. Best Effort Transport Class ID

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

Registry Group: BGP CT Parameters

Registry Name: Transport Class ID

Value Name

a mis en forme : En-tête

Commenté [BMI154]: Please add a link to the registry.

Commenté [BMI155]: I guess this TBD1

Commenté [BMI156]: Thsi corresponds to TBD2

a mis en forme : Pied de page

a mis en forme : En-tête

-----------------+-------------------------------

0 Best Effort Transport Class ID

22. Security Considerations

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

When redistributing between SAFI 4 and SAFI 76 Classful Transport routes, there is a possibility of SAFI 4 routes mixing with SAFI 1 service routes. To avoid such scenarios, it is RECOMMENDED that implementations support keeping SAFI 4 routes in a separate transport RIB, distinct from service RIB that contain SAFI 1 service routes.

In scenarios where MPLS is enabled on link to a device in an untrusted domain, e.g., a PE-CE link or ASBR-ASBR inter-AS link, security can be provided against MPLS label spoofing by using MPLS context tables as described in MPLS enabled CE (Section 14.2). Such that only MPLS traffic with labels advertised to the BGP speaker are allowed to forward. However, the PE may not be able to perform any checks based on inner payload in the MPLS packet since it performs label swap forwarding. Such ‘inner ~~payload~~ payload'-based checks may be offloaded to a downstream node that forwards and processes inner payload, e.g., an IP FIB router. These security aspects should be considered when using MPLS enabled CE devices.

Commenté [BMI157]: IP router :-)

a mis en forme : Anglais (États-Unis)

23. References

23.1. 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>.

Commenté [BMI158]: Is this normative ?

[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>.

[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>.

[RFC4456] Bates, J., Ed., Chen, Ed., and Chandra, Ed., "BGP Route

a mis en forme : Pied de page

Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", April 2006, <https://datatracker.ietf.org/doc/html/rfc4456#section-9>.

[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>.

[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>.

[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>.

[RFC8664] Sivabalan, S., Filsfils, C., Tantsura, J., Henderickx, W., and J. Hardwick, "Path Computation Element Communication Protocol (PCEP) Extensions for Segment Routing", RFC 8664, DOI 10.17487/RFC8664, December 2019, <https://www.rfc-editor.org/info/rfc8664>.

[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>.

[RFC8986] Filsfils, C., Ed., Camarillo, P., Ed., Leddy, J., Voyer, D., Matsushima, S., and Z. Li, "Segment Routing over IPv6

a mis en forme : En-tête

Commenté [BMI159]: I don't think this is normative

a mis en forme : Pied de page

(SRv6) Network Programming", RFC 8986, DOI 10.17487/RFC8986, February 2021, <https://www.rfc-editor.org/info/rfc8986>.

[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>.

[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>.

[SRTE] Talaulikar, Ed. and S. Previdi, "Advertising Segment Routing Policies in BGP", 28 January 2023, <https://tools.ietf.org/html/draft-ietf-idr-segment-routing-te-policy-20>.

23.2. Informative References

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

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

[Colorful-Prefix-Routing-SRv6] Wang, Ed., "BGP Colorful Prefix Routing for SRv6 based Services", 26 March 2023, <https://www.ietf.org/archive/id/draft-wang-idr-cpr-01.html>.

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

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

[MNH-ENCAP-DSCP] Vairavakkalai, Ed., "BGP MultiNexthop Attribute", 17 June 2023, <https://www.ietf.org/archive/id/draft-kaliraj-idr-

a mis en forme : En-tête

Commenté [BMI160]: Is this one normative ?

Commenté [BMI161]: I think this one is informative.

Commenté [BMI162]: This is provided as an implementation example. I tend to see this as informative.

a mis en forme : Pied de page

multinexthop-attribute-06.html#section-5.4.3.4>.

[MNH-ENCAP-MPLS]
Vairavakkalai, Ed., "BGP MultiNexthop Attribute", 17 June 2023, <https://www.ietf.org/archive/id/draft-kaliraj-idr-multinexthop-attribute-06.html#section-5.4.3.1>.

[MNH-EP] Vairavakkalai, Ed., "BGP MultiNexthop Attribute", 17 June 2023, <https://www.ietf.org/archive/id/draft-kaliraj-idr-multinexthop-attribute-06.html#section-5.4.1>.

