IDR Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 8, 2020

L. Chan
R. Szarecki, Ed.
Juniper Networks

July 7, 2019

# Inter-Domain Traffic Steering with BGP Labeled Colored Unicast (BGP-LCU)

draft-szarecki-idr-bgp-lcu-traffic-steering-00

## Abstract

This document describes technology that enables for Inter-Domain signaling of existence of E2E path that satisfy high-level traffic treatment behavior intent. The inter-domain path is built by the BGP protocol, as a concatenation of per TE-domain internal paths (segments), provisioned by one of existing intra-domain techniques. The traffic treatment behavior is encoded as an integer value called as "COLOR". The domain internal paths/tunnels are marked as satisfying given traffic treatment behavior. Then the tunnel destination and its COLOR are exchanged between TE-Domains using a new BGP LABELED-COLORED-UNICAST NLRI (BGP-LCU) defined in this document.

## Status of This Memo



## Copyright Notice

Table of Contents

## 1. Introduction

The networks of today grow to high 10,000's - 100,000's of nodes (routers) and beyond. This trend continues. To operate such a large topology, the common practice is to divide it into domains (see Figure 1) and integrate through layered routing protocol infrastructure in order to secure end-to-end (E2E) connectivity. Please see [I-D.ietf-mpls-seamless-mpls].

The nowadays critical and demanding applications rely on network infrastructure, and plain connectivity becomes an insufficient service level.

While the Differentiated Services architecture [RFC2475] allows for multiple service levels across same connectivity path, it does not address topological differentiation such as latency, non-fate-sharing, encryption or bandwidth. These challenges are addressed by existing Traffic Engineering (TE)techniques such RSVP, SR-TE or multi-topology IGPs (e.g. Maximally Redundant Tree [RFC7811], Segment Routing IGP FlexAlgo [I-D.ietf-lsr-flex-algo])in the scope of a limited size domain (TE-DOMAIN).

This document describes technology that enables signaling of existence of E2E path that satisfy high-level traffic treatment behavior intent. The inter-domain path is built by the BGP protocol, as a concatenation of per TE-domain internal paths (segments), provisioned by one of existing intra-domain techniques mentioned above. This way, inter-domain paths for a variety of traffic treatment intents are established without even need to expose the topology of any domain to any of the other domains.

```
                     BGP
   <-----------><-----------><--><--------->

   +-----------+ +----------+    +---------+
   |domain 1   | |  domain 2|----|domain 3 |
   |(e.g.      .-.     (e.g.|  /-|  (e.g.  |
   |FlexAlgo) ( R )    RSVP)|-/  |  SR-TE  |
   |           `-'          |----| w/ PCE) |
   +-----------+ +----------+    +---------+

   <--------->    <--------->    <--------->
   TE-domain      TE-domain      TE-domain
   <-------- cooperating domains ---------->
```

Cooperating TE-domains

Figure 1

The traffic treatment behavior (T-intent) is encoded as an integer value called as "COLOR". The TE-domain internal paths/tunnels are marked as satisfying given traffic treatment behavior as defined in Segment Routing Policy Architecture [I-D.ietf-spring-segment-routing-policy]. Then reachability of the tunnel destination and its COLOR are exchanged between TE-Domains using a new BGP LABELED-COLORED-UNICAST NLRI (BGP-LCU) defined in this document. The procedures of stitching/nesting intra domain tunnels advertised in BGP-LCU resulting in inter-domain E2E path is also specified in this document.

### 1.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

## 2. Conventions used in this document

TE-DOMAIN - Continuous set of links and nodes that allow establishing tunnels that satisfy T-intent between each edge node without using BGP-LCU (defined in this document). Typically TE-domain is 1:1 mapped to IGP area (flooding domain),and intra-TE-domain tunnels are instantiated by RSVP (w/ or w/o assistance of PCE), SR/SRv6 w/ IGP FlexAlgo, static or PCE controlled SRTE/SRv6TE policies. A deployment when TE-domain comprises few connected IGP flooding domains is also possible.

COLOR - the integer value of 32bits representing given traffic treatment behavior intent (T-intent).

BGP-LCU - BGP Labeled Colored Unicast. Name given to SAFI(s) that carries traffic treatment intent toward destination system together with label(s)used to forward traffic across TE-DOMAINS. Defined in this document.

