BESS WorkGroup
Internet-Draft
Intended status: Informational
Expires: August 13, 2021

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

February 9, 2021

BGP Color-Aware Routing Problem Statement
draft-dskc-bess-bgp-car-problem-statement-01

Abstract

This document explores the scope, use-cases and requirements for a BGP based routing solution to establish end-to-end intent-aware paths across a multi-domain service provider network environment.

Table of Contents

1. Introduction

1.1. Objective

This document explores the scope, use-cases and requirements for a BGP based routing solution to establish end-to-end intent-aware paths across a multi-domain service provider network environment.

The targeted design outcome is to define the technology and protocol extensions that may be required in a manner that addresses the widest application.

To introduce the problem space that the document focuses on, let us start with the BGP-based delivery of an intent across several SR-MPLS/MPLS domains.

```
          +-----+                +-----+     V/v via PE2  +-----+
   .......|S-RR1|<...............|S-RR2|<.................|S-RR3| <..
   :      +-----+                +-----+     Color C      +-----+   :
   :                                                                :
   :                                                                :
+--:----------+-------------+--------------+-------------+----------:--+
|  :          |             |              |             |          :  |
|  :          |             |              |             |          :  |
|  :        +---+         +---+          +---+         +---+        :  |
|  :        |121|         |231|          |341|         |451|        :  |
|  :        +---+         +---+          +---+         +---+        :  |
|----+        |             |              |             |        +----|
| PE1|        |             |              |             |        |PE2 |
|----+        |             |              |             |        +----|
|           +---+         +---+          +---+         +---+           |
|           |122|         |232|          |342|         |452|           |
|           +---+         +---+          +---+         +---+           |
|  Access     |   Metro     |   Core       |   Metro     |   Access    |
|  domain 1   |   domain 2  |   domain 3   |   domain 4  |   domain 5  |
+-------------+-------------+--------------+-------------+-------------+
iPE         iBRM          iBRC           eBRC           eBRM         ePE
```

Figure 1: Reference large-scale multi-domain network topology

The figure above shows a reference large-scale multi-domain network topology. PE1 and PE2 are PEs; the other nodes are border routers (BR) between domains in different tiers of the network. A VPN route is advertised via service RRs (S-RR) between an egress PE (PE2) and an ingress PE (PE1).

BGP must provide reachability from PE1 to PE2 based on various intent. For instance, BGP may provide reachability to PE2 using either low latency or best effort.

A VPN route having a requirement of low latency routing will select the BGP reachability information to PE2 that is based on low latency.

The problem space is then widened to include any intent (including NFV chains and their location), any dataplane and the application of

the intent-based routing to the Service/VPN routes. All of this is detailed in the rest of the document.

1.2. State-of-the-art

The following solution is widely deployed - [I-D.ietf-spring-segment-routing-policy]:

- o In reference figure above, an Egress PE PE2 advertises a BGP VPN route V/v with a BGP Color Extended Community C [I-D.ietf-idr-tunnel-encaps] to indicate the service intent that PE2 requests for the traffic bound to V/v. Note: The Color Extended Community may be applied to any BGP service route. For simplicity in this document, we will use a VPN route example.

- o An ingress PE1 steers V-destined packets onto an SR Policy bound to (C, PE2).

- o C may express any of the following requirements:

  - * Minimization of a cost metric vs a latency metric.

  - * Exclusion/Inclusion of SRLG and/or Link Affinity.

    - + In inter-domain context, exclusion/inclusion of entire domains.

  - * Inclusion of virtual network function chains [I-D.ietf-spring-sr-service-programming].

- o An SR-PCE (or a set of them) computes the end-to-end path and installs it at PE1 as an SR Policy. The end-to-end path may seamlessly cross multiple domains.

The SR-PCE solution being defined at the IETF [RFC8664]and being widely deployed is reminded in this introduction as a useful "state-of-the-art" context to consider when defining the BGP-based alternative solution.

1.2.1. Color

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

1.2.2. Colored vs Color-Aware

The solution must support the ability to distinguish BGP routes that require the usage of a particular intent from BGP routes that are actually satisfying a particular intent. Hence, this document defines the notion of colored and color-aware routes.

- o Colored: Egress PE PE2 colored its BGP VPN route V/v to indicate the intent that it requests for the traffic bound to V/v.