[MNH-TC] Vairavakkalai, Ed., "BGP MultiNexthop Attribute", 17 June 2023, <https://www.ietf.org/archive/id/draft-kaliraj-idr-multinexthop-attribute-06.html#section-5.4.2.2>.

[MPLS-NAMESPACES]
Vairavakkalai, Ed., "BGP signalled MPLS namespaces", 7 August 2023, <https://datatracker.ietf.org/doc/html/draft-kaliraj-bess-bgp-sig-private-mpls-labels-05#section-6.1>.

[MULTI-NH-ATTR]
Vairavakkalai, Ed., "BGP MultiNexthop Attribute", 17 June 2023, <https://datatracker.ietf.org/doc/html/draft-kaliraj-idr-multinexthop-attribute-06>.

[PCEP-RSVP-COLOR]
Rajagopalan, Ed. and Pavan. Beeram, Ed., "Path Computation Element Protocol(PCEP) Extension for RSVP Color", 3 January 2023, <https://datatracker.ietf.org/doc/html/draft-ietf-pce-pcep-color-00>.

[RTC-Ext] Zhang, Z., Ed. and Haas, Ed., "Generic Route Constraint Distribution Mechanism for BGP", 26 June 2023, <https://tools.ietf.org/html/draft-zzhang-idr-bgp-rt-constrains-extension-03#section-2>.

[SRV6-INTER-DOMAIN]
K A, Ed., "SRv6 inter-domain mapping SIDs", 23 June 2023, <https://datatracker.ietf.org/doc/html/draft-salih-spring-srv6-inter-domain-sids-03>.

Appendix A. Applicability to ~~Intra-~~Intra-AS and ~~different~~ Different Inter AS ~~deployments~~Deployments~~.~~

As described in ~~BGP VPN~~Section 10 of [RFC4364] ~~Section 10~~, in an ~~o~~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 (SAFI 4) or BGP CT (SAFI 76).

Illustration of CT Procedures (Section 19) illustrates how constructs of BGP CT work in an Inter AS option-C deployment. The BGP CT constructs: SAFI 76, Transport Class and Resolution Scheme are used in an option-C deployment.

a mis en forme : En-tête

a mis en forme : Pied de page

In Intra AS and Inter AS ~~option-A, option-B~~Options A&B scenarios, SAFI 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~~ intra AS and Inter AS Options ~~option~~ A~~,~~ and B deployment scenarios.

A.1. Intra AS U~~u~~se ~~case~~Case

A.1.1. Topology

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

           10.21.21.21 ---- Traffic Direction ----> 10.31.31.31
```

Figure 12: BGP CT Intra-AS.

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

A.1.2. ~~Transport~~ IP/MPLS Layer

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

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

Following tunnels exist for Gold ~~transport~~ Transport ~~class~~Class~~.~~:

PE11_to_PE12_gold - RSVP-TE tunnel

PE12_to_PE11_gold - RSVP-TE tunnel

Following tunnels exist for Bronze Transport Class:~~transport class.~~

PE11_to_PE12_bronze - RSVP-TE tunnel

PE11_to_PE12_bronze - RSVP-TE tunnel

These tunnels are provisioned to belong to Transport Class:~~transport class~~ 100 or 200.

a mis en forme : En-tête

Commenté [BMI163]: Please use the documentation address blocks.

Commenté [BMI164]: TSV poeple will be disturbed by this naming.

a mis en forme : Pied de page

A.1.3. Service Layer ~~route ~~Route E~~e~~xchange

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

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

CE31 advertises a route for example prefix 10.31.31.31 with nexthop self to PE12. CE31 can attach a Mapping Community Color:0:100 on this route, to indicate its request for Gold SLA. Or, PE11 can attach the same using locally configured policies.

Consider, CE31 is getting VPN service from PE12. The RD:10.31.31.31 route is readvertised in SAFI 128 by PE12 with nexthop self (10.12.12.12) and label V-L1, to RR11 with the Mapping Community Color:0:100 attached. This SAFI 128 route reaches PE11 via RR11 with the nexthop unchanged as PE12 and label V-L1. Now PE11 can resolve the PNH 10.12.12.12 using PE11_to_PE12_gold RSVP TE LSP.

The IP FIB at PE11 VRF will have a route for 10.31.31.31 with a nexthop when resolved using Resolution Scheme belonging to the mapping community Color:0:100, points to a PE11_to_PE12_gold tunnel.

BGP CT SAFI 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.

A.2. Inter AS ~~option~~Option~~-~~ A U~~u~~se C~~e~~ase

A.2.1. Topology

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

10.31.31.31          ---- Traffic Direction ---->          10.41.41.41
```

Figure 13: BGP CT Inter-AS option A.