<COLOR,DESTINATION> - colored BGP-LCU prefix, where COLOR is integer encoding traffic treatment intent and DESTINATION is IPv4 or IPv6 subnet address (not necessary host address).

[Label1,Label2,<COLOR,DESTINATION>] - notation used for the labeled colored unicast NLRI

SR-DOMAIN - continuous set of nodes and links that support SR and have at minimum single, shared prefix SID space. So, prefix SID (incl. Node SID and Anycast SID) values are unique in SR-DOMAIN.

BSID - Binding SID. The local label allocated for TE tunnel (RSVP-LSP, SR Policy, etc)

## 3. Traffic treatment behavior intent (T-intent)

The service traffic, while traversing network(s) consumes resources from those networks. The path provided by network to service traffic could be optimized according to needs of the service. A simple example is a real-time communication application that would benefit from being placed on low-latency path. On the other hand, video streaming would best benefit from a low-loss path. Another example is sensitive data like personal health data, which would benefit from a taking path over encrypted links.

It is granted that ability of network to provide distinct path (tunnels) that satisfy treatment intended by application (or class of application) would provide best possible balance between application performance and network resource utilization.

The T-intent is high-level description of traffic treatment. Examples of T-intent are: "low-latency transport", "transport over encrypted infrastructure", "transport path that is topologically disjoined then other path", "transport path over encrypted links/ segments", etc. It is up to the discretion of the network operator (or co-operating operators) to define a set of T-intents that have sense for them.

### 3.1. COLOR

The T-intents defined by operator are encoded in control plane as 32-bit integer value called COLOR, in such way that color-to-T-intent mapping is of monotonic. Therefore, based on COLOR value the

T-intent could be identified without ambiguity. The designation and mapping of COLOR value used for inter-domain operation to T-intent requires agreement of all operators of cooperating domains.

COLOR value of zero (0x00000000) is restricted and MUST NOT be used.

### 3.2. COLOR name spaces

The concept of COLOR as defined above is not specific to inter-domain network slicing, and it actually was introduced in [I-D.ietf-spring-segment-routing-policy] and is used by SR-TE and SR IGP FlexAlgo (called there algorithm) in scope of single TE-domain.

Authors recognizes possibility that color-code values used inside given TE-domain may be not the same as agreed between TE-domains. Furthermore, it is possible that same color value is mapped to different T-intents inside TE-domain and for inter-TE-domain context.

It is recommended for network designers to adjust both color-code schemas to be identical in order to simplify operation. It is assumed in this specification, that color-code schema used for inter-TE-domain as well in each TE-domain is identical

## 4. Scaling Consideration

The BGP-LCU path scale grow with product of number of COLORS supported by multi-domain network system and number of DESTINATIONS in this system. It become obvious that for some network there is a risk of exhausting available MPLS label space.

For large deployments, stacking of labels would be necessary to achieve desired scalability.

## 5. BGP labeled-colored-unicast NLRI

This document defines new SAFI for labeled, colored, unicast (IPv4 and IPv6), and corresponding BGP NLRI that carries label(s) sequence binding to colored prefix - the <COLOR,DESTINATION> tuple. The SAFI value is [[TBD]].

For easy reading BGP instance/session supporting above new SAFI, we will reference it as "BGP-LCU" (Labeled-Colored-Unicast).

### 5.1. BGP capability negotiation

In addition to AFI/SAFI negotiation on the opening of BGP session, in order to send NLRI with more than one label on stack the Multiple Labels Capability (MLC) MUST be successfully negotiated for the

session in order to carry multiple label sequence in BGP-LCU NLRI. If MLC is not negotiated or negotiation failed, BGP-LCU NLRI MUST carry only one label.

The Multiple Labels Capability is defined in chapter 2.1 of BGP Labeled Unicast [RFC8277]. The BGP speaker supporting BGP-LCU MUST follow procedure defined there.

Implementation SHOULD send withdraw of <COLOR,DESTINATION> if length of labels sequence in (to be advertised) NLRI would exceed peers capability.

### 5.2. BGP UPDATE message MP_REACH_NLRI

The procedure described in BGP Labeled Unicast[RFC8277] is used to encode the <COLOR,DESTINATION> tuple as prefix into NLRI with exception of NLRI length field. The Length field is encoded in one or two octets, in order to accommodate large sequence of labels. The Length field encoding follows BGP FlowSpec [RFC5575] encoding of length.

