

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 20 October 2022

H. Chen
Futurewei
M. Toy
Verizon
A. Wang
China Telecom
Z. Li
China Mobile
L. Liu
Fujitsu
X. Liu
Volta Networks
18 April 2022

SR Path Ingress Protection
draft-chen-idr-sr-ingress-protection-06

Abstract

This document describes extensions to Border Gateway Protocol (BGP) for protecting the ingress node of a Segment Routing (SR) tunnel or path.

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].

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

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

This Internet-Draft will expire on 20 October 2022.

Copyright Notice

Copyright (c) 2022 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	2
2. Terminologies	3
3. SR Path Ingress Protection Example	4
4. Behavior after Ingress Failure	4
5. Extensions to BGP	5
5.1. SR Path Ingress Protection Sub-TLV	5
5.1.1. Primary Ingress Sub-TLV	6
5.1.2. Service Sub-TLV	7
5.1.3. Traffic Description Sub-TLVs	8
6. Backup Ingress Behavior	11
7. Security Considerations	12
8. Acknowledgements	13
9. IANA Considerations	13
9.1. BGP Tunnel Encapsulation Attribute Sub-TLVs	13
9.2. Ingress Protection Information Sub-TLVs	13
10. References	13
10.1. Normative References	13
10.2. Informative References	14
Authors' Addresses	15

1. Introduction

The fast protection of a transit node of a Segment Routing (SR) path or tunnel is described in [I-D.bashandy-rtgwg-segment-routing-ti-lfa] and [I-D.hu-spring-segment-routing-proxy-forwarding]. [RFC8424] presents extensions to RSVP-TE for the fast protection of the ingress node of a traffic engineering (TE) Label Switching Path (LSP). However, these documents do not discuss any protocol extensions for the fast protection of the ingress node of an SR path or tunnel.

This document fills that void and specifies protocol extensions to Border Gateway Protocol (BGP) for the fast protection of the ingress node of an SR path or tunnel. Ingress node and ingress, fast protection and protection as well as SR path and SR tunnel will be used exchangeably in the following sections.

2. Terminologies

The following terminologies are used in this document.

SR: Segment Routing

SRv6: SR for IPv6

SRH: Segment Routing Header

SID: Segment Identifier

CE: Customer Edge

PE: Provider Edge

LFA: Loop-Free Alternate

TI-LFA: Topology Independent LFA

TE: Traffic Engineering

BFD: Bidirectional Forwarding Detection

VPN: Virtual Private Network

L3VPN: Layer 3 VPN

FIB: Forwarding Information Base

PLR: Point of Local Repair

BGP: Border Gateway Protocol

IGP: Interior Gateway Protocol

OSPF: Open Shortest Path First

IS-IS: Intermediate System to Intermediate System

3. SR Path Ingress Protection Example

To protect against the failure of the (primary) ingress node of a (primary) SR path, a backup ingress node is configured or selected and is different from the (primary) ingress node. A backup SR path from the backup ingress node is computed and installed. Primary ingress and ingress as well as primary SR path and SR path will be used exchangeably.

Figure 1 shows an example of protecting ingress PE1 of a SR path, which is from ingress PE1 to egress PE3.

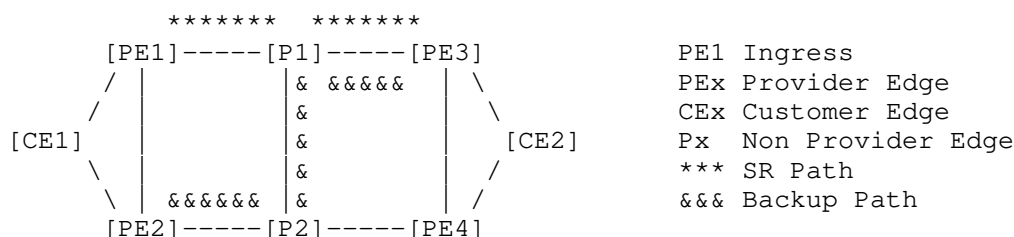


Figure 1: Protecting Ingress PE1 of SR Path PE1-P1-PE3

In normal operations, CE1 sends the traffic with destination PE3 to ingress PE1, which imports the traffic into the SR path.