- o Color-Aware: A new BGP solution which signals multiple "ways" to reach a given destination (e.g. PE2)

- o Steering a colored VPN route to a color-aware route

  * If PE2 signals a VPN route V/v with color C

  * If PE1 installs that VPN route

  * If PE1 learns about a BGP Color-Aware Route R/r to PE2 for color C

  * Then PE1 steers packets destined to V/v via R/r

- o Note the similarity with the state-of-the-art reference:

  * The steering onto an SR Policy bound to (C, PE2) is replaced by the steering on a Color-Aware BGP route (C, PE2)

  * The data model is the same "resolution via (C, PE2)"

  * The difference is how the (C, PE2) path is obtained: BGP signaling vs SR-PCE signaling

1.2.3. Per-Destination and Per-Flow Steering

Ingress PE PE1 steers packets destined for a service (VPN) route V/v via BGP Color-Aware Route R/r to PE2

- o Per-Destination Steering: Incoming packets on PE1 match BGP service route V/v to be steered based on the destination IP address of the packets.

- o Per-Flow Steering: Incoming packets on PE1 match BGP service route V/v to be steered based on the combination of the destination IP address and additional elements in the packet header (i.e., IP flow). Such a packet lookup may recurse on a forwarding array where some of the entries are BGP color-aware routes to PE2. A

given flow is mapped to a specific entry in this array i.e. via a specific BGP color-aware route to PE2.

1.3. Why a BGP-based alternative is needed

- An operator with an existing Seamless-MPLS/BGP-LU deployment [I-D.ietf-mpls-seamless-mpls] may consider a BGP based extension as a more incremental approach.

- There may be an expectation that BGP would support a larger scale.

- Opacity of a remote domain due to trust boundaries within an inter-domain construction.

1.4. Color Domains

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

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

- Color Domain: A color domain may represent a collection of one or more network domains with a single, consistent color/intent mapping.

- Color re-mapping may happen at color domain boundaries.

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

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

1.5. BGP Color-Aware Routing

A BGP solution that is solving this problem statement is called BGP Color-Aware Routing, and is referred to as BGP CAR in this document.

2. Intent bound to a Color

The BGP CAR solution must support the following intents bound to a color:

- Minimization of a cost metric vs a latency metric

      * Minimization of different metric types, static and dynamic

   o Exclusion/Inclusion of SRLG and/or Link Affinity and/or minimum MTU

   o In the inter-domain context, exclusion/inclusion of entire domains

   o Inclusion of one or several virtual network function chains

      * Located in a regional domain and/or core domain, in a DC

   o Localization of the virtual network function chains

      * Some functions may be desired in the regional DC or vice versa

   o Per-Destination and Per-Flow steering

3. BGP CAR Use-cases

The BGP CAR route may be a transport route or a service route (in this document, we use the term VPN instead of service for simplicity).

3.1. BGP Transport CAR

   o Transport Intent

      * Intent-aware routing between PEs connected across multiple transit domains

         + Set up BGP based end-to-end paths stitching intent-aware intra-domain segments

   o The network diagram below illustrates the reference network topology used in this section for Transport CAR:

```
        +-----+                +-----+                  +-----+
   .....|S-RR1| <............. |S-RR2| <............... |S-RR3| <...
   :    +-----+                +-----+                  +-----+    :
   :                                                               :
   :                                                               :
   :                                                               :
+--:--------------+       +-----------------+       +--------------:--+
|  :              |       |                 |       |              :  |
|  :              |       |                 |       |              :  |
|  :          +---|  D=20 |---+         +---|  D=25 |---+          :  |
|  :          |121|-------|211|         |231|-------|321|          :  |
|  :          +---| \   / |---+         +---| \   / |---+          :  |
|----+            |  \ /  |                 |  \ /  |            +----|
|PE11|            |   V   |                 |   V   |            |PE31|
|----+            |  / \  |                 |  / \  |            +----|
|             +---| /   \ |---+         +---| /   \ |---+             |
|             |122|-------|212|         |232|-------|322|             |
|             +---|  D=15 |---+         +---|  D=10 |---+             |
|                 |       |                 |       |                 |
|    Domain 1     |       |    Domain 2     |       |    Domain 3     |
+-----------------+       +-----------------+       +------------------+
```

Figure 2: Transport CAR Reference Topology

The following network design assumptions apply to the reference topology above, as an example:

* Independent ISIS/OSPF SR instance in each domain.

* eBGP peering link between ASBRs (121-211, 121-212, 122-211, 122-212, 231-321, 231-322, 232-321 and 232-322).

