

IDR WorkGroup
Internet-Draft
Intended status: Experimental
Expires: 17 May 2024

D. Rao, Ed.
S. Agrawal, Ed.
Cisco Systems
Co-authors
[Section 11](#)
14 November 2023

BGP Color-Aware Routing (CAR)
draft-ietf-idr-bgp-car-04

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 17 May 2024.

Copyright Notice

Copyright (c) 2023 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to [BCP 78](#) and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Revised BSD License text as described in Section 4.e of the [Trust Legal Provisions](#) and are provided without warranty as described in the Revised BSD License.

Table of Contents

1.	Introduction	4
1.1.	Terminology	5
1.2.	Illustration	7
1.3.	Requirements Language	9
2.	BGP CAR SAFI	9
2.1.	Data Model	9
2.2.	Extensible Encoding	10
2.3.	BGP CAR Route Origination	11
2.4.	BGP CAR Route Validation	11
2.5.	BGP CAR Route Resolution	11
2.6.	AIGP Metric Computation	12
2.7.	Native MultiPath Capability	13
2.8.	BGP CAR Signaling through different Color Domains	13
2.9.	Format and Encoding	15
2.9.1.	BGP CAR SAFI NLRI Format	15
2.9.2.	Color-Aware Route NLRI Type	16
2.9.3.	IP Prefix NLRI Type	21
2.9.4.	Local-Color-Mapping (LCM) Extended-Community	23
2.10.	LCM and BGP Color Extended-Community usage	24
2.11.	Error Handling	24
3.	Service Route Automated Steering on Color-Aware Path	26
4.	Filtering	27
4.1.	(E, C) Subscription and Filtering	27
5.	Scaling	28
5.1.	Ultra-Scale Reference Topology	28
5.2.	Deployment model	30
5.2.1.	Flat	30
5.2.2.	Hierarchical Design with Next-Hop-Self at Ingress Domain BR	31
5.2.3.	Hierarchical Design with Next-Hop-Unchanged at Ingress Domain BR	33
5.3.	Scale Analysis	34
5.4.	Scaling Benefits of the (E, C) BGP Subscription and Filtering	36
5.5.	Anycast SID	36
5.5.1.	Anycast SID for Transit Inter-domain Nodes	37
5.5.2.	Anycast SID for Transport Color Endpoints (e.g., PEs)	37
6.	Routing Convergence	37
7.	CAR SRV6	37
7.1.	Overview	37
7.1.1.	Routed Service SID	38
7.1.2.	Non-routed Service SID	38
7.2.	Deployment Options For CAR SRV6 Locator Reachability Distribution and Forwarding	40
7.2.1.	Hop by Hop IPv6 Forwarding for BGP SRV6 Prefixes	40

7.2.2.	Encapsulation between BRs for BGP SRv6 Prefixes . . .	40
7.3.	Operational Benefits of using CAR SAFI for SRv6 Locator Prefix Distribution	41
8.	CAR IP Prefix Route	42
9.	VPN CAR	43
9.1.	VPN CAR Type-1 NLRI	44
9.2.	VPN CAR IP Prefix NLRI Type	45
10.	IANA Considerations	46
10.1.	BGP CAR NLRI Types Registry	46
10.2.	BGP CAR NLRI TLV Registry	47
10.3.	Guidance for Designated Experts	47
10.4.	BGP Extended-Community Registry	47
11.	Manageability and Operational Considerations	48
12.	Security Considerations	48
13.	Co-authors	49
14.	Contributors	50
15.	Acknowledgements	51
16.	References	51
16.1.	Normative References	51
16.2.	Informative References	53
Appendix A.	Illustrations of Service Steering	55
A.1.	E2E BGP transport CAR intent realized using IGP FlexAlgo	55
A.2.	E2E BGP transport CAR intent realized using SR Policy	56
A.3.	BGP transport CAR intent realized in a section of the network	58
A.3.1.	Provide intent for service flows only in core domain running ISIS FlexAlgo	58
A.3.2.	Provide intent for service flows only in core domain over TE tunnel mesh	60
A.4.	Transit network domains that do not support CAR	62
A.5.	Resource Avoidance using BGP CAR and IGP Flex-Algo	63
A.6.	Per-Flow Steering over CAR routes	65
A.7.	Advertising BGP CAR routes for shared IP addresses	66
Appendix B.	Color Mapping Illustrations	68
B.1.	Single color domain containing network domains with N:N color distribution	68
B.2.	Single color domain containing network domains with N:M color distribution	68
B.3.	Multiple color domains	72
Appendix C.	CAR SRv6 Illustrations	73
C.1.	BGP CAR SRv6 locator reachability hop by hop distribution	73
C.2.	BGP CAR SRv6 locator reachability distribution with encapsulation	76
C.3.	BGP CAR (E, C) route distribution	78
Appendix D.	CAR SAFI NLRI update packing efficiency calculation	81

Authors' Addresses [85](#)

1. Introduction

BGP Color-Aware Routing (CAR) is a BGP based routing solution to establish end-to-end intent-aware paths across a multi-domain service provider transport network. BGP CAR distributes distinct routes to a destination network endpoint such as a PE router, for different intents or colors. Color is a 32-bit numerical value associated with a network intent (low-cost, low-delay, avoid some resources etc.) as defined in [[RFC9256](#)].

BGP CAR fulfills the transport and VPN problem statement and requirements described in [[I-D.hr-spring-intentaware-routing-using-color](#)].

For this purpose, this document specifies two new BGP SAFIs, called BGP CAR (SAFI 83) and VPN CAR (SAFI 84) that carry infrastructure routes to set up the transport paths. CAR SAFIs apply to IPv4 Unicast and IPv6 Unicast AFIs (AFI 1 and AFI 2).

BGP CAR SAFI is enabled on transport devices in a provider network (underlay) to set up color-aware transport/infrastructure paths across the provider network. The transport network may comprise of multiple BGP ASNs or multiple IGP domains (e.g., [[I-D.ietf-mpls-seamless-mpls](#)]). BGP CAR can also be enabled within a VRF on a PE router towards a peering CE router, and within a customer network. VPN CAR SAFI enables the distribution of intent-aware routes from different customers received on a PE across the provider network, maintaining the separation of the customer address spaces that may overlap. BGP CAR thus enables intent-aware routed transport paths to be set up across a multi-domain network that can span both customer and provider network domains.

The document also defines two CAR route types for this purpose.

The CAR Type-1 NLRI enables the generation and distribution of multiple color-aware routes to the same destination IP prefix (E, C) for different colors. This is applicable to situations where a transport node such as a PE has a common IP address (such as a loopback); this address is used as the BGP next hop for service routes and as the transport endpoint for the data plane path; and where multiple routes are needed for this same prefix or address to set up a unique path for each intent. One example is setting up MPLS/SR-MPLS LSPs to an egress PE, one per intent.

The CAR Type-2 NLRI enables the distribution of multiple color-aware routes to a transport node for the case where the operator specifies a unique network IP address block for a given intent, and the transport node gets assigned a unique IP prefix or address for each intent. An example use-case is SRv6 per-intent locators.

1.1. Terminology

Intent (in routing)	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). More specific concept w.r.t. routing beyond best-effort, compared to intent as declarative abstraction in [RFC9315] .
Color	A 32-bit numerical value associated with an intent (e.g. low-cost , low-delay, avoid some resources etc.) as defined in [RFC9256] .
Colored Service Route	An egress PE (e.g. E2) colors its BGP service (e.g. 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 [RFC9012] (used as per [RFC9256]), or in locator part of SRv6 Service SID [RFC9252] .
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 [RFC9256] , IGP Flex-Algo [RFC9350] , 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	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 (for example,

Color-aware path	IGP Flex-Algo first then SR Policy then BGP CAR.	
+-----+		
Color Domain	A set of nodes which share the same Color-to-Intent mapping, typically under single administration. This set can be organized in one or several IGP instances or BGP ASNs/domains. Color re-mapping may happen at color domain boundaries.	
+-----+		
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 [RFC9256], 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 maps to BGP CAR path (or other Color-Aware Routed Paths: e.g. SR Policy). On non availability of Color-Aware path, local policy may map to traditional routing/TE path (e.g. BGP LU, RSVP-TE, IGP/LDP).	
+-----+		
	Intra-Domain Resolution: BGP CAR route maps to intra-domain color aware path (e.g. SR Policy, IGP Flex-Algo, BGP CAR) or traditional routing/TE path (e.g. RSVP-TE, IGP/LDP, BGP-LU).	
+-----+		

Table 1

Abbreviations:

- * BR: An inter-domain Border Router, either for an IGP Area (ABR) or a BGP Autonomous System (ASBR).

- * RR: BGP Route Reflector.
- * AFI/SAFI: BGP Address-Family/Sub-Address-Family.
- * BGP-LU: BGP Labeled Unicast SAFI [[RFC8277](#)].
- * BGP-IP: BGP IPv4/IPv6 Unicast AFI/SAFIs [[RFC4271](#)], [[RFC4760](#)].

1.2. Illustration

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

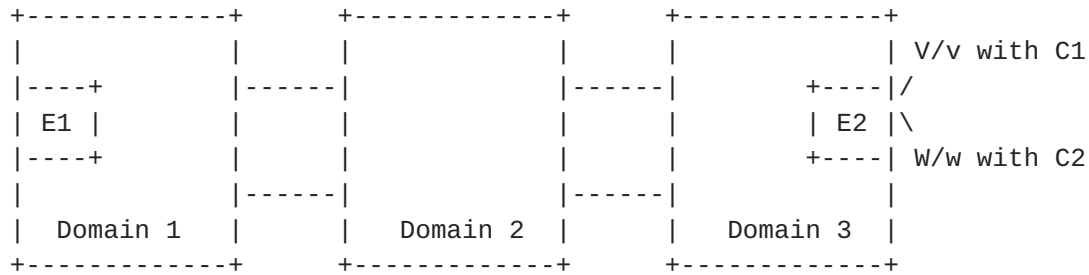


Figure 1

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 ([Section 5](#)). 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)
 - natural extension of BGP IP/LU data model (E)
 - 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.
- * Native support for multiple transport encapsulations (e.g., MPLS, SR, SRv6, IP)

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: Falls into two categories, to accommodate the use-cases described in the introduction:
 - Type-1: Key is IP Prefix and Color (E, C). Color in NLRI key distinguishes a color-aware route for a common IP prefix, one per intent. Color also indicates the intent associated with the route.

- Type-2: Key is IP Prefix (E). The unique IP prefix assigned for an intent (i.e IP Prefix == Intent or Color) distinguishes the color-aware route. Color is not necessary in NLRI key as a distinguisher.
- * 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 to represent the intent associated with the CAR route:
 - when a CAR route propagates between different color domains.
 - when a CAR route has a unique IP prefix for an intent.
- * BGP Color Extended-Community (Color-EC): Optional 32-bit Color value used to represent the intent associated with the CAR BGP next hop, as per [[RFC9256](#)].

The sections below describe the data model in detail. The sections that describe the protocol processing for CAR SAFI generally apply consistently to both route types (for instance, any operation based on color). The examples use (E, C) for simplicity.

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: e.g., SR Policy, IGP Flex-Algo, RSVP-TE, 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).

A path that is not valid MUST NOT be considered for BGP best path selection.

2.5. BGP CAR Route Resolution

A BGP color-aware route (E2, C1) from N is automatically resolved over a color-aware route (N, C1) by default. The color-aware route (N, C1) is provided by color aware mechanisms such as IGP Flex-Algo, SR policy or recursively by BGP CAR. When multiple producers of (N, C1) are available, the default preference is: IGP Flex-Algo, SR Policy, BGP CAR.

Local policy SHOULD provide additional control:

- * 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 a different color 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 to fallback via C2

- * The resolution of (N, C1) may be egress driven. In an SRv6 domain, node N selects and advertises an SRv6 SID from a locator for intent C1, with a BGP CAR route. In such a case, the ingress node resolves the SRv6 SID over a route for the intent aware locator of N or its summary provided by SRv6 Flex Algo or BGP CAR Type-2 route itself (e.g., [Appendix C.2](#)).
- * Resolution may be mapped to traditional mechanisms that are unaware of color or that provide best effort, such as RSVP-TE, IGP/LDP, BGP LU/IP (e.g., [Appendix A.3.2](#)).

Route resolution via a different color C2 can be automated by attaching BGP Color Extended-Community C2 to CAR route (E2, C1), leveraging Automated steering as described in [Section 8.4](#) of Segment Routing Policy Architecture [[RFC9256](#)] for BGP CAR routes. This mechanism is illustrated in section B.2. This mechanism SHOULD be supported.

