BESS WorkGroup
Internet-Draft
Intended status: Standards Track
Expires: January 6, 2023

D. Rao
S. Agrawal
C. Filsfils
Cisco Systems
D. Steinberg
Lapishills Consulting Limited
L. Jalil
Verizon
Y. Su
Alibaba, Inc
B. Decraene
Orange
J. Guichard
Futurewei
K. Talaulikar
K. Patel
Arrcus, Inc
H. Wang
Huawei Technologies
J. Uttaro
ATT

July 5, 2022

BGP Color-Aware Routing (CAR)
draft-dskc-bess-bgp-car-05

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).

This Internet-Draft will expire on January 6, 2023.

Table of Contents

1. Introduction

This document specifies a new BGP SAFI called BGP Color-Aware Routing (BGP CAR). BGP CAR fulfills the transport and VPN problem statement and requirements described in [dskc-bess-bgp-car-problem-statement].

1.1. Terminology

Intent	Any combination of the following behaviors: a/ Topology path selection (e.g. minimize metric, avoid resource), b/ NFV service insertion (e.g. service chain steering), c/ per-hop behavior (e.g. 5G slice).
Color	A 32-bit numerical value associated with an intent: e.g. low-cost vs low-delay vs avoiding some resources.
Colored Service Route	An egress PE E2 colors its BGP VPN route V/v to indicate the intent that it requests for the traffic bound to V/v. The color is encoded as a BGP Color Extended community [I-D.ietf-idr-tunnel-encaps].
Color-Aware Path to (E2, C)	A routed path to E2 which satisfies the intent associated with color C. Several technologies may provide a Color-Aware Path to (E2, C): SR Policy [I-D.ietf-spring-segment-routing-policy], IGP Flex-Algo [I-D.ietf-lsr-flex-algo], BGP CAR [specified in this document].
Color-Aware Route (E2, C)	A distributed or signaled route that builds a color-aware path to E2 for color C.
Service Route Automated Steering on Color-aware path	E1 automatically steers a C-colored service route V/v from E2 onto an (E2, C) path. If several such paths exist, a preference scheme is used to select the best path: E.g. IGP Flex-Algo first then BGP CAR then SR Policy.
Color Domain	A set of nodes which share the same Color-to-Intent mapping. This set can be organized in one or several IGP instances or BGP domains.
Resolution of a BGP CAR route (E, C)	An inter-domain BGP CAR route (E, C) from N is resolved on an intra-domain color-aware path (N, C) where N is the next-hop of the BGP CAR route.
Resolution vs Steering	In this document and consistently with the terminology of the SR Policy document [I-D.ietf-spring-segment-routing-policy], steering is used to describe the mapping of a service route onto a BGP CAR path while the term

	resolution is preserved for the mapping of an inter-domain BGP CAR route on an intra-domain color-aware path.	
	Service Steering: Service route -> BGP CAR path (or other Color-Aware Routed Paths: e.g., SR Policy)	
	Intra-Domain Resolution: BGP CAR route -> intra-domain color aware path (e.g. SR Policy, IGP Flex-Algo, BGP CAR)	

1.2. Illustration

Here is a brief illustration of the salient properties of the BGP CAR solution.

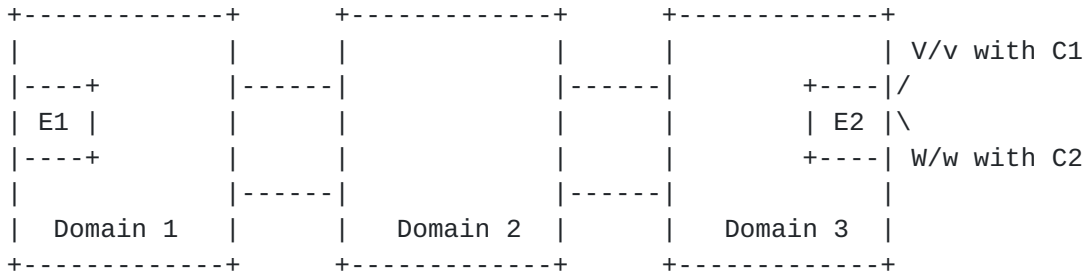


Figure 1

All the nodes are part of an interdomain network under a single authority and with a consistent color-to-intent mapping:

- C1 is mapped to "low-delay"
  - Flex-Algo FA1 is mapped to "low delay" and hence to C1
- C2 is mapped to "low-delay and avoid resource R"
  - Flex-Algo FA2 is mapped to "low delay and avoid resource R" and hence C2

E1 receives two service routes from E2:

- V/v with BGP Color Extended-Community C1
- W/w with BGP Color Extended-Community C2

E1 has the following color-aware paths:

- (E2, C1) provided by BGP CAR with the following per-domain support:
  - Domain1: over IGP FA1
  - Domain2: over SR Policy bound to color C1
  - Domain3: over IGP FA1
- (E2, C2) provided by SR Policy

E1 automatically steers the received service routes as follows:

- V/v via (E2, C1) provided by BGP CAR
- W/w via (E2, C2) provided by SR Policy

Illustrated Properties:

- Leverage of the BGP Color Extended-Community
  - The service routes are colored with widely-used BGP Color Extended-Community
- (E, C) Automated Steering
  - V/v and W/w are automatically steered on the appropriate color-aware path
- Seamless co-existence of BGP CAR and SR Policy
  - V/v is steered on BGP CAR color-aware path
  - W/w is steered on SR Policy color-aware path
- Seamless interworking of BGP CAR and SR Policy
  - V/v is steered on a BGP CAR color-aware path that is itself resolved within domain 2 onto an SR Policy bound to the color of V/v

Other properties:

- MPLS dataplane: with 300k PE's and 5 colors, the BGP CAR solution ensures that no single node needs to support a dataplane scaling

in the order of Remote PE * C. This would otherwise exceed the MPLS dataplane.

- Control-Plane: a node should not install a (E, C) path if it does not need it

- Incongruent Color-Intent mapping: the solution supports the signaling of a BGP CAR route across different color domains

The keys to this simplicity are:

- the leverage of the BGP Color Extended-Community to color service routes

- the definition of the automated steering: a C-colored service route V/v from E2 is steered onto a color-aware path (E2, C)

- the definition of the data model of a BGP CAR path: (E, C)

  * consistent with SR Policy data model

- the definition of the recursive resolution of a BGP CAR route: a BGP CAR (E2, C) via N is resolved onto the color-aware path (N, C) which may itself be provided by BGP CAR or via another color-aware routing solution: SR Policy, IGP Flex-Algo.

1.3. Requirements Language

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

2. BGP CAR SAFI

2.1. Data Model

The BGP CAR data model is:

- NLRI Key: IP Prefix, Color

- NLRI non-key encapsulation data: MPLS label stack, Label index, SRv6 SID list etc.

- BGP Next Hop

- AIGP Metric: accumulates color/intent specific metric across domains

- Local-Color-Mapping Extended-Community (LCM-EC): Optional 32-bit Color value used when a CAR route propagates between different color domains

2.2. Extensible encoding

Extensible encoding is ensured by:

- NLRI Route-Type field: provides extensibility to add new NLRI formats for new route-types

- Key length: field enables handling of unsupported route-types opaquely, enabling transitivity via RRs

- TLV-based encoding of non-key part of NLRI: enables flexible support for multiple encapsulations with efficient update packing

- AIGP Attribute provides extensibility via TLVs, enabling definition of additional metric semantics for a color as needed for an intent

2.3. BGP CAR Route Origination

A BGP CAR route may be originated locally (e.g., loopback) or through redistribution of an (E, C) color-aware path provided by another routing solution: SR Policy, IGP Flex-Algo, RSVP-TE or BGP-LU [RFC8277].

2.4. BGP CAR Route Validation

A BGP CAR path (E, C) from N with encapsulation T is valid if color-aware path (N, C) exists with encapsulation T available in dataplane.

A local policy may customize the validation process:

- the color constraint in the first check may be relaxed: instead N is reachable via alternate color(s) or in the default routing table

- the dataplane availability constraint of T may be relaxed, to use an alternate encapsulation

- a performance-measurement verification may be added to ensure that the intent associated with C is met (e.g. delay < bound)

2.5. BGP CAR Route Resolution

A BGP color-aware route (E2, C1) from N is resolved over a color-aware route (N, C1). The color-aware route (N, C1) may be provided recursively by BGP CAR or by other routing solutions: SR Policy, IGP Flex-Algo, RSVP-TE, BGP-LU.

When multiple resolutions are possible, the default preference should be: IGP Flex-Algo, SR Policy, RSVP-TE, BGP CAR, BGP LU.

Through local policy, a BGP color-aware route (E2, C1) from N may be resolved over a color-aware route (N, C2): i.e. the local policy maps the resolution of C1 over C2. For example, in a domain where resource R is known to not be present, the inter-domain intent C1="low delay and avoid R" may be resolved over an intra-domain path of intent C2="low delay". Another example is, if no (N, C1) path is available, and the user has allowed resolution via C2.

Resolution may also be automated using Color-EC as illustrated in Appendix B.2 .

The color-aware route (N, C1) may have a different dataplane encapsulation than the one of (E2, C1): e.g. a BGP CAR route (E2, C1) with SR-MPLS encapsulation may be transported over an intermediate SRv6 domain.

2.6. AIGP Metric Computation

The Accumulated IGP (AIGP) Attribute is updated 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. For 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 a penalty in accumulated IGP metric. The discontinuity is also indicated to upstream nodes via a bit in the AIGP TLV.

AIGP metric computation is recursive.

To avoid continuous IGP metric churn causing end to end BGP CAR churn, an implementation should provide thresholds to trigger AIGP update.

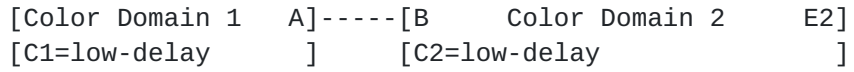
Additional AIGP extensions may be defined to signal state for specific use-cases: MSD along the BGP CAR advertisement, Minimum MTU along the BGP CAR advertisement.

2.7. Path Availability

The (E, C) route inherently provides availability of redundant paths at every hop. For instance, BGP CAR routes originated by two egress ABRs in a domain are advertised as multiple paths to ingress ABRs in the domain, where they become equal-cost or primary-backup paths. A failure of an egress ABR is detected and handled by ingress ABRs locally within the domain for faster convergence, without any necessity to propagate the event to upstream nodes for traffic restoration.

BGP ADD-PATH should be enabled for BGP CAR to signal multiple next hops through a transport RR.