This example 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 ~~f~~in this example~~or this traffic~~), when traversing these provider core networks.

a mis en forme : En-tête

a mis en forme : Pied de page

a mis en forme : En-tête

A.2.2. Transport Layer

Commenté [BMI165]: Idem as above

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

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

Following tunnels exist for Gold Transport Class:~~transport class.~~

PE11_to_ASBR11_gold - RSVP-TE tunnel

ASBR11_to_PE11_gold - RSVP-TE tunnel

PE21_to_ASBR21_gold - SRTE tunnel

ASBR21_to_PE21_gold - SRTE tunnel

Following tunnels exist for Bronze Transport Class:~~transport class.~~

PE11_to_ASBR11_bronze - RSVP-TE tunnel

ASBR11_to_PE11_bronze - RSVP-TE tunnel

PE21_to_ASBR21_bronze - SRTE tunnel

ASBR21_to_PE21_bronze - SRTE tunnel

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

A.2.3. Service Layer R~~r~~oute ~~exchange~~Exchange

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

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

CE41 advertises a route for example prefix 10.41.41.41 with nexthop 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:10.41.41.41 route is readvertised in SAFI 128 by PE21 with nexthop self (10.21.21.21) and label V-L1, to RR21 with the Mapping Community Color:0:100 attached. This SAFI 128 route reaches ASBR21 via RR21 with the nexthop unchanged as PE21 and label V-L1. Now ASBR21 can resolve the PNH 10.21.21.21 using ASBR21_to_PE21_gold SRTE LSP.

The IP FIB at ASBR21 VRF will have a route for 10.41.41.41 with a nexthop resolved using Resolution Scheme associated with mapping

a mis en forme : Pied de page

a mis en forme : En-tête

community Color:0:100, pointing to ASBR21_to_PE21_gold tunnel.

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

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

BGP CT SAFI 76 is not used in this Inter AS option-A deployment. But the Transport class and Resolution Scheme constructs are used to preserve end-to-end SLA.

A.3. Inter AS O~~o~~ption~~-~~ B ~~usecase~~Use Case

A.3.1. Topology

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

10.31.31.31          ---- Traffic Direction ---->            10.41.41.41
```

Figure 14: BGP CT Inter-AS option B.

This example shows two provider network Autonomous systems AS1, AS2. They serve L3VPN customers AS3, 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.

A.3.2. Transport Layer

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

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

Following tunnels exist for Gold Transport Class:~~transport class.~~

PE11_to_ASBR12_gold - RSVP-TE tunnel

ASBR12_to_PE11_gold - RSVP-TE tunnel

a mis en forme : Pied de page

PE22_to_ASBR21_gold - SRTE tunnel

ASBR21_to_PE22_gold - SRTE tunnel

Following tunnels exist for Bronze Transport Class:~~transport class.~~

PE11_to_ASBR12_bronze - RSVP-TE tunnel

ASBR12_to_PE11_bronze - RSVP-TE tunnel

PE22_to_ASBR21_bronze - SRTE tunnel

ASBR21_to_PE22_bronze - SRTE tunnel

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

A.3.3. Service Layer ~~route~~ Route ~~exchange~~Exchange

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

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

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

CE41 advertises a route for example prefix 10.41.41.41 with nexthop self to PE22 VRF. CE41 can attach a Mapping Community Color:0:100 on this route, to indicate its request for Gold SLA. Or, PE22 can attach the same using locally configured policies.

Consider, CE41 is getting VPN service from PE22. The RD:10.41.41.41 route is readvertised in SAFI 128 by PE22 with nexthop self (10.22.22.22) and label V-L1, to RR23 with the Mapping Community Color:0:100 attached. This SAFI 128 route reaches ASBR21 via RR23 with the nexthop unchanged as PE22 and label V-L1. Now ASBR21 can resolve the PNH 10.22.22.22 using ASBR21_to_PE22_gold SRTE LSP.

Next, ASBR21 readvertises the RD:10.41.41.41 route with nexthop self to ASBR12, with a newly allocated MPLS label, V-L2. Forwarding for this label is installed to Swap V-L1, and Push labels for ASBR21_to_PE22_gold tunnel.

ASBR12 further readvertises the RD:10.41.41.41 route via RR13 to PE11 with nexthop self 10.12.12.12. PE1 resolves the nexthop 10.12.12.12 over PE11_to_ASBR11_gold RSVP TE tunnel.

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

BGP CT SAFI 76 is not used in this Inter AS option-B deployment. But

a mis en forme : En-tête

a mis en forme : Pied de page

the Transport class and Resolution Scheme constructs are used to preserve end-to-end SLA.

Appendix B. Why reuse RFC 8277 and RFC 4364?

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

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

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

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

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

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

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

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

B.1. Update packing considerations

BGP CT carries transport class as an attribute. This means routes that don't share the same transport class cannot be packed into same Update message. Update packing in BGP CT will be similar to RFC 8277

a mis en forme : En-tête

a mis en forme : Pied de page

family routes carrying attributes like communities or extended communities. Service families like SAFI 128 have considerably more scale than transport families like SAFI 4 or SAFI 76, which carry only loopbacks. Update packing mechanisms that scale for SAFI 128 routes will scale similarly for SAFI 76 routes also.

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

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

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

Appendix C. Scaling using BGP MPLS Namespaces

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

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

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

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

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

a mis en forme : En-tête

a mis en forme : Pied de page

a mis en forme : En-tête

C.1. Illustration.

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

C.2. Topology