When CE1 detects the failure of ingress PE1, it switches the traffic to backup ingress PE2, which imports the traffic from CE1 into a backup SR path. The backup path is from the backup ingress PE2 to the egress PE3. When the traffic is imported into the backup path, it is sent to the egress PE3 along the path.

4. Behavior after Ingress Failure

After the failure of the ingress of an SR path happens, there are a couple of different ways to detect the failure. In each way, there may be some specific behavior for the traffic source (e.g., CE1) and the backup ingress (e.g., PE2).

In one way, the traffic source (e.g., CE1) is responsible for fast detecting the failure of the ingress (e.g., PE1) of an SR path. Fast detecting the failure means detecting the failure in a few or tens of milliseconds. The backup ingress (e.g., PE2) is ready to import the traffic from the traffic source into the backup SR path installed.

In normal operations, the source sends the traffic to the ingress of the SR path. When the source detects the failure of the ingress, it switches the traffic to the backup ingress, which delivers the traffic to the egress of the SR path via the backup SR path.

In another way, the backup ingress is responsible for fast detecting the failure of the ingress of an SR path.

In normal operations, the source (e.g., CE1) sends the traffic to the ingress (e.g., PE1) and may send the traffic to the backup ingress (e.g., PE2). It sends the traffic to the backup ingress (e.g., PE2) after the ingress fails.

The backup ingress does not import any traffic from the source into the backup SR path in normal operations. When it detects the failure of the ingress, it imports the traffic from the source into the backup SR path.

5. Extensions to BGP

For a SR path from a primary ingress node to an egress node, a backup ingress node is selected to protect the failure of the primary ingress node of the SR path. This section describes the extensions to BGP for representing the information for protecting the primary ingress node in a BGP UPDATE message and distributing the information to the backup ingress node. The information includes a SR backup path.

[I-D.ietf-idr-segment-routing-te-policy] specifies a way of representing a SR path in a BGP UPDATE message and distributing the SR path to the ingress node of the SR path.

This is extended to represent the information for protecting the primary ingress by defining a few of new Sub-TLVs.

5.1. SR Path Ingress Protection Sub-TLV

A new Sub-TLV, called SR Path Ingress Protection Sub-TLV, is defined. When a UPDATE message is sent to the backup ingress node for protecting the primary ingress node of a SR path, the message contains this Sub-TLV. Its format is illustrated below.

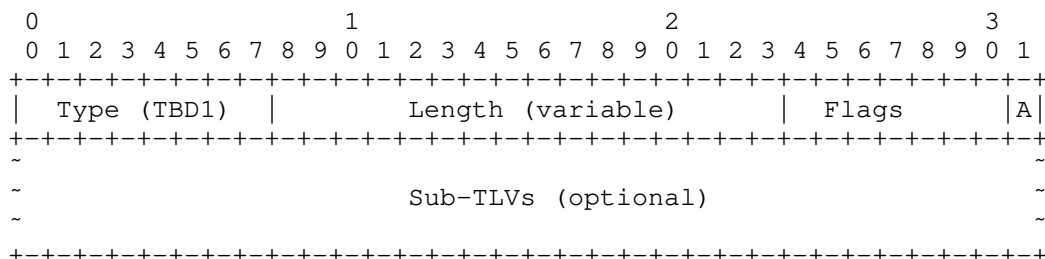


Figure 2: SR Path Ingress Protection Sub-TLV

Type: TBD1 is to be assigned by IANA.

Length: Variable.

Flags: 1 octet. One flag is defined.

Flag A: 1 bit. It is set to

- 1: request a backup ingress to let the forwarding entry for the backup SR path be Active.
- 0: request a backup ingress to let the forwarding entry for the backup SR path be inactive initially and to make the entry be active after detecting the failure of the primary ingress node of the primary SR path.

A few optional Sub-TLVs are defined, which are Primary Ingress Sub-TLV, Service Sub-TLV and Traffic Description Sub-TLV.

