BESS WorkGroup
Internet-Draft
Intended status: Standards Track
Expires: August 26, 2021

D. Rao
S. Agrawal
C. Filsfils
K. Talaulikar
Cisco Systems

February 22, 2021

BGP Color-Aware Routing(CAR)
draft-dskc-bess-bgp-car-00

Abstract

This document describes a BGP based routing solution to establish end-to-end intent-aware paths across a multi-domain service provider transport network. This solution is called BGP Color-Aware Routing (BGP CAR).

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 August 26, 2021.

Copyright Notice

Table of Contents

1. Introduction

1.1. Objectives

- Address the Transport problem statement and requirements described in [dskc-bess-bgp-car-problem-statement]

- Define an inter-domain BGP-based Color-Aware Routing proposal to steer traffic for a C-colored service route V/v from a PE onto a BGP color-aware path to (PE, C)

  * Provide an alternative to the SR-PCE based design [I-D.ietf-spring-segment-routing-policy]

1.2. Requirements Language

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

2. Concepts

A refresher on core concepts used in this document, some of which are described in [BGP-CAR-Problem-Statement]

2.1. Color

The solution must reuse the color concept defined in [I-D.ietf-spring-segment-routing-policy]. The color is a 32-bit numerical value that, today, associates an SR-policy with an intent (e.g., low latency).

2.2. Colored vs Color-Aware

- Colored: Egress PE PE2 colored its BGP VPN route V/v to indicate the intent that it requests for the traffic bound to V/v.
- Color-Aware: a new BGP solution which signals multiple "ways" to reach a given destination (e.g. PE2)
- Steering a colored VPN route to a color-aware route
  - If PE2 signals a VPN route V/v with color C
  - If PE1 installs that VPN route
  - If PE1 learns about a BGP Color-Aware Route R/r to PE2 for color C
  - Then PE1 steers packets destined to V/v via R/r

2.3. Color Domains

A domain (or network domain) generally refers to a unit of isolation or hierarchy in the network topology; for example, access, metro and or core domains. From a routing perspective, a domain may have a distinct IGP area or instance; or a distinct BGP ASN.

With the use of a 'Color' to represent intent, it is useful to describe the distinct concept of a color domain.

A color domain refers to a collection of one or more network domains with a single, consistent color-to-intent mapping.

When a route gets distributed into a domain with a different color-to-intent mapping scheme, the color associated with the route needs to be mapped to the locally assigned value in that domain.

Deployments under a single authority are expected to use the same color-to-intent mapping across all network domains.

A solution must distinguish the actual protocol boundaries (IGP, ASN) from the color domain boundaries.

2.4. BGP Color-Aware Routing

In the remainder of this document, the BGP Color-Aware Routing Solution is referred to as BGP CAR.

3. BGP extensions for CAR

This section analyzes the requirements for BGP CAR and proposes extensions, specifically for Transport Color-Aware-Routing

3.1. Why a new SAFI is required

- Existing BGP SAFI for BGP-LU (AFI 1 or 2 and SAFI 4) signals transport destination (likely PE loopback) with just an IP prefix in NLRI.
  - BGP CAR needs to signal multiple "ways" to reach a transport destination, each for a different intent or color; i.e., it needs a Color-Aware NLRI
- Hence, a new SAFI is needed for BGP Transport CAR which can encode IP prefix and Color

3.2. Data Model of New SAFI

The essential elements of the data model for the transport CAR SAFI are as follows:

- NLRI Key: E, C
  - E: IPv4/IPv6 prefix: Prefix is unique in inter-domain network.
  - Color: Distinguishes per-intent instances of a prefix. Additionally, it signals the intent provided by with the route in originator color domain. 32-bit value as per [I-D.ietf-spring-segment-routing-policy]
- NLRI non key data
  - To encode multiple encapsulations with efficient packing
    - MPLS label stack
    - Label Index (hint for label allocation from SRGB - same as BGP SR Prefix SID Attr Label Index TLV)
    - SRv6 SID(s)

+ Etc.

- Next-Hop
  - BGP Next-Hop
- AIGP Metric
  - To accumulate color/intent specific metric across domains
  - AIGP Attribute provides extensibility via TLVs, enabling definition of additional metric semantics for a color as needed for an intent
- Local-Color-Mapping Extended-Community (LCM-EC)
  - 32-bit Color value
  - Optional, used when a CAR route propagates across domains with different or inconsistent color-to-intent mapping schemes

The detailed protocol operations for these elements are described in later sections.

3.3. Extensible, future-proof encoding

Since a new SAFI is required, it is prudent to define an extensible encoding so that additional use-cases can be supported in future, without imposing limitations

Key design aspects for an extensible encoding:

Encode a NLRI (Route) Type field. This provides extensibility to add new NLRI formats for new route-types

Encode a key length field. This enables handling unsupported route-types opaquely, enabling transitivity via RRs

Define non-key NLRI data using TLVs. This enables flexible and efficient encoding of data such as multiple encapsulations

3.4. BGP CAR Family

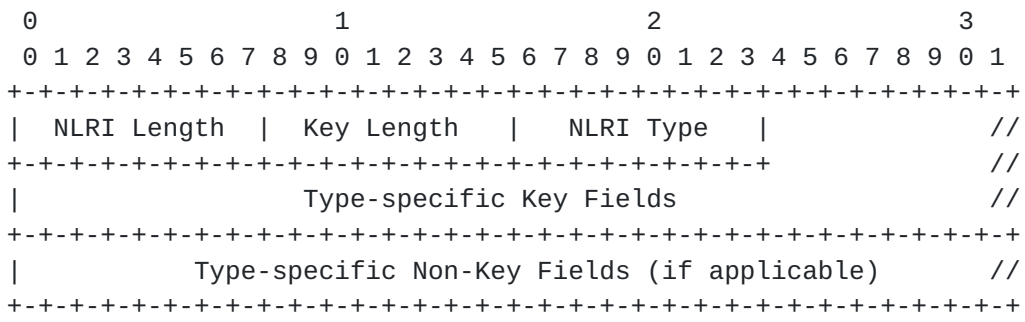
BGP CAR leverages the BGP multi-protocol extensions [RFC4760] and uses the MP_REACH_NLRI and MP_UNREACH_NLRI attributes for routes updates by using the SAFI value TBD1 along with AFI 1 for IPv4 prefixes and AFI 2 for IPv6 prefixes.

BGP speakers MUST use BGP Capabilities Advertisement to ensure support for processing of BGP CAR updates. This is done as specified in [RFC4760], by using capability code 1 (multi-protocol BGP), with AFI 1 and 2 (as required) and SAFI TBD1.

The sub-sections below specify the generic encoding of the BGP CAR NLRI followed by the encoding for specific NLRI types introduced in this document.

3.4.1. BGP CAR SAFI NLRI Format

The generic format for the BGP CAR Address-Family NLRI is shown below:



where:

- NLRI Length: 1 octet field that indicates the length in octets of the NLRI excluding the NLRI Length field itself.

- Key Length: 1 octet field that indicates the length in octets of the NLRI type-specific key fields. Key length MUST be at least 2 less than the NLRI length.

- NLRI Type: 1 octet field that indicates the type of the BGP Transport CAR NLRI.

- Type-Specific Key Fields: Depend on the NLRI type and of length indicated by the Key Length.

- Type-Specific Non-Key Fields: optional and variable depending on the NLRI type. The NLRI encoding allows for encoding of specific non-key information associated with the route (i.e. the key) as part of the NLRI for efficient packing of BGP updates.

The indication of the key length enables BGP Speakers to determine the key portion of the NLRI and use it along with the NLRI Type field in an opaque manner for handling of unknown or unsupported NLRI

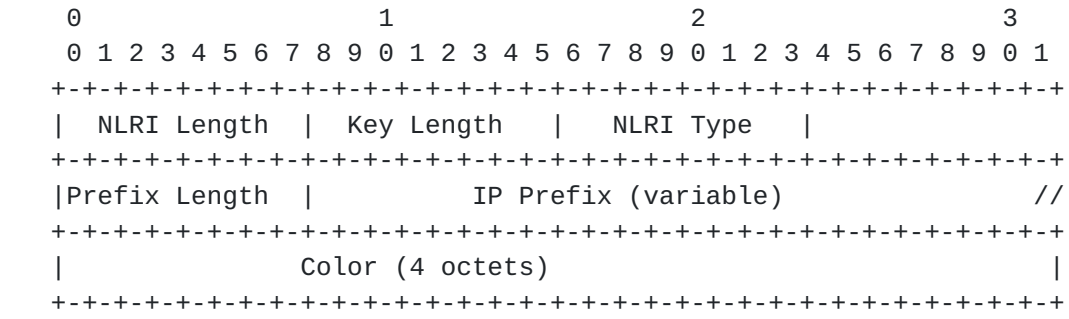
types. This can help Route Reflectors (RR) to propagate NLRI types introduced in the future in a transparent manner.

The NLRI encoding allows for encoding of specific non-key information associated with the route (i.e. the key) as part of the NLRI for efficient packing of BGP updates.

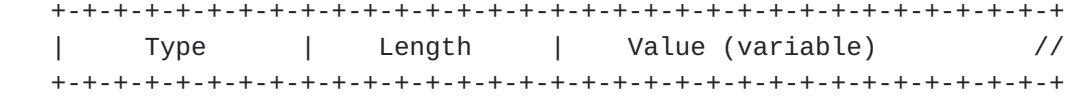
The non-key portion of the NLRI MUST be omitted while carrying it within the MP_UNREACH_NLRI when withdrawing the route advertisement.