2.8. BGP CAR signaling through different color domains



Let us assume a BGP CAR route (E2, C2) is signaled from B to A; two border routers of respectively domain 2 and domain 1. Let us assume that these two domains do not share the same color-to-intent mapping. Low-delay in domain 2 is color C2 while C1 in domain 1 (C1 <> C2).

The BGP CAR solution seamlessly supports this (rare) scenario while maintaining the separation and independence of the administrative authority in different color domains.

The solution works as follows:

- Within domain 2, the BGP CAR route is (E2, C2) via E2
- B signals to A the BGP CAR route as (E2, C2) via B with Local-Color-Mapping-Extended-Community (LCM-EC) of color C2
- A is aware (classic peering agreement) of the intent-to-color mapping within domain 2 ("low-delay" in domain 2 is C2)
- A maps C2 in LCM-EC to C1 and signals within domain 1 the received BGP CAR route as (E2, C2) via A with LCM-EC(C1)

- The nodes within the receiving domain 1 use the local color encoded in the LCM-EC for next-hop resolution and service steering

Salient properties:

- The NLRI never changes
- E is globally unique, which makes E-C in that order unique
- In the vast majority of the cases, the color of the NLRI is used for resolution and steering
- In the rare case of color incongruence, the local color encoded in LCM-EC takes precedence

Further illustrations are provided in Appendix B.

2.9. Format and Encoding

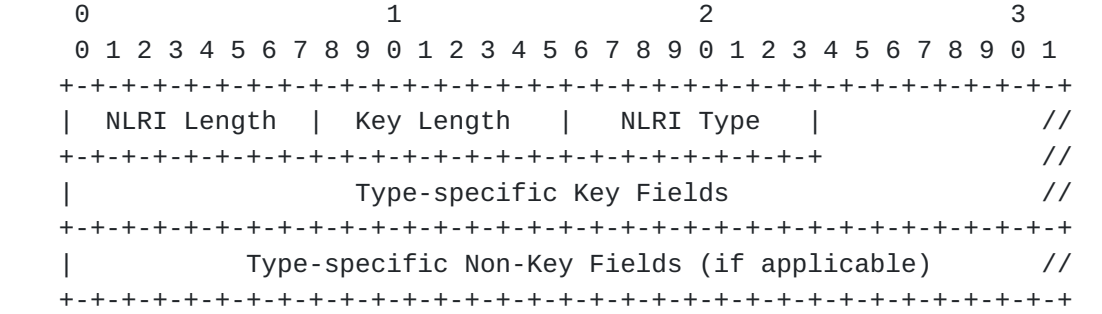
BGP CAR leverages the BGP multi-protocol extensions [RFC4760] and uses the MP_REACH_NLRI and MP_UNREACH_NLRI attributes for route updates by using the SAFI value TBD1 along with AFI 1 for IPv4 prefixes and AFI 2 for IPv6 prefixes.

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.

2.9.1. BGP CAR SAFI NLRI Format

The generic format for the BGP CAR SAFI 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 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 definition 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 deployed Route Reflectors (RR) to propagate NLRI types introduced in the future in a transparent manner.

A route (NLRI) can carry more than one non-key TLV (of different types). This provides significant benefits such as signaling multiple encapsulations simultaneously for the same route, each with a different value (label/SID etc). This enables simpler, efficient migrations with low overhead :

- avoids duplicate routes to signal different encapsulations

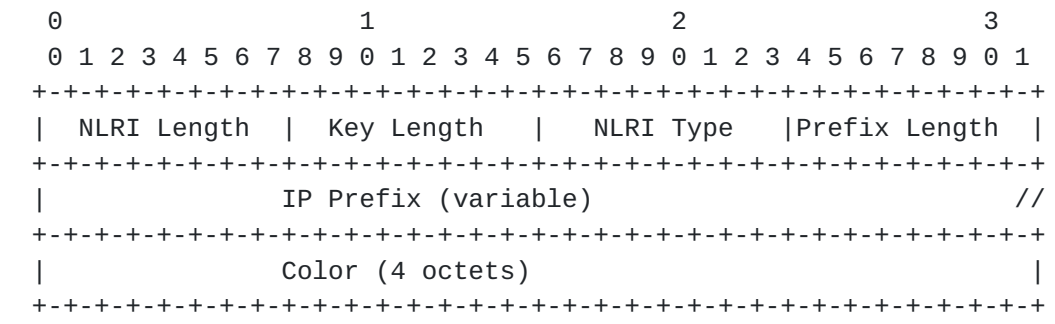
- avoids need for separate control planes for distribution

- preserves update packing

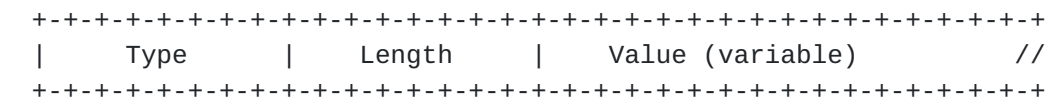
The non-key portion of the NLRI MUST be omitted while carrying it within the MP_UNREACH_NLRI when withdrawing the route advertisement.

2.9.2. Color-Aware Routes 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. It indicates the total length comprised of the Prefix Length field, IP Prefix field, and the Color field, as described below. For IPv4 (AFI=1), the minimum length is 5 and maximum length is 9. For IPv6 (AFI=2), the minimum length is 5 and maximum length is 21.

- NLRI Type: 1

- Type-Specific Key Fields: as below

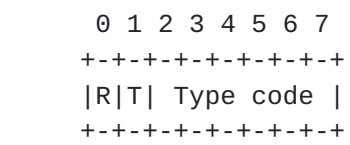
  * Prefix Length: 1 octet field that carries the length of prefix in bits. Length MUST be less than or equal to 32 for IPv4 (AFI=1) and less than or equal to 128 for IPv6 (AFI=2).

  * 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., 0 octet for prefix length 0, 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.

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

* Type: 1 octet that contains the type code and flags. It is encoded as shown below:



where:

+ R: Bit is reserved and MUST be set to 0 and ignored on receive.

+ T: Transitive bit, applicable to speakers that change the BGP CAR next hop

  - T bit set to indicate TLV is transitive. An unrecognized transitive TLV MUST be propagated by a speaker that changes the next hop

  - T bit unset to indicate TLV is non-transitive. An unrecognized non-transitive TLV MUST not be propagated by a speaker that changes next hop

  A speaker that does not change next hop SHOULD propagate all received TLVs.

+ Type code: Remaining 6 bits contain the type of the 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.

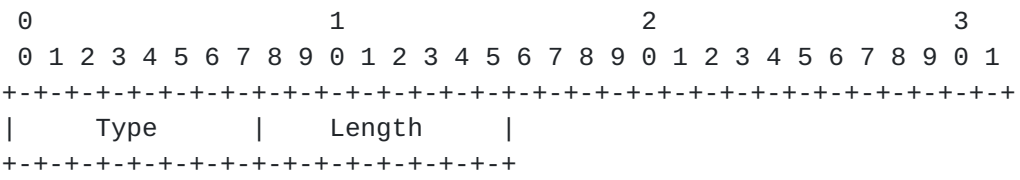
The prefix is routable across the administrative domain where BGP transport CAR is deployed. It is possible that the same prefix is originated by multiple BGP CAR speakers in the case of anycast addressing or multi-homing.

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.

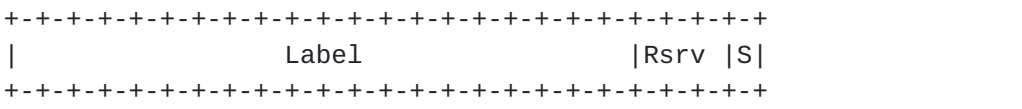
The following sub-sections specify the non-key TLVs associated with the Color-Aware Routes NLRI type.

2.9.2.1. 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:



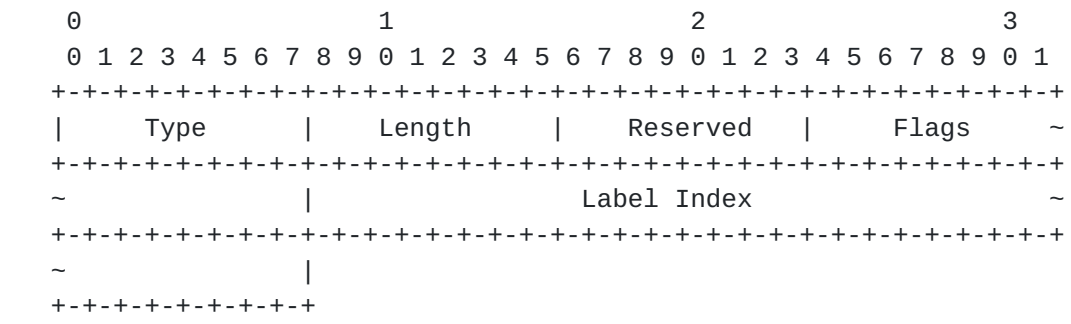
where:

- Type : Type code is 1. T bit MUST be unset
- 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).

2.9.2.2. 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 labeled color-aware routes and has the following format:



where:

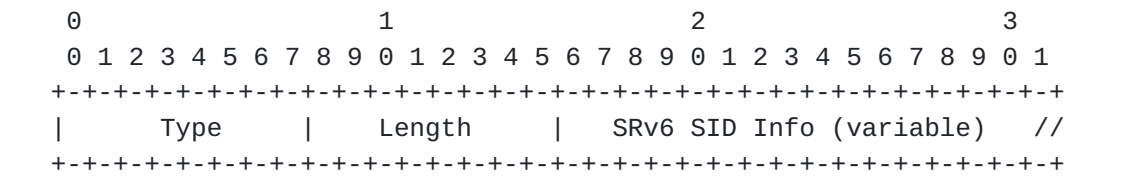
- Type : Type code is 2. T bit MUST be set
- Length: 7
- Reserved: 1 octet field that MUST be set to 0 and ignored on receipt.
- Flags: 2 octet field that maps to the Flags field of the Label-Index TLV of the BGP Prefix SID Attribute [RFC8669].
- Label Index: 4 octet field that maps to the Label Index field of the Label-Index TLV of the BGP Prefix SID Attribute [RFC8669].

This TLV provides the equivalent functionality as Label-Index TLV of [RFC8669] for Transport CAR route in SR-MPLS deployments. It provides much better packing efficiency by carrying label Index in NLRI instead of the BGP Prefix SID attribute. 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.

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].