Local policy takes precedence over color based automated resolution. For a CAR route, Color-EC color takes precedence over route NLRI color. When LCM-EC is present, Color-EC color takes precedence over LCM-EC color.

The color-aware route (N, C1) may be provided by BGP CAR itself in a hierarchical transport routing design. In such cases, based on the procedures described above, recursive resolution may occur over the same or different CAR route type. [Section 7.1.2](#) describes a scenario where CAR Type-1 route resolves over CAR Type-2.

CAR Type-2 route is allowed to be without color for best effort. In this case, resolution is based on BGP next hop N, or when present, a best-effort SRv6 SID advertised by node N.

A BGP CAR route may recursively resolve over a BGP route carrying Tunnel Encapsulation attribute (TEA). Procedures of [section 8 of RFC9012](#) apply in presence of BGP Color-EC in the CAR route. They are extended to use LCM-EC and Color in CAR route NLRI as per above and [Section 2.9.4](#) in absence of BGP Color-EC. Among other options, a BGP CAR BR may advertise a CAR route to an ingress BR with a specific BGP next hop per color, with a TEA or Tunnel Encapsulation EC.

[2.6](#). AIGP Metric Computation

The Accumulated IGP (AIGP) Attribute [[RFC7311](#)] 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 used to set the metric value.

If BGP CAR routes traverse across a discontinuity in the transport path for a given intent, a penalty is added in accumulated IGP metric (value set by user policy). For instance, when color C1 path is not available, and route resolves via color C2 path (e.g., [Appendix A.3](#)).

AIGP metric computation is recursive.

To avoid continuous IGP metric changes 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. This is out of scope for this document.

[2.7.](#) Native MultiPath Capability

The (E, C) route definition inherently provides availability of redundant paths at every BGP hop, identical to BGP-LU or BGP IP. For instance, BGP CAR routes originated by two or more 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 [[RFC7911](#)] SHOULD be enabled for BGP CAR to signal multiple next hops through a transport RR.

[2.8.](#) BGP CAR Signaling through different Color Domains

```
[Color Domain 1  A]-----[B      Color Domain 2      E2]
[C1=low-delay    ]       [C2=low-delay                ]
```


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 it is 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, even though the color-to-intent mapping 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

Operational considerations are in [Section 11](#). Further illustrations are provided in [Appendix B](#).

2.9. Format and Encoding

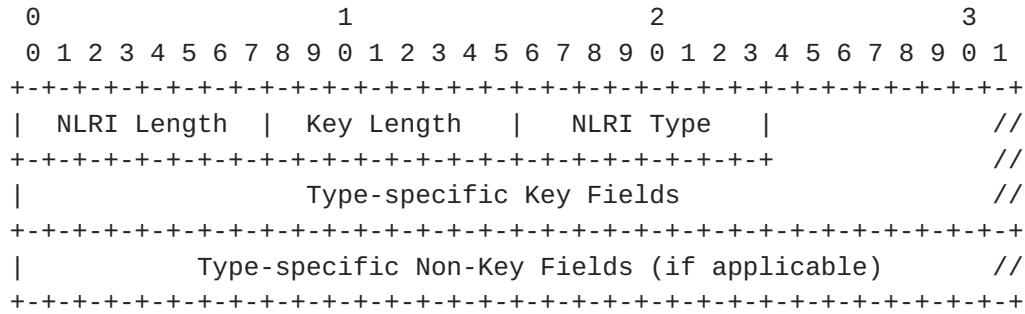
BGP CAR leverages the BGP multi-protocol extensions [RFC4760] and uses the MP_REACH_NLRI and MP_UNREACH_NLRI attributes for route updates by using the SAFI value 83 along with AFI 1 for IPv4 prefixes and AFI 2 for IPv6 prefixes.

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 83.

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.

It also helps make error handling more resilient and minimally disruptive as described in [Section 2.11](#).

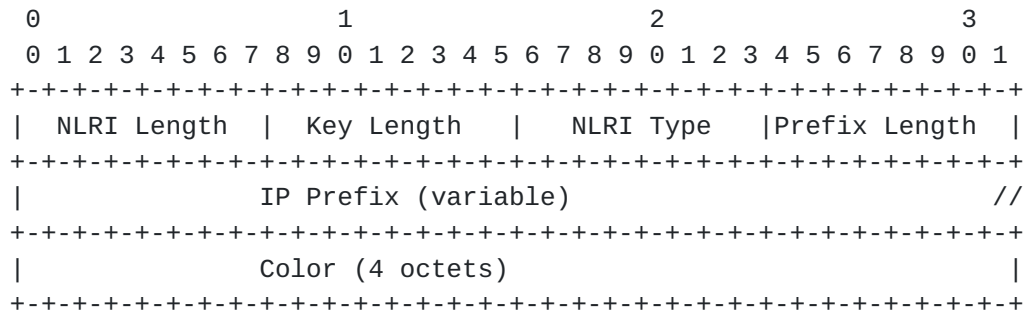
A route (NLRI) can carry more than one non-key TLV (of different types). This provides significant benefits such as signaling multiple encapsulations simultaneously for the same route, each with a different value (label/SID etc). This enables simpler, efficient migrations with low overhead :

- * avoids need for duplicate routes to signal different encapsulations
- * avoids need for separate control planes for distribution
- * preserves update packing (e.g. [Appendix D](#))

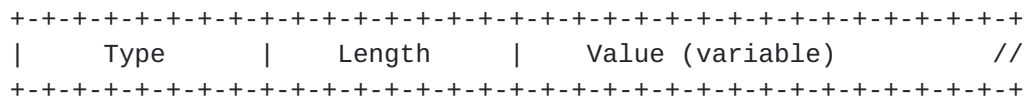
The non-key portion of the NLRI MUST be omitted while carrying it within the MP_UNREACH_NLRI when withdrawing the route advertisement.

[2.9.2](#). Color-Aware Route 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. Last octet has enough trailing bits to make the end of the field fall on an octet boundary. Note that the value of the trailing bits is irrelevant. 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:

```

    0 1 2 3 4 5 6 7
    +--+--+--+--+--+
    |R|T| Type code |
    +--+--+--+--+--+
  
```

where:

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

- o 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.

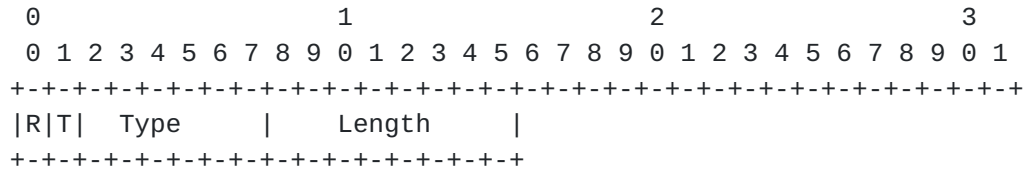
The prefix is routable across the administrative domain where BGP transport CAR is deployed. It is possible that the same prefix is originated by multiple BGP CAR speakers in the case of anycast addressing or multi-homing.

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.

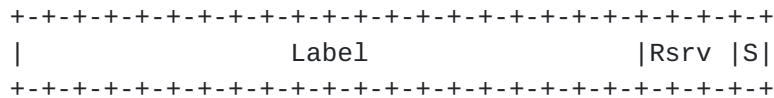
The following sub-sections specify the non-key TLVs associated with the Color-Aware Routes NLRI type.

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:



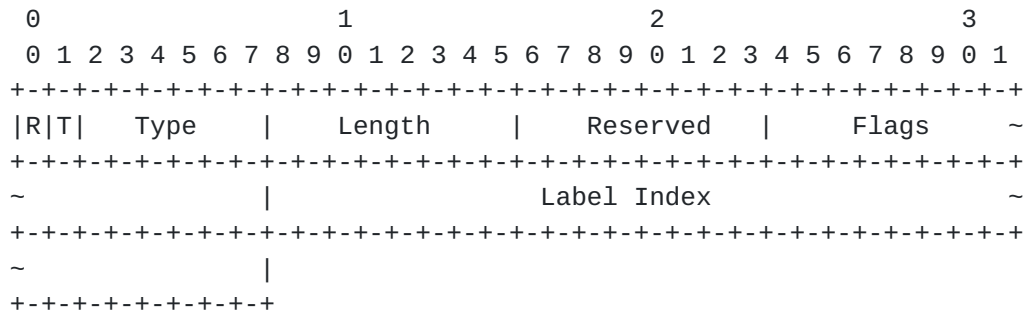
where:

- * 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 for usage as described in [RFC8277]. Number of labels is derived from length field. 3-bit Rsrv and 1-bit S field SHOULD be set to zero on transmission and MUST be ignored on reception.

When a BGP transport CAR speaker is propagating the route further after setting itself as the next hop, 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:

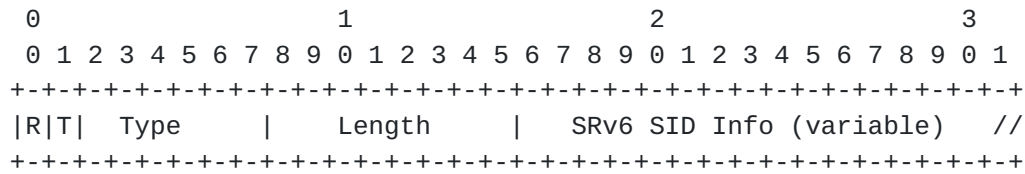
- * Type : Type code is 2. T bit MUST be set
- * Length: 7
- * Reserved: 1 octet field that MUST be set to 0 and ignored on receipt.
- * Flags: 2 octet field that maps to the Flags field of the Label-Index TLV of the BGP Prefix SID Attribute ([RFC8669](#)).
- * Label Index: 4 octet field that maps to 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 (Appendix D). 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 next hop, it allocates a local label for the specific prefix and color combination. When the received update has the CAR 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](#). [RFC8986](#) 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 [[RFC9252](#)]) of the SRv6 SID that MUST be of size in multiples of one octet and less than 16.

BGP CAR SRv6 SID TLV definitions provide the following benefits:

- * Native encoding of SIDs avoids robustness issue caused by overloading of MPLS label fields.
- * Simple encoding to signal Unique SIDs (non-transposition), maintaining BGP update prefix packing
- * Highly efficient transposition scheme (12-14 bytes saved per NLRI), also maintaining BGP update prefix packing

The BGP color-aware route update for SRv6 encapsulation MUST include the BGP Prefix-SID attribute along with the SRv6 L3 Service TLV carrying the SRv6 SID information as specified in [[RFC9252](#)]. When using the transposition scheme of encoding for packing efficiency of BGP updates [[RFC9252](#)], transposed part of SID is carried in SRv6 SID TLV and not limited by MPLS label field size.

[I-D.agrawal-spring-srv6-mpls-interworking] describes MPLS and SRv6 interworking procedures and extension to BGP CAR routes.

2.9.3. IP Prefix NLRI Type


```

0          1          2          3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
| NLRI Length | Key Length | NLRI Type |Prefix Length |
+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
|                IP Prefix (variable)                //
+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+

```

Followed by optional TLVs encoded as below:

```

+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
|R|T|  Type  | Length  | Value (variable) //
+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+

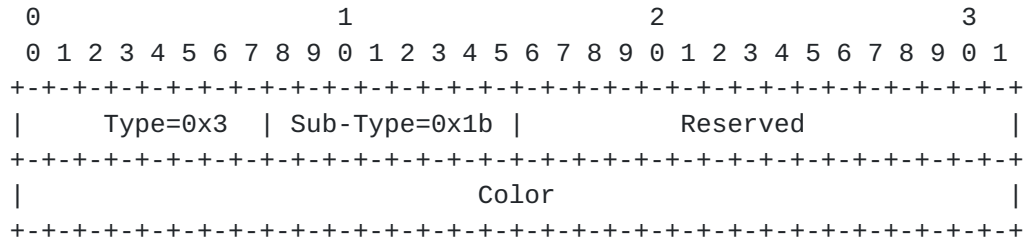
```

where:

- * NLRI Length: variable
- * Key Length: variable. It indicates the total length comprised of the Prefix Length field and IP Prefix field as described below. For IPv4 (AFI=1), the minimum length is 1 and maximum length is 5. For IPv6 (AFI=2), the minimum length is 1 and maximum length is 17.
- * NLRI Type: 2
- * 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. Last octet has enough trailing bits to make the end of the field fall on an octet boundary. Note that the value of the trailing bits is irrelevant. 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).
- * Type-Specific Non-Key Fields: Encoded as per Type-Specific Non-Key Fields of Color-Aware Routes NLRI Type in [Section 2.9.2](#).