```
                  [RR11]                            [RR31]
                     |                                |
                     |                                |
           [PE11]\   |  /[BN11]--+       +--[BN31]\   |   /[PE31]
                   \ | /          \     /          \  |  /
   [CE41]--[PE12]--[P11]           [BN21]           [P31]--[PE32]--[CE31]
            ..     /   \          /     \          /     \    ..
            ..    /     \[BN12]--+       +--[BN32]/       \   ..
           [PE11000]                                       [PE31000]
          |                  |                |                   |
    AS4   |    ..Domain1..   |  ..Domain2..   |    ..Domain3..    | AS3
          |                  |   (backbone)   |                   |

                     <---- Traffic Direction ----
```

Figure 15: BGP MPLS Namespaces.

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

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

At the Transport layer - in Domain1, PE11..PE11000 negotiate BGP families (SAFI 4, SAFI 76) with BN11, BN12. In Domain2, BN11 and BN12 similarly negotiate the transport families with BN21, which in turn peers with BN31 and BN32. In Domain3, BN31 and BN32 peer with PEs PE31..PE31000. Each of these BNs change BGP nexthop to self, when re advertising the SAFI 4, SAFI 76 transport routes.

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

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

a mis en forme : Pied de page

a mis en forme : En-tête

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

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

C.3. Context Protocol Nexthop Address (CPNH)

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

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

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

C.4. Service Forwarding Helper, and changes to transport layer.

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

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

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

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

C.5. BGP MPLS Namespace Address family (AFI:16399, SAFI:128)

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

a mis en forme : Pied de page

associated with CPNH1.

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

C.6. Changes to Service Layer route exchange

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

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

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

C.7. Analysis of forwarding behavior

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

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

Appendix D. BGP CT deployment in SRv6 networks

This section describes BGP CT deployment in SRv6 multi-domain network using Inter-AS Option C architecture.

D.1. SID stacking approach

This approach uses stacking of service SRv6 SID over transport SRv6 SID. Transport layer SIDs of types End, End.B6.Encaps defined in [RFC8986], and type END.REPLACE* defined in [SRV6-INTER-DOMAIN] are carried in SAFI 76. Service SID is carried in a service family like SAFI 1 or SAFI 128.

In this approach, the number of Service SIDs required at the egress SN is equal to service functions (e.g. Prefix, VRF or Nexthop) and the number of Transport SIDs are equal to the number of transport classes.

```
                AS1                     AS2
```

a mis en forme : En-tête

a mis en forme : Pied de page

```
            ---gold--->           ----gold-->
  CE1---[PE1---P---ASBR1]-----[ASBR2---P---PE2]---CE2
            --bronze-->           --bronze-->

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

Figure 16: BGP CT in SRv6 Only Data plane

In the above topology, there are two AS domains, AS1 and AS2. These are pure IPv6 domains, with no MPLS enabled. Inter-AS links between AS1 and AS2 are also enabled with IPv6 forwarding.