2.9.2.3. 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:



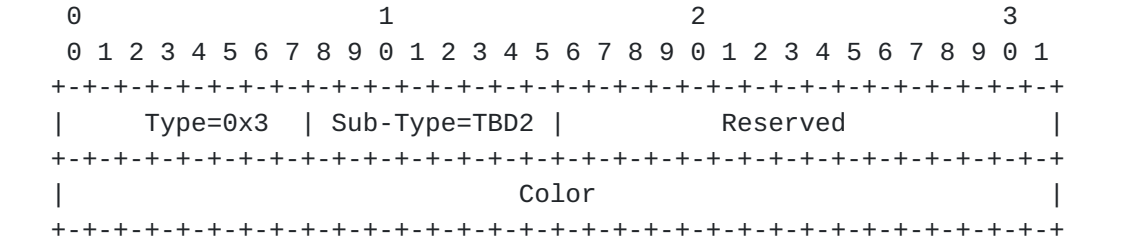
where:

- Type : Type code is 3. T bit MUST be unset
- 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.

2.9.3. 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 a CAR route crosses the originator color domain's boundary, LCM EC is added. LCM EC conveys the local color mapping for the intent (e.g. low latency) into transit or remote color domains.

An implementation SHOULD NOT send more than one instance of the LCM EC. However, if more than one instance is received, an implementation MUST disregard all instances other than the one with the numerically highest value.

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

2.10. Error Handling

The error handling 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, 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.

Following errors result in 'AFI/SAFI disable' or 'session reset':

- Minimum NLRI length check error.

- NLRI length conflict with key length.

- Key length encoding errors (such as minimum, maximum and conflict with prefix length).

There can be cases where the NLRI length value is in conflict with the enclosed non-key TLVs, which themselves carry length values. Either the length of a TLV would cause the NLRI length to be exceeded when parsing the TLV, or fewer than 2 bytes remain when beginning to parse the TLV.

In either of these cases, an error condition exists and the "treat-as-withdraw" approach MUST be used (unless some other, more severe error is encountered dictating a stronger approach), and the NLRI Length MUST be relied upon to enable the beginning of the next NLRI field to be located. The above recommendations follow the principle defined in section 4 of [RFC7606].

Type-Specific Non-Key TLV handling

- If multiple instances of same type are encountered, all but the first instance MUST be ignored.
- Type specific length constraints should be verified. The TLV is discarded if there is an error.
- A TLV is not considered malformed because of failing any semantic validation of its Value field.
- Speaker modifying the BGP next-hop MUST recognize at least one of the forwarding information TLV (such as label and SRv6 SID). If it is not able to, such NLRI is considered invalid and not eligible for best path selection.

3. Service route Automated Steering on Color-Aware path

E1 automatically steers a C-colored service route V/v from E2 onto an (E2, C) color-aware path. If several such paths exist, a preference scheme is used to select the best path: E.g. IGP Flex-Algo first then BGP CAR then SR Policy.

This is consistent with the automated service route steering on SR Policy (a routing solution providing color-aware path) defined in [I-D.ietf-spring-segment-routing-policy]. All the steering variations defined in [I-D.ietf-spring-segment-routing-policy] are applicable to BGP CAR color-aware path: on-demand steering, per-destination, per-flow, CO-only. For brevity, in this revision, we refer the reader to the [I-D.ietf-spring-segment-routing-policy] text.

Salient property: Seamless integration of BGP CAR and SR Policy.

Service steering via BGP CAR is applicable to any BGP SAFI, including SAFIs for IPv4/IPv6, L3VPN, PW, EVPN, FlowSpec, and BGP-LU.

Appendix A provides illustrations of service route automated steering.

4. Intents

The widely deployed color-aware path SR Policy solution demonstrates that the following intents can easily be associated with a color:

1. Minimization of a cost metric vs a latency metric
    * Minimization of different metric types, static and dynamic
2. Exclusion/Inclusion of SRLG and/or Link Affinity and/or minimum MTU/number of hops
3. Bandwidth management
4. In the inter-domain context, exclusion/inclusion of entire domains, and border routers
5. Inclusion of one or several virtual network function chains
    * Located in a regional domain and/or core domain, in a DC
6. Localization of the virtual network function chains
    * Some functions may be desired in the regional DC or vice versa
7. Per-Destination and Per-Flow steering

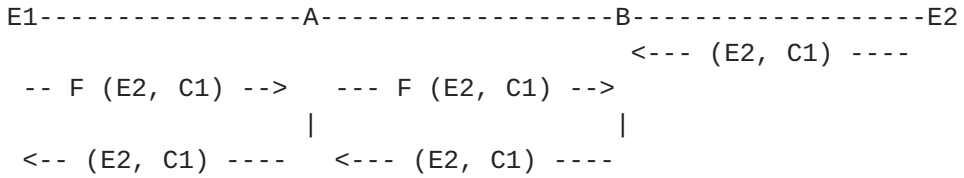
It is straightforward to note that the BGP CAR color-aware alternative supports intents 1, 2, 4 and 7.

Future revisions of this document will analyze the BGP CAR supports for 3, 5 and 6.

5. (E, C) Subscription and Filtering

This section defines an (E, C) BGP subscription model that allows to filter the (E, C) routes learned by a BGP CAR node.

5.1. Illustration



o BGP CAR route (E2, C1) advertised by E2 is not unconditionally distributed beyond a certain point (e.g., B)

- E1 subscribes to (E2, C1) by advertising a filter route F (E2, C1) to its upstream peer A
- If A has (E2, C1) in its BGP RIB, it will advertise (E2, C1) to E1
- If A does not have (E2, C1), it will advertise F (E2, C1) to its peer B
- B will advertise (E2, C1) to A, which will distribute it to E1

E1 may trigger a subscription for BGP CAR route (E2, C1) as a result of receiving a C1-colored service route V/v from E2, for on-demand steering via (E2, C1).

5.2. Definition

future version of this document

6. Scaling

This section analyses the key scale requirement of [ref:dskc-bess-bgp-car-problem-statement], 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

Figure 2 provides an ultra-scale reference topology. Section 6.2 presents three design models to deploy BGP CAR in the reference topology. Section 6.3 analyses the scaling properties of each model. Section 6.4 illustrates the scaling benefits of the (E, C) BGP subscription and filtering.

6.1. Ultra-Scale Reference Topology

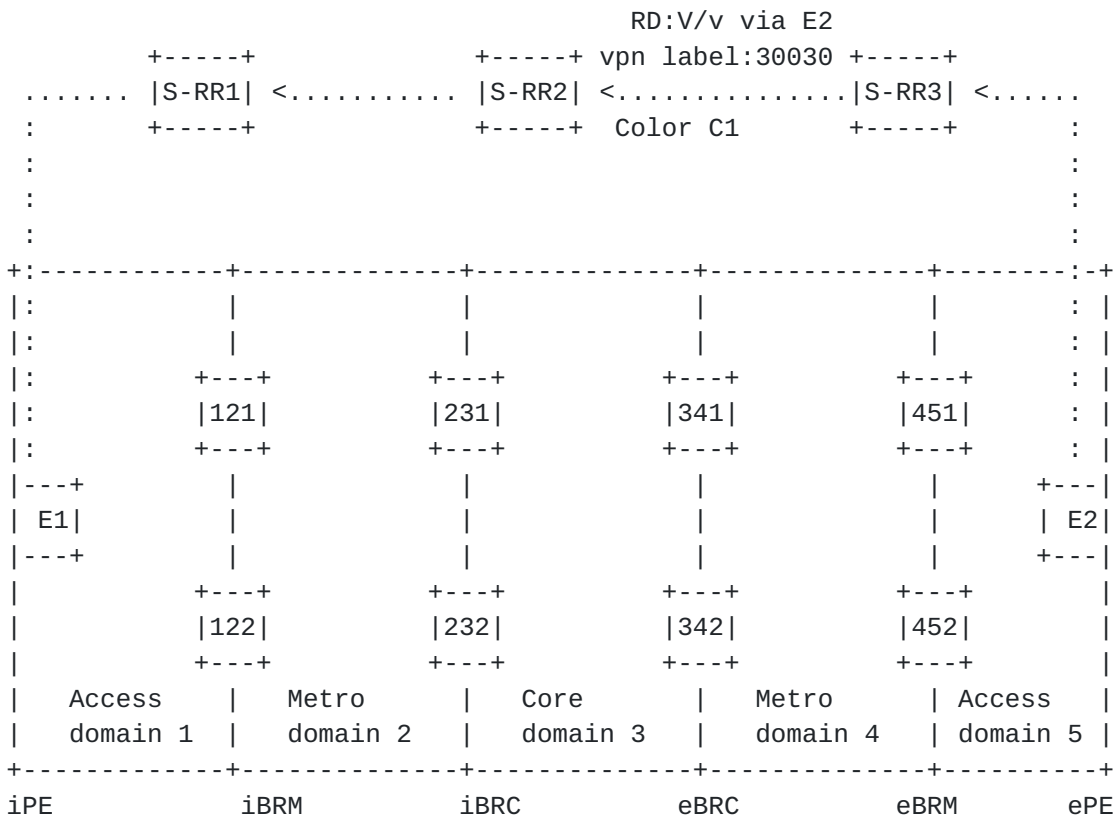


Figure 2: Ultra-Scale Reference Topology

The following applies to the reference topology above:

- Independent ISIS/OSPF SR instance in each domain.
- Each domain has Flex Algo 128. Prefix SID for a node is SRGB 168000 plus node number.
- A BGP CAR route (E2, C1) is advertised by egress BRM node 451.The route is sourced locally from redistribution from IGP-FA 128.
- Not shown for simplicity, node 452 will also advertise (E2, C1).
- When a transport RR is used within the domain or across domains, ADD-PATH is enabled to advertise paths from both egress BRs to it's clients.
- Egress PE E2 advertises a VPN route RD:V/v with BGP Color extended community C1 that propagates via service RRs to ingress PE E1.
- E1 steers V/v prefix via color-aware path (E2,C1) and VPN label 30030