* Peering links have equal cost metric.

* Peering links have delay configured or measured as shown by "D". D=50 for cross peering links.

* VPN service is running from PE31 to PE11 via service RRs (S-RRn in figure).

o The following sections illustrate a few examples of intent use-cases applicable to transport routes.

3.1.1. Use-case of minimization of a cost metric vs a latency metric

o In the reference topology of Figure 2

   Each domain has Algo 0 and Flex Algo 128

Algo 0 is for minimum cost metric(cost optimized).

Flex Algo 128 definition is for minimum delay (low latency).

- o Cost Optimized

  * Color C1 - Minimum cost intent. (Here, a BGP CAR route with Color C1 is being used, instead of BGP-LU.)

  * On PE11, VPN routes colored with C1 are steered via (C1, PE31) BGP CAR route

    + BGP CAR for C1 sets up path(s) between PEs for end-to-end minimum cost.

    + (2) These paths traverse over intra-domain Algo 0 in each domain and account for the peering link cost between ASBRs.

    + Example: PE11 learns (C1, PE31) CAR route via several equal paths:

      1. One such path is through FA0 to node 121, links 121-211, FA0 to 231, link 231-321, FA0 to PE31

      2. Another such path is through FA0 to node 122, link 122-212, FA0 to 232, link 232-322, FA0 to PE31.

- o Minimize latency

  * Color C2 - Minimum latency intent.

  * On PE11, VPN routes colored with C2 are steered via (C2, PE31) BGP CAR route.

    + BGP CAR for C2 advertises paths between PEs for minimum end-to-end delay.

    + (2) These paths traverse over intra-domain Flex Algo 128 in each domain and account for peering link delay between ASBRs.

    + (3) Example: PE11 learns (C2, PE31) BGP CAR route and best path is through FA128 to node 122, link 122-212, FA128 to 232, link 232-322, FA128 to PE31.

3.1.2. Use-case of exclusion/inclusion of link affinity

- o Color C3 - Intent to Minimize cost metric and avoid purple links

- o In the reference topology of Figure 2

  Each domain has Flex Algo 129 and some links have purple affinity.

  Flex Algo 129 definition is set to minimum cost metric and avoid purple links (within domain).

  Peering cross links are colored purple by policy.

- o On PE11, VPN routes colored with C3 are steered via (C3, PE31) BGP CAR route.

  * BGP CAR for C3 sets up paths between PEs for minimum end-to-end cost and avoiding purple link affinity.

  * These paths traverse over intra domain Flex Algo 129 in each domain and accounts for peering link cost between ASBR and avoiding purple links.

  * Example: PE11 learns (C3, PE31) BGP CAR route via 2 paths.

    1. First path is through FA 129 to node 121, link 121-211, FA129 to 231, link 231-321, FA129 to PE31.

    2. Second path is through FA129 to node 122, link 122-212, FA129 to 232, link 232-322, FA129 to PE31.

3.1.3. Use-case of exclusion/inclusion of domains

```
         +-----+                +-----+                 +-----+
     ....|S-RR1| <............. |S-RR2| <.............. |S-RR3| <....
     :   +-----+                +-----+                 +-----+     :
     :                                                              :
     :                      +----------------+                      :
     :                      |                |                      :
  +--:--------------+       |---+        +---|       +--------------:--+
  |  :              |   |---|211|        |241|---|   |              :  |
  |  :              |   |   |---+        +---|   |   |              :  |
  |  :          +---|   |   |    Domain 2    |   |   |---+          :  |
  |  :          |121|---|   +----------------+   |---|421|          :  |
  |  :          +---|                                |---+          :  |
  |----+            |                                |            +----|
  |PE11|            |                                |            |PE41|
  |----+            |                                |            +----|
  |             +---|                                |---+             |
  |             |131|---|   +----------------+   |---|431|             |
  |             +---|   |   |                |   |   |---+             |
  |                 |   |   |---+        +---|   |   |                 |
  |    Domain 1     |   |---|311|        |341|---|   |      Domain 4   |
  +-----------------+       |---+        +---|       +-----------------+
                            |   Domain 3     |
                            +----------------+
```

Figure 3

Color C4 - Avoid sending selected traffic via Domain 3

o VPN routes advertised from PEs with Color C4

o BGP CAR for Color C4 should only set up paths between PE11 and PE41 that exclude Domain 3