5.1.1. Primary Ingress Sub-TLV

A Primary Ingress Sub-TLV indicates the IP address of the primary ingress node of a primary SR path. It has two formats: one for primary ingress node IPv4 address and the other for primary ingress node IPv6 address, which are illustrated below.

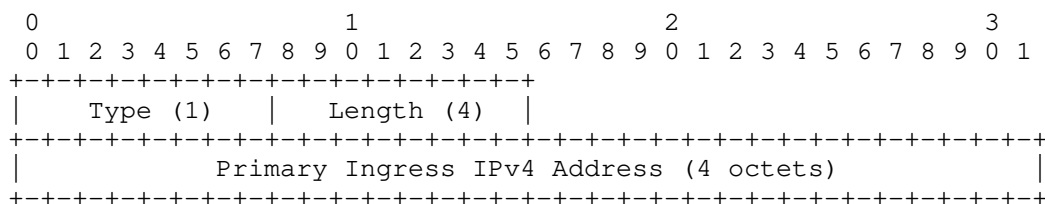


Figure 3: Primary Ingress IPv4 Address Sub-TLV

Type: Its value (1 suggested) is to be assigned by IANA.

Length: 4.

Primary Ingress IPv4 Address: 4 octets. It represents an IPv4 host address of the primary ingress node of a primary SR path.

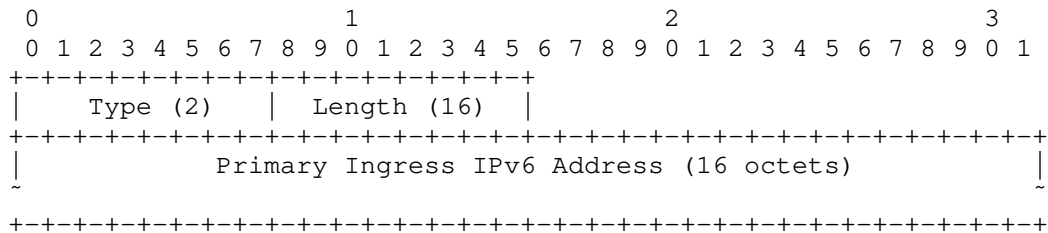


Figure 4: Primary Ingress IPv6 Address Sub-TLV

Type: Its value (2 suggested) is to be assigned by IANA.

Length: 16.

Primary Ingress IPv6 Address: 16 octets. It represents an IPv6 host address of the primary ingress node of a primary SR path.

5.1.2. Service Sub-TLV

A Service Sub-TLV contains a service ID or label to be added into a packet to be carried by a SR path. It has three formats: the first one for the service identified by a label, the second one for the service identified by a service identifier (ID) of 32 bits, and the third one for the service identified by a service identifier (ID) of 128 bits. Their formats are illustrated below.

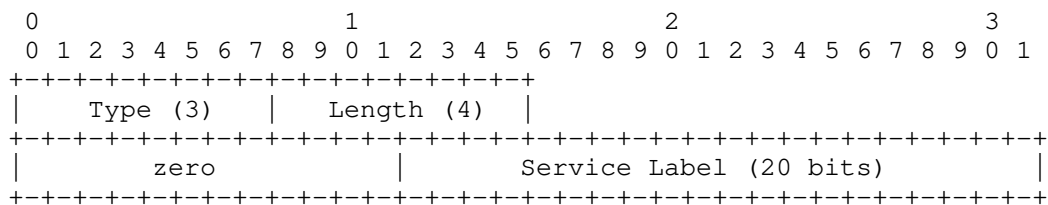


Figure 5: Service Label Sub-TLV

Type: Its value (3 suggested) is to be assigned by IANA.

Length: 4.

Service Label: the least significant 20 bits. It represents a label of 20 bits.

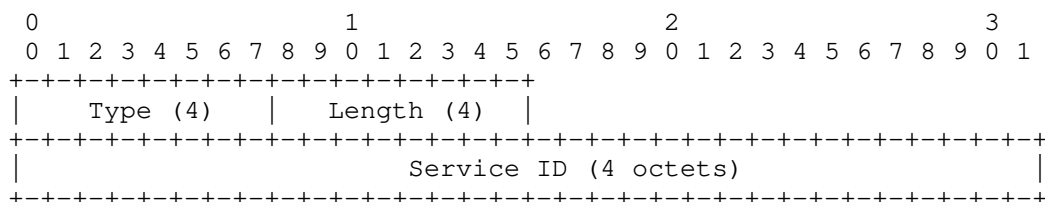


Figure 6: 32 Bits Service ID Sub-TLV

Type: Its value (4 suggested) is to be assigned by IANA.