2.9.4. 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: 0x1b.
- * 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's color domain boundary, LCM-EC is added. LCM-EC conveys the local color mapping for the intent (e.g. low latency) in other (transit or destination) color domains.

For Type-2 routes, it is also added in the originator color domain to indicate the color associated with the IP prefix.

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.

If two BGP paths for a route have different LCM values, it is considered an error and the route is not considered for bestpath selection.

If present, LCM-EC is the effective intent of a BGP CAR route.

LCM-EC Color is used instead of the Color in CAR route NLRI for procedures described in earlier sections such as route validation, resolution, AIGP calculation and steering.

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. LCM and BGP Color Extended-Community usage

There are 2 distinct requirements to be supported as stated in [[I-D.hr-spring-intentaware-routing-using-color](#)]" :

1. Domains with different intent granularity ([section 6.3.1.9](#))
2. Network domains under different administration, i.e. color domains ([section 6.3.1.10](#))

Requirement 1 is the case where within the same administrative or color domain, BGP CAR routes for N end-to-end intents may need to traverse across an intermediate transit domain where only M intents are available, $N \geq M$. Example: a multi-domain network is designed as Access-Core-Access. The core may have the most granular N intents, whereas the access only has fewer M intents. So, the BGP next-hop resolution for a CAR route in the access domain must be via a color-aware path for one of these M intents. As described in [Section 2.5](#) and [Appendix B.2](#), BGP Color Extended-Community is used to automate the CAR route resolution.

For requirement 2, where CAR routes traverse across different color domains, LCM-EC is used to carry the local color mapping for the NLRI color in other color domains as already described in [Section 2.8](#) and [Appendix B.3](#).

Both LCM-EC and BGP Color Extended-Community may be present at the same time with a BGP CAR route. Example: BGP CAR route (E, C1) from color domain D1, with LCM-EC C2 in color domain D2, may also carry Color-EC C3 and next hop N in a transit network domain within D2 where C2 is being resolved via an available intra-domain intent C3 ([Appendix B.2](#) and [Appendix B.3](#) combined).

Default order of processing for resolution in presence of LCM-EC is local policy, then BGP Color-EC color, and finally LCM-EC color.

2.11. Error Handling

The error handling actions as described in [[RFC7606](#)] are applicable for handling of BGP update messages for BGP-CAR. In general, as indicated in [[RFC7606](#)], the goal is to minimize the disruption of a session reset or 'AFI/SAFI disable' to the extent possible.

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.

The CAR NLRI definition encodes NLRI length and key length explicitly. The NLRI length MUST be relied upon to enable the beginning of the next NLRI field to be located. Key length MUST be relied upon to extract the key and perform 'treat-as-withdraw' for malformed information.

A sender MUST ensure that the NLRI and key lengths are number of actual bytes encoded in NLRI and key fields respectively, regardless of content being encoded.

Given NLRI length and Key length MUST be valid, failures in following checks result in 'AFI/SAFI disable' or 'session reset':

- * Minimum NLRI length (must be atleast 2, as key length and NLRI type are required fields)
- * Key Length MUST be at least two less than NLRI Length

NLRI Type specific error handling:

- * By default, a speaker SHOULD discard unrecognized or unsupported NLRI type and move to next NLRI.
- * Key length and key errors of known NLRI type SHOULD result in discard of NLRI similar to unrecognized NLRI type.(This MUST be logged for trouble shooting).

Transparent propagation of unrecognized NLRI type:

- * Key length allows unrecognized route types to transit through RR transparently without a software upgrade. Such RR does not need to interpret key portion of NLRI and works on opaque key of given length. An implementation SHOULD provide a knob that controls the RR unrecognized route type propagation behavior and possibly at granularity of route type values allowed. This gives ability to operator to allow specific route type transparent reflection based on client speaker support.

- * In such a case RR may reflect NLRIs with NLRI type specific key length and field errors. Clients of such RR that consume the route for installation will perform the key error handling of known NLRI type or discard unrecognized type. This prevents propagation of routes with NLRI errors any further in network.

Type-Specific Non-Key TLV handling:

- * 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
- * Type specific length constraints should be verified. The TLV MUST be discarded if there is an error.
- * If multiple instances of same type are encountered, all but the first instance MUST be ignored.
- * 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 TLVs (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 SR Policy then BGP CAR.

An egress PE may request intent through the transport for service routes in the BGP Color Extended-Community [[RFC9012](#)]. An ingress PE steers service traffic over CAR Type-1 NLRI using service route next hop and BGP Color extended community.

This is consistent with the automated service route steering on SR Policy (a routing solution providing color-aware path) defined in [[RFC9256](#)]. All the steering variations defined in [[RFC9256](#)] are applicable to BGP CAR color-aware path: on-demand steering, per-destination, per-flow, CO-only. For brevity, please refer to [[RFC9256](#)] [Section 8](#).

[Appendix A](#) provides illustrations of service route automated steering over BGP CAR.

An egress PE may request intent through the transport for service routes by allocating the SRv6 Service SID from a routed intent-aware locator prefix ([Section 3.3 of \[RFC8986\]](#)). Steering at an ingress PE is via resolution of the Service SID over CAR Type-2 IP Prefix NLRI. Service Steering over BGP CAR SRv6 transport is described in [Section 7](#).

Service steering via BGP CAR is applicable to any BGP SAFI, including SAFIs for IPv4/IPv6 (SAFI 1), L3VPN (SAFI 128), PW, EVPN (SAFI 70), FlowSpec, and BGP-LU (SAFI 4).

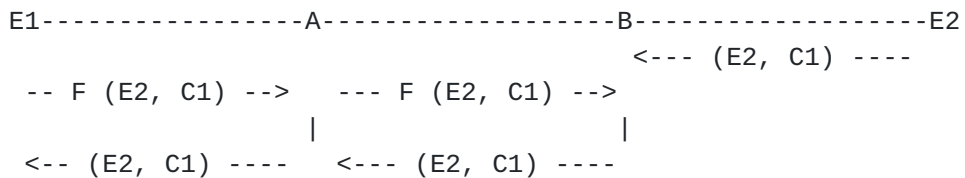
4. Filtering

PE and BRs may support filtering of CAR routes, for instance to only accept routes of locally configured colors.

RTC [\[RFC4684\]](#) may also be applied to the CAR SAFI, where Route Target ECs [\[RFC4360\]](#) can be used to constrain distribution of CAR routes. RT assignment may be via user policy, for example an RT value can be assigned to all routes of a specific color.

4.1. (E, C) Subscription and Filtering

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



- * 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).

On-demand subscription and filtering procedures are outside the scope of this document.

5. Scaling

This section analyses the key scale requirement of [\[I-D.hr-spring-intentaware-routing-using-color\]](#), 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

Two key principles used to address the scaling requirements are a hierarchical network and routing design, and on-demand route subscription and filtering.

Figure 2 provides an ultra-scale reference topology. [Section 5.2](#) presents three design models to deploy BGP CAR in the reference topology, including hierarchical options. [Section 5.3](#) analyses the scaling properties of each model. [Section 5.4](#) illustrates the scaling benefits of the (E, C) BGP subscription and filtering.

5.1. Ultra-Scale Reference Topology

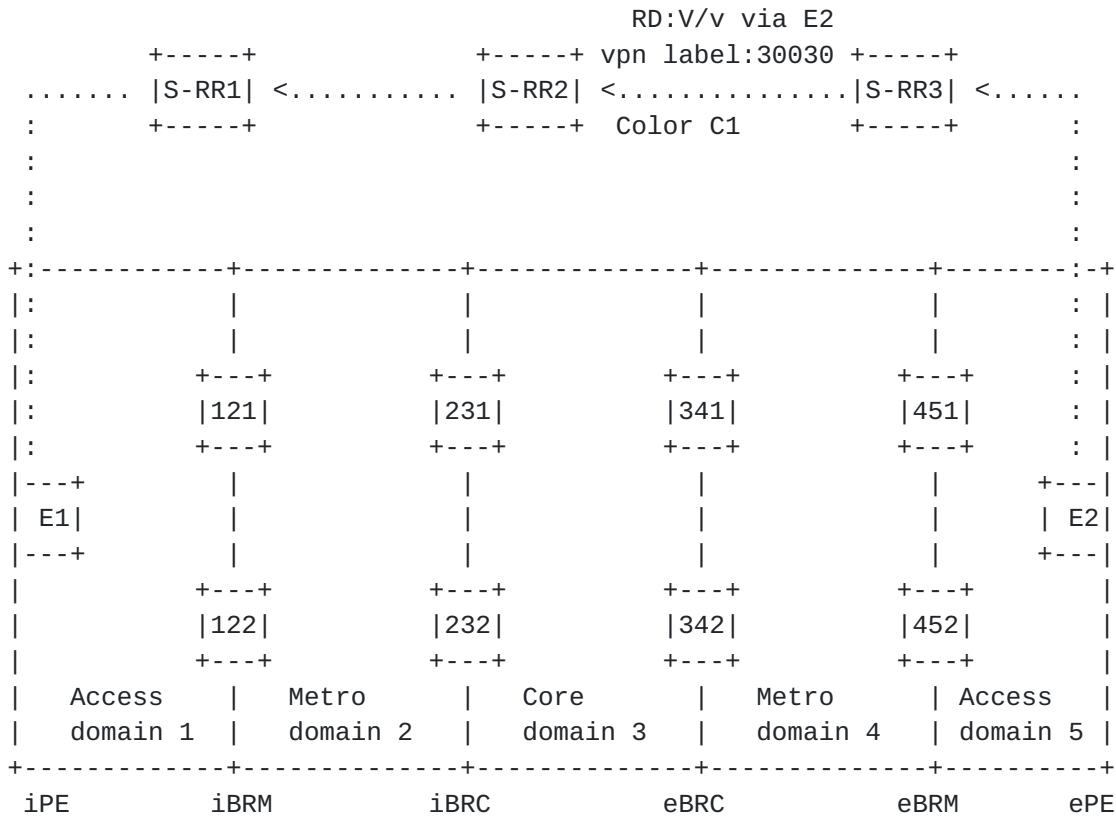


Figure 2: Ultra-Scale Reference Topology

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

5.2. Deployment model

5.2.1. Flat

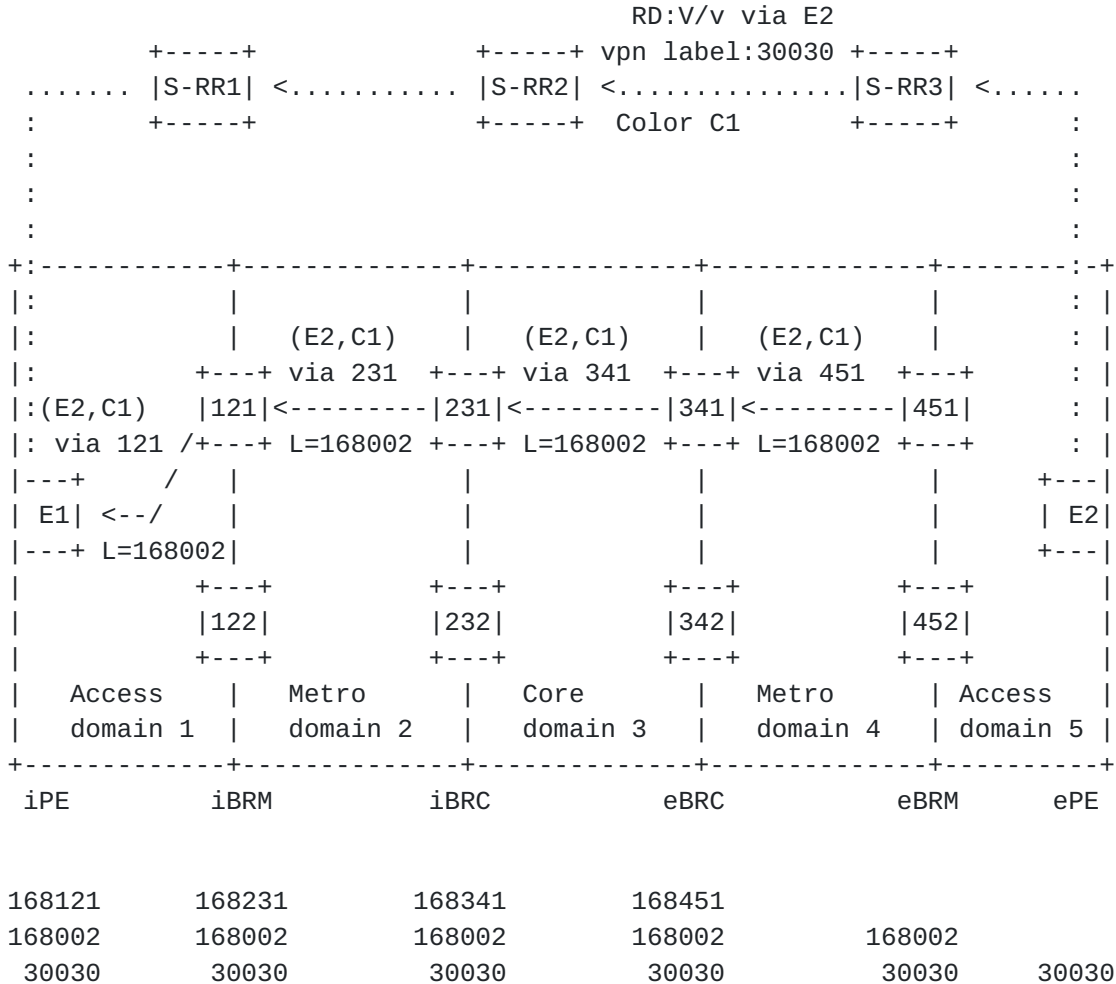


Figure 3

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)