3.1.4. Use-case of virtual network function chains in local and core domains

```
          ____                      ____
         /    \                    /    \
        | NFV1 |                  | NFV2 |
         \    /                    \    /
  +---------------------+  +--------------------+  +-------------------+
  |      |E11|          |  |       |E21|        |  |                   |
  |      +---+          |  |       +---+        |  |                   |
  |                     |  |                    |  |                   |
  |                     |  |                    |  |                   |
  |----+              +------+              +------+            +----|
  |PE11|              | 121  |              | 231  |            |PE31|
  |----+              +------+              +------+            +----|
  |                     |  |                    |  |                   |
  |                     |  |                    |  |                   |
  |                     |  |                    |  |                   |
  |                     |  |                    |  |                   |
  |      Domain 1       |  |      Domain 2      |  |     Domain 3      |
  +---------------------+  +--------------------+  +-------------------+
```

Figure 4

- o Color intent

  - * C5 - Routing via min-cost paths

  - * C6 - Routing via a local NFV service chain situated at E11

  - * C7 - Routing via a centrally located NFV service chain situated at E21

- o Forwarding of packets from PE11 towards PE31:

  - * (C5, PE31) mapped packets are sent via nodes 121, 231 to PE31

  - * (C6, PE31) mapped packets are sent to E11 and then post-service chain, via 121, 231 to PE31

  - * (C7, PE31) mapped packets are sent via 121 to E21 and then post-service chain, via 231 to PE31

3.2. BGP VPN CAR

- o VPN (Service layer) intent

  - * Extend the signaling of intent awareness end-to-end: CE site to CE site across provider networks

+ Provide ability for a CE to select paths through specific PEs for a given intent

  - Example-1: Certain intent in transport not available via specific PEs

  - Example-2: Certain CE-PE connection does not support specific intent

  - Example-3: Site access via certain CE does not support specific intent. For instance, link connecting a specific CE to a DC hosting loss-sensitive service may have better quality than a link from another CE

+ Provide ability for a CE to send traffic indicating a specific intent (via suitable encapsulation) to the PE for optimal steering.

* Intent aware routing support for multiple service (VPN) interworking models

  + Beyond options such as iBGP or Inter-AS Option C that inherently extend from PE to PE

    1. Inter-AS Option A

    2. Inter-AS Option B

    3. GW based interworking(L3VPN, EVPN)

  + Interworking with existing L3VPN deployments, both PEs and CEs

o The network diagram below illustrates the reference network topology used in this section for VPN CAR.

```
        +-------------------+           +-------------------+
        |   ....|S-RR|....  |           |  ....|S-RR|.....  |
        |   :   +----+   :  |           |  :   +----+    :  |
        |   :    :  :    :  |           |  :    :  :     :  |
        |----+   :  :   +---|   D=20    |---+   :  :   +----|
       /|PE11|   :  :   |121|-----------|211|   :  :   |PE21|\
 D=25/  |----+   :  :   +---| X       X |---+   :  :   +----|  \ D=25
   /    |        :  :       |  X   X    |       :  :        |   \   V/24
CE1     |        :  :       |    X D=50|       :  :        |    CE2<---
   X    |        :  :       |  X   X    |       :  :        |   X
D=15 X  |----+   :  :   +---| X       X |---+   :  :   +----|  X D=10
       X|PE12|...:  :...|212|-----------|232|...:  :...|PE22|X
        |----+          +---|   D=10    |---+          +----|
        |                   |           |                   |
        |     AS 1          |           |     AS 2          |
        +-------------------+           +-------------------+
```

Figure 5: VPN CAR reference topology

The following network design assumptions apply to the reference topology above, as an example:

* Independent ISIS/OSPF SR instance in each AS.

* eBGP peering link between VPN ASBRs 121-211, 121-212, 122-211, 122-212.

* VPN service is running between PEs via service RRs in each AS to local ASBRs. Between ASBRs, its Option-B i.e. next hop self for VPN SAFI.

* CE1 is dual homed to PE11 and PE12. Similarly, CE2 is dual homed to PE21 and PE22.

* Peering links have equal cost metric

* Peering links have delay configured or measured as shown by "D".

* CE2 advertises prefix V/24 to CE1. It is advertised as RD:V/24 between PEs, including color-awareness

o The following sections illustrate a few examples of intent use-cases applicable to VPN (service) routes.

3.2.1. Use-case of minimization of a cost metric vs a latency metric

o In the reference topology of Figure 5

Each AS has Flex Algo 0 and 128.

