IDR Working Group
INTERNET-DRAFT
Intended status: Experimental
Expires: Feb 23, 2022

Louis Chan

Juniper Netwo

Aug 23, 2021

Color Operation with BGP Label Unicast
draft-chan-idr-bgp-lu2-04.txt

Abstract

This document specifies how to carry colored path advertisement via an enhan
to the existing protocol BGP Label Unicast. It would allow backward compatib
with RFC8277.

The targeted solution is to use stack of labels advertised via BGP Label Uni
2.0 for end to end traffic steering across multiple IGP domains. The operati
similar to Segment Routing.

This proposed protocol will convey the necessary reachability information to
ingress PE node to construct an end to end path.

Another two problems addressed here are the interworking with Flex-Algo, and
MPLS label space limit problem.

Please note that there is a major change of protocol format starting from ve
01 draft. Except the optional BGP capability code, these rest of BGP attribu
used in this draft are defined in previous RFC or in use today in other scen

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of
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 Int
Drafts. The list of current Internet-Drafts is at
http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and ma
updated, replaced, or obsoleted by other documents at any time. It is
inappropriate to use Internet-Drafts as reference material or to cite them o
than as "work in progress."

This Internet-Draft will expire on Feb 23, 2022.

Copyright Notice

Table of Contents

1. Introduction

The proposed protocol is aimed to solve interdomain traffic steering, with

different transport services in mind. One application is low latency service multiple IGP domains, which could scale up to 100k or more routers network.

BGP is a flexible protocol. With additional of color attribute to BGP Label Unicast, a path with specific color would be given a meaning in application latency path, a fully protected path, or a path for diversity.

The stack of labels would mean an end to end path across domains through eac or ASBR. Each ABR or ASBR will take one label from the stack, and hence pick forwarding path to next ABR, ASBR, or the final destination.

And the label in the stack may be derived from any of the below

- Prefix SID
- Binding SID for RSVP LSP
- Binding SID for SR-TE LSP
- Local assigned label

The enhancement to the original RFC8277 is to add color extended community, multiple advertisement allowed. The result is similar to multi-topology BGP- different colors.

With Add-path [RFC7911] feature, non color RIB and colored RIB could be adve to the BGP neighbors without new additional attributes. Add-path capability required advertise multiple paths with same prefix but different colors.

A new [BGP-CAP] should be required to enforce such slicing operation during negotiation.

On the other hand, to enable the service prefixes to be mapped accordingly, L3VPN, L2VPN, EVPN and IP prefix with BGP signaling, the color extended comm is also added there. In the PE node, the service prefixes with color will be matched to a transport tunnel with the same color.

The following is an example. Between PE1 and PE2, there is a VPN service run with label 16, which is associated with color 100.

PE1----ABR1-----ABR2-----PE2

PE1 will send the following labels with a color 100 path plus VPN label

[2001 13001 801 16], where

2001 - SR label to reach ABR1

13001 - a Binding-SID label for ABR1-ABR2 tunnel. Underlying tunnel type is

801 - a Binding-SID label for ABR2-PE2 tunnel. Underlying tunnel type is SR-

16 - a VPN label, which is signaled via other means

[2001 13001 801] - denotes the label stack for this color 100 path to reach

The document here is going to describe how PE1 gains enough information to b
this label stack across routing domains.

If PE1 wants to reach PE2 with another colored path, say color 200, the labe
could be different.

At the same time, this architecture is also controller friendly, since all t
notation is Segment Routing compatible, like use of Binding-SID.

The above architecture could be used in conjunction with Flex-Algo [FLEXAGLO
one color could represent a Flex Algorithm. e.g. color 128 equals to Algo 12

When using with Flex Algo in huge network, there could be label space limit.
MPLS label 20 bits long and the maximum label space is around 1 million. In
to represent more IPv4 or IPv6 nodes, label stacking method is recommended.
loopback address could be represent by one or more labels. In this case, (20
n) of label address space is possible.

2. Conventions used in this document

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

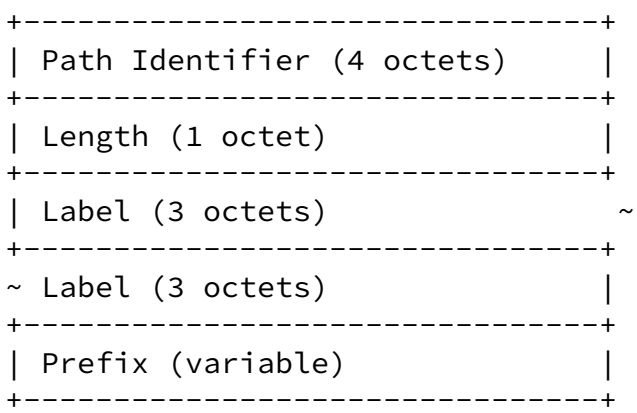
In this document, these words will appear with that interpretation only when
CAPS. Lower case uses of these words are not to be interpreted as carrying
significance described in RFC 2119.