Intra-AS nodes in AS1 and AS2 speak IBGP CT (AFI: 2, SAFI: 76) and ISIS-SRv6 between them. The Inter-AS nodes ASBR1, ASBR2 speak EBGP CT (AFI: 2, SAFI:76) between them. Transport Classes Gold (100) and Bronze (200) are provisioned in all PEs and ASBRs. All BGP CT advertisements in this example carry a MPLS label value of 3 (Implicit Null) in the NLRI encoding.

Reachability between PE1 and PE2 is formed using BGP CT family. Service families like IPv4 unicast (AFI: 1, SAFI: 1) and L3VPN (AFI: 2, SAFI: 128) is negotiated on multihop EBGP session between PE1 and PE2. These service routes carry service SID to identify service functions at the advertising PE, and mapping community to identify the desired Intent.

The following SRv6 locators are provisioned:

PE2-SRv6~~-~~: SRv6 Locator for PE2 best effort transport class

PE2-SRv6-gold-loc~~-~~: SRv6 Locator for PE2 gold transport class

PE2-SRv6-bronze-loc~~-~~: SRv6 Locator for PE2 bronze transport class

ASBR1-SRv6-loc~~-~~: SRv6 Locator for ASBR1 best effort transport class

ASBR1-SRv6-gold-loc~~-~~: SRv6 Locator for ASBR1 gold transport class

ASBR1-SRv6-bronze-loc~~-~~: SRv6 Locator for ASBR1 bronze transport class

ASBR2-SRv6-loc~~-~~: SRv6 Locator for ASBR2 best effort transport class

ASBR2-SRv6-gold-loc~~-~~: SRv6 Locator for ASBR2 gold transport class

ASBR2-SRv6-bronze-loc~~-~~: SRv6 Locator for ASBR2 bronze transport class

The following transport layer SRv6 End SIDs are provisioned or dynamically allocated on demand:

PE2-SRv6-gold~~-~~: PE2 End SID from PE2-SRv6-gold-loc, for gold transport class.

PE2-SRv6-bronze~~-~~: PE2 End SID from PE2-SRv6-bronze-loc, for bronze transport class.

a mis en forme : En-tête

a mis en forme : Pied de page

ASBR2-SRv6-PE2-gold-Replace~~:~~: at ASBR2 End.B6.Encaps SID for PE2, gold transport class.

ASBR2-SRv6-PE2-bronze-Replace~~:~~: at ASBR2 End.B6.Encaps SID for PE2, bronze transport class.

ASBR1-SRv6-gold~~:~~: ASBR1 End SID from ASBR1-SRv6-gold-loc, for gold transport class.

ASBR1-SRv6-PE2-gold-Replace~~:~~: at ASBR1 End.REPLACE SID for PE2, gold transport class.

ASBR1-SRv6-bronze~~:~~: ASBR1 End SID from ASBR1-SRv6-bronze-loc, for bronze transport class.

ASBR1-SRv6-PE2-bronze-Replace~~:~~: at ASBR1 End.REPLACE SID for PE2, bronze transport class.

Architecturally, the forwarding semantic of End.REPLACE SID operation is similar to Label SWAP operation in MPLS data plane. When a route received with End SID (e.g. PE2-SRv6-gold or PE2-SRv6-bronze transport SIDs) is readvertised with nexthop self, a IPv6 forwarding entry is emitted with a forwarding semantic of End.B6.Encaps operation, which means: Update IPv6 DA with Next Segment in SRH, and Encapsulate SRv6 SID corresponding to the correct transport class. This can be seen in IPv6 FIB of ASBR2 during "BGP CT processing at ASBR2" in the following illustration:

The following service layer SRv6 End.DT4 SIDs are provisioned:

PE2-SRv6-S1-DT4~~:~~: PE2 End.DT4 SID for service S1

The locators for above provisioned SRv6 SIDs will be advertised via ISIS between Intra-AS nodes and the established SRv6 tunnel to the node's loopback will be installed into the corresponding TRDB based on color.