3.4.2. CAR NLRI Type

The Color-Aware Routes NLRI Type is used for advertisement of color-aware routes and has the following format:



Followed by optional TLVs encoded as below:



where:

- NLRI Length: variable
- Key Length: variable
- NLRI Type: 1
- Type-Specific Key Fields: as below
  - Prefix Length: 1 octet field that carries the length of prefix in bits. Length MUST be less than or equal to 32 for IPv4 (AFI=1) and less than or equal to 128 for IPv6 (AFI=2).
  - IP Prefix: IPv4 or IPv6 prefix (based on the AFI). A variable size field that contains the most significant octets of the prefix, i.e., 1 octet for prefix length 1 to 8, 2 octets for

prefix length 9 to 16, 3 octets for prefix length 17 up to 24, 4 octets for prefix length 25 up to 32, and so on. The size of the field MUST be less than or equal to 4 for IPv4 (AFI=1) and less than or equal to 16 for IPv6 (AFI=2).

* Color: 4 octets that contains color value associated with the prefix. It distinguish different instances of a prefix. Additionally, it signals the intent associated with the route in originator color domain.

o Type-Specific Non-Key Fields: specified in the form of optional TLVs as below:

* Type: 1 octet field that contains the type of the non-key TLV

* Length: 1 octet field that contains the length of the value portion of the non-key TLV in terms of octets

* Value: variable length field as indicated by the length field and to be interpreted as per the type field.

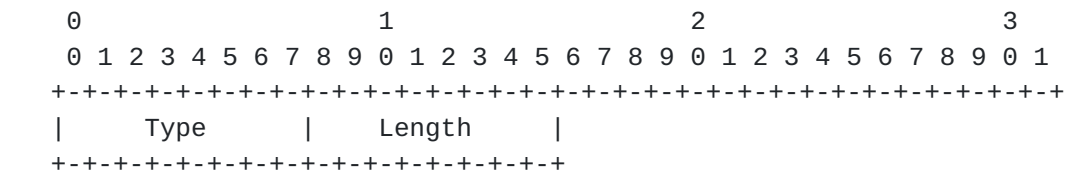
The prefix is routable across the administrative domain where BGP Transport CAR is deployed. It is possible that the same prefix is originated by multiple BGP Transport CAR speakers in the case of anycast addressing or multi-homing.

The Color is introduced to enable multiple route advertisements for the same prefix. The color is associated with an intent (e.g. low-latency) in originator color-domain.

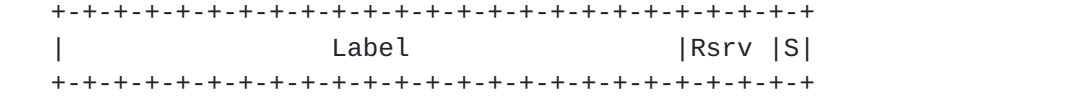
The following sub-sections specify the non-key TLVs associated with the Color-Aware Routes NLRI type.

3.4.2.1. Label TLV

The Label TLV is used for advertisement of color-aware routes along with their MPLS labels and has the following format:



Followed by one (or more) Labels encoded as below:



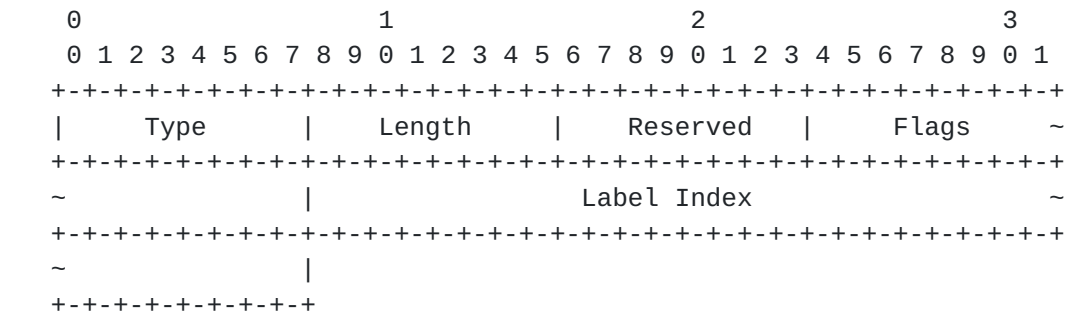
where:

- Type : 1
- Length: variable, MUST be a multiple of 3
- Label Information: multiples of 3 octet fields to convey the MPLS label(s) associated with the advertised color-aware route. It is used for encoding a single label or a stack of labels as per procedures specified in [RFC8277].

When a BGP Transport CAR speaker is propagating the route further after setting itself as the nexthop, it allocates a local label for the specific prefix and color combination which it updates in this TLV. It also MUST program a label cross-connect that would result in the label swap operation for the incoming label that it advertises with the label received from its best-path router(s).

3.4.2.2. Label Index TLV

The Label Index TLV is used for advertisement of Segment Routing MPLS (SR-MPLS) Segment Identifier (SID) [RFC8402] information associated with the labelled color-aware routes and has the following format:



where:

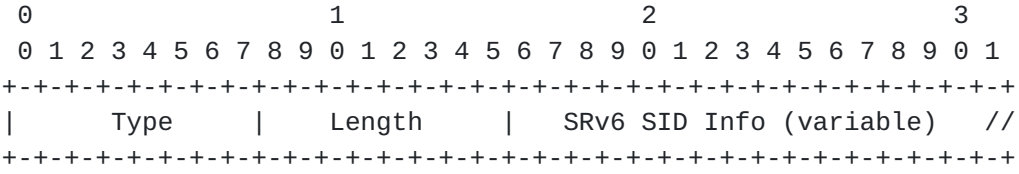
- Type : 2
- Length: 7
- Reserved: 1 octet field that MUST be set to 0 and ignored on receipt.
- Flags: 2 octet field that maps to the Flags field of the Label-Index TLV of the BGP Prefix SID Attribute [RFC8277].
- Label Index: 4 octet field that maps to the Label Index field of the Label-Index TLV of the BGP Prefix SID Attribute [RFC8277].

This TLV provides the equivalent functionality as Label-Index TLV of [RFC8669] for Transport CAR in SR-MPLS deployments. The BGP Prefix SID Attribute SHOULD be omitted from the labeled color-aware routes when the attribute is being used to only convey the Label Index TLV for better BGP packing efficiency.

When a BGP Transport CAR speaker is propagating the route further after setting itself as the nexthop, it allocates a local label for the specific prefix and color combination. When the received update has the Label Index TLV, it SHOULD use that hint to allocate the local label from the SR Global Block (SRGB) using procedures as specified in [RFC8669].

3.4.2.3. SRv6 SID TLV

BGP Transport CAR can be also used to setup end-to-end color-aware connectivity using Segment Routing over IPv6 (SRv6) [RFC8402]. [I-D.ietf-spring-srv6-network-programming] specifies the SRv6 Endpoint behaviors (e.g. End PSP) which MAY be leveraged for BGP CAR with SRv6.The SRv6 SID TLV is used for advertisement of color-aware routes along with their SRv6 SIDs and has the following format:



where:

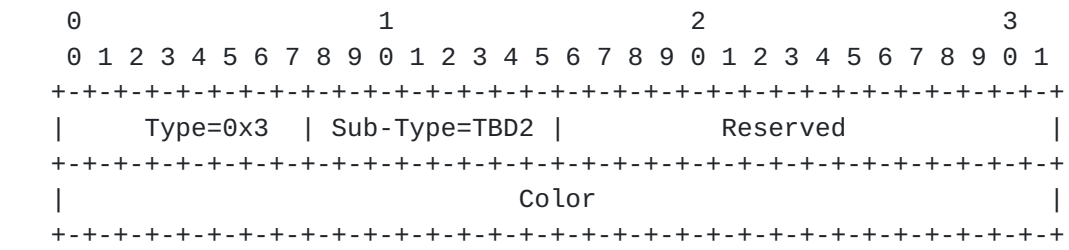
- Type : 3
- Length: variable, MUST be either less than or equal to 16, or be a multiple of 16

- SRv6 SID Information: field of size as indicated by the length that either carries the SRv6 SID(s) for the advertised color-aware route as one of the following:
  - A single 128-bit SRv6 SID or a stack of 128-bit SRv6 SIDs
  - A transposed portion (refer [I-D.ietf-bess-srv6-services]) of the SRv6 SID that MUST be of size in multiples of one octet and less than 16.

The BGP color-aware route update for SRv6 MUST include the BGP Prefix-SID attribute along with the TLV carrying the SRv6 SID information as specified in [I-D.ietf-bess-srv6-services] when using the transposition scheme of encoding for packing efficiency of BGP updates.

3.4.3. Local-Color-Mapping (LCM) Extended Community

This document defines a new BGP Extended Community called "LCM". The LCM is a Transitive Opaque Extended Community with the following encoding:



where:

- Type: 0x3
- Sub-Type: TBD2.
- Reserved: 2 octet of reserved field that MUST be set to zero on transmission and ignored on reception.
- Color: 4-octet field that carries the 32-bit color value.

When CAR route crosses the original color domain boundary, LCM EC is added. LCM EC associate the local color mapping for the intent (e.g. low latency) in transit or remote color domain. Note: reminder "BGP CAR needs to signal multiple "ways" to reach a transport destination, each for a different intent or color". Original BGP CAR route (E, C) still signal multiple "ways" to reach E, but once LCM EC is added, intent is carried in it and not by C in NLRI.