Length: 4.

Service ID: 4 octets. It represents a Service Identifier (ID) of 32 bits.

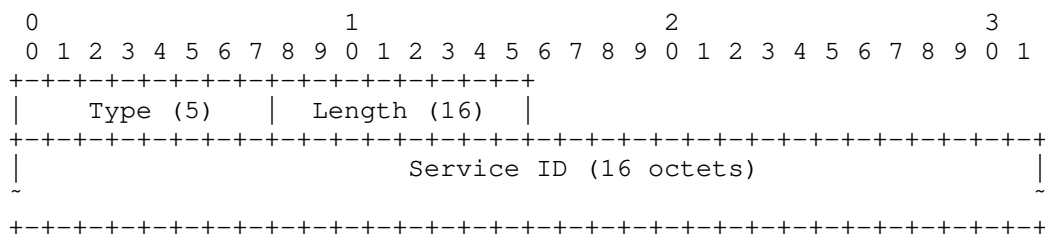


Figure 7: 128 Bits Service ID Sub-TLV

Type: Its value (5 suggested) is to be assigned by IANA.

Length: 16.

Service ID: 16 octets. It represents a Service Identifier (ID) of 128 bits.

5.1.3. Traffic Description Sub-TLVs

A Traffic Description Sub-TLV describes the traffic to be imported into a backup SR path. Five Traffic Description Sub-TLVs are defined. Two of them are FEC Sub-TLVs and the others are interface Sub-TLVs.

Two FEC Sub-TLVs are IPv4 and IPv6 FEC Sub-TLVs. Their formats are illustrated below.

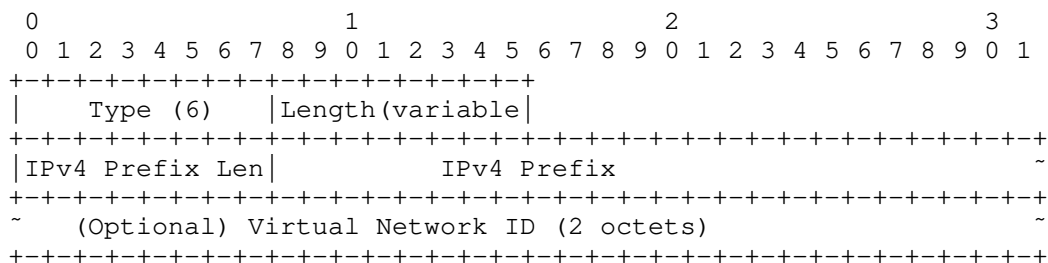


Figure 8: IPv4 FEC Sub-TLV

Type: Its value (6 suggested) is to be assigned by IANA.

Length: Variable.

IPv4 Prefix Len: Indicates the length of the IPv4 Prefix.

IPv4 Prefix: IPv4 Prefix rounded to octets.

Virtual Network ID: 2 octets. This is optional. It indicates the ID of a virtual network.

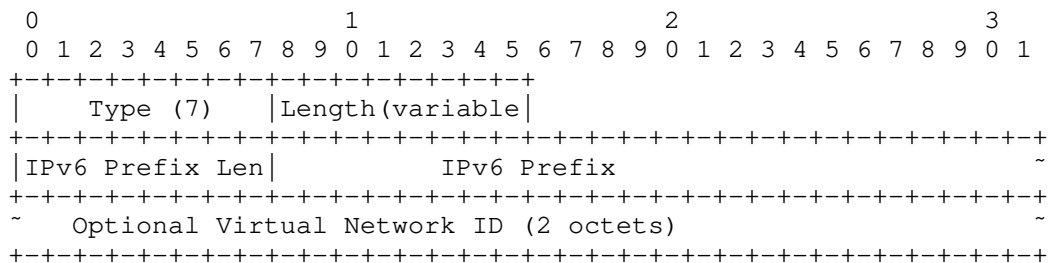


Figure 9: IPv6 FEC Sub-TLV

Type: Its value (7 suggested) is to be assigned by IANA.