The <COLOR,DESTINATION> tuple form colored-IPv4 or colored-IPv6 prefix. The new sub-address family of SAFI [[TBD]] is allocated for labeled <COLOR,DESTINATION>. The AFI 1 and 2 are used for destination of IPv4 and IPv6 families respectively. The NLRI structure of SAFI [[TBD]] is shown below on Figure 2 for single label and Figure 3 for multiple labels (note: the color and destination elements of prefix are shown explicitly).

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Length     |                 Label                 |Rsrv |S|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                          color                                |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       destination                             ~
~                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

NLRI with One Label.

Figure 2

```
    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
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |   Length(1 or 2 octetx) ~
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                 Label                 |Rsrv |S~
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   ~                 Label                 |Rsrv |S|
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                          color                                |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                       destination                             ~
   ~                                                               |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

NLRI encoding with more then one label bind

Figure 3

- o Length: The Length field consists of a single or two octets. It specifies the length in bits of the remainder of the NLRI field. Note that for each label, the length is increased by 24 bits. The length of color is fixed and is always 32bits. In an MP_REACH_NLRI attribute whose AFI/SAFI is 1/[[TBD]], the length of destination element of prefix will be 32 bits or less. In an MP_REACH_NLRI attribute whose AFI/SAFI is 2/[[TBD]], the length of destination element of prefix will be 128 bits or less. For NLRI shorter than 240 bits (30 octets) the Length is encoded is single octet. For NLRI of 240 bits or longer, two octets are used and the firs nibble is set to value 0xF. Therefore, maximum size of NLRI is 4095b.See [RFC5575]. As specified in MP-BGP [RFC4760], the actual length of the NLRI field will be the number of bits specified in the Length field rounded up to the nearest integral number of octets.

- o Label: The Label field is a 20-bit field containing an MPLS label value (see MPLS Label Encoding [RFC3032]). The null labels (values: 0, 2, 3) are allowed only as last label (or as only labels) in NLRI.

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

- o S: In all labels except the last (i.e., in all labels except the one immediately preceding the prefix), the S bit MUST be 0. In the last label, the S bit MUST be 1. Note that failure to set the S bit in the last label will make it impossible to parse the NLRI correctly. See Section 3, paragraph j of Revised Error Handling

for BGP UPDATE Messages [RFC7606] for a discussion of error handling when the NLRI cannot be parsed.

Note that the UPDATE message not only advertises the binding between the <COLOR,DESTINATION> and the label(s), it also advertises a path to the prefix via the node identified in the Next Hop field of the MP_REACH_NLRI attribute.

If the procedures of BGP ADD-PATHs [RFC7911] are being used, a four-octet "path identifier" (as defined in Section 3 of [RFC7911]) is part of the NLRI and precedes the Length field.

## 5.3. BGP explicit WITHDRAWN message

The withdrawal methodology follows the one described in chapter 2.4 of [RFC8277].  For convenience short description is given below.

The label(s) binding to <COLOR,DESTINATION> could be explicitly withdrawn by sending BGP UPDATE message with MP_UNREACH_NLRI attribute.  The NLRI field of MP_UNREACH_NLRI is encoded as follows:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Length     |         Compatibility                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            color                              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        destination                            ~
~                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

NLRI for Withdrawal

Figure 4

Where:

o  Compatibility: Compatibility field SHOULD be set to 0x800000. Upon reception, the value of the Compatibility field MUST be ignored.

If the procedures of [RFC7911] are being used, a four-octet "path identifier" (as defined in Section 3 of [RFC7911]) is part of the NLRI and precedes the Length field.

### 5.4. Some BGP attribute considerations

#### 5.4.1. BGP Next-Hop

The next-hop network address field in LABELED-COLORED-UNICAST SAFI updates may be either a IPv4 address or a IPv6 address(es) independent of the LABELED-COLORED-UNICAST AFI. This is in accordance to existing specfication in [RFC4760], MP-BGP for IPv6[RFC2545] and IPv4 NLRI with IPv6 Next-Hop[RFC5549]

#### 5.4.2. Prefix SID

In the deployment when multiple TE-domains forms single SR-domain, and therefore prefix SIDs (incl. Node SIDs and Anycast SIDs) are unique in entire multi-domain scope, BGP prefix SID attribute [I-D.ietf-idr-bgp-prefix-sid] may be attached to BGP-LCU NLRI, and SHOULD be honored.

Implementation SHOULD allow for disabling prefixSID processing by local configuration, and in such case tread this attribute as unsupported (therefore advertised without modification, since BGP prefix SID attribute is of transitive optional type). Implementation SHOULD allow, via local configuration, for removing BGP prefix SID attribute from BGP path.

#### 5.4.3. Color Extended Community

The Color Extended Community, defined in Tunnel Encapsulation Attribute[I-D.ietf-idr-tunnel-encaps], MAY be attached to BGP-LCU NLRI.

The purpose of attaching this community is to provide a hint to BGP-LCU update receiver on how BGP Next-Hop attribute shall be resolved. Giving such hint could be useful e.g. for case when colors values used for given T-intent for inter-domain and intra-domain contexts are not equal (see chapter 3.2. ). Exact procedure to handle this case is out of scope of this specification.

In order to avoid ambiguity and simplify implementation, it is recommended to do not attach more than one Color Extended Community.

#### 5.4.4. Tunnel encapsulation

The tunnel encapsulation attribute (23)[I-D.ietf-idr-tunnel-encaps] SHOULD NOT be attached to BGP-LCU NLRI.

If tunnel encapsulation attribute is attached, it MUST NOT conflict with intent of particular BGP path and its NLRI.

It is recommended for implementation to not process this attribute and pass it, if needed, as unsupported optional transitive attribute.

# 6. BGP operation

## 6.1. Injection of labeled colored unicast route to BGP

### 6.1.1. Injecting from colored routes

The ingress border node (BN12)of given TE-domain (domain2 on Figure 5)is provisioned by means out of scope of this document with multiple colored tunnels to endpoints in this domain.