5.2.2. Hierarchical Design with Next-Hop-Self at Ingress Domain BR

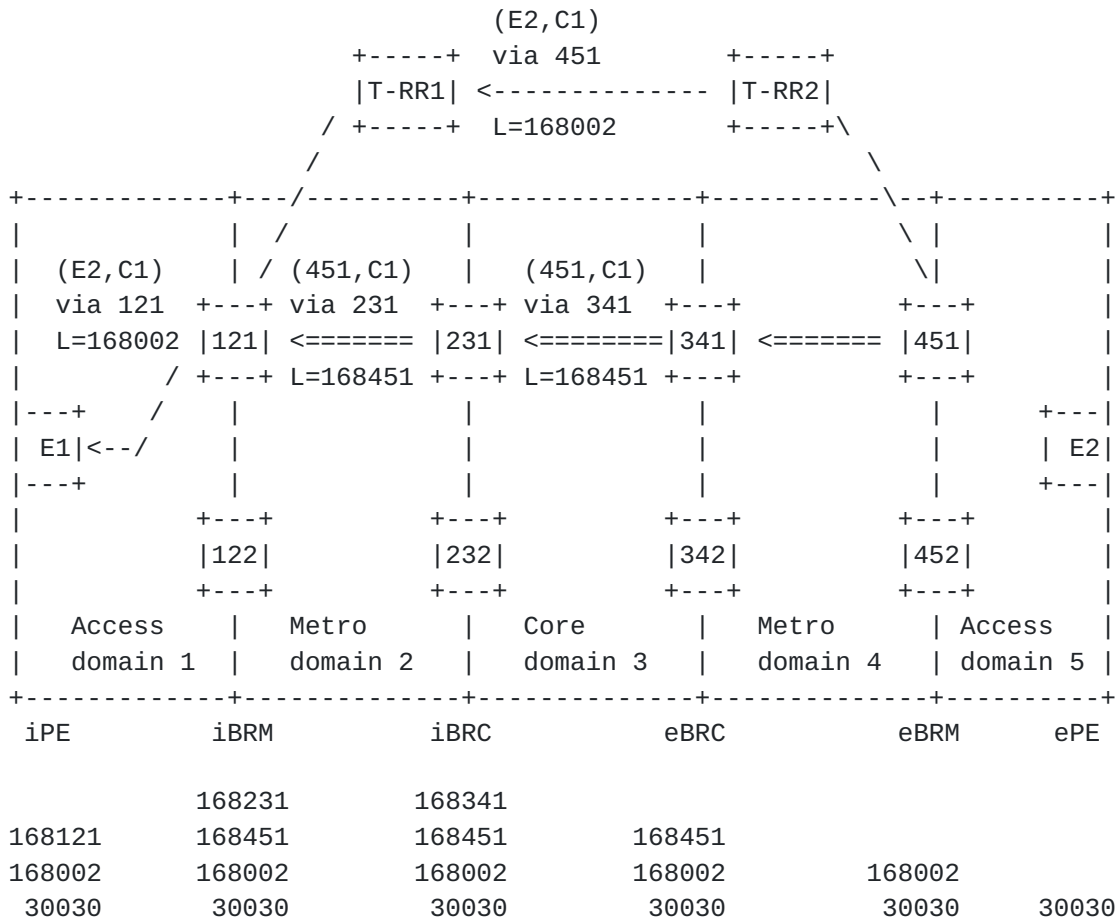


Figure 4: Heirarchical BGP transport CAR, Next-Hop-Self (NHS) at iBR

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

5.2.3. Hierarchical Design with Next-Hop-Unchanged at Ingress Domain BR

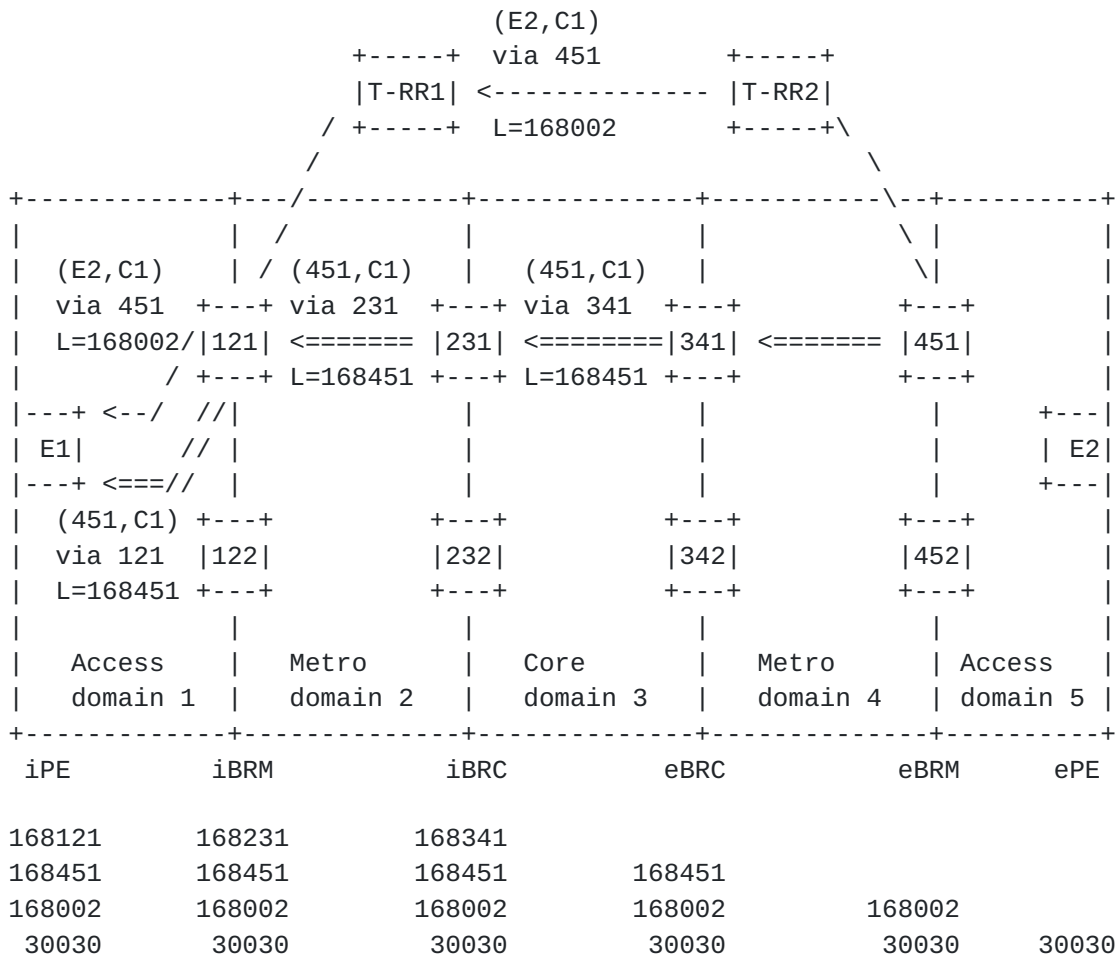


Figure 5: Heirarchical BGP transport CAR, Next-Hop-Unchanged (NHU) at iBR

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)

5.3. 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 (341,C)
		(451,C) via (231,C)	
H.NHU	(E2,C) via (451,C)	(451,C) via (121,C)	(451,C) via (341,C)
		(451,C) via (231,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
- They also aid faster convergence by allowing the PE routes to be distributed via out-of-band RRs that can be scaled independent of the transport BRs.

- * 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.

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

The (E, C) subscription scheme from [Section 4.1](#) provides the following scaling benefits for the models in [Section 5.2](#)

- * 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 its 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

5.5. Anycast SID

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

[5.5.1.](#) 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.

[5.5.2.](#) 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.

[6.](#) Routing Convergence

BGP CAR leverages existing well-known design techniques to provide fast convergence.

[Section 2.7](#) describes how BGP CAR provides localized convergence within a domain for BR failures, including originating BRs, without propagating failure churn into other domains.

Anycast SID techniques described in [Section 5.5](#) can provide further convergence optimizations for BR and PE failures deployed in redundant designs.

[7.](#) CAR SRv6

[7.1.](#) Overview

Two distinct cases apply to steering services over SRv6 based intent-aware multi-domain transport paths, as described in [Section 5 of \[RFC9252\]](#).

7.1.1. Routed Service SID

The SRv6 Service SID that is advertised with a service route is allocated by an egress PE from a routed intent-aware locator prefix ([Section 3.3 of \[RFC8986\]](#)). Service steering at an ingress PE is via resolution of the Service SID signaled with the service route ([\[RFC9252\]](#)).

The intent-aware transport path to the locator of the egress PE is provided by underlay IP routing, for instance, IGP-FlexAlgo [\[RFC9350\]](#) within a domain, and BGP-CAR across multiple IGP domains or ASNs.

An SRv6 locator is assigned for a given intent or color. This locator prefix is distributed using BGP-CAR to ingress PEs in a remote domain. The locator prefix may also be summarized on a border node along the path and a summary route distributed to ingress PEs. An IP Prefix CAR route (Type-2) is defined for this purpose described in [Section 2.9.3](#) and [Section 8](#).

The SRv6 locator may be shared with an IGP FlexAlgo, or may be assigned specific to BGP for a given intent. A BGP CAR advertised SRv6 locator prefix may also be used for resolution of the SRv6 service SID advertised for best effort connectivity.

[Appendix C.1](#) and [Appendix C.2](#) illustrates the control and forwarding behaviors for this case.

[Section 7.2](#) describes the deployment options.

[Section 7.3](#) describes operational considerations related to BGP CAR vs BGP IPv6 SAFI for inter-domain route distribution.

7.1.2. Non-routed Service SID

The SRv6 Service SID allocated by an egress PE is not routed. The service route is advertised by the egress PE with a Color Extended-Community C ([\[RFC9252\] section 5](#)).

The intent-aware path within an egress domain is provided by an SR-TE or similar policy to the egress PE (E, C) [\[RFC9256\]](#). This (E, C) policy is distributed into the multi-domain network from egress BRs using a BGP-CAR Type-1 route, towards ingress PEs in other domains. This signaling is the same as for SR-MPLS, as described in earlier sections.

The (E, C) BGP CAR Type-1 route is advertised from a BR with an SRv6 transport SID allocated from a locator assigned for the intent C. An SR-PCE or local configuration may ensure multiple BRs in the egress domain that originate the (E, C) route advertise the same SRv6 transport SID.

An ingress PE in a remote domain steers a received service route with Color C via this (E, C) BGP CAR route, as described in [Section 3](#).

Additionally, the ingress PE resolves the transport SID received in the (E, C) CAR route via another underlay intent-aware route. BGP-CAR also provides the underlay intent-aware inter-domain reachability to this transport SID:

- * BRs in the egress domain advertise an IP Prefix Type-2 CAR route for the locator prefix that covers the transport SID allocated by the egress BR for this (E, C) route. This IP Prefix CAR route is distributed across BGP hops in the underlay towards the ingress PE similar to previous case and may be summarized.
- * Thus, the service route's service SID resolves via the Type-1 route, which in turn carries a transport SID that resolves via the Type-2 route. Multiple Type-1 routes may resolve via a single Type-2 route.

Note: This is the typical resolution order as the Type-2 route provides intent-aware reachability to the BRs that advertise the Type-1 specific routes for each egress PE. However, there can be use-cases where a Type-2 route may resolve via a Type-1 route.