The SRv6 tunnel ingress routes are published in the Gold and Bronze TRDBs at ASBR2 as shown below:

```
  Gold TRDB routes at ASBR2

       [ISIS SRv6] PE2-LPBK
           NH:  Encap "Gold-SRv6-Tunnel-to-PE2" tunnel

       [ISIS SRv6] PE2-SRv6-gold
           NH:  Encap "Gold-SRv6-Tunnel-to-PE2" tunnel

  Bronze TRDB routes at ASBR2

       [ISIS SRv6] PE2-LPBK
           NH: Encap "Bronze-SRv6-Tunnel-to-PE2" tunnel

       [ISIS SRv6] PE2-SRv6-bronze:
           NH: Encap "Bronze-SRv6-Tunnel-to-PE2" tunnel

  ASBR2: IPv6 FIB for SRv6
```

a mis en forme : En-tête

a mis en forme : Pied de page

```
       [ISIS SRv6] PE2-SRv6-gold,
         NH: Encap "Gold-SRv6-Tunnel-to-PE2"

       [ISIS SRv6] PE2-SRv6-bronze,
         NH: Encap "Bronze-SRv6-Tunnel-to-PE2"
```

Figure 17: TRDBs at ASBR2

The illustrations that follow, show how the BGP CT route for gold transport plane is originated, import processing done and propagated through this network. Similar processing is followed for the bronze transport plane route as well.

Firstly, PE2 originates BGP CT route for its transport layer endpoints like Loopback address with SRv6 SID information to ASBR2 as shown below.

```
   IBGP CT routes from PE2 to ASBR2

       RD1:PE2-LPBK,
         transport-target:0:100,
         Prefix-SID: PE2-SRv6-gold
         NH: PE2-LPBK

       RD2:PE2-LPBK,
         transport-target:0:200,
         Prefix-SID: PE2-SRv6-bronze
         NH: PE2-LPBK

   PE2: IPv6 FIB for SRv6

       [BGP CT] PE2-SRv6-S1-DT4
         NH: Decap, Perform service S1
```

Figure 18: BGP CT advertisements from PE2 to ASBR2

When ASBR2 receives the IBGP CT advertisement for gold route from PE2, it performs import processing and nexthop resolution for the endpoint PE2-LPBK in the gold TRDB based on its transport-target:0:100. This would resolve over the ISIS-SRv6 route in gold TRDB and pick "Gold-SRv6-Tunnel-to-PE2" tunnel.

On successful resolution, a IPv6 transit route for ASBR2-SRv6-PE2-gold-replace/128 is installed in the global IPv6 FIB with "Gold-SRv6-Tunnel-to-PE2" tunnel as nexthop, enabling SRv6 forwarding for gold SLA. The BGP CT routes for RD1:PE2-LPBK is further advertised towards ASBR1 via EBGP CT as shown below. During this readvertisement, the nexthop is set to self, and SID is rewritten to ASBR2-SRv6-gold-Replace.

```
 EBGP CT routes from ASBR2 to ASBR1

     RD1:PE2-LPBK,
       transport-target:0:100,
       Prefix-SID: ASBR2-SRv6-PE2-gold-Replace,
       NH: ASBR2_InterAS_Link

     RD2:PE2-LPBK,
```

```
        transport-target:0:200,
        Prefix-SID: ASBR2-SRv6-PE2-bronze-Replace,
        NH: ASBR2_InterAS_Link

  ASBR2: IPv6 FIB for SRv6

      [BGP CT] ASBR2-SRv6-PE2-gold-Replace
        NH: UpdateIPv6DA(SRH.NextSegment), Encap "Gold-SRv6-Tunnel-to-
PE2"

      [BGP CT] ASBR2-SRv6-PE2-bronze-Replace
        NH: UpdateIPv6DA(SRH.NextSegment), Encap "Bronze-SRv6-Tunnel-to-
PE2"
```

Figure 19: BGP CT processing at ASBR2

When ASBR1 receives this EBGP CT advertisement from ASBR2, an IPv6 route for ASBR1-SRv6-gold-Replace/128 is installed with a nexthop of ASBR1_InterAS_Link in the global IPv6 FIB, enabling SRv6 forwarding for gold SLA. The BGP CT route for RD1:PE2-LPBK is further advertised to PE1 via IBGP CT, with nexthop set to self, and SID rewritten to ASBR1-SRv6-gold-Replace.

```
  IBGP CT routes from ASBR1 to PE1

      RD1:PE2-LPBK,
        transport-target:0:100,
        Prefix-SID: ASBR1-SRv6-PE2-gold-Replace,
        NH: ASBR1-LPBK

      RD2:PE2-LPBK,
        transport-target:0:200,
        Prefix-SID: ASBR1-SRv6-PE2-bronze-Replace,
        NH: ASBR1-LPBK

  ASBR1: IPv6 FIB for SRv6

      [BGP CT] ASBR1-SRv6-PE2-gold-Replace,
        NH: ASBR2_InterAS_Link
        SID op: ReplaceSID(ASBR2-SRv6-PE2-gold-Replace)

      [BGP CT] ASBR1-SRv6-PE2-bronze-Replace,
        NH: ASBR2_InterAS_Link
        SID op: ReplaceSID(ASBR2-SRv6-PE2-bronze-Replace)
```