```
+------------+   +----------------+   +---------+
|           ++---+-+  domain 2  +-+---++        |
|           |      |----red---->|      |        |
|           | BN12 |----blue--->| BN2n |        |
|  domain 1 |      |--blue--+   |      |        |
|           |      |-red-+  |   |      |        |
|           ++---+-+     |  |   +-+---++        |
|            |   |       v  v     |   |domain N |
|            |   |     +------+   |   |         |
+------------+   +-----+ PE2  +---+   +---------+
                       +------+
```

Injection form colored routes

Figure 5

The tunnel type is irrelevant for further discussion as long as MPLS frames can be encapsulated over it.  Tunnels could be of MPLS, MPLS-SR, SRv6, MPLSoUDP, etc.  Each of above tunnels has associated one or more intra-domain colors encoding traffic treatment provided by given tunnel, as per [I-D.ietf-spring-segment-routing-policy].

The color-code schema used for Inter-domain is assumed to be the same as one used internally by TE-domain.

Let assume following color schema for domain 2, domain 1 and Inter-domain as shown in Table 1.

| | COLOR |
|---|---|
| T-intent 1 | red |
| T-intent 2 | blue |

Table 1: COLOR code schema - intra- and inter-TE-domain

The BGP speaker (BN12 in Figure 5 ) injects into BGP-LCU four routes with the NLRI fields and BGP attributes values as follow:

- o NLRI DESTINATION := intra-domain tunnel destination IP prefix address (e.g. IP of loopback of BN2n and PE2 in Figure 5).
- o NLRI COLOR := the color code value for T-intent the original tunnel satisfy inside given domain (e.g. red or blue).
- o Exactly one label of value derived according to procedure describe in chapter 6.6. In this case this label MUST be non-null label.
- o S:=1
- o Length := 56 + length of tunnel destination prefix
- o The BGP Next-Hop attribute is set to "self".

Other BGP attributes may also be added as needed by network configuration.

The BGP speaker may crates also MPLS forwarding entries for local label values advertised in NLRI of they do not exist previously.

Please note:

- o This operation does not create any IP RIB entry nor <COLOR,DESTINATION> RIB entry.
- o This operation does not create any IP entry in FIB
- o This operation may create one or more MPLS entry in FIB if needed. The entry's key would be local label allocated as described in chapter 6.6. and advertised in NLRI. The associated action depends on tunnel type but could be generalized as popping label, pushing header(s) of tunnel given BGP route is originating form, forwarding trough egress interface of this tunnel.

### 6.1.2. Injections from non-colored labeled routes

The injection of <COLOR,DESTINATION> into BGP from non-colored routes is similar to one from labeled colored routes, except there is no COLOR of original route to inherit. Therefore, local configuration MUST provide COLOR value that is used for NLRI construction.

- o Implementation MUST support specification of one or more COLOR(s) that would be used for all DESTINATIONS when injected to BGP as LABELED-COLORED-UNICAST NLRI (of SAFI [[TBD]]). If multiple colors are specified, multiple NLRI is injected into BGP.

- o Implementation MAY support specification of COLOR in dependency on (original) route destination, attributes and/or session on which given <COLOR,DESTINATION> are injected to BGP as LABELED-COLORED-UNICAST NLRI (of SAFI [[TBD]].

Similarly, to case described in chapter 6.1.1. , label value MUST be non-null label.

Please note:

- o This operation does not create any IP RIB entry nor <COLOR,DESTINATION> RIB entry.

- o This operation does not create any IP entry in FIB

- o This operation creates one or more MPLS entry in FIB. The entry's key would be local label allocated as described in chapter 6.6. and advertised in NLRI. The associated action depends on tunnel type but could be generalized as popping label, pushing header(s) of tunnel given BGP route is originating form, forwarding trough egress interface of this tunnel.