6.2. Deployment model

6.2.1. Flat

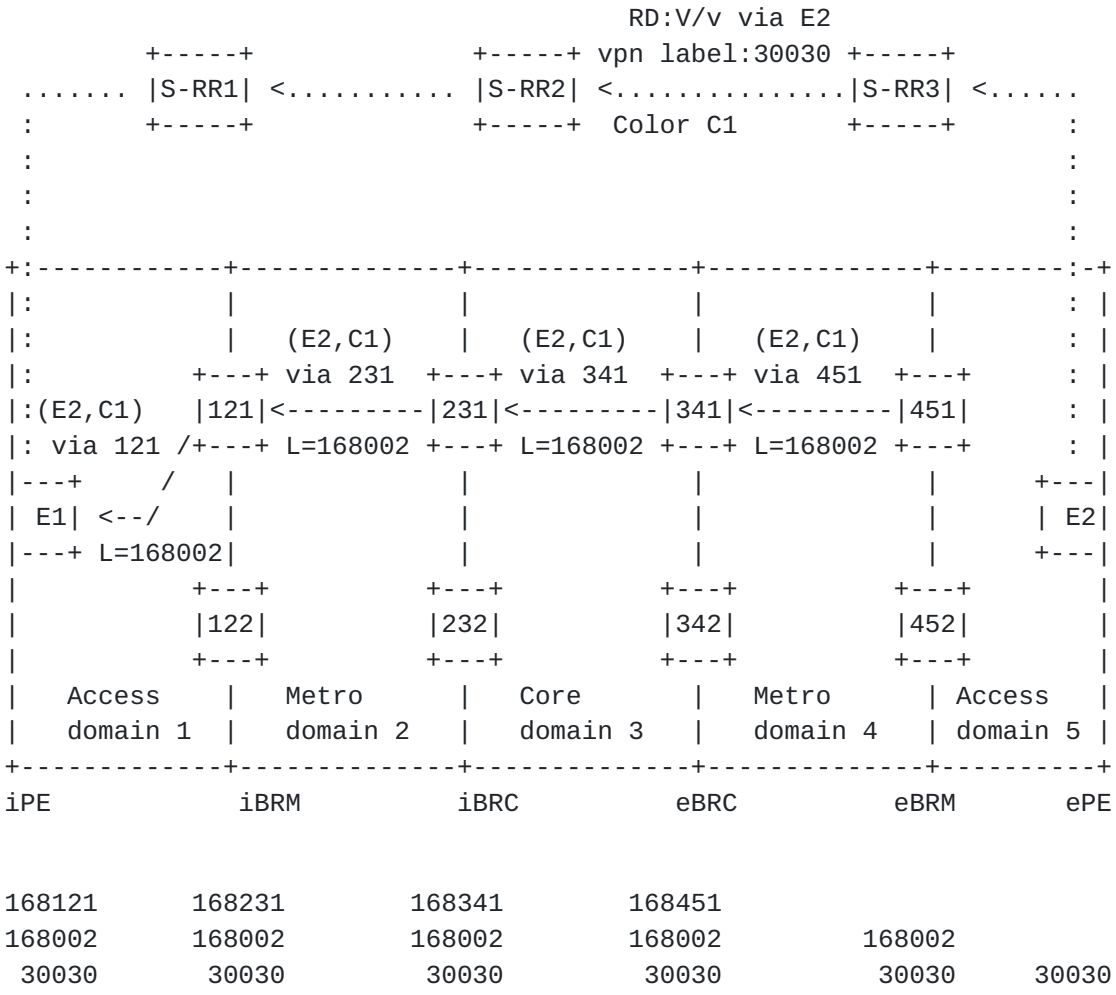


Figure 3

1. Node 451 advertises BGP CAR route (E2, C1) to 341, from which it goes to 231 then to 121 and finally to E1

2. Each BGP hop allocates local label and programs swap entry in forwarding for (E2, C1)

3. E1 receives BGP CAR route (E2, C1) via 121 with label 168002

    1. Let's assume E1 selects that path

4. E1 resolves BGP CAR route (E2, C1) via 121 on color-aware path (121, C1)

1. Color-aware path (121, C1) is FA128 path to 121 (label 168121)

5. E1's imposition color-aware label-stack for V/v is thus

    1. 30030 <=> V/v

    2. 168002 <=> (E2, C1)

    3. 168121 <=> (121, C1)

6. Each BGP hop performs swap operation on 168002 bound to color-aware path (E2,C1)

6.2.2. Hierarchical Design with next-hop-self at ingress domain BR

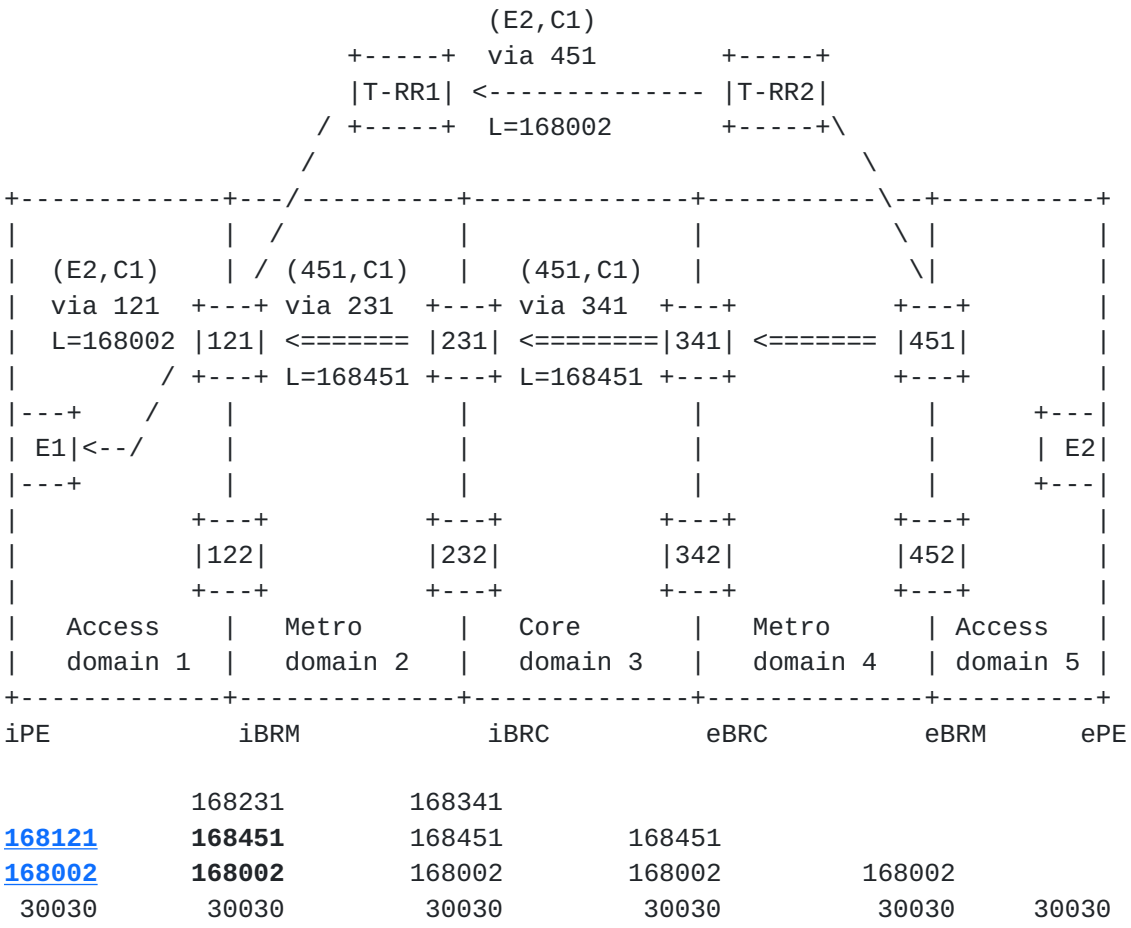


Figure 4: Heirarchical BGP transport CAR, NHS at iBR

1. Node 451 advertises BGP CAR route (451, C1) to 341, from which it goes to 231 and finally to 121

2. Each BGP hop allocates local label and programs swap entry in forwarding for (451, C1)

3. 121 resolves received BGP CAR route (451, C1) via 231 (label 168451) on color-aware path (231, C1)

    1. Color-aware path (231, C1) is FA128 path to 231 (label 168231)

4. 451 advertises BGP CAR route (E2, C1) via 451 to Transport RR T-RR2, which reflects it to T-RR1, which reflects it to 121

5. 121 receives BGP CAR route (E2, C1) via 451 with label 168002

    1. Let's assume 121 selects that path

6. 121 resolves BGP CAR route (E2, C1) via 451 on color-aware path (451, C1)

    1. Color-aware path (451, C1) is BGP CAR path to 451 (label 168451)

7. 121 imposition of color-aware label stack for (E2, C1) is thus

    1. 168002 <=> (E2, C1)

    2. 168451 <=> (451, C1)

    3. 168231 <=> (231, C1)

8. 121 advertises (E2, C1) to E1 with next hop self (121) and label 168002

9. E1 constructs same imposition color-aware label-stack for V/v via (E2, C1) as in the flat model:

    1. 30030 <=> V/v

    2. 168002 <=> (E2, C1)

    3. 168121 <=> (121, C1)

10. 121 performs swap operation on 168002 with hierarchical color-aware label stack for (E2, C1) via 451 from step 7

11. Nodes 231 and 341 perform swap operation on 168451 bound to color-aware path (451, C1)

12. 451 performs swap operation on 168002 bound to color-aware path (E2, C1)