The LCM EC MAY be used for filtering of BGP CAR routes and/or for applying routing policies for the intent.

3.5. BGP transport CAR Route Origination

- BGP CAR routes may be originated from a node via local injection (e.g., loopback)
  - Routes will be advertised with Implicit-NULL (or equivalent), and optionally may include Label-Index
- BGP Transport CAR routes may also be originated from a node, sourced from another mechanism
  - IGP Flexible Algorithm(FA) [I-D.ietf-lsr-flex-algo] redistribution
    - FA identifier mapping to BGP transport CAR color or vice versa by local policy. This will allow redistribution of prefixes, prefix SID between FA and BGP CAR
  - SR Policy [I-D.ietf-spring-segment-routing-policy]
    - An SR Policy is identified through the tuple (color, E) where color is a 32-bit numerical value that associates the SR Policy with an intent (e.g. low-latency). When color of SR policy maps directly into BGP CAR color because of same intent or through some local configuration, endpoint of policy can be advertised in BGP Transport CAR to rest of network for end to end color-aware transport connectivity.
  - BGP-LU [RFC8277]
    - Redistribution between BGP-LU and BGP CAR color table and vice versa. Most likely(but not limited) color represents best effort intent in BGP CAR domain. This provide connectivity between BGP-LU only domain and BGP CAR domain with best effort color-awareness.

3.6. BGP CAR Next-Hop Processing

3.6.1. Validation

- Validation of BGP Next-Hop: Reachability verified via underlying routing control plane. Local policy should be provided to verify it
  - Strictly within intent of BGP CAR route i.e "color"

    * Default routing table

    * Skip it when updates are propagated out of band

- Validation of Encapsulation: Validate data-plane availability of encapsulation before using and propagating further.

- Validation of the intent: Validate the intent provided by the underlying transport (e.g., via OAM), where applicable.

3.6.2. Resolution

BGP color-aware routes may be resolved over various intra-domain and inter-domain mechanisms that provide connectivity to the BGP next-Hop with the desired intent

- Leverage the notion of "color" in NLRI or LCM-EC to determine the matching intent-aware mechanism and instance.

- Leverage ODN/AS mechanisms where needed, for instance to use SR-PCE for an SR-policy to the BGP next-hop

- Flexible for all encapsulations

    * (SR-)MPLS

    * SRv6, IPv4/IPv6, etc.

- Flexible over various underlay mechanisms

    * SR Policy: Color from BGP CAR route and policy endpoint from BGP CAR Next hop

    * IGP Flexible Algorithm: Color from BGP CAR mapped to Flex Algo by configuration.

    * IGP/BGP best effort (SR, LDP, RSVP-TE, BGP-LU etc.)

    * BGP CAR in hierarchical CAR design

- Support selection preference among available mechanisms

- Fallback to a different color or best effort path

3.7. AIGP Metric Computation

- BGP CAR nodes update the Accumulated IGP (AIGP) Attribute as the BGP CAR route propagates across the network.
- The value set (or appropriately incremented) in the AIGP TLV corresponds to the metric associated with the underlying intent of the color. Example. when the color is associated with a low-latency path, the metric value is set based on the delay metric.
  - Information regarding the metric type used by the underlying intra-domain mechanism can also be set
- If BGP CAR routes traverse across a discontinuity in the transport path for a given intent, add penalty in accumulated IGP
- If BGP CAR routes traverse across a discontinuity in the transport path for a given intent, the AIGP TLV is used to indicate this e.g. with a discontinuity bit.
- AIGP metric computation is recursive.
- To avoid continuous IGP metric churn causing end to end BGP CAR churn, implementation should provide thresholds to trigger AIGP update.
- Additional AIGP extensions may be defined to signal state for specific use-cases.
  - MSD along the BGP CAR advertisement.
  - Minimum MTU along the BGP CAR advertisement.

3.8. Multiple color domains

- When BGP CAR routes get distributed to a domain with a different color-to-intent mapping, the color signaled must be re-mapped to the local color being used within the receiving domain
- A key requirement to consider is the separation and independence of the administrative authority in different color domains.
  - Each color domain needs to use its own local color. The route can traverse multiple such color domains where the color mappings change
- This requirement is addressed by the following steps :

    * The NLRI of the CAR route is never changed

        + E is globally unique. Hence even if C is local-domain significant, E-C in that order is globally unique

    * Each color domain needs to use its own local color. The route can traverse multiple such color domains where the color mappings change

        + To address this requirement, a border node in a color domain encodes its local color mapping in a Local-Color-Mapping Extended-Community when sending the route to a peer in a different color domain

        + The border routers within the receiving domain map the received LCM-EC Color value to a local color assigned for that intent and rewrite the LCM-EC

        + The nodes within the receiving domain use the local color encoded in the LCM-EC for next-hop resolution and BGP CAR route installation

- The LCM-EC is only used when a CAR route needs to be distributed across a color domain boundary. The likely case (color consistency) is supported with the simplest and most efficient scheme (E, C) key and no LCM-EC.

- Example: When going from a domain D1 to a domain D2 where D1 uses the color scheme is the NLRI but D2 uses another color scheme, then on the peering session from D1 to D2, D1 on egress or D2 on ingress inserts the LCM-EC which carries the mapped local color that will be used in D2. When the route travels from D2 to a domain D3 which uses the color scheme in the NLRI then either the LCM-EC is kept but its internal C is remapped to the color scheme of D3 or the LCM-EC is removed

- Color intent encoded in the service routes in the Color Ext-community should also be re-mapped consistently

- A color boundary is typically well-defined, at a BGP peering session on a border Router, and at a service/transport RR.

- A color domain may extend across one or more BGP ASNs

4. Steering a Colored Service Route onto an (E, C) BGP CAR route

BGP colored service routes (i.e., containing Color extended community [I-D.ietf-idr-tunnel-encaps]) resolve over BGP transport CAR routes i.e. (E, C), conceptually identical to the steering mechanism used with SR Policies.

All steering options are supported: Automated, on-demand steering, per-destination, per-flow, CO-only

Co-existence with SR-policy based steering is also supported

By default, when BGP CAR is enabled, a BGP CAR route will be preferred.

Similarly, if an IGP Flex-Algo route exists, typically for an intra-domain endpoint, it is preferred over a BGP CAR route to the same endpoint.

A node may support a local policy to set the preferences between different mechanisms.

The following sub-sections illustrate example scenarios of Colored Service Route Steering over E2E BGP CAR resolving over different intra-domain mechanisms

4.1. E2E BGP transport CAR intent realized using IGP FA

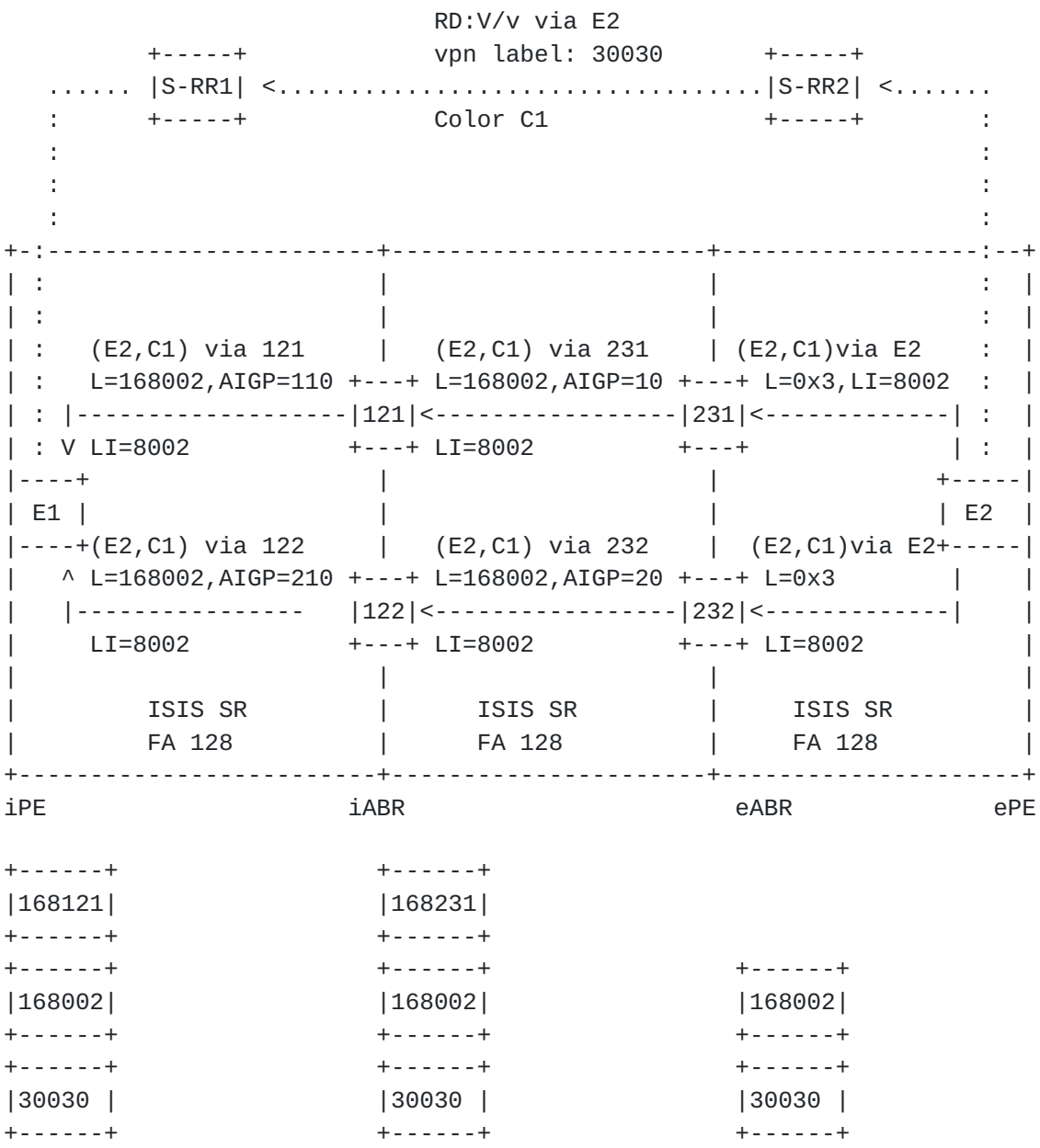


Figure 1: BGP FA Aware transport CAR path

Use case: Provide end to end intent for service flows.

o With reference to the topology above:

  * IGP FA 128 is running in each domain.