An ingress PE via this resolution builds the packet encapsulation that contains the Service SID and the received (E, C) route's transport SID in the SID-list.

[Appendix C.3](#) contains an example that illustrates the control plane distribution, recursive resolution and forwarding behaviors described above.

Note: An SR-policy may also be defined for multi-domain end to end [[RFC9256](#)], independent of BGP CAR. In that case, both BGP CAR and SR-TE inter-domain paths may be available at an ingress PE for an (E, C) route ([Section 1.2](#)).

7.2. Deployment Options For CAR SRv6 Locator Reachability Distribution and Forwarding

Since an SRv6 locator (or summary) is an IPv6 prefix, it will be installed into the IPv6 forwarding table on a BGP router, such as an ABR or ASBR for forwarding. A few options to forward packets for BGP SRv6 prefixes ([\[I-D.agrawal-spring-srv6-mpls-interworking\]](#)) apply to BGP CAR as follows.

7.2.1. Hop by Hop IPv6 Forwarding for BGP SRv6 Prefixes

- * Hop by hop IPv6 lookup and forwarding on both BRs and P nodes in a domain
 - No tunnel encapsulation between BRs in a domain
 - No per-PE SID allocation and installation on any BGP hop
- * P nodes need to learn BGP SRv6 routes. With summarization, route scale requirements can be minimized.
- * BGP routing is enabled on all internal nodes (iBGP)
- * BRs distribute external SRv6 routes to internal peers
 - Next-hop unchanged with recursive resolution via IGP at each hop.
- * Similar to Internet / BGP IP routing well-known model
 - Can support large scale route distribution

Illustration in [Appendix C.1](#)

7.2.2. Encapsulation between BRs for BGP SRv6 Prefixes

- * IPv6 lookup and forwarding for BGP SRv6 prefixes only on BGP BRs
 - P nodes do not learn or install these prefixes
- * SRv6 (or other) encapsulation to reach the BGP SRv6 next hop
 - Not needed for connected next hops, such as eBGP single-hop
- * SRv6 outer encapsulation may be H.Encaps.Red or H.Insert.Red
- * BGP route distribution between BRs (via RRs, or directly if single-hop eBGP)

- * An egress BR sets itself as BGP NH, selects and advertises an appropriate SID for SRv6 based encapsulation towards itself
 - BGP NH and SID for specific intent within domain
- * An ingress BR encapsulates SRv6 egress PE destined packets with encapsulation to BGP NH, ie. Egress BR
- * If SRv6 encapsulation, then SID from egress BR is common SID, shared by multiple BGP SRv6 prefixes
 - No per-PE SID allocation and installation on any BGP hop
- * If MPLS/SR-MPLS transport, route will carry label/prefix-SID allocated by NH, may be shared

Illustration in [Appendix C.2](#)

7.3. Operational Benefits of using CAR SAFI for SRv6 Locator Prefix Distribution

When reachability to an SRv6 SID is provided by distribution of a locator prefix via underlay routing, BGP IPv6 SAFI (AFI/SAFI=2/1) may also be used for inter-domain distribution of these IPv6 prefixes as described in [[I-D.agrawal-spring-srv6-mpls-interworking](#)] ([Section 7.1.2](#)) or [[I-D.wang-idr-cpr](#)].

Using the BGP CAR SAFI provides significant operational advantages:

- * CAR SAFI is a separate BGP SAFI used for underlay transport intent-aware routing. It avoids overloading of BGP IPv6 SAFI, which also carries Internet (service) prefixes. Using CAR SAFI provides:
 - Automatic separation of SRv6 locator (transport) routes from Internet (service) routes,
 - o Preventing inadvertent leaking of routes
 - o Avoiding need to configure specific route filters for locator routes
 - Priority handling of infrastructure prefixes over Internet prefixes
- * CAR SAFI also supports inter-domain distribution of (E, C) routes sourced from SR-Policy, in addition to SRv6 locator IPv6 prefixes.

- * CAR SAFI may also be used for best-effort routes in addition to intent-aware routes as described in the next section.

8. CAR IP Prefix Route

An IP Prefix CAR route is a route type (Type-2) that carries a routable IP prefix. Type-2 routes are installed in the default routing and forwarding table and provide longest-prefix-match forwarding. This is unlike Type-1 routes, where it is the signaled forwarding data such as labels/SIDs that are installed in the forwarding table to create end to end paths.

Type-2 routes may be originated into BGP CAR SAFI either from an egress PE or from a BR in a domain.

As described in [Section 2.1](#), it is used for cases where a unique routable IP prefix is assigned for a given intent or color. It may also be used for routes providing best-effort connectivity.

A few applicable example use-cases:

- * SRV6 locator prefix with color for specific intents
- * SRV6 locator prefix without color for best effort
- * Best effort transport reachability to a PE/BR without color

For specific intents, color may be signaled with the CAR Type-2 route for purposes such as intent-aware SRv6 SID or BGP next-hop selection at each transit BR, color based routing policies and filtering, and intent-aware next-hop resolution ([Section 2.5](#)). For such purposes, color associated with the CAR IP Prefix route is signaled using LCM-EC.

Reminder: LCM-EC conveys end-to-end intent/color associated with route/NLRI. When traversing network domain(s) where a different intent/color is used for next-hop resolution, BGP Color-EC may additionally be used as in [Section 2.10](#).

A special case of intent is best effort which may be represented by a color and follow above procedures. But to be compatible with traditional operational usage, CAR Type-2 route is allowed to be without color for best effort. In this case, the routes will not carry an LCM-EC. Resolution is described in [Section 2.5](#).

As described in [Section 7.3](#), infrastructure prefixes are intended to be carried in CAR SAFI instead of SAFIs that also carry service routes such as BGP-IP ([RFC 4271](#)) and BGP-LU ([RFC 8277](#)). However, if

such routes are also distributed in these SAFIs, a router may receive both BGP CAR SAFI paths and IP/LU SAFI paths. By default, CAR SAFI transport path is preferred over BGP IP or BGP-LU SAFI path.

A BGP transport CAR speaker that supports packet forwarding lookup based on IPv6 prefix route (such as a BR) will set itself as next hop while advertising the route to peers. It will also install the IPv6 route into forwarding with the received next hop and/or encapsulation. If such a transit router does not support this route type, it will not install this route and will not set itself as next hop, hence will not propagate the route any further.

9. VPN CAR

This section illustrates the extension of BGP CAR to address the VPN CAR requirement stated in Section 6.1.2 of [\[I-D.hr-spring-intentaware-routing-using-color\]](#), using MPLS transport as an example.



- * 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)
or 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)
or connected OIF
- 6.a. Label L3 is installed as the imposition label for (V, CC)

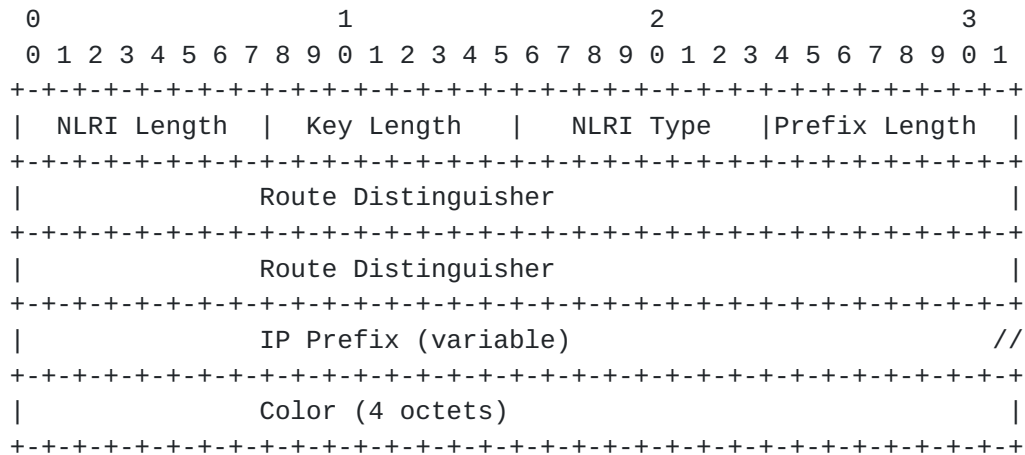
VPN CAR distribution for (RD, V, CC) requires a new SAFI that follows same VPN semantics as defined in [\[RFC4364\]](#), the difference being that the advertised routes carry CAR NLRI defined in [Section 2.9.2](#) and [Section 2.9.3](#) of this document. Procedures defined in [\[RFC4364\]](#) and [\[RFC4659\]](#) apply to VPN CAR SAFI.

Further, all CAR SAFI procedures described in [Section 2](#) above apply to CAR SAFI enabled within a VRF. VPN CAR SAFI routes follow color based steering as described in [Section 3](#) and illustrated in example above.

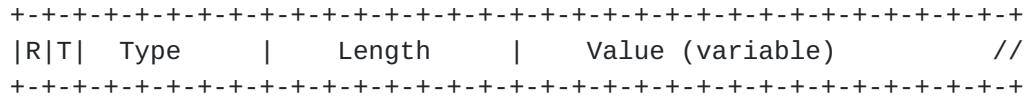
CAR routes distributed in VPN CAR SAFI are infrastructure routes advertised by CEs in different customer VRFs on a PE. The VPN RD distinguishes CAR routes of different customers being advertised by the PE.

[9.1.](#) VPN CAR Type-1 NLRI

VPN CAR NLRI with RD has the format shown below



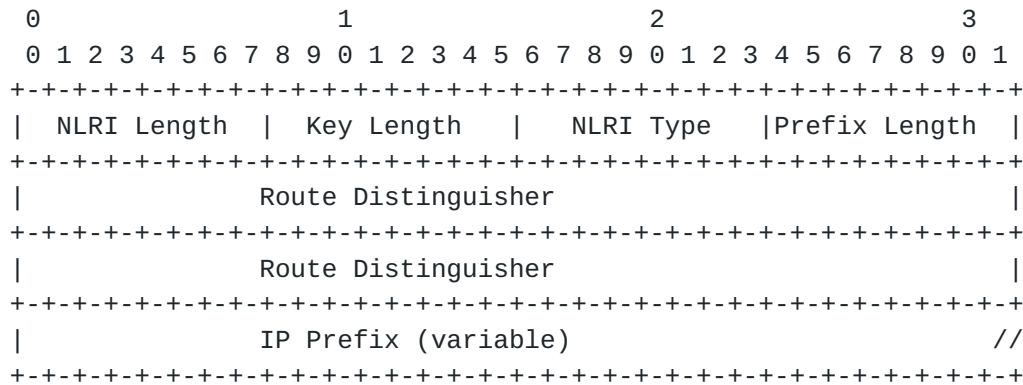
Followed by optional TLVs encoded as below:



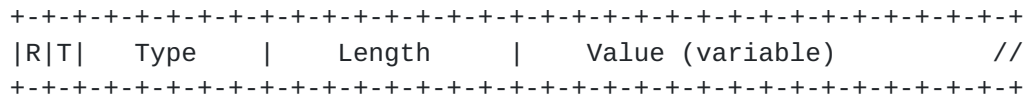
where:

- * All fields are encoded as per [Section 2.9.2](#).
- * Key Length: It indicates the total length comprised of the RD, Prefix Length field, IP Prefix field, and the Color field
- * Route Distinguisher: 8 octet field encoded according to [[RFC4364](#)]

9.2. VPN CAR IP Prefix NLRI Type



Followed by optional TLVs encoded as below:



where:

- * All fields are encoded as per [Section 2.9.3](#).
- * Key Length: It indicates the total length comprised of the RD, Prefix Length field and IP Prefix field
- * Route Distinguisher: 8 octet field encoded according to [[RFC4364](#)]

10. 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.

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 Route NLRI	[This document]
2	IP Prefix NLRI	[This document]
3-255	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]
4-64	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 has assigned the sub-type 0x1b for "Local Color Mapping (LCM)" under the "BGP Transitive Opaque Extended-Community" registry under the "BGP Extended-Community" parameter registry.

11. Manageability and Operational Considerations

Color assignments in a multi-domain network operating under a common or cooperating administrative control (i.e., a 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. This is a logical best practice in a single color or administrative domain, which is the most typical deployment scenario.

When color-aware routes propagate across a color domain boundary, there is typically no need for coordinating color assignments, since the IP prefix is unique in the transport network, and hence makes the color scope also unique and non-conflicting. The color only needs to be re-mapped into a local color assigned for the same intent (which is carried in the LCM-EC).