Note: E1 does not need the BGP CAR (451, C1) route

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

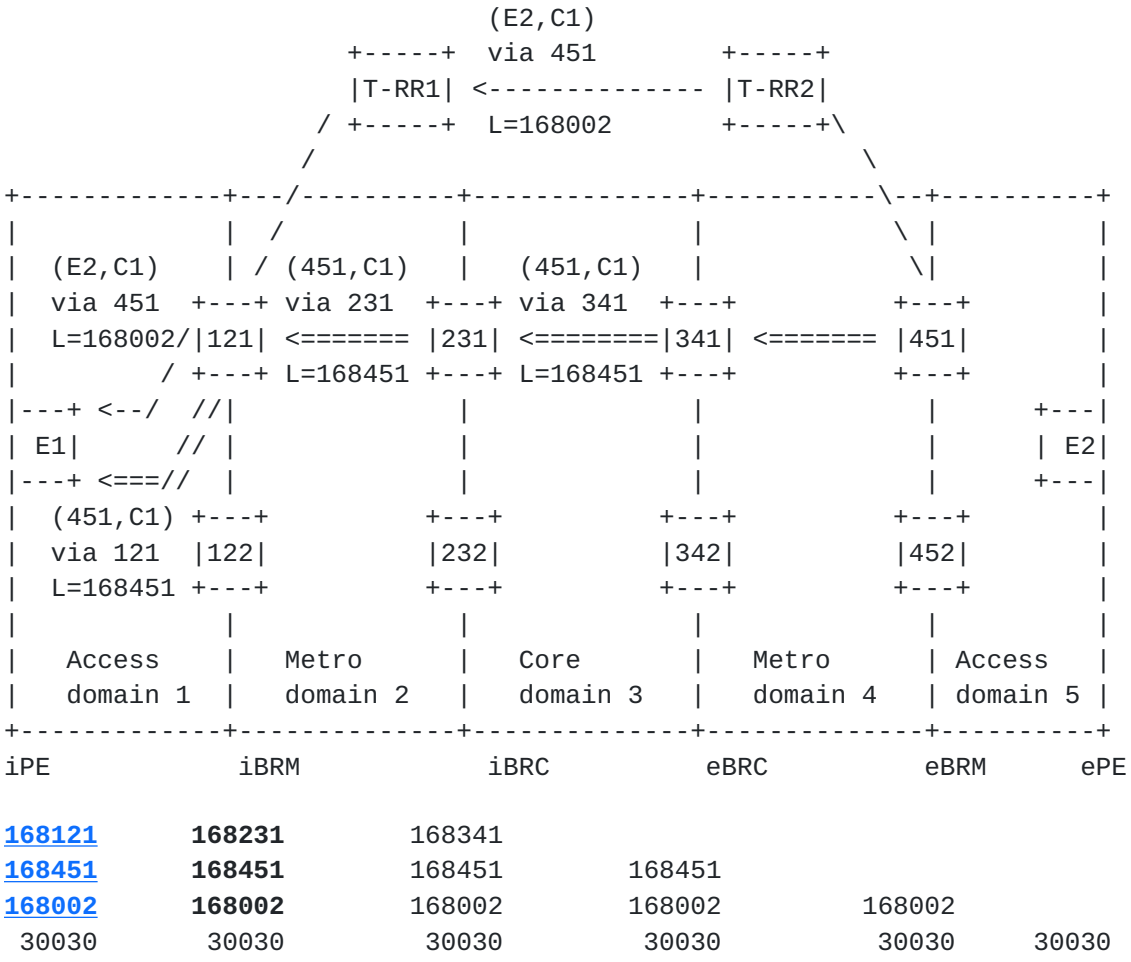


Figure 5: Heirarchical BGP transport CAR, NHU at iBR

1. Nodes 341, 231 and 121 receive and resolve BGP CAR route (451, C1) the same as in the previous model

2. Node 121 allocates local label and programs swap entry in forwarding for (451, C1)

3. 451 advertises BGP CAR route (E2, C1) to Transport RR T-RR2, which reflects it to T-RR1, which reflects it to 121

4. Node 121 advertises (E2, C1) to E1 with next hop as 451 i.e. next-hop unchanged

5. 121 also advertises (451, C1) to E1 with next hop self (121) and label 168451

6. E1 resolves BGP CAR route (451, C1) via 121 on color-aware path (121, C1)

    1. Color-aware path (121, C1) is FA128 path to 121 (label 168121)

7. E1 receives BGP CAR route (E2, C1) via 451 with label 168002

    1. Let's assume E1 selects that path

8. E1 resolves BGP CAR route (E2, C1) via 451 on color-aware path (451, C1)

    1. Color-aware path (451, C1) is BGP CAR path to 451 (label 168451)

9. E1's imposition color-aware label-stack for V/v is thus

    1. 30030 <=> V/v

    2. 168002 <=> (E2, C1)

    3. 168451 <=> (451, C1)

    4. 168121 <=> (121, C1)

10. Nodes 121, 231 and 341 perform swap operation on 168451 bound to (451, C1)

11. 451 performs swap operation on 168002 bound to color-aware path (E2, C1)

6.3. Scale Analysis

The following two tables summarize the control-plane and dataplane scale of these three models:

	E1	121	231
FLAT	(E2,C) via (121,C)	(E2,C) via (231,C)	(E2,C) via (341,C)
H.NHS	(E2,C) via (121,C)	(E2,C) via (451,C) (451,C) via (231,C)	(451,C) via (341,C)
H.NHU	(E2,C) via (451,C) (451,C) via (121,C)	(451,C) via (231,C)	(451,C) via (341,C)

	E1	121	231
FLAT	V -> 30030 168002 168121	168002 -> 168002 168231	168002 -> 168002 168341
H.NHS	V -> 30030 168002 168121	168002 -> 168002 168451 168231	168451 -> 168451 168341
H.NHU	V -> 30030 168002 168451 168121	168451 -> 168451 168231	168451 -> 168451 168341

- The flat model is the simplest design, with a single BGP transport level. It results in the minimum label/SID stack at each BGP hop. However, it significantly increases the scale impact on the core BRs (e.g. 341), whose FIB capacity and even MPLS label space may be exceeded.
  - 341's dataplane scales with (E2,C) where there may be 300k E's and 5 C's hence 1.5M entries > 1M MPLS dataplane
- The hierarchical models avoid the need for core BRs to learn routes and install label forwarding entries for (E, C) routes.
  - Whether NH self or unchanged at 121, 341's dataplane scales with (451,C) where there may be thousands of 451's and 5 C's hence well under the 1M MPLS dataplane
- The next-hop-self option at ingress BRM (e.g. 121) hides the hierarchical design from the ingress PE, keeping its outgoing label programming as simple as the flat model. However, the ingress BRM requires an additional BGP transport level recursion, which coupled with load-balancing adds dataplane complexity. It

needs to support a swap and push operation. It also needs to install label forwarding entries for the egress PEs that are of interest to its local ingress PEs.

- With the next-hop-unchanged option at ingress BRM (e.g. 121), only an ingress PE needs to learn and install output label entries for egress (E, C) routes. The ingress BRM only installs label forwarding entries for the egress ABR (e.g. 451). However, the ingress PE needs an additional BGP transport level recursion and pushes a BGP VPN label and two BGP transport labels. It may also need to handle load-balancing for the egress ABRs. This is the most complex dataplane option for the ingress PE.

6.4. Scaling Benefits of the (E, C) BGP Subscription and Filtering

The (E, C) subscription scheme from Section 5 provides the following scaling benefits for the models in Section 6.2

- An ingress PE (E1) only learns (E, C) routes that it needs to install into data plane for service route automated steering

- An ingress BRM (121) only learns (E, C) routes that it needs to install into data plane (for Next-Hop-Self), or that it needs to distribute towards it's ingress PEs (inline RR with Next-Hop-Unchanged)

- An ingress BRM or a transport RR only needs to distribute the necessary subset of (E, C) routes to each client (subscriber); this minimizes their processing load for generating updates

- As a result, withdrawal of (E, C) routes when a remote node fails (E2), may also be faster, aiding better convergence

6.5. Anycast SID

This section describes how Anycast SID complements and improves the scaling designs above.

6.5.1. Anycast SID for transit inter-domain nodes

- Redundant BRs (e.g. two egress BRMs, 451 and 452) advertise BGP CAR routes for a local PE (e.g., E2) with the same SID (based on label-index). Such egress BRMs may be assigned a common Anycast SID, so that the BGP next-hops for these routes will also resolve via a color-aware path to the Anycast SID.

- The use of Anycast SID naturally provides fast local convergence upon failure of an egress BRM node. In addition, it decreases the

recursive resolution and load-balancing complexity at an ingress BRM or PE in the hierarchical designs above.

6.5.2. Anycast SID for transport color endpoints (e.g., PEs)

The common Anycast SID technique may also be used for a redundant pair of PEs that share an identical set of service (VPN) attachments.

- For example, assume a node E2' paired with E2 above. Both PEs should be configured with the same static label/SID for the services (e.g., per-VRF VPN label/SID), and will advertise associated service routes with the Anycast IP as BGP next-hop.

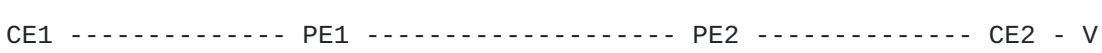
- This design provides a convergence and recursive resolution benefit on an ingress PE or ABR similar to the egress ABR case in the previous section. But its applicability is limited to cases where the constraints above can be met.

7. Routing Convergence

This section will analyze routing convergence.

8. VPN CAR

This section illustrates the extension of BGP CAR to address the VPN CAR requirement stated in Section 3.2 of [dskc-bess-bgp-car-problem-statement].



- BGP CAR is enabled between CE1-PE1 and PE2-CE2

- BGP VPN CAR is enabled between PE1 and PE2

- Provider publishes intent 'low-delay' is mapped to color CP on its inbound peering links

- Within its infrastructure, Provider maps intent 'low-delay' to color CPT

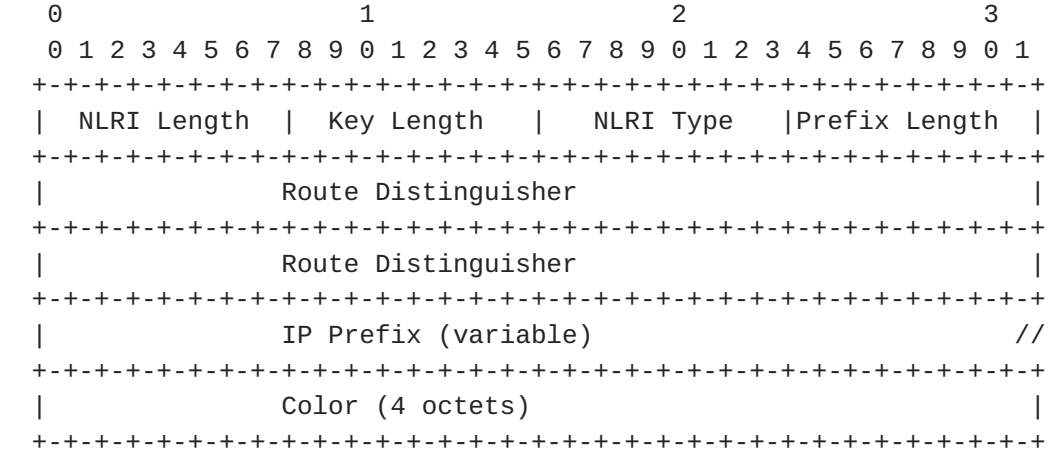
- On CE1 and CE2, intent 'low-delay' is mapped to CC

(V, CC) is a Color-Aware route originated by CE2

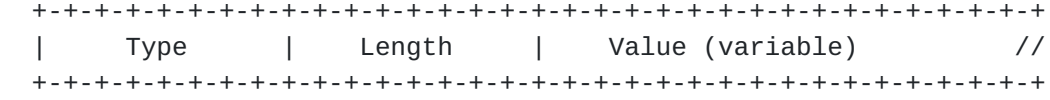
1. CE2 sends to PE2 : [(V, CC), Label L1] via CE2 with LCM (CP)
2. PE2 installs in VRF A: [(V, CC), L1] via CE2 which resolves on (CE2, CP)
/ connected OIF
2.a. PE2 allocates VPN Label L2 and programs swap entry for (V, CC)
3. PE2 sends to PE1 : [(RD, V, CC), L2] via PE2 with regular Color Extended
Community (CPT)
4. PE1 installs in VRF A: [(V, CC), L2] via (PE2, CPT) steered on (PE2, CPT)
4.a. PE1 allocates Label L3 and programs swap entry for (V, CC)
5. PE1 sends to CE1 : [(V, CC), L3] via PE1 without any LCM
6. CE1 installs : [(V, CC), L3] via PE1 which resolves on (PE1, CC)
/ connected OIF
6.a. Label L3 is installed as the imposition label for (V, CC)

VPN CAR distribution for (RD, V, CC) requires a new SAFI that follows same VPN semantics as defined in [RFC4364], the difference being that the advertised routes carry CAR NLRI defined in Section 2.9.2 of this document.