  * Egress PE E2 advertises a VPN route RD:V/v colored with (color extended community) C1 to steer traffic to BGP transport CAR (E2, C1). VPN route propagates via service RRs to ingress PE E1.

* BGP CAR route (E2, C1) with next-hop, label-index and label as shown above are advertised through border routers in each domain.

* Local policy on each hop maps intent C1 to resolve CAR route next-hop over IGP FA 128 of the domain. AIGP attribute influences BGP CAR route best path decision as per [RFC7311]. BGP CAR label swap entry is installed that goes over FA 128 LSP to next-hop providing intent in each IGP domain. Update AIGP metric to reflect FA 128 metric to next-hop.

* Ingress PE E1 learns CAR route (E2, C1). It steers colored VPN route RD:V/v into (E2, C1)

o Important:

* IGP FA 128 top label provides intent in each domain.

* BGP CAR label (e.g. 168002) carries end to end intent. Thus stitches intent over intra domain FA 128.

4.2. E2E BGP transport CAR intent realized using SR Policy

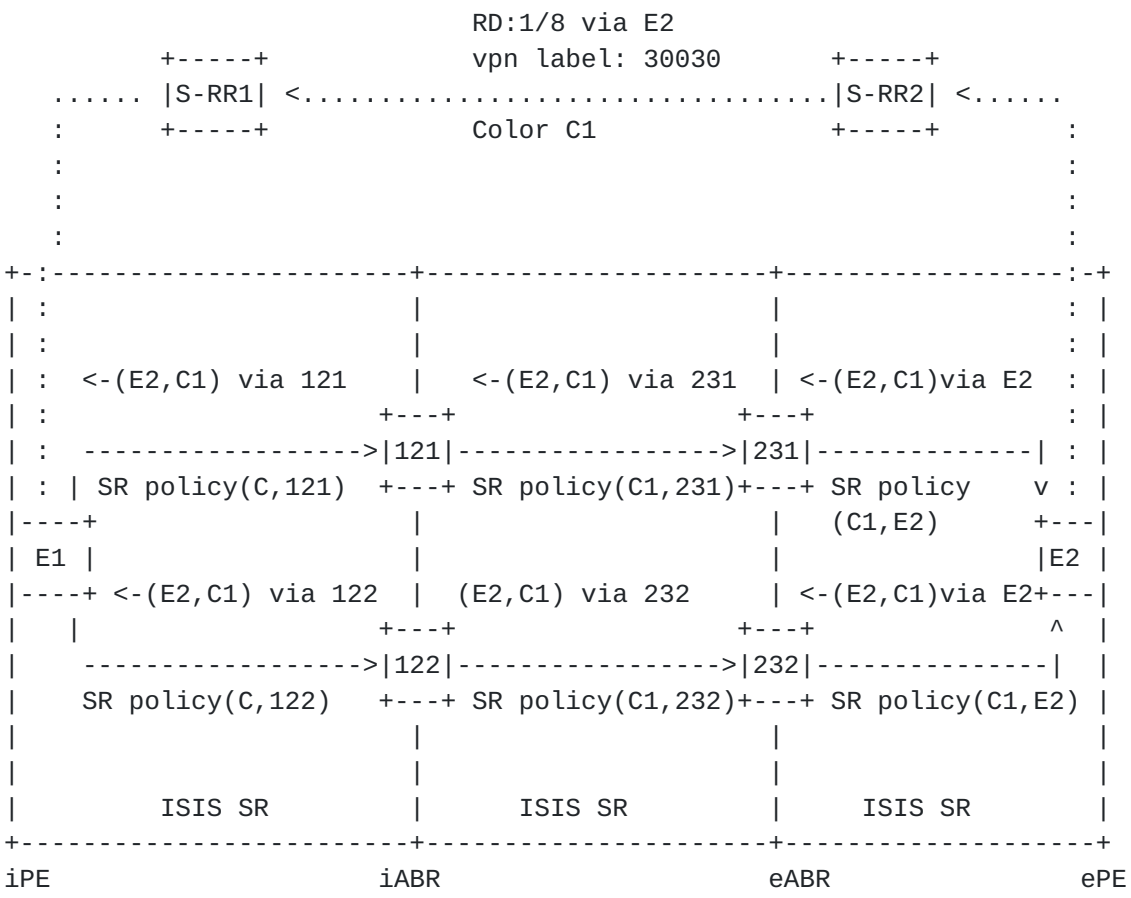


Figure 2: BGP SR policy Aware transport CAR path

Use case: Provide end to end intent for service flows

- With reference to the topology above:
  - SR Policy provide intra domain intent.
  - Egress PE E2 advertises a VPN route RD:V/v colored with (color extended community) C1 to steer traffic to BGP transport CAR (E2, C1). VPN route propagates via service RRs to ingress PE E1.
  - BGP CAR route (E2, C1) with next-hop, label-index and label as shown above are advertised through border routers in each domain.
  - Local policy on each hop maps intent C1 to resolve CAR route next-hop over an SR policy(C1, next-hop). BGP CAR label swap entry is installed that goes over SR policy segment list.

    * Ingress PE E1 learns CAR route (E2, C1). It steers colored VPN route RD:V/v into (E2, C1).

- Important:

    * SR policy provides intent in each domain.

    * BGP CAR label (e.g. 168002) carries end to end intent. Thus stitches intent over intra domain SR policies.

4.3. BGP transport CAR intent realized in a section of the network

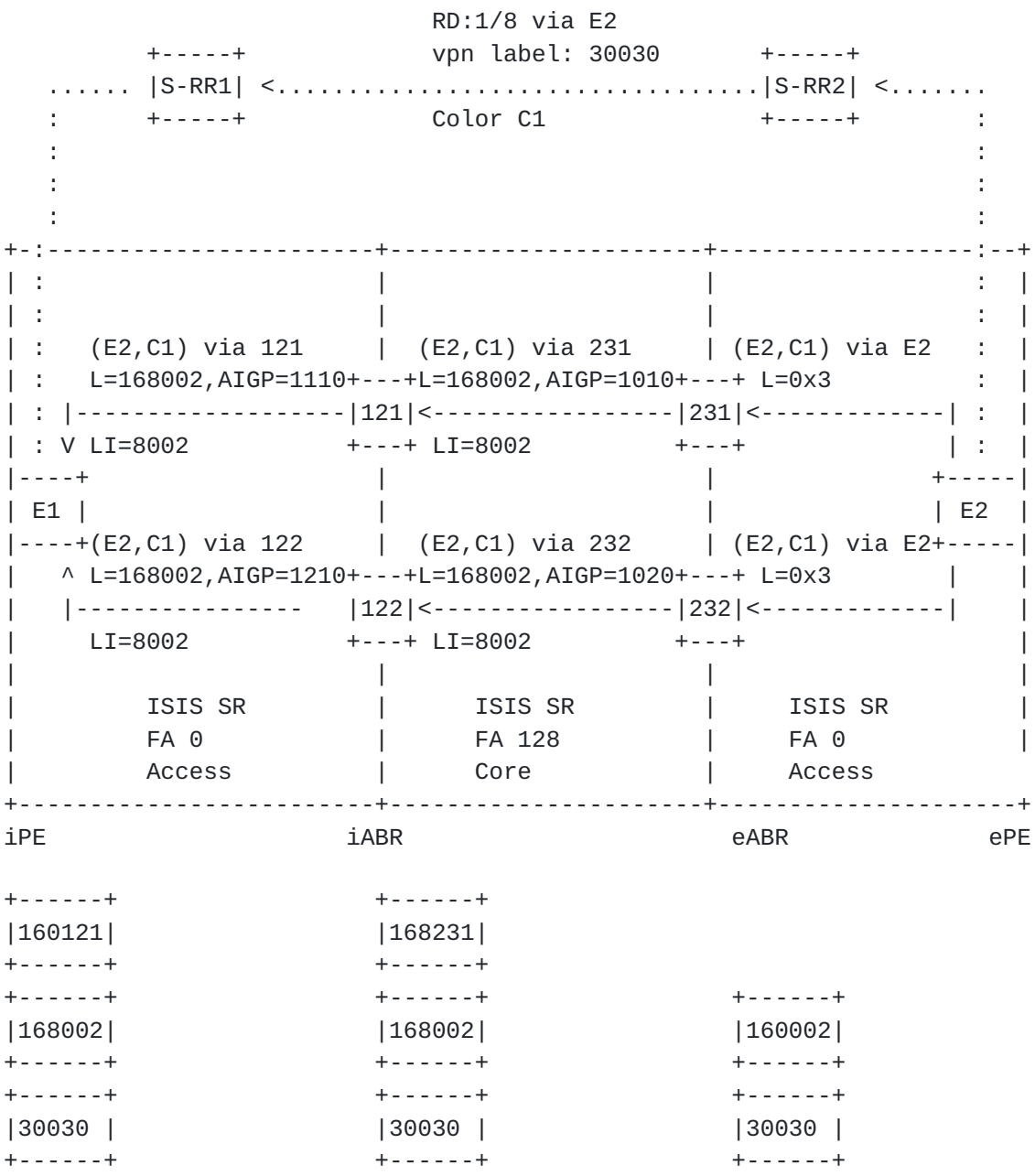


Figure 3: BGP Hybrid FA Aware transport CAR path

Use case: Provide intent for service flows only in Core domain.

o With reference to the topology above:

  * IGP FA 128 is only enabled in Core (e.g. WAN network). Access only has base algo 0.