3. Carrying Label Mapping Information with Color and Label Stack

3.1. Use of Add-path to advertise multiple color paths

The use of Path Identifier is to allow multiple advertisement of the same pr
but with different colors or null color.

The extended NLRI format would be like this



3.2. Color extended community for BGP Labeled Unicast

The addition of Color Extended Community is an opaque extended community fro
RFC4360 and RFC5512. The draft allows multiple color values advertisement.

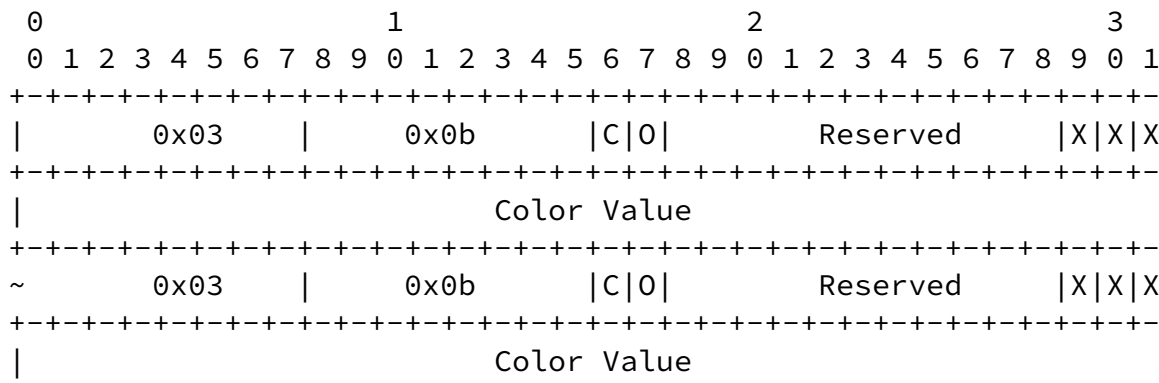


Figure 1: Color value advertisement format

Both in BGP update and MP_UNREACH_NLRI message, multiple color extended comm could be included. It means that multiple colors, indicating different kind services, could share the same label stack. With the use of Path-ID, the mul colors are considered as one bundled update. Any subsequent update is based Path-ID.

If color extended community is not present in a BGP update message, it would treated as normal BGP-LU without any color.

3 bits of XXX is reserved here for the draft.

The meaning for XXX is interpreted as sub-slice of color, with 0 to 7 in dec or 000b and 111b in binary. These sub-slice could be used in either of the following case.

a) Primary path and fallback paths in order of preference
  0 - primary path
  1 - first and most preferred backup path
  ....
  7 - least preferred backup path

b) ECMP paths up to 8, since all paths should be active in forwarding plane.

Color value 0 is reserved for future interoperability purpose.

Color value 1 - 31 are not recommended to use, and this range is reserved fo future use.

3.3. Color extended community for service prefixes

The same format of color extended community is advertised with service prefi which could be VPN prefixes or IP prefixes. The order of the color extended community could be interpreted as

- Order of primary and fallback colors

- Or, ECMP of equal split between color paths

The above would be interpreted by the receiving PE upon its local configurat

It is optional to enable sub-slice notation.

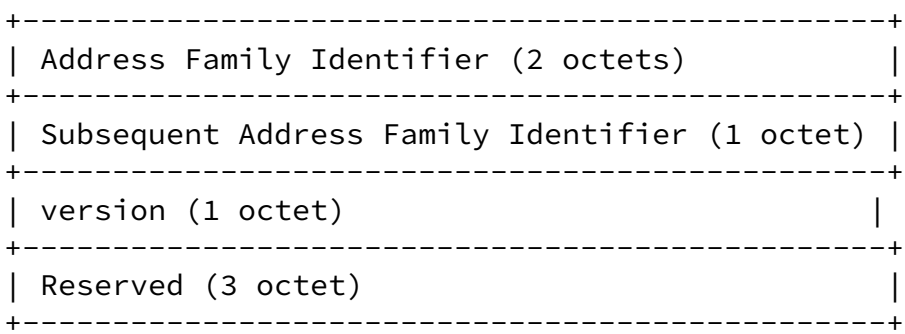
But if sub-slice bits are used, it will be used to map directly to each of t
slice path. If sub-slice path is not available for mapping, it should just f
to resolving by color.