VPN CAR NLRI with RD has the format shown below



Followed by optional TLVs encoded as below:



where:

Route Distinguisher: 8 octet field encoded according to [RFC4364]

9. IANA Considerations

IANA has assigned SAFI value 83 (BGP CAR) and SAFI value 84 (BGP VPN CAR) from the "SAFI Values" sub-registry under the "Subsequent Address Family Identifiers (SAFI) Parameters" registry with this document as a reference.

9.1. 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]).

9.2. 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]).

9.3. 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.

9.4. 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.

10. Manageability Considerations

Color assignments in a multi-domain network operating under a common or cooperating administrative control (i.e., color domain) should be managed similar to transport layer IP addresses, and ensure a unique and non-conflicting color allocation across the different network domains in that color domain.

If networks under different administrative control establish a shared transport service between them, where the same transport IP address is co-ordinated and shared across the two networks, then the color assignments associated with that IP address should also be co-ordinated to avoid any conflicts in either network.

11. Acknowledgements

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

12. References

12.1. Normative References

[I-D.ietf-bess-srv6-services]
Dawra, G., Talaulikar, K., Raszuk, R., Decraene, B., Zhuang, S., and J. Rabadan, "SRv6 BGP based Overlay Services", draft-ietf-bess-srv6-services-15 (work in progress), March 2022.

[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. V. D., Sangli, S. R., and J. Scudder, "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-22 (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-20 (work in progress), May 2022.

[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-22 (work in progress), March 2022.

[I-D.ietf-spring-srv6-network-programming]
Filsfils, C., Garvia, P. C., Leddy, J., Voyer, D., Matsushima, S., and Z. Li, "Segment Routing over IPv6 (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>.

12.2. 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>.

[RFC5462] Andersson, L. and R. Asati, "Multiprotocol Label Switching (MPLS) Label Stack Entry: "EXP" Field Renamed to "Traffic Class" Field", RFC 5462, DOI 10.17487/RFC5462, February 2009, <https://www.rfc-editor.org/info/rfc5462>.

[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>.

Appendix A. Illustrations of Service Steering

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

The examples use MPLS/SR for the transport data plane. Scenarios specific to other encapsulations will be added in subsequent versions.

A.1. E2E BGP transport CAR intent realized using IGP FlexAlgo

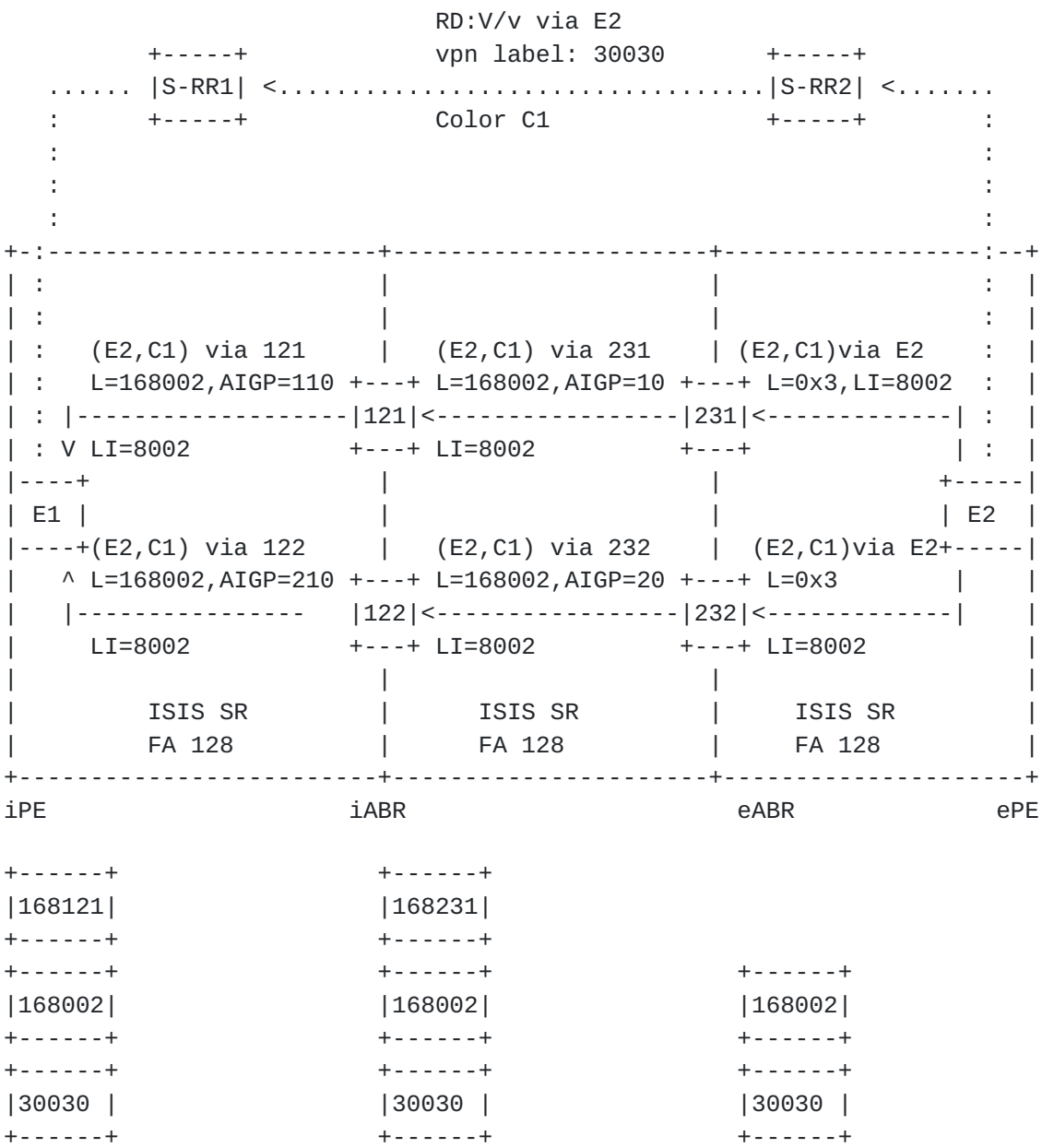


Figure 6: BGP FA Aware transport CAR path

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. When a RR is used in the domain, ADD-PATH is enabled to advertise multiple available paths.

  * 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)

- 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.

A.2. E2E BGP transport CAR intent realized using SR Policy

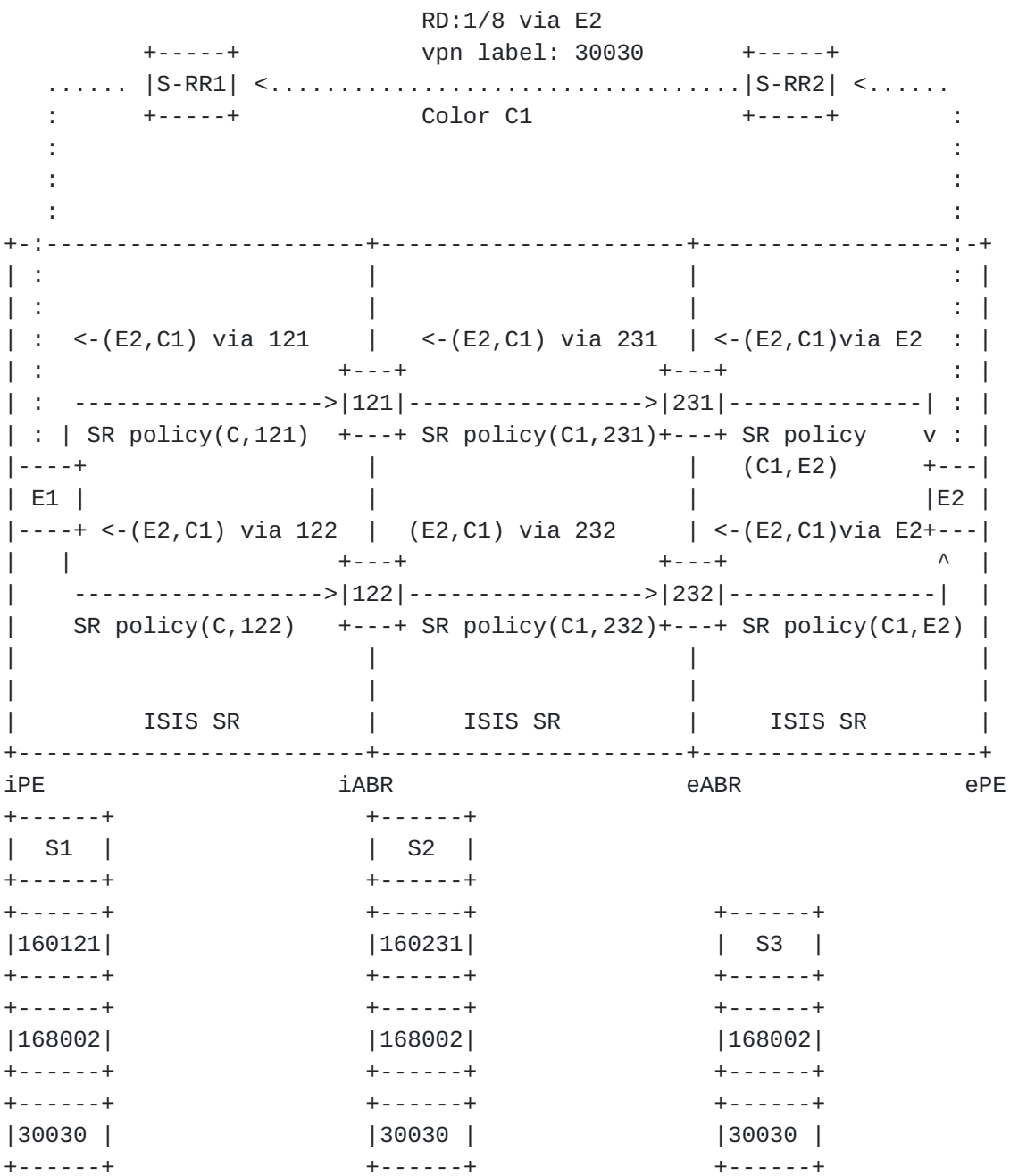


Figure 7: BGP SR policy Aware transport CAR path

Use case: Provide end to end intent for service flows

- With reference to the topology above:
  - SR Policy provide intra domain intent. Below are example SID lists of SR policies in each domain corresponding to label stack in Figure 7

+ SR policy (C,121) segments <S1, 121>

+ SR policy (C,231) segments <S2, 231>

+ SR policy (C,E2) segments <S3, E2>

* 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. When a RR is used in the domain, ADD-PATH is enabled to advertise multiple available paths.