  * Egress PE E2 advertises a VPN route RD:V/v colored with (color extended community) C1 to steer traffic to BGP transport CAR

(E2, C1). VPN route propagates via service RRs to ingress PE E1.

* BGP CAR route (E2, C1) with next-hop, label-index and label as shown above are advertised through border routers in each domain.

* Local policy on 231 and 232 maps intent C1 to resolve CAR route next-hop over IGP base algo 0 in right access domain. BGP CAR label swap entry is installed that goes over algo 0 LSP to next-hop. Update AIGP metric to reflect algo 0 metric to next-hop most likely with additional penalty.

* Local policy on 121 and 122 maps intent C1 to resolve CAR route next-hop learnt from Core domain over IGP FA 128. BGP CAR label swap entry is installed that goes over FA 128 LSP to next-hop providing intent in Core IGP domain.

* Ingress PE E1 learns CAR route (E2, C1). It maps intent C1 to resolve CAR route next-hop over IGP base algo 0. It steers colored VPN route RD:V/v into (E2, C1)

o Important:

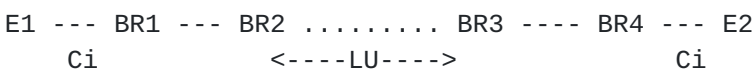
* IGP FA 128 top label provides intent in Core domain.

* BGP CAR label (e.g. 168002) carries intent from PEs which is realized in core domain

4.4. Transit network domains that do not support CAR

o In a brownfield deployment, color-aware paths between two PEs may need to go through a transit domain that does not support CAR. Example include an MPLS LDP network with IGP best-effort; or a BGP-LU based multi-domain network. MPLS LDP network with best effort IGP can adopt above scheme. Below is the example for BGP LU.

o Reference topology:



* Network between BR2 and BR3 comprises of multiple BGP-LU hops (over IGP-LDP domains).

* E1, BR1, BR4 and E2 are enabled for BGP CAR, with Ci colors

  * BR1 and BR2 are directly connected; BR3 and BR4 are directly connected

- BR1 and BR4 form an over-the-top peering (via RRs as needed) to exchange BGP CAR routes
- BR1 and BR4 also form direct BGP-LU sessions to BR2 and BR3 respectively, to establish labeled paths between each other through the BGP-LU network
- BR1 recursively resolves the BGP CAR next-hop for CAR routes learnt from BR4 via the BGP-LU path to BR4
- BR1 signals the transport discontinuity to E1 via the AIGP TLV, so that E1 can prefer other paths if available
- BR4 does the same in the reverse direction
- Thus, the color-awareness of the routes and hence the paths in the data plane are maintained between E1 and E2, even if the intent is not available within the BGP-LU island
- A similar design can be used for going over network islands of other types

5. Color Mapping Scenarios

There are a variety of deployment scenarios that arise w.r.t different color mappings in an inter-domain environment. This section attempts to enumerate them to provide clarity into the usage of the color related protocol constructs.

5.1. Single color domain containing network domains with N:N color distribution

All network domains (ingress, egress and all transit domains) are enabled for the same N colors

A color may of course be realized by different technologies in different domains as described above

The N intents are both signaled end-to-end via BGP CAR routes; as well as realized in the data plane

Section 4.1 is an example of this case

5.2. Single color domain containing network domains with N:M color distribution

Certain network domains may not be enabled for some of the colors, but may still be required to provide transit.

When a (E, C) route traverses a domain where color C is not available, the operator may decide to use a different intent of color c that is available in that domain to resolve the next-hop and establish a path through the domain

- The next-hop resolution may occur via paths of any intra-domain protocol or even via paths provided by BGP CAR
- The next-hop resolution color c may be defined as a local policy at ingress or transit nodes of the domain
- It may also be automatically signaled from egress border nodes by attaching a color extended community with value c to the BGP CAR routes

Hence, routes of N colors may be resolved via a smaller set of M colored paths in a transit domain, while preserving the original intent end-to-end.

Any ingress PE that installs a service (VPN) route with a color C, must have C enabled locally to install IP routes to (E, C) and resolve the service route next-hop

A degenerate case of these scenario is where a transit domain does not support any color. Section 4.3 describes an example of this case

5.3. Multiple color domains

When the routes are distributed between domains with different color-to-intent mapping schemes, both N:N and N:M ratios are possible, although an N:M mapping is more likely to occur.

Reference topology:

```
D1 ----- D2 ----- D3
 C1      C2       C3
```

- C1 in D1 maps to C2 in D2 and to C3 in D3
- BGP CAR is enabled in all three domains

The reference topology above is used to elaborate on the design described in Section-X

When the route originates in color domain D1 and gets advertised to a different color domain D2, following procedures apply:

The original intent in BGP CAR route is preserved; i.e. route is (E, C1)

A BR of D1 attaches LCM-EC with value C1 when advertising to a BR in D2

A BR in D2 receiving (E, C1) maps C1 in received LCM-EC to local color, say C2

Within D2, this LCM-EC value of C2 is used instead of the Color in CAR route NLRI (E, C1). This applies to all procedures described in the earlier section for a single color domain, such as next-hop resolution and route installation.

A colored service route V/v originated in domain D1 with next-hop E and color C1 will also have its color extended-community value re-mapped to C2, typically at a service RR

On an ingress PE in D2, V/v will resolve via C2

When a BR in D2 advertises the route to a BR in D3, a similar process is followed

6. Intent Use-cases

This section will describe how BGP CAR addresses the various intent use-cases described in [ref:dskc-bess-bgp-car-problem-statement]. Details will be added in a later revision of the document.

7. Scaling

A key requirement of [ref:dskc-bess-bgp-car-problem-statement] is scale, specifically:

- No intermediate node dataplane should need to scale to (Colors * PEs)
- No node should learn and install a BGP CAR route to (E,C) if it does not install a Colored service route to E
  - An intermediate node may learn a BGP CAR route to (E, C) in control plane if it is an inline RR to an ingress PE

* An intermediate node may learn and install a BGP CAR route to (E, C) if it is set up to be the next-hop for an ingress PE that installs the BGP CAR route

7.1. Data plane does not have to scale to Colors * PEs

Depending on the scale of the network as well as the constraints associated with the nodes at different tiers, an appropriate design should be adopted. Three design variations are illustrated below.

7.1.1. Inter-Domain Hop by hop BGP CAR for PE routes

Reference topology is shown below, with the BGP signaling and the resulting BGP and example IGP label stack at different hops