### 6.1.3. Injections from non-colored non-labeled routes

The injection of <COLOR,DESTINATION> into BGP from non-labeled, non-colored routes is similar to one from labeled non-colored routes, except that explicit or implicit null label shall be used in advertisement.

Please note:

- o This operation does not create any IP RIB entry nor <COLOR,DESTINATION> RIB entry.

- o This operation does not create any IP entry in FIB

- o This operation does not create any MPLS entry in FIB, since explicit null labels are already pre-programmed in FIB.

## 6.2. Receiving BGP-LCU from eBGP (single hop)

The path for <COLOR,DESTINATION> received is experiencing normal BGP process - the sanity is checked first, then configured policies. Finally, path is installed in BGP Loc-RIB and path selection process kick in. Since BGP Next Hop attribute value is IP address of connected subnet, it is used w/o further processing (resolution).

Please note:

- o This operation does create <COLOR,DESTINATION> entry in RIB.

- o This operation does not create any IP RIB entry.

- o This operation does not create any IP entry in FIB

- o This operation does not create any MPLS entry in FIB

## 6.3. Receiving BGP-LCU from iBGP or multihop-eBGP

The path for <COLOR,DESTINATION> received is experiencing normal BGP process - the sanity is checked first, then configured policies. Finally, path is installed in BGP Loc-RIB and path selection process kick in. Since BGP Next Hop attribute value is not a IP address of connected subnet, it needs to be resolved. Since the intention is to provide continuous transport that satisfy T-intent encoded in COLOR, the intra-domain tunnel used for resolution need also satisfy this T-intent. Therefore:

1. If BGP route for <COLOR,DESTINATION> is carrying Color Extended Community, The BGP NextHop attribute shall be resolved by tunnel of color carried in this community (which may be different then value of COLOR carried in NLRI. See chapter 3.2. above). Please see [I-D.ietf-spring-segment-routing-policy]

2. ElseIf BGP route for <COLOR,DESTINATION> is NOT caring Color Extended Community, The BGP NextHop attribute shall be resolved over tunnel of color equal to COLOR carried in NLRI

The fallback to resolution over other tunnels - other color or non-colored - is subject of local configuration policy on the node and/or value of "CO" bits of Color Extended Community.

Please note:

- o This operation does create <COLOR,DESTINATION> entry in RIB.
- o This operation does not create any IP RIB entry.
- o This operation does not create any IP entry in FIB
- o This operation does not create any MPLS entry in FIB

### 6.4. Advertising BGP-LCU over eBGP session and iBGP session with BGP NH changed (NH-self)

Whenever BGP path to <COLOR,DESTINATION> is re-advertised and BGP Next Hop attribute is changed, the label(s) portion of NLRI is modified. On the Next-Hop-change the BGP speaker replaces all label(s) in NLRI by single local label. The local label identifies <COLOR, DESTINATION>. The value of local labels is derived as described in chapter 6.6.

Any BGP speaker supporting LABELED-COLORED-UNICAST (SAFI=[[TBD]]) MUST support above behavior on Next-hop-change.

Please note:

- o This operation does not create <COLOR,DESTINATION> entry in RIB.
- o This operation does not create any IP RIB entry.
- o This operation does not create any IP entry in FIB
- o This operation may create or modify MPLS entry in RIB and FIB.
  - * New RIB and FIB entries are created if no label was allocated to <COLOR,DESTINATION> previously.
  - * The RIB and FIB entries are modified if given path is best and active. (or 2nd to best and BGP PIC EDGE is enabled)
  - * The RIB entry is modified if given path is best.

### 6.5. Advertising BGP-LCU over iBGP session when BGP NH remain unchanged.

Whenever BGP path to <COLOR,DESTINATION> is re-advertised but BGP Next-Hop attribute remains unchanged, the label(s) portion of NLRI MUST NOT be modified.

### 6.6. Label value assignment procedure

The selection of local label value MUST follow below procedure.

1. If BGP speaker is provided (e.g. by local configuration) with explicit label value binding for given <COLOR, DESTINATION>, it SHOULD be honored and used.