3.4. Color Slicing Capability

The Color Slicing Capability is a BGP capability [RFC5492], with Capability
(TBD).

The color slicing capability is an optional but preferred to have capability
could be configurable parameters at both side of BGP session but with assump
BGP add-path support [RFC7911]. If the specific BGP capability is not negoti
it is assumed version 0 without sub-slice notation. In this case, multiple p
with color attribute are advertised through BGP add-path.

The Capability Length field of this capability is variable. The Capability
field consists of one or more of the following tuples:



The meaning and use of the fields are as follows:

Address Family Identifier (AFI):

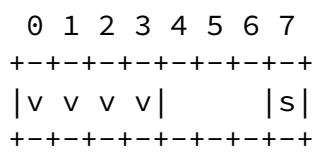
This field is the same as the one used in [RFC4760].

Subsequent Address Family Identifier (SAFI):

This field is the same as the one used in [RFC4760].

Version:

This field is for capability negotiation.



Each of 4 bits of v represents a flag of version from 0 to 4, where LSB deno
support of version 1, and MSB denotes version 4. Version 0 is the default mo
operation, which is described in this document. To determine the common capa
between the two BGP PEER, logical AND function to use determine the highest
denominator of protocol version.

For example, if BGP receive 0b0110 from its peer and perform AND function wi
own capability 0b0010, the result is 0b0010. Version 2 is selected.

The other examples are
- 0b0110 AND 0b0110, version 3 is selected
- 0b0100 AND 0b0010, version 0 is selected

Version 1 (0b0001) is reserved.

S-flag is the indication of use of sub-slice. Set to 1 if sub-slice notat
enforced. If either side is set to 0 for S-flag, sub-slice is not in use.

Reserved:

This field is reserved for future use.

4. Uniqueness of path entries

a) Use of color can be considered to slice into multiple BGP Label Unicast R
Therefore, it should be treated as unique entries for the <path-id, color(s)
prefix>.

e.g. <path-id, color(s), prefix>, [labels]

<123, 100, 10.1.1.1/32>, [1000 2000]

<124, 200, 10.1.1.1/32>, [1000 2000]

<222, {300,400}, 10.1.1.1/32>, [1000 2000]

<223, null, 10.1.1.1/32>, [1000 2000]

All these 4 NLRI are considered different but valid entries for different co
instances.

b) With sub-slice notation
<path-id, color-sub, prefix>, [labels]

<901, 100-0, 10.1.1.1/32>, [1000 2000]

<902, 100-1, 10.1.1.1/32>, [1001 3000]

<903, 100-7, 10.1.1.1/32>, [1002 4000]

These 3 NLRI are distinct, and the second and third NLRI could be used for
backup or ECMP purpose.

5. AIGP consideration

AIGP (RFC7311) would be also used in here to embed certain metric across.

6. Explicit Withdraw of a <path-id, color(s), prefix>

According to RFC8277, MP_UNREACH_NLRI can be used to remove binding of a <pa
color(s), prefix>.

If a path-id is associated with a prefix with multiple colors, the withdrawa
be applied to all associated colors.

To withdraw color(s) partially from the same path-id advertisement, BGP upda
should be used instead.

7. Error Handling Procedure

If BGP receiver could not handle the NLRI, it should silently discard with e
logging.

8. Controller Compatibility

The proposed architecture is compatible with controller for end to end provisioning. Persistent label, like Binding-SID is recommended to be used. controller could learn these labels from the network, and program specific e end path.

In this case, BGP-LU2 will provide a second best path to an ingress PE node, a controller, with more external information, could provide a best path from overall perspective.

Controller could also be deployed based on domain by domain perspective. e.g Optimizing latency of a RSVP LSP, or maintain the bandwidth and loading betw TE LSPs.

9. Interworking with Flex Algo

Flex Algo is a way of network slicing, but it is only an IGP protocol. In or scale across different domains, BGP is recommended as the method to distribu information across.

With color notation in this proposal, one router can distribute to another d via BGP.

There are two ways of mapping Flex-Algo to color attribute in BGP-LU2

a) Color 128 equals Flex Algo 128
b) Or, Color 400 is mapped to Flex Algo 128

10. Label stacking to increase label space

Due to the use of Flex-Algo [FLEXALGO], the MPLS label space might run into Each node will need extra labels for each Algo.