Figure 20: BGP CT processing at ASBR1

When PE1 receives this IBGP CT advertisement from ASBR1, it resolves the nexthop ASBR1-LPBK in the gold TRDB based on its transport-target:0:100. This would resolve over the ISIS-SRv6 route in gold TRDB and pick "Gold-SRv6-Tunnel-to-ASBR1".

This forms the end-to-end Gold SLA path from PE1 to PE2. The gold BGP CT route for PE2-LPBK is installed in gold TRDB, and can be used for resolving service route nexthops. The Transport layer SIDs are replaced at each border node, which reduces the number of SID decaps required at the egress PE.

a mis en forme : En-tête

a mis en forme : Surlignage

a mis en forme : Pied de page

```
    Gold TRDB routes at PE1

        [BGP CT] PE2-LPBK,
          NH: ASBR1-SRv6-gold
          SID op: EncapSID(ASBR1-SRv6-PE2-gold-Replace)

    Bronze TRDB routes at PE1

        [BGP CT] PE2-LPBK,
          NH: ASBR1-SRv6-bronze
          SID op: EncapSID(ASBR1-SRv6-PE2-bronze-Replace)

    PE1: IPv6 FIB for SRv6

        [BGP CT] PE2-LPBK,
          NH: ASBR1-SRv6-gold
          SID op: EncapSID(ASBR1-SRv6-PE2-gold-Replace)

        [BGP CT] PE2-LPBK,
          NH: ASBR1-SRv6-bronze
          SID op: EncapSID(ASBR1-SRv6-PE2-bronze-Replace)

        [ISIS SRv6] ASBR1-SRv6-gold,
          NH: Encap "Gold-SRv6-Tunnel-to-ASBR1"

        [ISIS SRv6] ASBR1-SRv6-bronze,
          NH: Encap "Bronze-SRv6-Tunnel-to-ASBR1"

                   Figure 21: BGP CT processing at PE1
```

Furthermore, any service routes received with nexthop as PE2-LPBK and Mapping Community as Color:0:100 indicating Gold SLA will use the Resolution Scheme associated with its Mapping Community to resolve over the PE2-LPBK CT route installed in the gold TRDB, and push the SRv6-gold SID stack to reach PE2.

Similarly, any service routes received with nexthop as PE2-LPBK and Mapping Community as Color:0:200 indicating Bronze SLA will use the Resolution Scheme associated with its Mapping Community to resolve over the PE2-LPBK CT route installed in the bronze TRDB, and push the SRv6-bronze SID stack to reach PE2. This is shown below.

```
BGP Service routes advertisement from PE2 to PE1:

     SVC_PFX1,
       color:0:100,
       Prefix-SID: PE2-SRv6-S1-DT4,
       NH: PE2-LPBK

     SVC_PFX2,
       color:0:200,
       Prefix-SID: PE2-SRv6-S1-DT4,
       NH: PE2-LPBK

PE1: Service routes FIB

     [BGP INET] SVC_PFX1, color:0:100
```

NH: EncapSID "PE2-SRv6-S1-DT4, ASBR1-SRv6-gold-Repace, Gold-SRv6-Tunnel-to-ASBR1(outer)"

[BGP INET] SVC_PFX2, color:0:200
NH: EncapSID "PE2-SRv6-S1-DT4, ASBR1-SRv6-bronze-Replace, Bronze-SRv6-Tunnel-to-ASBR1(outer)"

Figure 22: Service layer processing

The operational, scaling and convergence aspects of this approach are similar to the aspects of applying BGP CT procedures to the MPLS data plane.

D.2. ~~Color~~ Color-encoded Service SID (CPR) ~~approach~~Approach

CPR is defined in the document: Colorful Prefix Routing for SRv6 based services [Colorful-Prefix-Routing-SRv6], and uses IPv6 Unicast (AFI/SAFI = 2/1) as a transport family. CPR mechanism does not use BGP CT (SAFI 76) address family.