2. If BGP speaker is injecting <COLOR,DESTINATION> into BGP-LCU from other protocol or family, and Binding SID (BSID) as per Segment Routing Architecture [RFC8402] is assigned to original tunnel, then local label SHOULD be set to be equal to BSID value.

3. If BGP path carries BGP prefix-SID attribute, and given BGP speaker is enabled to process this attribute (e.g. by mean of local configuration), then this BGP speaker SHOULD allocate local label form it's SRGB [RFC8402].

4. If, given BGP speaker has local label already allocated for given <COLOR, DESTINATION> as result of processing earlier routing events, this same value MUST be used.

5. Else, BGP speaker allocates label from free labels of it's dynamic label block.

Please note that above procedure could result with local label value shared among multiple <COLOR,DESTINATION> prefixes, or unique label value for each <COLOR,DESTINATION>. It depends on particular network scenario and both possibilities are valid and legitimate.

## 7. Deployment and Operation Consideration

### 7.1. Building label stack

#### 7.1.1. Purpose of multiple-label stack

Due to potential large scale of colored prefixes, the BGP-LCU speaker may run out of label space, if 1:1 relationship between <COLOR:DESTINATION> and local label would be established.

Sharing label among multiple <COLOR:DESTINATION> prefixes could be not always possible and reduction of needed labels is hard to predict and is changing together with intra-domain tunnels path changes.

To predictably address this scaling challenge, the topmost label of packet incoming on ASBR/ABR shall represents immediate downstream intra-domain tunnel in the connected TE-domain rather than entire

end-to-end path.  Consequently, ingress PE need to push appropriate label stack on outgoing data packets.

This chapter describes how BGP-LCU could be configured and used on various nodes of multi-domain network system to instruct ingress PE to build and push label stack onto outgoing packets.

If network scale, in terms of number of DESTINATIONS and COLORS, do not requires usage of label stack, it is perfectly valid design to simply swap label in NLRI on every domain border and use one label on ingress PE for inter-TE-domain tunnel.

### 7.1.2. Ingress recursive resolution

The Recursive resolution of BGP-LCU NH attributes on ingress PE provides ability to construct label stack and relief transit BGP speakers (ASBRs and ABRs)label space pressure.  Recursive resolution is matter of network design and ingress PE capability and is inherently supported by BGP-LCU.

The below description is provided to the reader for convenience.

To provide ingress PE with sufficient information for building and pushing label stack onto packet, in addition to signal path for every <COLOR,EGRESS-PE> combination, would require signaling (in BGP-LCU) also path for <COLOR,ASBR/ABR> combination.  Please note that typically number of ASBRs/ABRs is two or three orders of magnitude lower than PEs.  Also, note that if given node is ASBR and PE, is should not be double-counted.  Therefor impact on BGP-LCU path scale is expected to be < 1%. and therefore negligible.

Please note that when BGP-LCU path is re-advertised to another BGP-LCU session, BGP Next-Hop attribute is changed, or not, according to following rules.  This rules do not represent default BGP behavior but could be implemented via local configuration of BGP speaker.

1. If path is advertised to eBGP and has AS-PATH empty, then BGP Next-Hop attribute MUST be changed.  This is default BGP behavior.

2. If path is learned from eBGP from AS that originated <COLOR,DESTINATION> prefix (is last on AS-PATH), then Next-Hop attribute should be changed.  This is observed common practice to change BGP Next-Hop attribute to self in this scenario.

3. In every other case, including re-advertising to eBGP sessions, BGP-LCU Next-Hop attribute, and consequently label(s) sequence in NLRI, should stay unmodified.

Example below (Figure 6) shows BGP-LCU update flow across domains. The BGP Next-Hop attribute manipulation and resolution are also shown in Table 2. Finally, MPLS FIB entries are also displayed.