However, if networks under different administrative control establish a shared transport service between them, where the same transport service 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.

It should be noted that the color assignments coordination are only necessary for routes to the shared service IP. Colors used for intra-domain or for inter-domain intents associated with the unique IP addresses do not need any coordination.

Extended communities (LCM-EC/Color-EC) carried in BGP CAR and Service routes must not be filtered, otherwise the desired intent will not be achieved.

12. Security Considerations

This extension defines a new SAFI within BGP and therefore does not change the underlying security issues inherent in the existing BGP protocol, such as those described in [[RFC4271](#)] and [[RFC4272](#)].

The extensions defined in this document allows BGP to carry color aware routes and their associated attributes within a separate BGP SAFI which is expected to be configured manually by an operator. As part of configuring a new SAFI, it is implied that the necessary policy filtering is configured on this SAFI to filter routing information by the routers participating in this network. Also, given that this SAFI and these mechanisms can only be enabled through configuration of routers within a single network, standard security measures should be taken to restrict access to the management interface(s) of routers that implement these mechanisms.

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

Since CAR SAFI is a separate BGP SAFI that carries transport or infrastructure routes for routers in the operator network, it provides automatic separation between infrastructure routes and service routes that are carried in existing BGP SAFIs such as BGP IPv4/IPv6 (SAFI=1), and BGP-LU (SAFI=4). It thus provides better security than would be obtained by distributing the infrastructure routes in existing SAFIs.

BGP CAR distributes label binding similar to [[RFC8277](#)] and hence its security considerations apply. Similarly, BGP CAR distributes infrastructure IPv6 prefixes and SRv6 SID for SRv6 based CAR and hence security considerations of [section 9.3 of \[RFC9252\]](#) apply.

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

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

13. Co-authors

Clarence Filsfils
Cisco Systems
Belgium
Email: cfilsfil@cisco.com

Bruno Decraene
Orange
France
Email: bruno.decraene@orange.com

Luay Jalil
Verizon
USA
Email: luay.jalil@verizon.com

Yuanchao Su
Alibaba, Inc
Email: yitai.syc@alibaba-inc.com

Jim Uttaro
ATT
USA
Email: ju1738@att.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

Jie Dong
Huawei Technologies
China
Email: jie.dong@huawei.com

14. Contributors

Dirk Steinberg
Lapishills Consulting Limited
Germany
Email: dirk@lapishills.com

Israel Means
AT&T
USA
Email: im8327@att.com

Reza Rokui
Ciena
USA
Email: rrokui@ciena.com

15. Acknowledgements

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

16. References

16.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", [BCP 14](#), [RFC 2119](#), DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [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>>.
- [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>>.
- [RFC8986] Filsfils, C., Ed., Camarillo, P., Ed., Leddy, J., Voyer, D., Matsushima, S., and Z. Li, "Segment Routing over IPv6 (SRv6) Network Programming", [RFC 8986](#), DOI 10.17487/RFC8986, February 2021, <<https://www.rfc-editor.org/info/rfc8986>>.
- [RFC9012] Patel, K., Van de Velde, G., Sangli, S., and J. Scudder, "The BGP Tunnel Encapsulation Attribute", [RFC 9012](#), DOI 10.17487/RFC9012, April 2021, <<https://www.rfc-editor.org/info/rfc9012>>.
- [RFC9252] Dawra, G., Ed., Talaulikar, K., Ed., Raszuk, R., Decraene, B., Zhuang, S., and J. Rabadan, "BGP Overlay Services Based on Segment Routing over IPv6 (SRv6)", [RFC 9252](#), DOI 10.17487/RFC9252, July 2022, <<https://www.rfc-editor.org/info/rfc9252>>.

- [RFC9256] Filsfils, C., Talaulikar, K., Ed., Voyer, D., Bogdanov, A., and P. Mattes, "Segment Routing Policy Architecture", [RFC 9256](#), DOI 10.17487/RFC9256, July 2022, <<https://www.rfc-editor.org/info/rfc9256>>.
- [RFC9350] Psenak, P., Ed., Hegde, S., Filsfils, C., Talaulikar, K., and A. Gulko, "IGP Flexible Algorithm", [RFC 9350](#), DOI 10.17487/RFC9350, February 2023, <<https://www.rfc-editor.org/info/rfc9350>>.

16.2. Informative References

- [I-D.agrawal-spring-srv6-mpls-interworking]
Agrawal, S., Ali, Z., Filsfils, C., Voyer, D., Dawra, G., Li, Z., Hegde, S., and S. R. Sangli, "SRv6 and MPLS interworking", Work in Progress, Internet-Draft, [draft-agrawal-spring-srv6-mpls-interworking-12](#), 10 July 2023, <<https://datatracker.ietf.org/doc/html/draft-agrawal-spring-srv6-mpls-interworking-12>>.
- [I-D.hr-spring-intentaware-routing-using-color]
Hegde, S., Rao, D., Uttaro, J., Bogdanov, A., and L. Jalil, "Problem statement for Inter-domain Intent-aware Routing using Color", Work in Progress, Internet-Draft, [draft-hr-spring-intentaware-routing-using-color-03](#), 23 October 2023, <<https://datatracker.ietf.org/doc/html/draft-hr-spring-intentaware-routing-using-color-03>>.
- [I-D.ietf-mpls-seamless-mpls]
Leymann, N., Decraene, B., Filsfils, C., Konstantynowicz, M., and D. Steinberg, "Seamless MPLS Architecture", Work in Progress, Internet-Draft, [draft-ietf-mpls-seamless-mpls-07](#), 28 June 2014, <<https://datatracker.ietf.org/doc/html/draft-ietf-mpls-seamless-mpls-07>>.
- [I-D.wang-idr-cpr]
Wang, H., Dong, J., Talaulikar, K., hantao, Chen, R., Xie, J., and X. Chen, "BGP Colored Prefix Routing (CPR) for SRv6 based Services", Work in Progress, Internet-Draft, [draft-wang-idr-cpr-03](#), 20 October 2023, <<https://datatracker.ietf.org/doc/html/draft-wang-idr-cpr-03>>.
- [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>>.
- [RFC4659] De Clercq, J., Ooms, D., Carugi, M., and F. Le Faucheur, "BGP/MPLS IP Virtual Private Network (VPN) Extension for IPv6 VPN", [RFC 4659](#), DOI 10.17487/RFC4659, September 2006, <<https://www.rfc-editor.org/info/rfc4659>>.
- [RFC5082] Gill, V., Heasley, J., Meyer, D., Savola, P., Ed., and C. Pignataro, "The Generalized TTL Security Mechanism (GTSM)", [RFC 5082](#), DOI 10.17487/RFC5082, October 2007, <<https://www.rfc-editor.org/info/rfc5082>>.
- [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>>.
- [RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", [RFC 5925](#), DOI 10.17487/RFC5925, June 2010, <<https://www.rfc-editor.org/info/rfc5925>>.
- [RFC6811] Mohapatra, P., Scudder, J., Ward, D., Bush, R., and R. Austein, "BGP Prefix Origin Validation", [RFC 6811](#), DOI 10.17487/RFC6811, January 2013, <<https://www.rfc-editor.org/info/rfc6811>>.
- [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>>.
- [RFC8205] Lepinski, M., Ed. and K. Sriram, Ed., "BGPsec Protocol Specification", [RFC 8205](#), DOI 10.17487/RFC8205, September 2017, <<https://www.rfc-editor.org/info/rfc8205>>.
- [RFC9315] Clemm, A., Ciavaglia, L., Granville, L. Z., and J. Tantsura, "Intent-Based Networking - Concepts and Definitions", [RFC 9315](#), DOI 10.17487/RFC9315, October 2022, <<https://www.rfc-editor.org/info/rfc9315>>.

Figure 6: BGP FA Aware transport CAR path

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

* With reference to the topology above:

- IGP FA 128 is running in each domain, and mapped to Color C1
- 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.
- On each BGP hop, (E2, C1) next hop is resolved 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 within each domain.
- BGP CAR label (e.g. 168002) carries end to end intent. Thus it stitches intent over intra domain FA 128.

[A.2.](#) E2E BGP transport CAR intent realized using SR Policy

- o SR policy (C1,121) segments <S1, 121>
 - o SR policy (C1,231) segments <S2, 231>
 - o SR policy (C1,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.
 - On each BGP hop, CAR route (E2, C1) next hop is resolved 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 within each domain.
 - BGP CAR label (e.g. 168002) carries end to end intent. Thus it 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

- 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 (+1000).
- On 121 and 122, CAR route (E2, C1) next hop learnt from Core domain is resolved 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)

* 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

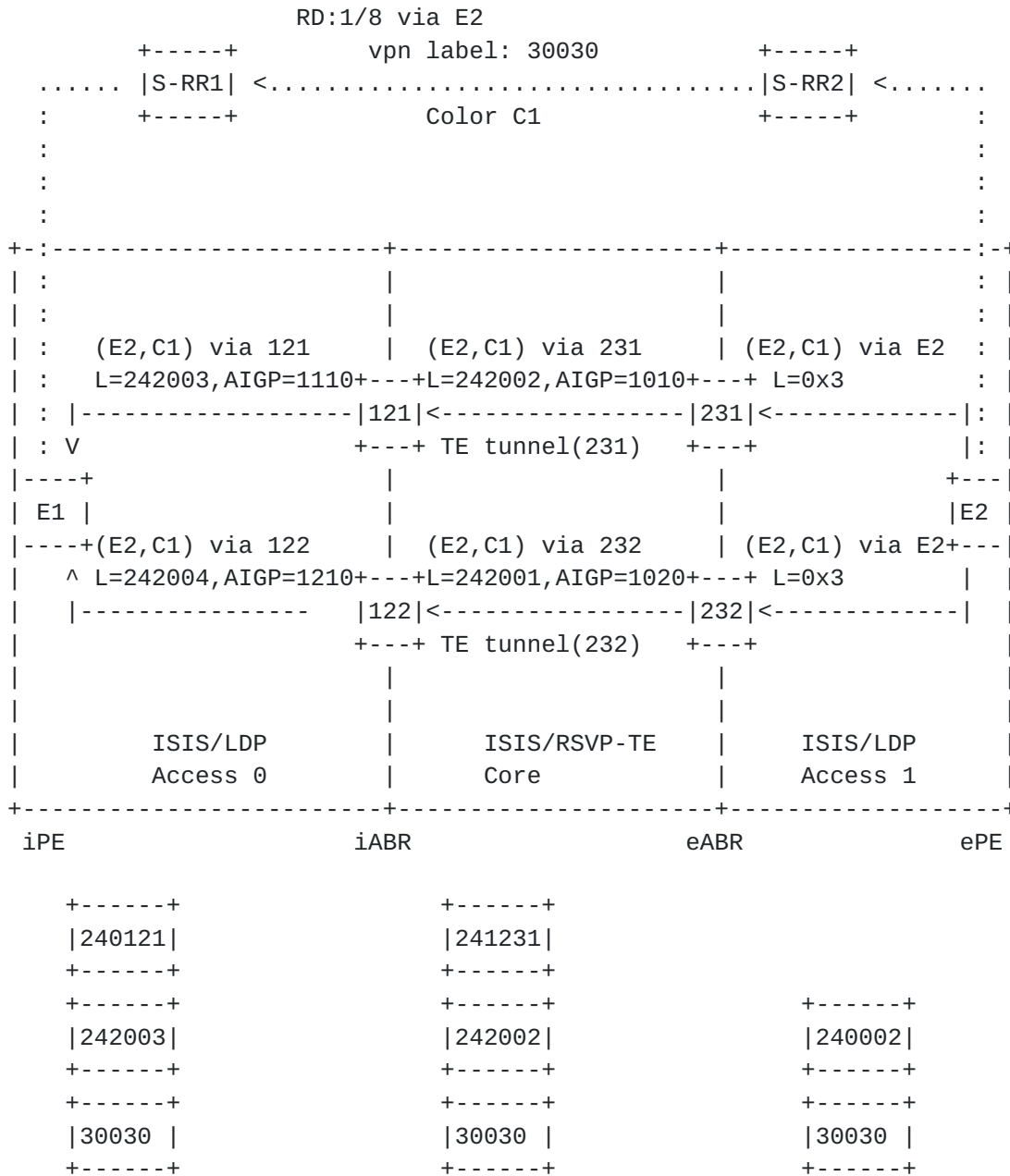


Figure 9: BGP CAR over TE tunnel mesh in core network

- * 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 (+1000).
- 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's next hop over best effort LDP LSP in Access domain 0. It steers colored VPN route RD:V/v via (E2, C1).