Flex Algo 0 is for minimum cost metric(cost optimized).

Flex Algo 128 definition is for minimum delay (low latency).

o Cost Optimized

* Color C1 - Minimum cost intent.

* On CE1, flows requiring cost optimized paths to V/24 are steered over (C1, V/24) route.

  + BGP CAR for C1 sets up paths between CEs for minimum end-to-end cost.

  + This advertisement needs BGP CAR between PE-CE for V/24 prefix and color C1 awareness.

  + It also needs BGP VPN CAR between PEs and ASBRs for RD:V/24 prefix and color C1 awareness (C1, RD:V/24).

  + Paths traverse over PE-CE links, intra-domain Flex Algo 0 in each AS and peering links between ASBRs, minimizing cost for VPN.

  + Example: CE1 learns (C1, V/24) CAR route through several equal cost paths:

    1. One path is through link CE1-PE11, FA0 to 121, link 121-211, FA0 to PE21 and link PE21-CE2.

    2. Another such path is through CE1-PE12, FA0 to node 122, link 122-212, FA0 to PE22, link PE22-CE2.

o Minimize latency

* Color C2 - Minimum latency intent

* On CE1, flows requiring low latency paths to prefix V/24 are steered over (C2, V/24) CAR route.

  + BGP CAR for C2 sets up paths between CEs for minimum end-to-end delay.

+ This advertisement needs BGP CAR between PE-CE for V/24 prefix and color C2 awareness.

+ It also needs BGP VPN CAR between PEs and ASBR for RD:V/24 prefix and color C2 awareness (C2, RD:V/24).

+ Paths traverse over intra-domain Flex Algo 128 in each AS and accounts for inter ASBR link delays and PE-CE link delays for the VPN.

+ Example: CE1 learns (C2, V/24) CAR best route through link CE1-PE12, FA128 to 122, link 122-212, FA128 to PE22 and link PE22-CE2.

3.2.2. Use-case of exclusion/inclusion of link affinity

o Color C3 - Intent to Minimize cost metric and avoid purple links

o In the reference topology of Figure 5

Each AS has Flex Algo 129 and some links have purple affinity.

Flex Algo 129 definition is set to minimum cost metric and avoid purple links (within AS).

ASBR cross links are colored purple by policy. Bottom PE-CE links are colored purple as well by policy

o On CE1, flows requiring minimum cost path avoiding purple links to V/24 are steered over (C3, V/24) BGP CAR route.

* BGP CAR for C3 setup paths between CEs for minimum end-to-end cost and avoiding purple link affinity.

* This advertisement needs BGP CAR between PE-CE for V/24 prefix and color C3 awareness

* It also needs BGP VPN CAR between PEs and ASBRs for RD:V/24 prefix and color C3 awareness (C3, RD:V/24).

* The path avoids purple PE-CE links, traverses over intra-domain Flex Algo 129 in each AS and avoids purple links between VPN ASBRs.

* Example: CE1 learns (C3, V/24) CAR route through link CE1-PE11, FA129 to 121, link 121-211, FA129 to PE21 and link PE21-CE2.

3.2.3. Use-case of virtual network function chains in local and core domains

```
                               +-----+
       ........................|S-RR | <.................
       :                       +-----+ <...........     :
       :                                          :     :
       :                                          :     :
       :  ___     ___           ___               :     :
       : /   \   /   \         /   \              :     :
       :| S1  | | S2  |       | S3  |             :     :
       : \   /   \   /         \   /              :     :
     +-:---------------+  +--------------+  +------:----:--+
     | :  |E11|  |E12| |  |    |E21|     |  |      :    :  |
     | :  +---+  +---+ |  |    +---+     |  |      :  +----|<-(V1/24,C1)
     | :               |  |              |  |      :  |PE31|--CE2
     | V               |  |              |  |      :  +----|
     |----+          +-------+         +-------+   :       |
CE1--|PE11|          | 121   |         | 231   |   :       |
     |----+          +-------+         +-------+   :  +----|<-(V2/24/C2)
     |                 |  |              |  |      :..|PE32|--CE3
     |                 |  |              |  |         +----|
     |                 |  |              |  |              |
     |                 |  |              |  |              |
     |    Domain 1     |  |   Domain 2   |  |   Domain 3   |
     +-----------------+  +--------------+  +--------------+
```

Figure 6

- o Color intent

  * C1 - Routing via NFV service chain comprising of [S1, S2] attached to E11 and E12

  * C2 - Routing via NFV service [S3] attached to E21

- o CE1, CE2, CE3 are sites of VPN1.