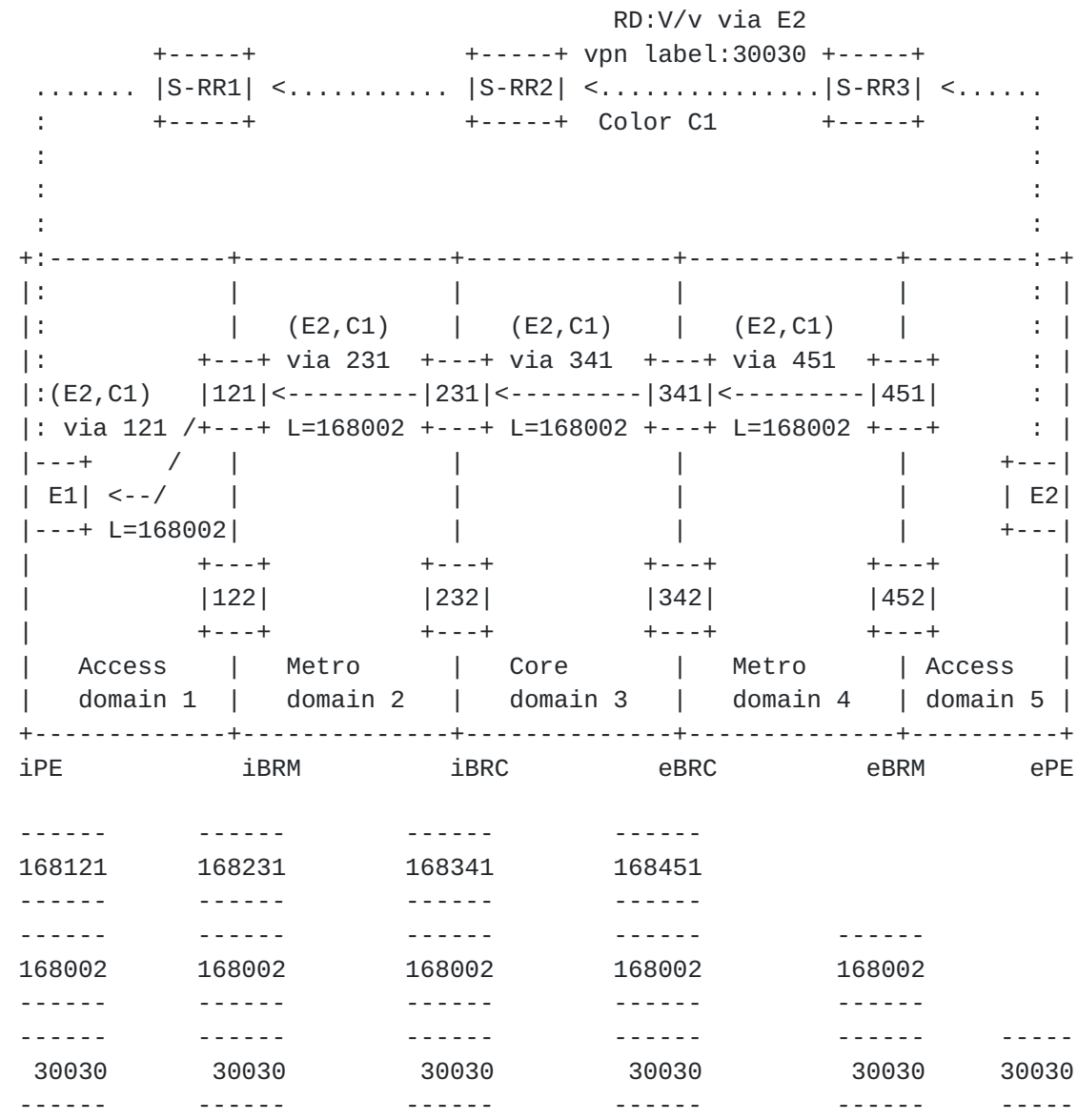


Figure 4: Single BGP transport CAR level

- With reference to the topology above:

  - Consider egress PE E2 advertises a VPN (service) route RD:V/v that propagates via service RRs to ingress PE E1.

  - A BGP CAR route (E2, C1) is advertised by egress BRM node 451. The route may be sourced locally, for instance by redistribution from an IGP-FA, and is distributed hop-by-hop through egress Metro, Core, ingress Metro to Access

  * Node 451, 341, 231 and 121 learns BGP CAR route (E2, C1). Each allocate local label and program swap entry in forwarding and set itself as next-hop.

  * E1 receives route. It recursively resolves (E2, C1) to build an outgoing label/SID stack to forward via nodes 121

- This is the simplest design, with a single BGP transport CAR level

- This results in the minimum label/SID stack at each inter-domain hop. However, it can significantly build up the scale overhead on the core BRs, and can easily exceed the FIB capacity as well as the MPLS label space on these nodes.

- A subscription based Emulated-Pull solution is required with this flat design to enable all the intermediate nodes to be able to avoid learning and installing all the (PE, C) entries in the network.

7.1.2. Hierarchical Design with Next-hop self at ingress domain BR

Reference topology is shown below, with the BGP signaling and the resulting BGP and example IGP label stack at different hops

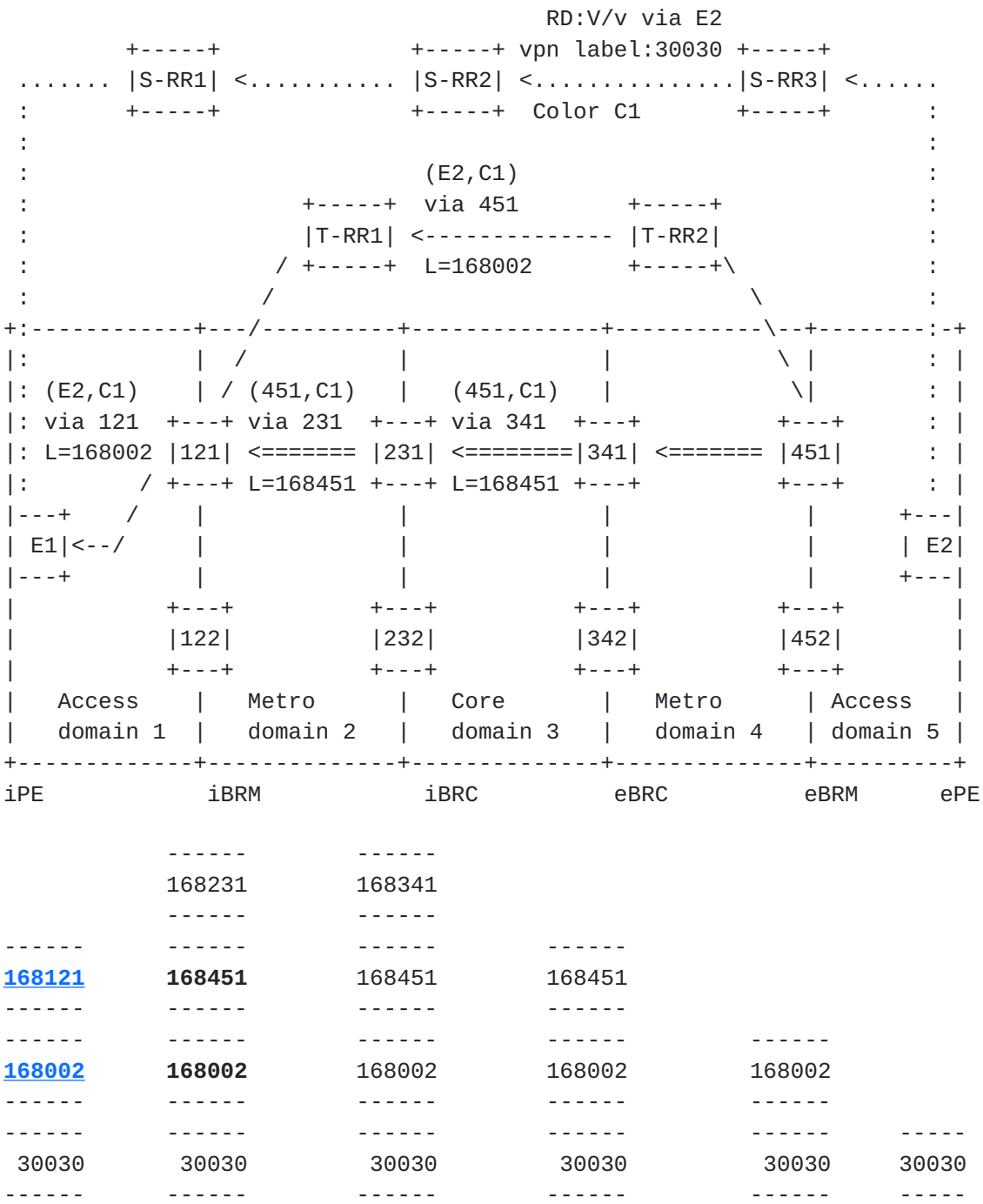


Figure 5: Heirarchical BGP transport CAR, NH at iBR

- With reference to the topology above:
  - Consider egress PE E2 advertises a VPN (service) route RD:V/v that propagates via service RRs to ingress PE E1.

  * A BGP CAR route (E2, C1) is also advertised by egress BRM node 451. The route may be sourced locally, for instance by redistribution from an IGP-FA, and is distributed via a Transport RR plane.

  * Ingress BRM node 121 learns about BGP CAR route (E2, C1) via node 451.

  * Node 121 also learns about BGP CAR route (451, C1) via node 231.

  * Node 121 advertise (E2, C1) received from T-RR to E1 with next-hop as it-self. It recursively resolves (E2, C1) to build an outgoing label/SID stack to forward traffic to (E1, C1) via (451, C1)

  * (451, C1) is not advertised to node 121

  * E1 receives route. It recursively resolves (E2, C1) to build an outgoing label/SID stack to forward via nodes 121

  * Ingress BRM node 121 needs to install data plane entry for (451, C1), and for (E2, C1).

- This hierarchical design avoids the need for core BRs to learn and install entries for (PE, C)

- An ingress BR (e.g., node 121) advertises the received remote (PE, C) routes to it's local ingress PE, setting next-hop to itself

  * Hence, the ingress BR need to install (PE, C) entries for egress PEs that it's local ingress PEs have installed BGP CAR routes for, as well as support a swap and push operation.

- This design keeps simple label programming on the ingress PE i.e. like single BGP transport CAR level. It is not exposed to hierarchical BGP CAR design at ingress BRM

- A subscription based Emulated-Pull model should be used with this design if the ingress BR has limited FIB capacity, and should only learn and install the necessary subset of (PE, C) routes.

7.1.3. Hierarchical Design with Next Hop Unchanged at ingress domain BR

Reference topology is shown below, with the BGP signaling and the resulting BGP and example IGP label stack at different hops.