```
<-------AS 1-------->   <---AS 2--->  <---AS 3--->
  +------+ +------+        +------+      +------+
  |domain| |domain|        |domain|      |domain|
 .+. 1   .-.  2   .-.    .-.  3 .-.    .-.  3  .+.
( I )   ( A )    ( B )--( C )  ( D )--( E )   ( Z )
 '+'     '-'      '-'    '-'    '-'    '-'     '+'
  +------+ +------+        +------+      +------+
```

Label stack build on ingress - topology

Figure 6

The Table 2 below, shows flow BGP-LCU Updates for DESTINATION "I"

| From | to | NLRI | BGP NH | AS-path |
|---|---|---|---|---|
| A | B | [L1 ,<red,I>] | A | |
| B | C | [L1' ,<red,I>] | B | [AS1] |
| C | D | [L1",<red,I>] | C | [AS1] |
| D | E | [L1" ,<red,I>] | C | [AS1 AS2] |
| | | [L2 ,<red,C>] | E | [AS2] |
| E | Z | [L1",<red,I>] | C | [AS1 AS2] |
| | | [L2',<red,C>] | E | [AS2] |

Table 2: COLOR code schema - intra- and inter-TE-domain

The Table 3 below shows RIB entry on node "Z" after recursive resolution

| Prefix/key | encap. operation | egress interface |
|---|---|---|
| <red,I> | push: L1", L2', [red-tunnel-to-E] | X |

Table 3: Ingress label stack build - RIB entry

Please note that ASBRs "D" and "E" do not modify BGP Next-Hop attribute for prefix <red,I>, therefore no label is changed. Consequently there is no MPLS FIB entry created for this prefix.

The above described method allows to build label stack on ingress PE, thus address high scale of <COLOR,DESTINATION> prefix while reducing data-plane states on domains border nodes.

### 7.2. Handling BGP-LCU ingress PEs with limited label imposition depth capabilities

The consequence of design in which inter-domain tunnel is represented as multiple labels stack, is that ingress PE would need to push even more labels onto the packet:

1. service label,

2. perhaps ELI/EL or FAT label(s)

3. sequence of labels for inter-domain tunnel (learned form BGP-LCU and recursively resolved as per chapter 7.1. above)

4. and finally, sequence of one or more labels used by given ingress PE to reach egress ASBR/ABR while satisfying T-intent. The sequence of label could be significantly long if SRTE policy is used.

Authors of this document acknowledges that currently there is equipment in field and in development, that have limited capability in pushing deep label stack (Legacy-PE).

To support such devices in ingress role, egress ASBR/ABR (node "E" on Figure 6) of ingress TE-DOMAIN comprising such PE (node "Z" on Figure 6)have to "reduce" stack depth.

Provided that egress ASBR (node "E") learns all BGP-LCU <COLOR,DESTINATION> prefixes (e.g from Route Server), it advertises this BGP-LCU path to iBGP session toward (set of) ingress Legacy-PE, with BGP Next-Hop attribute change to self. As result, path would be re-advertised with only one label. This reduce required label push depth on legacy ingress PE.

In the very high scale environment, by doing above, egress ASBR/ABR would consume large number of labels. Therefore, network designer needs to take this into consideration and if needed take appropriate action, which could be for example:

- o filter colored prefixes that are send to (all) legacy ingress PEs to smaller subset. This technique is specifically effective if ingress PEs are part of backhaul solution and provide transport to limited set of centralized service-aware nodes (vEPC, BNG, Video caches)

o replace ingress PE hardware or software to enable deeper label push.

## 8. Contributors

The following people have contributed to this document:

Jeff Haas, Juniper Networks

Shraddha Hedge, Juniper Networks

Santosh Kolenchery, Juniper Networks

Shihari Sangli, Juniper Networks

Krzysztof Szarkowicz, Juniper Networks

## 9. IANA Considerations

This document defines a new SAFI in the registry "Subsequent Address Family Identifiers (SAFI) Parameters" that has been assigned by IANA:

| Codepoint | Description | Reference |
|---|---|---|
| [[TBD]] | Labeled colored unicast SAFI | This document |

Table 4

## 10. Security Considerations

The security considerations of BGP (as specified in BGP-4 [RFC4271]) apply.

This document specifies that certain data packets be "tunneled" from one BGP speaker to another across single TE-domain. This requires that the packets be encapsulated while in flight. This document does not specify the encapsulation to be used, except it need to be able to carry MPLS packet as payload. However, if a particular encapsulation is used, the security considerations of that encapsulation are applicable.