The idea is to use multiple labels to represent a single node. In this case,

label space becomes (2^20)^n, depending on n stacking level.

For IPv6 address, there would be enough label space even if running with SR-

For example, for node 1.1.1.1, 2 consecutive labels are used to represent th

Algo 0: [100101 100001]

Algo 128: [400101 400001]

How the forwarding plane treats the stacked labels is out of the discussion

11. Tunneling SRv6 packet via MPLS

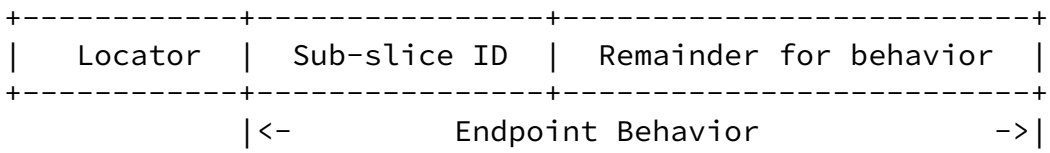
PE1-----ABR1-----ABR2-----PE2

In a SRv6 network, PE1 and PE2 is using SRv6 for VPN service. Between ABR1 a ABR2, it is capable of MPLS only. The use of BGP-LU2 would be a method to pr locator route mapping to MPLS tunnel between ABRs.

At ABR1, the mapping options could be

a) Use of color attribute associated with the VPN advertisement and map to t desired tunnel.
b) Up to the locator route. For example, use first 48 bits of SRv6 header FC00:0000:nnnn::/48 ; where nnnn is the locator portion

c) Making use of sub-slice information as defined in [SRV6-SUBSLICE]



Sub-slice ID could be used for mapping to different color path in MPLS. Fo example,

FC00:0000:nnnn:ssss::/64 ; where ssss is a sub-slice ID

ABR2 advertises a/64 prefix route inclusive of sub-slice ID via BGP-LU2 in ABR1. Hence, traffic will be redirected to a MPLS tunnel from ABR1.

d) With the format described in [SRV6-SUBSLICE], a mapping could be made bet sub-slice ID and <color, sub-slice> mentioned in section 3.2.

12. Security Considerations

TBD

13. IANA Considerations

TBD. It will require a new BGP capability code to enable such color operatio

New SAFI might be required as well.

14. References

14.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Le
BCP 14, RFC 2119, March 1997.

14.2. Informative References

[RFC3107] Rekhter, Y. and E. Rosen, "Carrying Label Information in BGP-4", R
3107, DOI 10.17487/RFC3107, May 2001,

<https://www.rfc-editor.org/info/rfc3107>.

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended
Communities Attribute", RFC 4360, February 2006

<https://www.rfc-editor.org/info/rfc4360>.

[RFC5512] Mohapatra, P. and E. Rosen, "The BGP Encapsulation Subsequent Addr
Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute"
5512, April 2009.

<https://www.rfc-editor.org/info/rfc5512>.

[RFC5575] Marques, P., Sheth, N., Raszuk, R., Greene, B., Mauch, J., and D.
McPherson, "Dissemination of Flow Specification Rules", RFC 5575, DO
10.17487/RFC5575, August 2009,

<http://www.rfc-editor.org/info/rfc5575>.

[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>.

[RFC7911] Walton, D., Retana, A., Chen, E., and J. Scudder, "Advertisement o Multiple Paths in BGP", RFC 7911, DOI 10.17487/RFC7911, July 2016,

<https://www.rfc-editor.org/info/rfc7911>.

[RFC8277] Rosen, E., "Using BGP to Bind MPLS Labels to Address Prefixes", RF DOI 10.17487/RFC8277, October 2017,

<https://www.rfc-editor.org/info/rfc8277>.

[BGP-CAP] Chandra, R. and J. Scudder, "Capabilities Advertisement

with BGP-4", RFC 2842, May 2000.

[FLEXAGLO] S. Hegde, P. Psenak and etc, IGP Flexible Algorithm

https://datatracker.ietf.org/doc/draft-ietf-lsr-flex-algo

[SRV6-SUBSLICE] Louis Chan, Sub-slicing for SRv6

https://datatracker.ietf.org/doc/draft-chan-srv6-sub-slice/

15. Acknowledgments

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

Yimin Shen, Juniper Networks

Author Address

Louis Chan (editor)
Juniper Networks
2604, Cityplaza One, 1111 King's Road
Taikoo Shing
Hong Kong

Phone: +85225876659
Email: louisc@juniper.net