CPR uses color encoded SRv6 service SIDs to determine the intent-aware transport paths for the service, without a separate transport SRv6 SID. It routes using "Colorful Prefix" locators in the transport layer, which are carried in the IPv6 Unicast BGP family.

A Nexthop Resolution Scheme similar to that of BGP CT Section 6 is used on IPv6 Unicast family to resolve “Colorful Prefix” locator routes that carry a mapping community to intent-aware paths in each domain.

By virtue of the CPR SID allocation scheme, the service SIDs inherit the Intent of the corresponding Colorful Prefix route just by performing longest prefix match in forwarding plane.

D.2.1. Analysis of CPR ~~approach~~Approach

The CPR approach can be used to support intent driven routing while minimizing SRv6 encapsulation overhead, at the cost of careful SID numbering and planning. The state in the transport network is a function of total number of Colorful Prefixes.

In the CPR approach, typically one service SID is allocated for each service function (e.g. VRF) which is associated with a specific intent. In some special scenarios, for example, when different service routes in the same VRF are with different intents, a unique service SID would need to be allocated for each intent associated with the VRF.

However, the CPR mechanism preserves BGP PIC (Prefix scale Independent Convergence) for the egress SN failure scenario where only Colorful Prefix routes need to be withdrawn.

CPR achieves strict Intent based forwarding for the service routes. Fallback to best effort transport class is achieved by numbering all SRv6 Colorful Prefix locators at the egress SN to fall in the same subnet as the SRv6 locator that uses best effort transport class. Customized intent fallback between different color transport classes may be achieved by allocating a CPR prefix for each such intent

a mis en forme : En-tête

a mis en forme : Pied de page

a mis en forme : En-tête

fallback policy, and advertising that CPR prefix with an appropriate mapping community, that maps to a customized resolution scheme. Alternatively, the intent fallback policy may be provisioned on the ingress nodes directly.

Further, IPv6 Unicast family is widely deployed to carry Internet Service routes. Repurposing IPv6 Unicast family to carry Transport routes also may impact the operational complexity and security aspects in the network.

Contributors

Co-Authors

Israel Means
AT&T
2212 Avenida Mara,
Chula Vista, California 91914
United States of America
Email: israel.means@att.com
Csaba Mate
KIFU, Hungarian NREN
Budapest
35 Vaci street,
1134
Hungary
Email: ietf@nop.hu

Deepak J Gowda
Extreme Networks
55 Commerce Valley Drive West, Suite 300,
Thornhill, Toronto, Ontario L3T 7V9
Canada
Email: dgowda@extremenetworks.com

Other Contributors

Balaji Rajagopalan
Juniper Networks, Inc.
Electra, Exora Business Park~Marathahalli - Sarjapur Outer Ring Road,
Bangalore 560103
KA
India
Email: balajir@juniper.net

Reshma Das
Juniper Networks, Inc.
1133 Innovation Way,
Sunnyvale, CA 94089
United States of America
Email: dreshma@juniper.net

a mis en forme : Anglais (États-Unis)

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

a mis en forme : Anglais (États-Unis)

a mis en forme : Pied de page

Email: mrajesh@juniper.net
Chaitanya Yadlapalli
AT&T
200 S Laurel Ave,
Middletown,, NJ 07748
United States of America
Email: cy098d@att.com

Gyan Mishra
Verizon Inc.
13101 Columbia Pike
Silver Spring, MD 20904
United States of America
Email: gyan.s.mishra@verizon.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
Capitalonline~~.~~
Beijing
China
Email: xiaohu.xu@capitalonline.net

Acknowledgements

The authors thank Jeff Haas, John Scudder, Susan Hares, Dongjie (Jimmy), Moses Nagarajah, Jeffrey (Zhaohui) Zhang, Joel Harpern, Jingrong Xie, 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 Trip, 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 a suggestion made by Richard Roberts and Krzysztof Szarkowicz.

Authors' Addresses

Kaliraj Vairavakkalai (editor)
Juniper Networks, Inc.
1133 Innovation Way,
Sunnyvale, CA 94089

a mis en forme : En-tête

a mis en forme : Anglais (États-Unis)

a mis en forme : Anglais (États-Unis)

a mis en forme : Anglais (États-Unis)

a mis en forme : Pied de page

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

a mis en forme : En-tête

a mis en forme : Anglais (États-Unis)

a mis en forme : Pied de page