- o Prefix V1/24 colored with C1 from CE2, and advertised as RD:V1/24 with C1 by PE31 to PE11 via S-RR

- o Prefix V2/24 colored with C2 from CE3, and advertised as RD:V2/24 with C2 by PE32 to PE11 via SS-RR

- o From PE11:

* [V1/24, C1] mapped packets are sent via S1, S2 and then routed to PE31, CE2

* [V2/24, C2] mapped packets are sent via S3 and then routed to PE32, CE3

4. Deployment Requirements

- Co-existence, compatibility and interworking with currently deployed SR-PCE based multi-domain color-aware solution

- Support different multi-domain deployment designs

  * Multiple IGP domains within a single AS (Seamless MPLS)

    + Inter-connect at node level (ABR)

  * Multiple BGP AS domains

    + Inter-connect via peering links (ASBR)

- Support end-to-end path crossing transport domains with different technologies and encapsulations

  * LDP-MPLS

  * RSVP-TE-MPLS

  * SR-MPLS

  * SRv6

  * IPv4/IPv6

- Support interworking between domains with different encapsulations (e.g, SR-MPLS and SRv6)

- Support multiple transport encapsulations within a domain for co-existence and migration

- Provide a BGP-based control-plane solution for the use-case illustrated in [RFC8604] together with deployment design guidelines for the leverage of anycast and binding SIDs.

5. Scalability

5.1. Scale Requirements

- o Support for massive scaled transport network

  - * Number of Remote PE's: >= 300k

  - * Number of Colors C: >= 5

- o Scalable MPLS dataplane solution

  - * With one label per (C, Remote PE), the 1M MPLS dataplane does not work.

  - * A notion of hierarchy or segment list is required.

    - + E.g. the SR-PCE builds the end-to-end path as a list of segments such that no single node needs to support a data-plane scaling in the order of (Remote PE * C)

    - + The solution is thus not a direct extension of BGP-LU

  - * Additionally, PE and transit nodes (ABRs) may be devices with limited forwarding table space

  - * Devices may have constraints on packet processing (e.g., label operations, number of labels pushed) and performance

- o Ability to abstract the topology from remote domains - for scale, stability and faster convergence

  - * Abstracting PE and/or ABR related state and network events

- o Support for an Emulated-PULL model for the BGP signaling

  - * The SR-PCE solution natively supports a PULL model: when PE1 installs a VPN route V/v via (C, PE2), PE1 requests its serving SR-PCE to compute the SR Policy to (C, PE2). I.e. PE1 does not learn unneeded SR policies.

  - * BGP Signaling is natively a PUSH model.

  - * Emulated-PULL refers to the ability for a BGP CAR node PE1 to "subscribe" to (C, PE2) route such that only the related paths are signaled to PE1.

* The subscription and related filtering solution must apply to any BGP CAR node

  + Transport CAR routes

    1. Ability for a node (PE/ABR/RR) to signal interest for routes of specific colors.

    2. PEs only learn routes that they need - remote VPN endpoints (PEs/ASBRs) or transit nodes (ABRs, ASBRs).

    3. ABRs also only learn and propagate routes they need locally in domain

  + Service/VPN CAR routes

    1. Ability for a node (PE) to signal interest for a specific (Egress PE, Color) transport route

    2. CEs learn routes that they need - interested colors

    3. PEs learn routes that they need - interested VPNs, colors

  + Automation of the subscription/filter route

    1. Similar to the SR-PCE solution, when an ingress PE1 installs VPN V/v via (C, PE2), PE1 originates its subscription/filter route for (C, PE2).

  + Efficient propagation and processing of subscription/filter routes.

  + Ability to perform aggregation and suppression of subscription/filter routes at nodes in the route propagation path to reduce explosion and churn in propagation of the filter routes themselves.

  + The solution may be optional for networks that do not have the large scaling requirements

5.2. Scale Analysis

It is useful to analyze the multiple scaling requirements and specifically the data plane constraints in the context of a few common reference designs and use-cases.