* 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

* In a brownfield deployment, color-aware paths between two PEs may need to go through a transit domain that does not support CAR. Examples 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.

* 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. The sessions may be eBGP or iBGP.
- * 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.

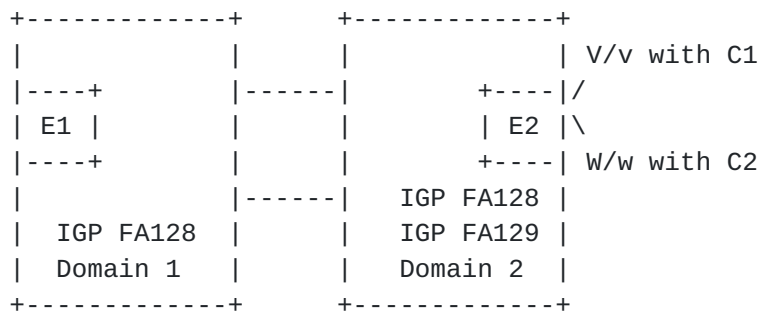


Figure 10: BGP CAR resolution over IGP FLex-Algo for resource avoidance in a domain

- * 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:
 - o Domain1: over IGP FA128
 - o Domain2: over IGP FA128
 - (E2, C2) provided by BGP CAR with the following per-domain resolution:
 - o Domain1: over IGP FA128
 - o 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 \[RFC9256\]](#) 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

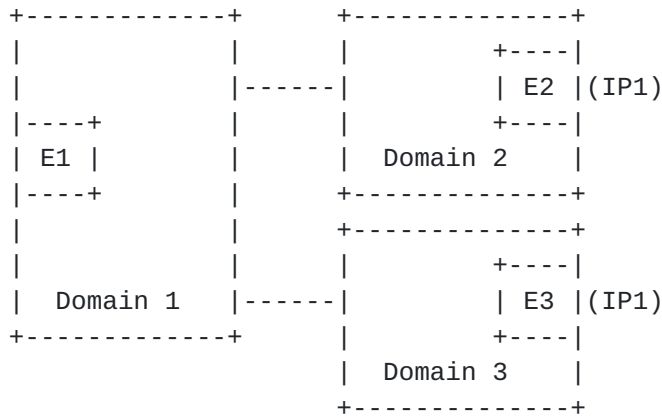


Figure 11: BGP CAR advertisements for shared IP addresses

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

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

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 for redundancy.

Example-1: Anycast with forwarding to nearest

- * Both E2 (in egress domain 2) and E3 (in egress domain 3) advertise Anycast (shared) IP (IP1, C1) with same label L1
- * An ingress PE E1 receives by default the best path(s) for (IP1, C1) propagated through BGP hops across the network.
- * The paths to (IP1, C1) from E2 and E3 may merge at a common node along the path to E1, forming equal cost multipaths or active-backup paths at that node
- * Service route V/v is advertised from egress domains D2 and D3 with color C1 and next hop IP1.
- * Traffic for V/v steered at E1 via (IP1, C1) is forwarded to either E2 or E3 (or both) as determined by routing along 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. C1 and C2 are colors assigned to distinguish the egress domains originating the routes to 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 CAR routes (IP1, C1) and (IP1, C2) do not get merged at any intermediate node, providing E1 control over path selection and load-balancing of traffic across these two routes. Each route may itself provide multipathing or Anycast to a set of egress nodes.
- * Service route V/v advertised from egress domains D2 and D3 with colors C1 and C2 respectively, but with same next hop IP1.
- * E1 will resolve and steer V/v path from D2 via (IP1, C1) and path from D3 via (IP2, C2). E1 will load-balance traffic to V/v across the two paths as determined by a local load-balancing policy.
- * Traffic for colored service routes steered at E1 is forwarded to either E2 or E3 (or load-balanced across both) as determined by E1.

In above example, D2 and D3 belonged to the same color or administrative domain. If D2 and D3 belonged to different color domains, the domains will coordinate the assignment of colors to be used with shared IP IP1 such that they do not cause conflicts. For instance, in Example-1 :

- * D2 and D3 may both use C1 for the same intent when they originate CAR route for IP1.
 - In this case, neither D2 nor D3 will reuse C1 for some other intent
- * Alternatively, D2 may use C2 and D3 may use C3 for originating a CAR route for IP1 for the same intent.
 - In this case, D2 will not use C3 for originating CAR route for IP1 for some other intent. Similarly, D3 will not use C2 for originating CAR route for IP1 for some other intent.

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 used for end-to-end intents, but may still be required to provide transit for routes of those colors.
- * When a (E, C1) route traverses a domain where color C1 is not available, the operator may decide to use a different intent of color C2 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 C2 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 C2 to the BGP CAR routes.
- * Hence, routes of N end-to-end colors may be resolved over paths from a smaller set of M colors 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 C1, must have C1 enabled locally to install IP routes to (E, C1) and resolve the service route's 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.

Illustration for N end to end intents over fewer M intra domain intents:

- o FA129 mapped to C20
 - o FA130 mapped to C30
 - o FA131 mapped to C40
 - Access domain provides 2 intra domain intents
 - o FA132 mapped to C1
 - o FA133 mapped to C2
 - Operator defines 4 BGP CAR end to end intents as below
 - o CAR color C100 that resolves on C1 in access and C10 in core domain
 - o CAR color C200 that resolves on C1 in access and C20 in core domain
 - o CAR color C300 that resolves on C2 in access and C30 in core domain
 - o CAR color C400 that resolves on C2 in access and C40 in core domain
 - E2 may originate BGP CAR routes with multiple BGP Color-ECs as shown above. At each hop, CAR route's next hop is resolved over the available intra-domain color. For example (E2,C100) with BGP color ECs C1, C10 resolves over C1 at ABR 231, C10 at ABR 121 and C1 at E1.
 - Egress PE E2 advertises a VPN route RD:V/v colored with BGP Color-EC C100 to steer traffic through FA 132 in access and FA 128 in core. It also advertises another VPN route RD:W/w colored with BGP Color-EC C200 to steer traffic through FA 132 in access and FA 129 in core.
- * Important:
- End-to-end (BGP CAR) colors can be decoupled from intra-domain transport colors.
 - Each BGP CAR color is a combination of various intra-domain colors or intents.

- Combination can be expressed by local policy at ABRs or by attaching multiple BGP Color-ECs at origination point of BGP CAR route.
- Service traffic is steered into suitable CAR color to use the most granular intent in a domain multiple hops away from ingress PE.
- Consistent reuse of standard color based resolution mechanism at both service and transport layers.

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.

[Appendix C](#). CAR SRv6 Illustrations

[C.1](#). BGP CAR SRv6 locator reachability hop by hop distribution

- * Transport RR (TRR) peers with P, BR and PE clients within an AS to propagate CAR prefixes. AddPath is enabled to propagate multiple paths.
- * ISIS (IGP) FlexAlgo 128 for SRv6 is running in each AS (AS may consist of multiple IGP domains)
 - Prefix B:C11::/32 summarizes Flex-Algo 128 block in AS1 for the given intent. Node locators in the egress domain are sub-allocated from the block for the given intent
 - Similarly, Prefix B:C12::/32 summarizes Flex-Algo 128 block in AS2
 - Per Flex-Algo external subnets for eBGP next hops IP1 and IP2 are distributed in ISIS within AS2
- * BGP CAR prefix route B:C11::/32 with LCM C1 is originated by AS1 BRs 231 and 232 on eBGP sessions to AS2 BRs 121 and 122.
- * ASBR 121 and 122 propagate the route in AS2 to all the P, ABRs and PEs through transport RR
- * Every router in AS2 resolves BGP CAR prefix B:C11::/32 next hops IP1 and IP2 in ISISv6 FlexAlgo 128 and programs B:C11::/32 prefix in global IPv6 forwarding table
- * AIGP attribute influences BGP CAR route best path decision
- * Egress PE E2 advertises a VPN route RD:V/v with SRv6 service SID B:C11:2:DT4::. Service SID is allocated by E2 from its locator of color C1 intent
- * Ingress PE E1 learns (via service RRs S-RR1 and S-RR2) VPN route RD:V/v with SRv6 SID B:C11:2:DT4::
- * Service traffic encapsulated with SRv6 Service SID B:C11:2:DT4:: is natively steered hop by hop along IPv6 routed path to B:C11::/32 provided by BGP CAR in AS2
- * Encapsulated service traffic is natively steered along IPv6 routed path to B:C11::/32 provided by ISISv6 FlexAlgo 128 in AS1
- * Design applies to multiple ASNs. BGP next hop is rewritten across a eBGP hop.

Important:

- * No tunneling/encapsulation on Ingress PE and BRs for BGP CAR provided transport.
- * Uses longest prefix match of SRv6 service SID to BGP CAR IP prefix. No mapping to labels/SIDs, instead use of simple IP based forwarding.

Packet forwarding

```
@E1: IPv4 VRF V/v => H.Encaps.red <B:C11:2:DT4::> => forward based on
                                     B:C11::/32
@P*: IPv6 table: B:C11::/32 => forward to interface, NH
@121: IPv6 Table: B:C11::/32 => forward to interface, NH
@231: IPv6 table: B:C11:2::/48 :: => forward via ISISv6 FA path to E2
@231: IPv6 Table B:C11:2::/48 => forward via ISISv6 FA path to E2
@E2: My SID table B:C11:2:DT4:: =>pop the outer header and lookup the
                                     inner DA in the VRF
```

[C.2.](#) BGP CAR SRv6 locator reachability distribution with encapsulation

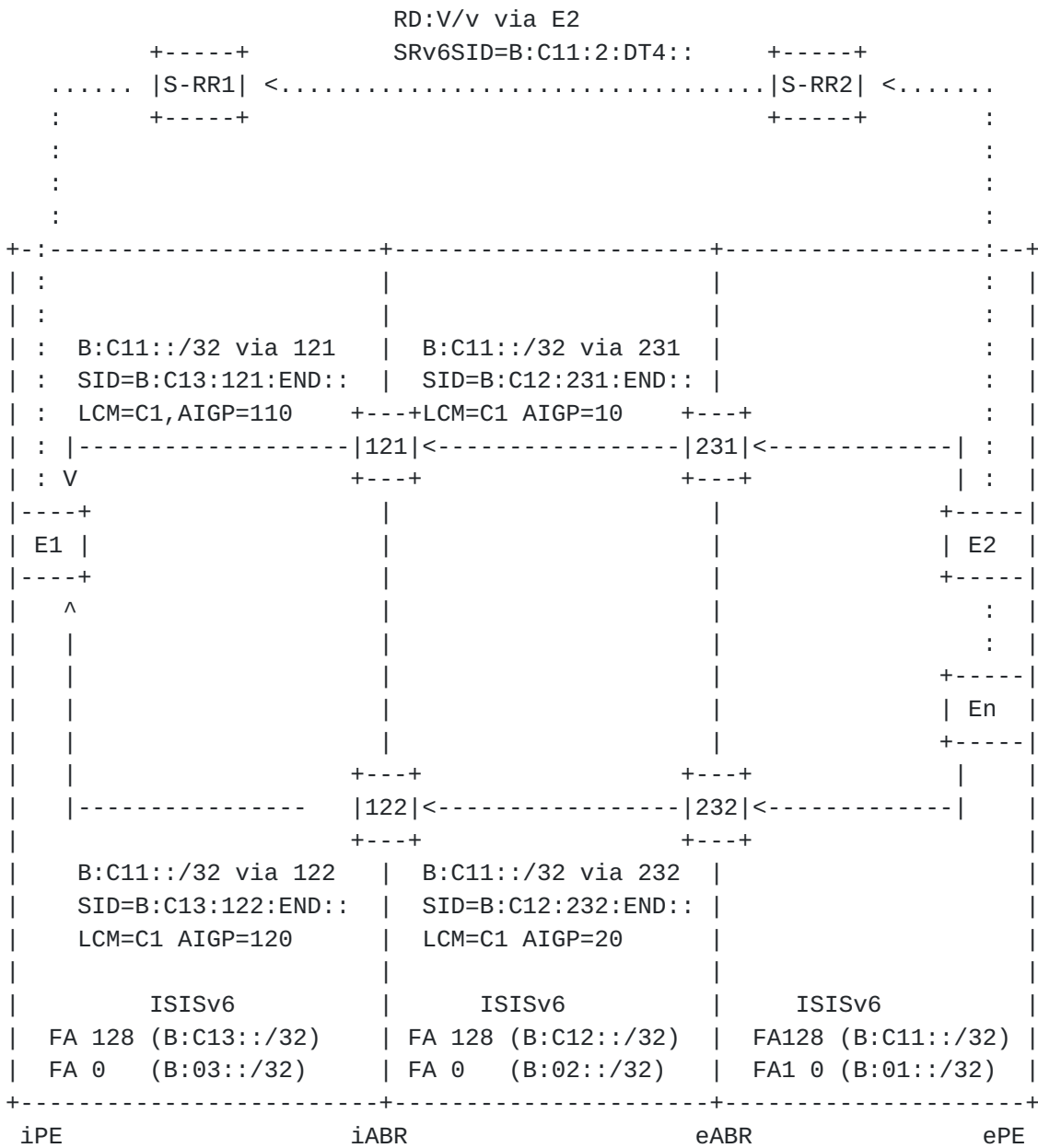


Figure 14

The topology above is an example to illustrate the BGP CAR SRv6 locator prefix route based design (Routed Service SID: [Section 7.1.1](#)), with intra-domain encapsulation. The example shown is iBGP, but also applies to eBGP (multi-AS).