If a particular intra-TE-domain tunnel encapsulation does not provide integrity and authentication, it is possible that a data packet's label stack can be modified, through error or malfeasance, while the packet is in flight. This can result in misdelivery of the packet. It should be that the tunnel encapsulation (MPLS), expected to be

most commonly used in deployments of this specification, does not provide integrity or authentication.

There are various techniques one can use to constrain the distribution of BGP UPDATE messages. If a BGP UPDATE advertises the binding of a particular label or set of labels to a particular address <COLOR,DESTINATION>, such techniques can be used to control the set of BGP speakers that are intended to learn of that binding. However, if BGP sessions do not provide privacy, other routers may learn of that binding.

When a BGP speaker processes a received MPLS data packet whose top label it advertised, there is no guarantee that the label in question was put on the packet by a router that was intended to know about that label binding. If a BGP speaker is using the procedures of this document, it may be useful for that speaker to distinguish its "internal" interfaces from its "external" interfaces and to "remember" label binding advertised over each "external" interfaces. Then, a data packet received on give "external" interface can be discarded if its top label was not advertised over this "external" interface. This reduces the likelihood of forwarding packets whose labels have been "spoofed" by untrusted sources.

## 11. References

### 11.1. Normative References

[I-D.ietf-idr-bgp-prefix-sid]
Previdi, S., Filsfils, C., Lindem, A., Sreekantiah, A., and H. Gredler, "Segment Routing Prefix SID extensions for BGP", draft-ietf-idr-bgp-prefix-sid-27 (work in progress), June 2018.

[I-D.ietf-idr-tunnel-encaps]
Patel, K., Velde, G., Ramachandra, S., and E. Rosen, "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-12 (work in progress), May 2019.

[I-D.ietf-spring-segment-routing-policy]
Filsfils, C., Sivabalan, S., daniel.voyer@bell.ca, d., bogdanov@google.com, b., and P. Mattes, "Segment Routing Policy Architecture", draft-ietf-spring-segment-routing-policy-03 (work in progress), May 2019.

[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>.

[RFC2545] Marques, P. and F. Dupont, "Use of BGP-4 Multiprotocol Extensions for IPv6 Inter-Domain Routing", RFC 2545, DOI 10.17487/RFC2545, March 1999, <https://www.rfc-editor.org/info/rfc2545>.

[RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, DOI 10.17487/RFC3032, January 2001, <https://www.rfc-editor.org/info/rfc3032>.

[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>.

[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>.

[RFC5549] Le Faucheur, F. and E. Rosen, "Advertising IPv4 Network Layer Reachability Information with an IPv6 Next Hop", RFC 5549, DOI 10.17487/RFC5549, May 2009, <https://www.rfc-editor.org/info/rfc5549>.

[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>.

### 11.2. Informative References

[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-03 (work in progress), July 2019.

[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.

[RFC2475] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, DOI 10.17487/RFC2475, December 1998, <https://www.rfc-editor.org/info/rfc2475>.

[RFC5575] Marques, P., Sheth, N., Raszuk, R., Greene, B., Mauch, J., and D. McPherson, "Dissemination of Flow Specification Rules", RFC 5575, DOI 10.17487/RFC5575, August 2009, <https://www.rfc-editor.org/info/rfc5575>.

[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>.

[RFC7811] Enyedi, G., Csaszar, A., Atlas, A., Bowers, C., and A. Gopalan, "An Algorithm for Computing IP/LDP Fast Reroute Using Maximally Redundant Trees (MRT-FRR)", RFC 7811, DOI 10.17487/RFC7811, June 2016, <https://www.rfc-editor.org/info/rfc7811>.

[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

Louis Chan
Juniper Networks
Cityplaza One, 1111 King's Road
Taikoo Shing
Hong Kong

Phone: +8522587665
Email: louisc@juniper.net

Rafal J. Szarecki (editor)
Juniper Networks
1133 Innovation Way
Sunnyvale, CA 94089
United States of America

Phone: +14089365629
Email: rafal@juniper.net