* 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).

o 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.

A.3. BGP transport CAR intent realized in a section of the network

A.3.1. Provide intent for service flows only in core domain running ISIS FlexAlgo

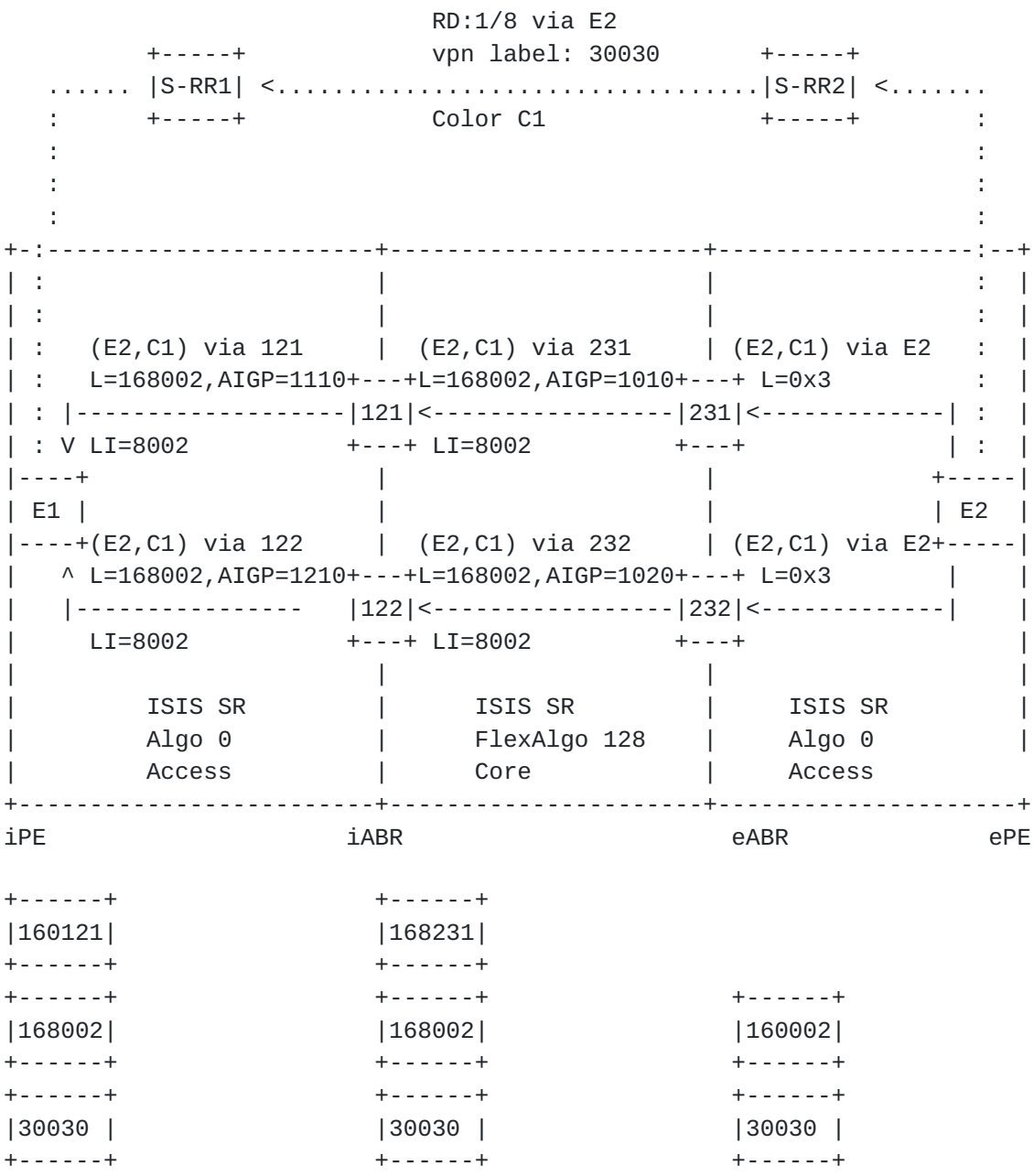


Figure 8: BGP Hybrid FlexAlgo Aware transport CAR path

- 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 via 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. When a RR is used in the domain, ADD-PATH is enabled to advertise multiple available paths.

* 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 with an 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.

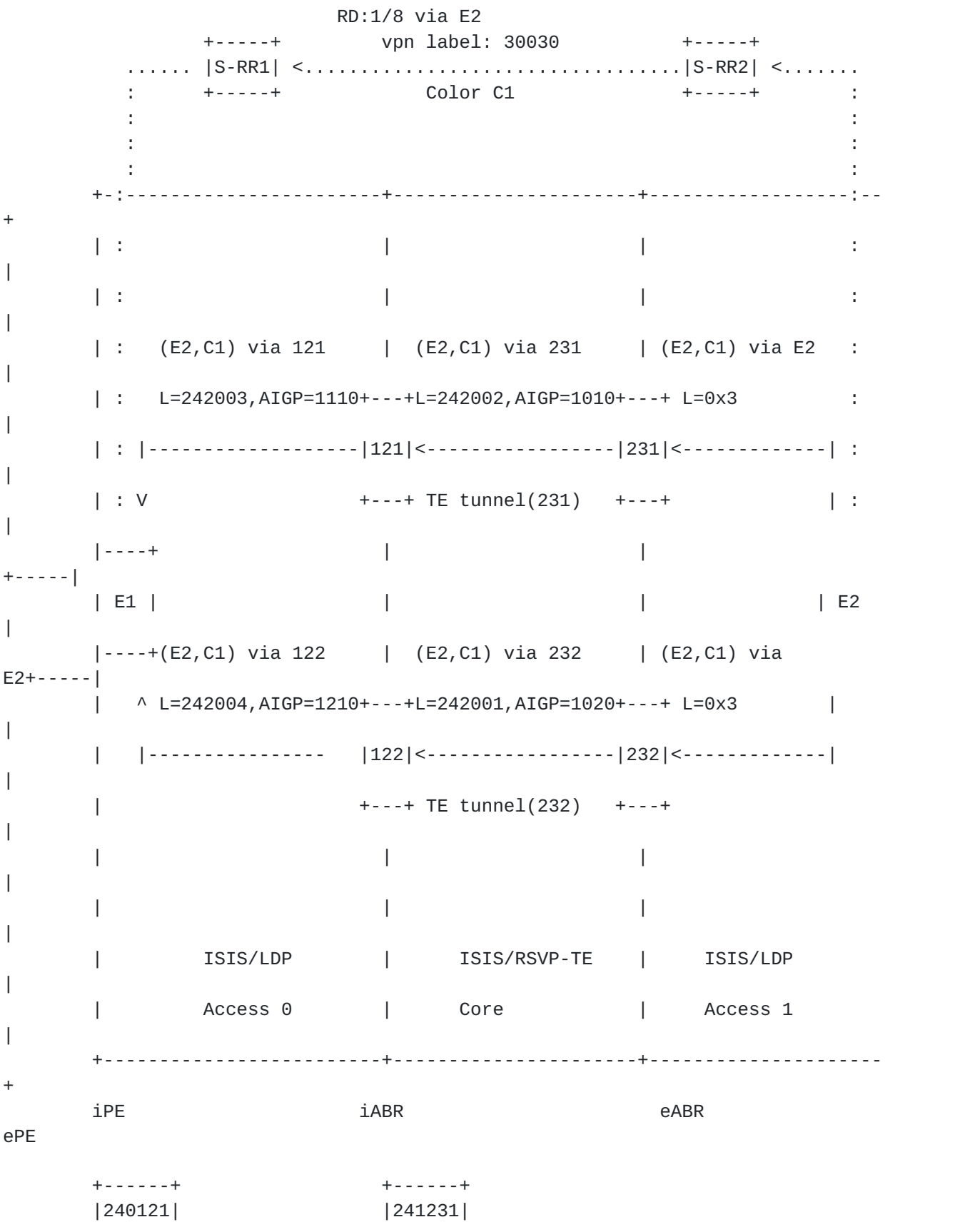
* Ingress PE E1 learns CAR route (E2, C1). It maps intent C1 to resolve CAR route next-hop over IGP base algo 0. It steers colored VPN route RD:V/v via (E2, C1)

o Important:

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

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

A.3.2. **Provide intent for service flows only in core domain over TE** tunnel mesh


RD:1/8 via E2
vpn label: 30030
S-RR1
S-RR2
Color C1
(E2,C1) via 121
L=242003,AIGP=1110
(E2,C1) via 231
L=242002,AIGP=1010
(E2,C1) via E2
L=0x3
121
231
TE tunnel(231)
E1
(E2,C1) via 122
L=242004,AIGP=1210
(E2,C1) via 232
L=242001,AIGP=1020
L=0x3
122
232
TE tunnel(232)
E2
(E2,C1) via E2
ISIS/LDP
Access 0
ISIS/RSVP-TE
Core
ISIS/LDP
Access 1
iPE
iABR
eABR
ePE
240121
241231

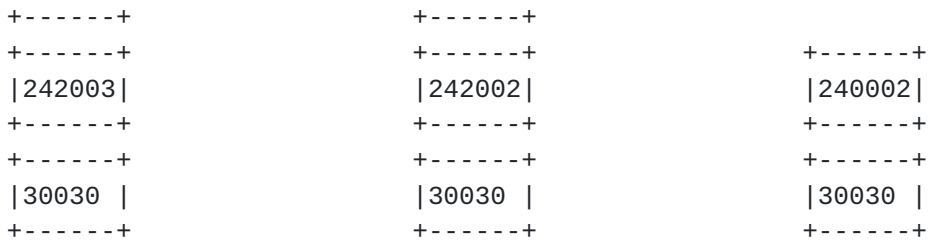


Figure 9: BGP CAR over TE tunnel mesh in core network

- With reference to the topology above:
  - RSVP-TE MPLS tunnel mesh is configured only in core (e.g. WAN network). Access only has ISIS/LDP. (Figure does not show all TE tunnels)
  - Egress PE E2 advertises a VPN route RD:V/v colored with (color extended community) C1 to steer traffic via BGP transport CAR

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

* BGP CAR route (E2, C1) with next-hops and labels as shown above is advertised through border routers in each domain. When a RR is used in the domain, ADD-PATH is enabled to advertise multiple available paths.