* IGP FlexAlgo 128 is running in each domain

- Prefix B:C11::/32 summarizes Flex-Algo 128 block in egress domain for the given intent. Node locators in the egress domain are sub-allocated from the block

- Prefix B:C12::/32 summarizes FA128 block in transit domain
- Prefix B:C13::/32 summarizes FA128 block in ingress domain
- * BGP CAR route B:C11::/32 is originated by ABRs 231 and 232 with LCM C1. Along the propagation path, border routers set next-hop-self and appropriately update the intra-domain encapsulation information for the C1 intent For example, 231 and 121 signal SRv6 SID allocated from their respective locators for the C1 intent. (Note, IGP FlexAlgo is shown for intra-domain path, but SR-Policy may also provide the path.)
- * AIGP attribute influences BGP CAR route best path decision
- * Egress PE E2 advertises a VPN route RD:V/v with SRv6 service SID B:C11:2:DT4::. Service SID is allocated by E2 from its locator of color C1 intent.
- * Ingress PE E1 learns CAR route B:C11::/32 and VPN route RD:V/v with SRv6 SID B:C11:2:DT4::
- * Traffic encapsulated with SRv6 Service SID B:C11:2:DT4:: is steered along IPv6 routed path provided by BGP CAR IP prefix route to locator B:C11::/32

Important

- * Uses longest prefix match of SRv6 service SID to BGP CAR prefix. No mapping labels/SIDs, instead simple IP based forwarding.
- * Originating domain PE locators of the given intent can be summarized on transit BGP hops eliminating per PE state on border routers.

Packet forwarding

```
@E1:  IPv4 VRF V/v => H.Encaps.red <B:C13:121:END::, B:C11:2:DT4::>
@121: My SID table: B:C13:121:END:: => Update DA with B:C11:2:DT4::
@121: IPv6 Table: B:C11::/32 => H.Encaps.red <B:C12:231:END::>
@231: My SID table: B:C12:231:END:: => Update DA with B:C11:2:DT4::
@231: IPv6 Table B:C11:2::/48 => forward via ISISv6 FA path to E2
@E2: My SID table B:C11:2:DT4:: =>pop the outer header and lookup the
      inner DA in the VRF
```

C.3. BGP CAR (E, C) route distribution

propagates through border routers. At each BGP hop, BGP CAR prefix next-hop resolution triggers intra-domain transit SR policy (C2, CAR next hop). For example:

- SR policy (231, C2) with segments <B:02:y:END::, B:02:231:END::>
 - SR policy (121, C2) with segments <B:03:x:END::, B:03:121:END::>
 - x and y are node ids within the respective domains
- * Egress PE E2 advertises a VPN route RD:V/v with BGP color extended community C2
 - * Ingress PE E1 steers VPN route from E2 onto BGP CAR route (E2, C2) that results in H.Encaps of SRv6 transport SID B:C21:2:B6:: and SRv6 service SID in as last segment in IPv6 header.
 - * Longest prefix match on CAR prefix B:C21::/32 of the IPv6 destination B:C21:2:B6:: steers the packet along intent aware path from ingress PE through ABRs in transit domain
 - * IPv6 packet destination B:C21:2:B6:: lookup in mySID table on ABR 231 or 232 results in END.B6 behavior i.e. push of policy segments to E2.

Important

- * Ingress PE steers services via (E,C) CAR route as per [[RFC9256](#)]
- * In data plane (E,C) resolution results in IPv6 header destination being SRv6 SID of END.B6 behavior whose locator is of given intent on originating ABRs.
- * CAR IP prefix route along the transit path provides simple LPM IPv6 forwarding along the transit BGP hops.
- * CAR NLRI Type-2 prefix summarizes binding SIDs of all SR policies on originating ABR of a given intent to different PEs in egress domain. This eliminates per PE state on transit routers

Packet forwarding


```
@E1: IPv4 VRF V/v => H.Encaps.red <<SR policy (C2,121) sid list>,
      B:C21:2:B6::, B:0:E2:DT4::>
@121: My SID table: B:03:121:END:: => Update DA with B:C21:2:B6::
@121: IPv6 Table: B:C21::/32 => H.Encaps.red <SR Policy (C2,231) sid
      list>
@231: My SID table: B:02:231:END:: => Update DA with B:C21:2:B6::
@231: MySIDtable B:C21:2:B6:: => H.Encaps.red <SR Policy (C2,E2) sid
      list>
@E2: IPv6 Table B:0:2:DT4:: =>pop the outer header and lookup the
      inner DA in the VRF
```

Appendix D. CAR SAFI NLRI update packing efficiency calculation

CAR SAFI NLRI encoding is optimized for update packing. It allows per route information (example label, label index and SRv6 SID encapsulation data) to be carried in non-key TLV part of NLRI. This allows multiple NLRIs to be packed in single update message when other attributes are shared. Analysis below shows comparison of total BGP data on the wire for CAR SAFI and [\[RFC8277\]](#) style encoding in MPLS label (case a), SR extension with MPLS (per-prefix label-index in Prefix-SID attribute) [\[RFC8669\]](#) (case b) and SRv6 SID (case c) cases. Scenarios considered are ideal packing (maximum number of routes packed to update message limit of 4k bytes), practical deployment case with average packing (5 routes share set of BGP path attributes and hence packed in single update message) and worst-case of no packing (each route in separate update message).

Summary of ideal, practical and no-packing BGP data in each case

Encoding	BGP CAR NLRI	RFC-8277 style NLRI	Result
-----+-----+-----+-----			
case a: Label (Ideal)	27.5 MB	26 MB	
(Practical)	86 MB	84 MB	No degradation from RFC8277 like encoding
(No packing)	325 MB	324 MB	
-----+-----+-----+-----			
case b: Label & Label-index (Ideal)	42 MB	339 MB Packing not possible	CAR SAFI encoding more efficient by 88% in best case and 71% in average case over
(Practical)	99 MB	339 MB Packing not possible	RFC8277 style encoding (which precludes packing)
(No packing)	339 MB	339 MB	
-----+-----+-----+-----			
case c: SRV6 SID (Ideal)	49 MB	378 MB	Results are similar to SR MPLS case. Transposition provides
(Practical)	115 MB	378 MB	further 20% reduction in BGP data.
(No packing)	378 MB	378 MB	
-----+-----+-----+-----			

Analysis considers 1.5 million routes (5 colors across 300k endpoints)

case a: BGP data exchanged for non SR MPLS case

Consider 200 bytes of shared attributes

CAR SAFI signal Label in non-key TLV part of NLRI

Each NLRI size for AFI 1 = 12(key) + 5(label) = 17 bytes

Ideal packing:

number of NLRIs in 4k update size = 223 (4k-200/17)

number of update messages of 4k size = 1.5 million/223 = 6726

Total BGP data on wire = 6726 * 4k = ~27.5MB

Practical packing (5 routes in update message)

size of update message = (17 * 5) + 200 = 285

Total BGP data on wire = 285 * 300k = ~86MB

No-packing case (1 route per update message)

size of update message = 17 + 200 = 217

Total BGP data on wire = 217 * 1.5 million = ~325MB

SAFI 128 8277 style encoding with label in NLRI

Each NLRI size for AFI 1 = 13(key) + 3(label) = 16 bytes

Ideal packing:

number of NLRIs in 4k update size = 237 (4k-200/16)

number of update messages of 4k size = 1.5 million/237 = ~6330

Total BGP data on wire = 6330 * 4k = ~25.9MB

Practical packing (5 routes in update message)

size of update message = (16 * 5) + 200 = 280

Total BGP data on wire = 280 * 300k = ~84MB

No-packing case (1 route per update message)

size of update message = 16 + 200 = 216

Total BGP data on wire = 216 * 1.5 million = ~324MB

case b: BGP data exchanged for SR label-index

Consider 200 bytes of shared attributes

CAR SAFI signal Label in non-key TLV part of NLRI

Each NLRI size for AFI 1

$$= 12(\text{key}) + 5(\text{label}) + 9(\text{Index}) = 26 \text{ bytes}$$

Ideal packing:

$$\text{number of NLRIs in 4k update size} = 146 \text{ (4k-200/26)}$$

$$\text{number of update messages of 4k size} = 1.5 \text{ million}/146 = 6726$$

$$\text{Total BGP data on wire} = 10274 * 4k = \sim 42\text{MB}$$

Practical packing (5 routes in update message)

$$\text{size of update message} = (26 * 5) + 200 = 330$$

$$\text{Total BGP data on wire} = 330 * 300k = \sim 99\text{MB}$$

No-packing case (1 route per update message)

$$\text{size of update message} = 26 + 200 = 226$$

$$\text{Total BGP data on wire} = 226 * 1.5 \text{ million} = \sim 339\text{MB}$$

SAFI 128 8277 style encoding with label in NLRI

Each NLRI size for AFI 1 = 13(key) + 3(label) = 16 bytes

Ideal packing

Not supported as label index is encoded in prefix SID attribute

Practical packing (5 routes in update message)

Not supported as label index is encoded in prefix SID attribute

No-packing case (1 route per update message)

$$\text{size of update message} = 16 + 210 = 226$$

$$\text{Total BGP data on wire} = 216 * 1.5 \text{ million} = \sim 339\text{MB}$$

case c: BGP data exchanged with 128 bit single SRv6 SID

Consider 200 bytes of shared attributes

CAR SAFI signal Label in non-key TLV part of NLRI

Each NLRI size for AFI 1 = 12(key) + 18(Srv6 SID) = 30 bytes

Ideal packing:

number of NLRIs in 4k update size = 126 (4k-200/30)

number of update messages of 4k size = 1.5 million/126 = ~12k

Total BGP data on wire = 12k * 4k = ~49MB

Practical packing (5 routes in update message)

size of update message

= (30 * 5) + 236 (including prefix SID) = 386

Total BGP data on wire = 386 * 300k = ~115MB

No-packing case (1 route per update message)

size of update message = 12 + 236 (SID in prefixSID) = 252

Total BGP data on wire = 252 * 1.5 million = ~378MB

SAFI 128 8277 style encoding with label in NLRI (No transposition)

Each NLRI size for AFI 1 = 13(key) + 3(label) = 16 bytes

Ideal packing

Not supported as label index is encoded in prefix SID

attribute

Practical packing (5 routes in update message)

Not supported as label index is encoded in prefix SID

attribute

No-packing case (1 route per update message)

size of update message = 16 + 236 = 252

Total BGP data on wire = 252 * 1.5 million = ~378MB

BGP data exchanged with SRv6 SID 4 bytes transposition into SRv6 SID TLV

Consider 200 bytes of shared attributes

CAR SAFI signal Label in non-key TLV part of NLRI

Each NLRI size for AFI 1 = 12(key) + 6(Srv6 SID) = 18 bytes

Ideal packing:

number of NLRIs in 4k update size = 211 (4k-200/18)

number of update messages of 4k size = 1.5 million/211 = ~7110

Total BGP data on wire = 7110 * 4k = ~29MB

Practical packing (5 routes in update message)

size of update message

= (18 * 5) + 236 (including prefix SID) = 326

Total BGP data on wire = 326 * 300k = ~98MB

No-packing case (1 route per update message)

size of update message

= 12 + 236 (SID in prefix SID attribute) = 252

Total BGP data on wire = 252 * 1.5 million = ~378MB

Authors' Addresses

Dhananjaya Rao (editor)
Cisco Systems
United States of America
Email: dhrao@cisco.com

Swadesh Agrawal (editor)
Cisco Systems
United States of America
Email: swaagraw@cisco.com

Co-authors

[Section 11](#)

Email: dhananjaya.rao@gmail.com