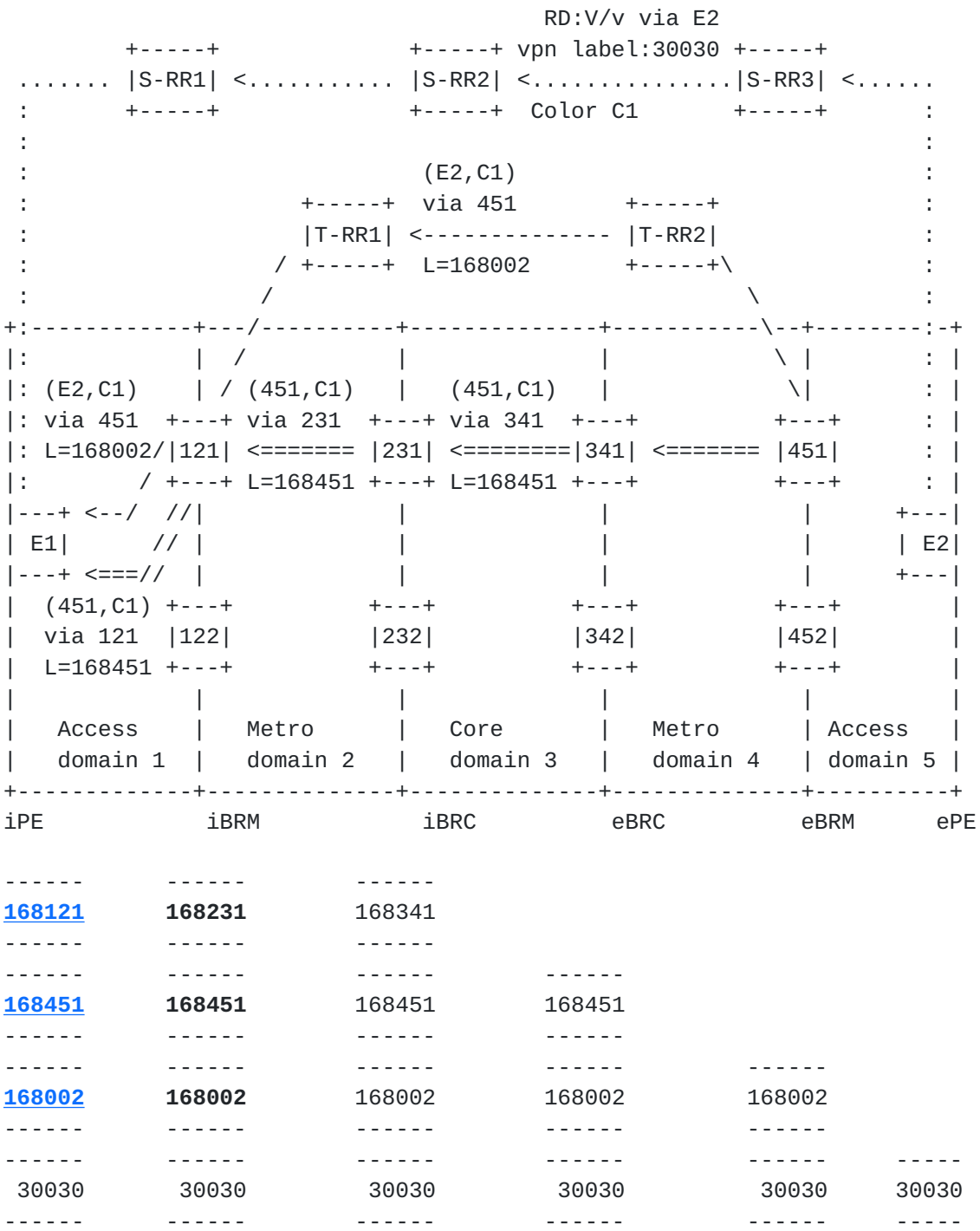


Figure 6: Heirarchical BGP transport CAR, NHU at iBR

- With reference to the topology above:
  - Consider egress PE E2 advertises a VPN (service) route RD:V/v that propagates via service RRs to ingress PE E1.

  * A BGP CAR route (E2, C1) is also advertised by egress BRM node 451. The route may be sourced locally, for instance by redistribution from an IGP-FA, and is distributed via a Transport RR plane.

  * Ingress BRM node 121 learns about BGP CAR route (E2, C1) via node 451.

  * Node 121 also learns about BGP CAR route (451, C1) via node 231.

  * Node 121 advertises both routes to E1.

  * (E2, C1) is advertised with NH via node 451; i.e., next-hop unchanged

  * (451, C1) is advertised with next-hop 121 i.e., next-hop self and local label 16451

  * Hence, E1 receives both routes. It recursively resolves (E2, C1) to build an outgoing label/SID stack to forward traffic to E1, via nodes 121 and 451.

  * Ingress BRM node 121 only needs to install data plane entry for (451, C1), and not for (E2, C1).

- In summary, with this design:

  * Only E1 needs to learn and install (E2, C1) because it has to install a service route RD:V/v with next-hop E2, and associated with a Color C1

  * However, E1 incurs additional complexity to perform the additional recursion to build and program the label stack. The complexity increases when there are multiple paths to be load-balanced across.

7.2. Automated Emulated-Pull Model to learn BGP CAR (PE, C)

From [BGP-CAR-Problem-Statement], we remind:

- The SR-PCE solution natively supports a PULL model: when E1 installs a VPN route V/v via (E2, C1), E1 requests its serving SR-PCE to compute the SR Policy to (E2, C1). I.e. E1 does not learn unneeded SR policies.

- BGP Signaling is natively a PUSH model.

- Emulated-PULL refers to the ability for a BGP CAR node E1 to "subscribe" to (E2, C1) route such that only the related paths are signaled to E1.

7.2.1. Subscription based BGP CAR Signaling

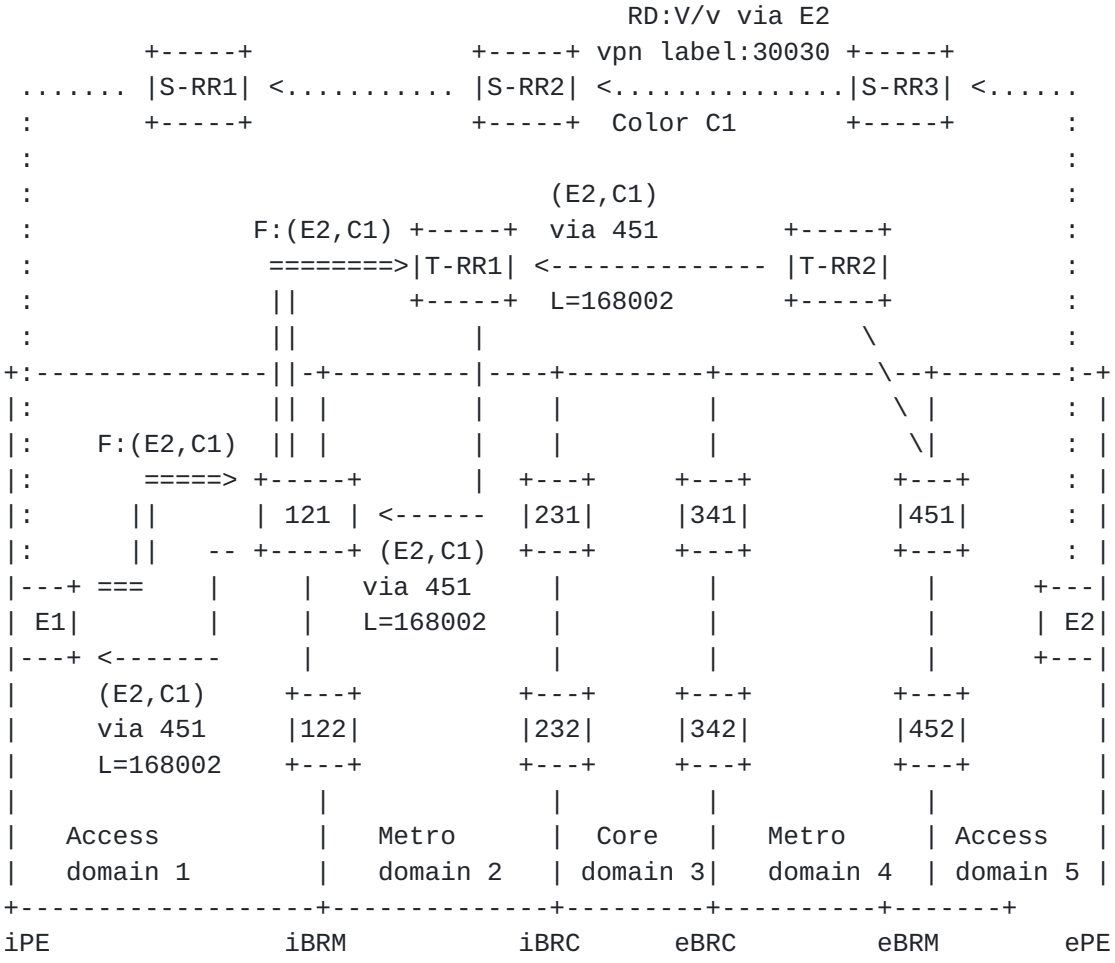


Figure 7: BGP transport CAR route Subscription

- Using the reference figure above that illustrates the use-case in section Figure 6

  - Ingress PE E1 subscribes to (E2, C1) using a BGP CAR filter route F (E2, C1), sent via Ingress BRM node 121

    + node 121 may act as an RR to E1

  - Node 121 propagates F(E2, C1) to Transport-RR T-RR1.

  - Assume Transport-RR has learnt routes for all PEs in network.

  * Based on received F(E2, C1), T-RR1 selectively sends (E2, C1) route to node 121, with Next-Hop of node 451 (i.e., egress BRM).

  * node 121 propagates the received (E2, C1) route to E1 that subscribed for it, with Next-Hop of node 451 (i.e., with BGP Next-Hop unchanged), and received label 168002.

  * Hence E1 learns (E2, C1) that it needs for resolving the received VPN route next-hop for colored route RD:V/v.

  * Note, redundant control flows that exist, for instance via node 122, are not shown above for simplicity.