* Local policy on 231 and 232 maps intent C1 to resolve CAR route next-hop over best effort LDP LSP in access domain 1. BGP CAR label swap entry is installed that goes over LDP LSP to next-hop. AIGP metric is updated to reflect best effort metric to next- hop with an additional penalty.

* Local policy on 121 and 122 maps intent C1 to resolve CAR route next-hop in Core domain over TE tunnels. BGP CAR label swap entry is installed that goes over a TE tunnel to next-hop providing intent in Core domain. AIGP metric is updated to reflect TE tunnel metric.

* Ingress PE E1 learns CAR route (E2, C1). It maps intent C1 to resolve CAR route next-hop over best effort LDP LSP in Access domain 0. It steers colored VPN route RD:V/v via (E2, C1).

o Important:

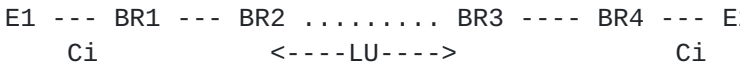
* TE tunnel LSP provides intent in Core domain.

* Dynamic BGP CAR label carries intent from PEs which is realized in core domain by resolution via TE tunnel.

A.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

A.5. Resource Avoidance using BGP CAR and IGP Flex-Algo

This example illustrates a case of resource avoidance within a domain for a multi-domain color-aware path.

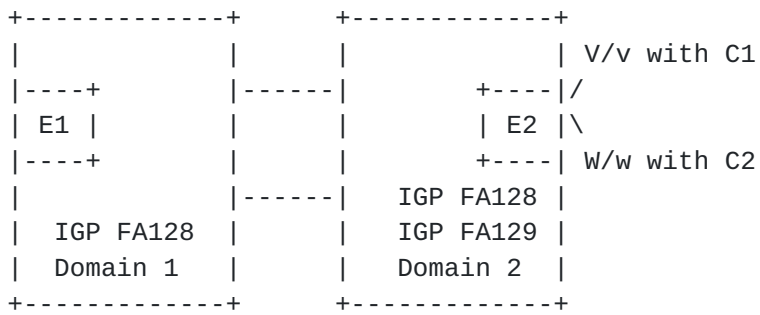


Figure 10: BGP CAR resolution over IGP FLex-Algo for resource avoidance in a domain

- C1 and C2 represent two unique intents in multi-domain network

- C1 is mapped to "minimize IGP metric"
- C2 is mapped to "minimize IGP metric and avoid resource R"

- Resource R represents link(s) or node(s) to be avoided
- Flex-Algo FA128 in Domain 2 is mapped to "minimize IGP metric" and hence to C1
- Flex-Algo FA129 in Domain 2 is mapped to "minimize IGP metric and avoid resource R" and hence to C2
- Flex-Algo FA128 in Domain 1 is mapped to "minimize IGP metric"
  - There is no resource R to be avoided in Domain 1, hence both C1 and C2 are mapped to FA128
- E1 receives two service routes from E2:
  - V/v with BGP Color Extended-Community C1
  - W/w with BGP Color Extended-Community C2
- E1 has the following color-aware paths:
  - (E2, C1) provided by BGP CAR with the following per-domain resolution:
    - Domain1: over IGP FA128
    - Domain2: over IGP FA128
  - (E2, C2) provided by BGP CAR with the following per-domain resolution:
    - Domain1: over IGP FA128
    - Domain2: over IGP FA129, avoiding resource R
- E1 automatically steers the received service routes as follows:
  - V/v via (E2, C1) provided by BGP CAR
  - W/w via (E2, C2) provided by BGP CAR

Observations:

- C1 and C2 are realized over a common intra-domain intent (FA128) in one domain and distinct intents in another domain as required

- 32-bit Color space provides flexibility in defining a large number of intents in a multi-domain network. They may be efficiently realized by mapping to a smaller number of intra-domain intents in different domains.

A.6. Per-Flow Steering over CAR routes

This section provides an example of ingress PE per-flow steering as defined in section 8.6 of [I-D.ietf-spring-segment-routing-policy] onto BGP CAR routes.

With reference to the Figure 6

- Ingress PE E1 learns best effort BGP LU route E2

- Ingress PE E1 learns CAR route (E2, C1), C1 is mapped to "low delay"

- Ingress PE E1 learns CAR route (E2, C2), C2 is mapped to "low delay and avoid resource R"

- Ingress PE E1 is configured to instantiate an array of paths to E2 where the entry 0 is the BGP LU path to N, color C1 is the first entry and color C2 is the second entry. The index into the array is called a Forwarding Class (FC). The index can have values 0 to 7, especially when derived from the MPLS TC bits [RFC5462]

- E1 is configured to match flows in its ingress interfaces (upon any field such as Ethernet destination/source/VLAN/TOS or IP destination/source/DSCP or transport ports etc.) and color them with an internal per-packet FC variable (0, 1 or 2 in this example).

- This array is presented as composite candidate path of SR policy (E2, C100) and acts as a container for grouping constituent paths of different colors/best effort. This representation provide automated steering for services colored with Color Extended Community C100 via paths of different colors. Note that color extended community C100 is used as indirection to the composite policy configured on ingress PE.

- Egress PE E2 advertises a VPN route RD:V/v with Color Extended community C100 to steer traffic via composite SR policy (E2, C100) i.e. FC array of paths.

E1 receives three packets K, K1, and K2 on its incoming interface. These three packets matches on VPN route which recurses on E2. E1 colors these 3 packets respectively with forwarding-class 0, 1, and 2.

As a result

- E1 forwards K along the best effort path to E2 (i.e., for MPLS data plane, it pushes the best effort label of E2).
- E1 forwards K1 along the (E2, C1) BGP CAR route
- E1 forwards K2 along the (E2, C2) BGP CAR route

A.7. Advertising BGP CAR routes for shared IP addresses

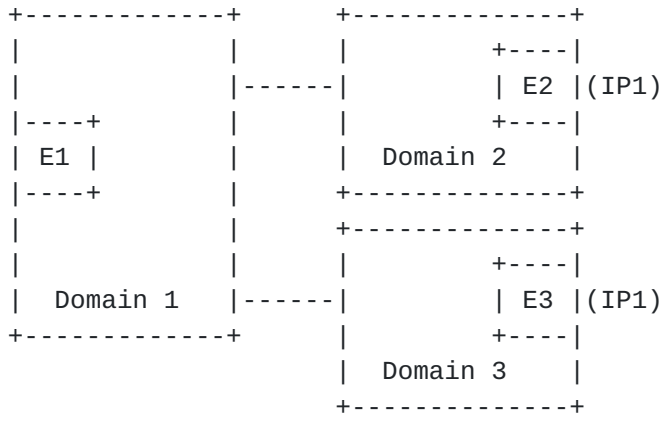


Figure 11: BGP CAR advertisements for shared IP addresses

This example describes a case where the same transport IP address is originated from multiple nodes in different domains.

One use of this scenario is an Anycast transport service, where packet encapsulation may terminate on any one among a set of nodes. All the nodes are capable of forwarding the inner payload, for instance via an IP lookup in the global table.

A couple of variations of the use-case are described in the example below.

One node is shown in each domain, but there will be multiple nodes in practice.

Example-1: Anycast with forwarding to nearest

- Both E2 and E3 advertise Anycast IP (IP1, C1) with same label L1
- An ingress PE E1 receives by default the best path(s) propagated through BGP hops across the network.
- The paths to (IP1, C1) from E2 and E3 may get merged at a common node along the path to E1.
- Traffic for colored service routes steered at E1 via (IP1, C1) is forwarded to either E2 or E3 (or both) as determined by the network (nodes in the path).

Example-2: Anycast with egress domain visibility at ingress PE

- E2 advertises (IP1, C1) and E3 advertises (IP1, C2) CAR routes for the Anycast IP IP1.
- An ingress PE E1 receives the best path(s) propagated through BGP hops across the network for both (IP1, C1) and (IP1, C2).
- The two CAR routes do not get merged at any intermediate node, providing E1 control over path selection and load-balancing of traffic across these routes.
- Traffic for colored service routes steered at E1 is forwarded to either E2 or E3 (or load-balanced across both) as determined by E1.

Appendix B. Color Mapping Illustrations

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 and provide clarity into the usage of the color related protocol constructs.

B.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.
- Appendix A.1 is an example of this case.

B.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 color-awareness 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 variation of this scenario is where a transit domain does not support any color. Appendix A.3 describes an example of this case.

B.3. Multiple color domains

When the routes are distributed between domains with different color-to-intent mapping schemes, both N:N and N:M cases 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 color domains

The reference topology above is used to elaborate on the design described in Section 2.8

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

- The original intent in the 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
  - A BR in D2 may receive (E, C1) from multiple D1 BRs which provide equal cost or primary/backup paths
- 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 service steering.
- A colored service route V/v originated in color 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, the same process repeats.

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

Dirk Steinberg
Lapishills Consulting Limited
Germany

Email: dirk@lapishills.com

Luay Jalil
Verizon
USA

Email: luay.jalil@verizon.com

Yuanchao Su
Alibaba, Inc

Email: yitai.syc@alibaba-inc.com

Bruno Decraene
Orange
France

Email: bruno.decraene@orange.com

Jim Guichard
Futurewei
USA

Email: james.n.guichard@futurewei.com

Ketan Talaulikar
Arrcus, Inc
India

Email: ketant.ietf@gmail.com

Keyur Patel
Arrcus, Inc
USA

Email: keyur@arrcus.com

Haibo Wang
Huawei Technologies
China

Email: rainsword.wang@huawei.com

Jim Uttaro
ATT
USA

Email: ju1738@att.com