A couple of example scenarios are listed below for reference.

o Seamless-MPLS design, with IGP Flex-Algo in each domain

```
                                    +-----+
     .............................. |S-RR | <........................
     :                              +-----+                          :
     :                                                               :
     :                                                               :
  +--:-----------------+  +--------------------+  +-----------------:--+
  |  :                 |  |                    |  |                 :  |
  |  :                 |  |                    |  |                 :  |
  |  :               +------+                +------+               :  |
  |  :               | 121  |                | 231  |               :  |
  |  V               +------+                +------+               :  |
  |----+                 |  |                    |  |               +----|
  |PE11|                 |  |                    |  |               |PE31|
  |----+                 |  |                    |  |               +----|
  |                  +------+                +------+                  |
  |                  | 122  |                | 232  |                  |
  |                  +------+                +------+                  |
  |                    |  |                    |  |                    |
  |      Domain 1      |  |      Domain 2      |  |     Domain 3       |
  +--------------------+  +--------------------+  +--------------------+
```

Figure 7

o Inter-AS Option C VPN design, with IGP Flex-Algo in each domain, and eBGP peering between domains

```
        +-----+                +-----+                  +-----+
    ....|S-RR1| <............. |S-RR2| <............... |S-RR3| <...
    :   +-----+                +-----+                  +-----+     :
    :                                                               :
    :                                                               :
    :                                                               :
 +--:---------------+      +------------------+      +--------------:--+
 |  :               |      |                  |      |              :  |
 |  :               |      |                  |      |              :  |
 |  :           +---|      |---+          +---|      |---+          :  |
 |  :           |121|------|211|          |231|------|321|          :  |
 |  :           +---|      |---+          +---|      |---+          :  |
 |----+              |      |                  |      |            +----|
 |PE11|              |      |                  |      |            |PE31|
 |----+              |      |                  |      |            +----|
 |              +---|      |---+          +---|      |---+             |
 |              |122|------|212|          |232|------|322|             |
 |              +---|      |---+          +---|      |---+             |
 |                  |      |                  |      |                 |
 |    Domain 1      |      |    Domain 2      |      |    Domain 3     |
 +------------------+      +------------------+      +-----------------+
```

Figure 8

6. Network Availability

- o The BGP CAR solution should provide high network availability for typical deployment topologies, with minimum loss of connectivity in different network failure scenarios.

- o The network failure scenarios, applicable technologies and design options described in [I-D.ietf-mpls-seamless-mpls] should be used as a reference.

- o In the Seamless-MPLS reference topology in previous section:

  * Failure of intra-domain links should limit loss of connectivity (LoC) to < 50ms. E.g., PE11 to a P node (not shown), 121 to a P node in Domain1 or Domain2)

  * Failure of an intra-domain node (P node in any domain) should limit LoC to < 50ms

  * Failure of an ABR node (e.g., 121, 231) should limit LoC to < 1sec

* Failure of a remote PE node (e.g., PE3) should limit LoC to < 1sec

o In the Inter-AS Option C VPN reference topology in previous section:

  * Failure of intra-domain links should limit LoC to < 50ms. E.g., PE11 to a P node (not shown), 121 to a P node in Domain1 or Domain2)

  * Failure of an intra-domain node (P node in any domain) should limit LoC to < 50ms

  * Failure of an ASBR node (e.g., 121, 211) should limit LoC to < 1sec

  * Failure of a remote PE node (e.g., PE3) should limit LoC to < 1sec

  * Failure of an external link (e.g., 121-211) should limit LoC to < 1sec

o The solution should explore and describe additional techniques and design options that are applicable to further improve handling of the failure cases listed above.

7. BGP Protocol Requirements

o Support signaling and distribution of different Color-Aware routes to reach a participating node, e.g., a PE. Intent should be indicated by the notion of a Color as defined in SR Policy Architecture.

  * Signal different instances of a prefix distinguished by color

  * Signal intent associated with a given route

o Support for a flexible NLRI definition to accommodate both efficiency of processing (e.g., packing) and future extensibility

  * Avoid limitations associated with existing SAFI NLRI definitions. For example, 24-bit label.

o Support for validation of paths

  * Reachability of next-hop in control plane

  * Availability and programming of encapsulation in data plane

   * Validation of intent

- Next-hop resolution for Color-Aware route

   * Flexibility to use different intra-domain and inter-domain mechanisms - IGP-FA, SR-TE, RSVP-TE, IGP, BGP-LU etc.

   * Recursive resolution over BGP Color-Aware routes

   * Ability to carry end-to-end cumulative metric for a given color

   * Support setting up an end-to-end Color-Aware path using a different/less preferred or best-effort paths in domains where a particular intent is not available

- Separation of transport and VPN service semantics.

   * Allow for different route distribution planes for service vs transport routes.