- In addition, the subscription can be recursive triggered (not shown in the reference diagram above):

  * Upon receiving (E2, C1), E1 further subscribes to (451, C1) using a BGP CAR filter route F (451, C1) sent via node 121

  * Node 121 may not have learnt (451, C1), and hence propagates F (451, C1) to node 231

  * Assuming node 231 has learnt (451, C1), it will selectively send (451, C1) to node 121

  * Node 121 propagates received (451, C1) route to E1, with next-hop set to self and local label 168451

  * Node 121 also installs a data plane entry in this case for label 168451 and BGP recursive next-hop 231

  * Hence, E1 also learns (451, C1) that it needs for resolving the next-hop for (E2, C1)

  * This recursive subscription procedure can be used to minimize state further on ingress BRM nodes, if necessary

- The subscription based selective route signaling technique minimizes the state learnt and installed on both the ingress PEs as well as transit nodes.

  * The solution applies to all the design variants described in section Section 7.1

- This subscription-based selective route signaling has another benefit

  * It minimizes routing state that nodes such as BRs or T-RRs need to push to each of their subscription clients

  * When a remote node such as an egress BR or egress PE fails, the withdrawal of these routes can also be faster as a result, leading to faster convergence

- Details regarding the subscription based signaling will be described in a later version.

7.3. Additional Design Options

Other related well-known techniques that may be used to complement the solution design or provide an alternative as needed

7.3.1. Anycast SID for transit inter-domain nodes

Redundant BRs (e.g. egress BRMs) advertise their local domain's PE routes with same SID (based on label-index)

Anycast SID assigned to the egress BRMs abstracts state and hence avoids necessity to propagate failure of an egress BRM to ingress BRMs and PEs.

It also avoids traffic convergence issues for traffic from a remote ingress PE

7.3.2. Anycast SID for transport color endpoints i.e PEs

Anycast SID may be assigned to a redundant pair of PEs that have a common, dedicated set of service (VPN) attachments

Used with Anycast SID/static labels for services (e.g., per-VRF VPN label/SID)

This technique, similarly, abstracts state for the egress PEs and hence failure events from remote ingress PEs.

7.4. Convergence

Both existing and additional techniques are used to provide fast convergence for various network failure and change events

BGP Add-Path should be enabled for BGP CAR to signal multiple next hops through RR for fast convergence.

8. Interworking Scenarios

Details regarding various interworking scenarios will be added in a later version.

9. Fault Handling

This the fault management actions as described in [RFC7606] are applicable for handling of BGP update messages for BGP-CAR.

When the error determined allows for the router to skip the malformed NLRI(s) and continue processing of the rest of the update message, then it MUST handle such malformed NLRIs as 'Treat-as-withdraw'. In other cases, where the error in the NLRI encoding results in the inability to process the BGP update message (e.g. length related encoding errors), then the router SHOULD handle such malformed NLRIs as 'AFI/SAFI disable' when other AFI/SAFI besides BGP-CAR are being advertised over the same session. Alternately, the router MUST perform 'session reset' when the session is only being used for BGP-CAR.

10. IANA Considerations

IANA is requested to assign SAFI value TBD1 (BGP CAR) from the "SAFI Values" sub-registry under the "Subsequent Address Family Identifiers (SAFI) Parameters" registry with this document as a reference.

10.1. BGP CAR NLRI Types Registry

IANA is requested to create a "BGP CAR NLRI Types" sub-registry under the "Border Gateway Protocol (BGP) Parameters" registry with this document as a reference. The registry is for assignment of the one octet sized code-points for BGP CAR NLRI types and populated with the values shown below:

Type	NLRI Type	Reference
0	Reserved (not to be used)	[This document]
1	Color-Aware Routes NLRI	[This document]
2-255	Unassigned	

Allocations within the registry are to be made under the "Specification Required" policy as specified in [RFC8126]).

10.2. BGP CAR NLRI TLV Registry

IANA is requested to create a "BGP CAR NLRI TLV Types" sub-registry under the "Border Gateway Protocol (BGP) Parameters" registry with this document as a reference. The registry is for assignment of the one octet sized code-points for BGP-CAR NLRI non-key TLV types and populated with the values shown below:

Type	NLRI Type	Reference
0	Reserved (not to be used)	[This document]
1	Label TLV	[This document]
2	Label Index TLV	[This document]
3	SRv6 SID TLV	[This document]
4-255	Unassigned	

Allocations within the registry are to be made under the "Specification Required" policy as specified in [RFC8126]).

10.3. Guidance for Designated Experts

In all cases of review by the Designated Expert (DE) described here, the DE is expected to ascertain the existence of suitable documentation (a specification) as described in [RFC8126]. The DE is also expected to check the clarity of purpose and use of the requested code points. Additionally, the DE must verify that any request for one of these code points has been made available for review and comment within the IETF: the DE will post the request to the IDR Working Group mailing list (or a successor mailing list designated by the IESG). If the request comes from within the IETF, it should be documented in an Internet-Draft. Lastly, the DE must ensure that any other request for a code point does not conflict with work that is active or already published within the IETF.

10.4. BGP Extended Community Registry

IANA is requested to allocate the sub-type TBD2 for "Local Color Mapping (LCM)" under the "BGP Transitive Opaque Extended Community" registry under the "BGP Extended Community" parameter registry.

11. Security Considerations

TBD

12. Acknowledgements

The authors would like to acknowledge the review and inputs from many people.TBD

13. References

13.1. Normative References

[I-D.ietf-bess-srv6-services]
Dawra, G., Filsfils, C., Talaulikar, K., Raszuk, R., Decraene, B., Zhuang, S., and J. Rabadan, "SRv6 BGP based Overlay services", draft-ietf-bess-srv6-services-05 (work in progress), November 2020.

[I-D.ietf-idr-bgp-ipv6-rt-constrain]
Patel, K., Raszuk, R., Djernaes, M., Dong, J., and M. Chen, "IPv6 Extensions for Route Target Distribution", draft-ietf-idr-bgp-ipv6-rt-constrain-12 (work in progress), April 2018.

[I-D.ietf-idr-tunnel-encaps]
Patel, K., Velde, G., Sangli, S., and J. Scudder, "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-21 (work in progress), January 2021.

[I-D.ietf-lsr-flex-algo]
Psenak, P., Hegde, S., Filsfils, C., Talaulikar, K., and A. Gulko, "IGP Flexible Algorithm", draft-ietf-lsr-flex-algo-13 (work in progress), October 2020.

[I-D.ietf-spring-segment-routing-policy]
Filsfils, C., Talaulikar, K., Voyer, D., Bogdanov, A., and P. Mattes, "Segment Routing Policy Architecture", draft-ietf-spring-segment-routing-policy-09 (work in progress), November 2020.

[I-D.ietf-spring-srv6-network-programming]
Filsfils, C., Camarillo, P., Leddy, J., Voyer, D., Matsushima, S., and Z. Li, "SRv6 Network Programming", draft-ietf-spring-srv6-network-programming-28 (work in progress), December 2020.

[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>.

[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>.

[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>.

[RFC5512] Mohapatra, P. and E. Rosen, "The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", RFC 5512, DOI 10.17487/RFC5512, April 2009, <https://www.rfc-editor.org/info/rfc5512>.

[RFC5701] Rekhter, Y., "IPv6 Address Specific BGP Extended Community Attribute", RFC 5701, DOI 10.17487/RFC5701, November 2009, <https://www.rfc-editor.org/info/rfc5701>.

[RFC7311] Mohapatra, P., Fernando, R., Rosen, E., and J. Uttaro, "The Accumulated IGP Metric Attribute for BGP", RFC 7311, DOI 10.17487/RFC7311, August 2014, <https://www.rfc-editor.org/info/rfc7311>.

[RFC7606] Chen, E., Ed., Scudder, J., Ed., Mohapatra, P., and K. Patel, "Revised Error Handling for BGP UPDATE Messages", RFC 7606, DOI 10.17487/RFC7606, August 2015, <https://www.rfc-editor.org/info/rfc7606>.

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

[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>.

[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>.

[RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <https://www.rfc-editor.org/info/rfc8402>.

[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>.

13.2. Informative References

[I-D.ietf-mpls-seamless-mpls] Leymann, N., Decraene, B., Filsfils, C., Konstantynowicz, M., and D. Steinberg, "Seamless MPLS Architecture", draft-ietf-mpls-seamless-mpls-07 (work in progress), June 2014.

[RFC3906] Shen, N. and H. Smit, "Calculating Interior Gateway Protocol (IGP) Routes Over Traffic Engineering Tunnels", RFC 3906, DOI 10.17487/RFC3906, October 2004, <https://www.rfc-editor.org/info/rfc3906>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <https://www.rfc-editor.org/info/rfc4271>.

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

[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>.

[RFC6952] Jethanandani, M., Patel, K., and L. Zheng, "Analysis of BGP, LDP, PCEP, and MSDP Issues According to the Keying and Authentication for Routing Protocols (KARP) Design Guide", RFC 6952, DOI 10.17487/RFC6952, May 2013, <https://www.rfc-editor.org/info/rfc6952>.

[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>.

Authors' Addresses

Dhananjaya Rao
Cisco Systems
USA

Email: dhrao@cisco.com

Swadesh Agrawal
Cisco Systems
USA

Email: swaagraw@cisco.com

Clarence Filsfils
Cisco Systems
Belgium

Email: cfilsfil@cisco.com

Ketan Talaulikar
Cisco Systems
India

Email: ketant@cisco.com