- Support signaling of different transport encapsulations

- Support for signaling multiple encapsulations for co-existence and migration

- Generation of BGP Color-Aware routes sourced from IGP-FA, SR-TE policies and BGP-LU from a domain

- Support signaling across domains with different color mappings for a given intent.

8. Future Considerations

Multicast service intent

9. Acknowledgements

Many people contributed to this document.

The authors would especially like to thank Jim Uttaro for his guidance on the work and feedback on many aspects of the problem statement. We would also like to thank Daniel Voyer, Luay Jalil and Robert Raszuk for their review and valuable suggestions.

We also express our appreciation to Bruno Decreane, Keyur Patel, Jim Guichard, Alex Bogdanov, Dirk Steinberg, Hannes Gredler and Xiaohu Hu for discussions on several topics that have helped provide input to the document.

The authors would like to thank Stephane Litkowski for his detailed review and for making valuable suggestions to improve the quality of the document. We would also like to thank Kamran Raza and Kris Michelson for their review and comments on the document and to Simon Spraggs, Jose Liste and Jiri Chaloupka for their early inputs on the problem statement.

10. References

10.1. Normative References

[I-D.agrawal-spring-srv6-mpls-interworking]
Agrawal, S., Ali, Z., Filsfils, C., Voyer, D., and Z. Li, "SRv6 and MPLS interworking", draft-agrawal-spring-srv6-mpls-interworking-03 (work in progress), August 2020.

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

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

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

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

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

[I-D.ietf-spring-sr-service-programming]
Clad, F., Xu, X., Filsfils, C., daniel.bernier@bell.ca, d., Li, C., Decraene, B., Ma, S., Yadlapalli, C., Henderickx, W., and S. Salsano, "Service Programming with Segment Routing", draft-ietf-spring-sr-service-programming-03 (work in progress), September 2020.

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

[I-D.voyer-pim-sr-p2mp-policy]
Voyer, D., Filsfils, C., Parekh, R., Bidgoli, H., and Z. Zhang, "Segment Routing Point-to-Multipoint Policy", draft-voyer-pim-sr-p2mp-policy-02 (work in progress), July 2020.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, DOI 10.17487/RFC4360, February 2006, <https://www.rfc-editor.org/info/rfc4360>.

[RFC4684] Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk, R., Patel, K., and J. Guichard, "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, DOI 10.17487/RFC4684, November 2006, <https://www.rfc-editor.org/info/rfc4684>.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <https://www.rfc-editor.org/info/rfc4760>.

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

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

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

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

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

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

[RFC8277] Rosen, E., "Using BGP to Bind MPLS Labels to Address Prefixes", RFC 8277, DOI 10.17487/RFC8277, October 2017, <https://www.rfc-editor.org/info/rfc8277>.

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

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

[RFC8669] Previdi, S., Filsfils, C., Lindem, A., Ed., Sreekantiah, A., and H. Gredler, "Segment Routing Prefix Segment Identifier Extensions for BGP", RFC 8669, DOI 10.17487/RFC8669, December 2019, <https://www.rfc-editor.org/info/rfc8669>.

10.2. Informative References

[I-D.filsfils-spring-sr-policy-considerations]
Filsfils, C., Talaulikar, K., Krol, P., Horneffer, M., and P. Mattes, "SR Policy Implementation and Deployment Considerations", draft-filsfils-spring-sr-policy-considerations-06 (work in progress), October 2020.

[I-D.ietf-idr-performance-routing]
Xu, X., Hegde, S., Talaulikar, K., Boucadair, M., and C. Jacquenet, "Performance-based BGP Routing Mechanism", draft-ietf-idr-performance-routing-03 (work in progress), December 2020.

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

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

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

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

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <https://www.rfc-editor.org/info/rfc4364>.

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

[RFC7911] Walton, D., Retana, A., Chen, E., and J. Scudder, "Advertisement of Multiple Paths in BGP", RFC 7911, DOI 10.17487/RFC7911, July 2016, <https://www.rfc-editor.org/info/rfc7911>.

Authors' Addresses

Dhananjaya Rao
Cisco Systems
USA

Email: dhrao@cisco.com

Swadesh Agrawal
Cisco Systems
USA

Email: swaagraw@cisco.com

Clarence Filsfils
Cisco Systems
Belgium

Email: cfilsfil@cisco.com

Ketan Talaulikar
Cisco Systems
India

Email: ketant@cisco.com