Length: Variable.

IPv6 Prefix Len: Indicates the length of the IPv6 Prefix.

IPv6 Prefix: IPv6 Prefix rounded to octets.

Virtual Network ID: 2 octets. This is optional. It indicates the ID of a virtual network.

An Interface sub-TLV indicates the interface from which the traffic is received and imported into the backup SR path/tunnel. It has three formats: one for interface index, the other two for IPv4 and IPv6 address, which are illustrated below.

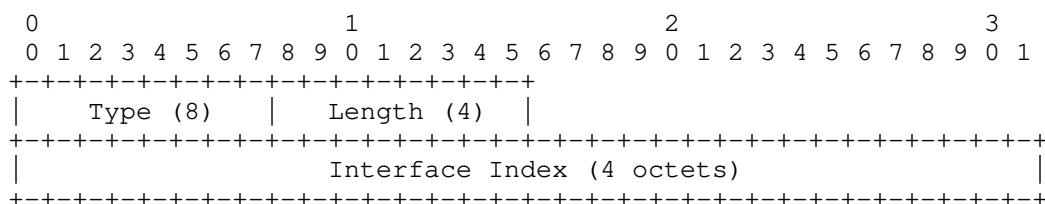


Figure 10: Interface Index Sub-TLV

Type: Its value (8 suggested) is to be assigned by IANA.

Length: 4.

Interface Index: 4 octets. It indicates the index of an interface.

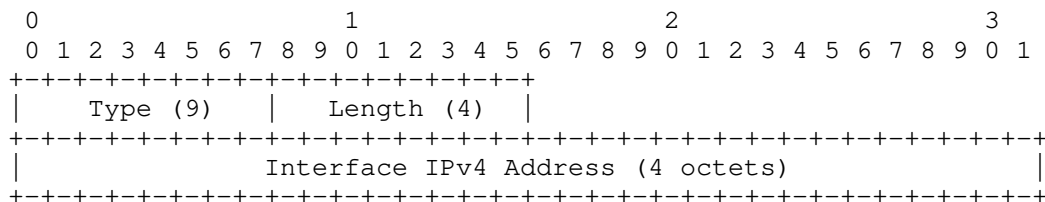


Figure 11: Interface IPv4 Address Sub-TLV

Type: Its value (9 suggested) is to be assigned by IANA.

Length: 4.

Interface IPv4 Address: 4 octets. It represents the IPv4 address of an interface.

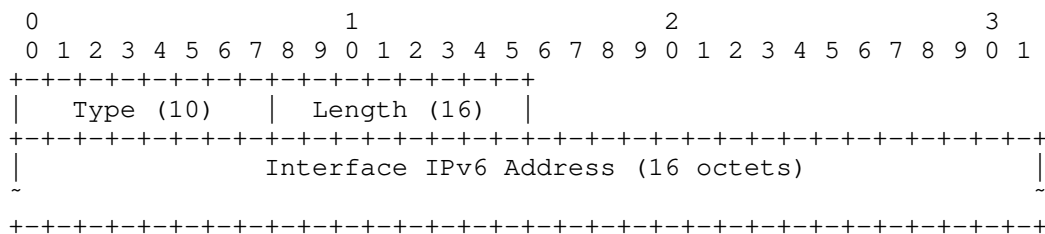


Figure 12: Interface IPv6 Address Sub-TLV

Type: Its value (10 suggested) is to be assigned by IANA.

Length: 16.

Interface IPv6 Address: 16 octets. It represents the IPv6 address of an interface.

6. Backup Ingress Behavior

When a backup ingress node receives a UPDATE message containing the information for protecting the primary ingress node of a SR path, it installs a forwarding entry in its FIB based on the information. The information is encoded in a SR policy of the following structure:

SR Policy SAFI NLRI: <Distinguisher, Policy-Color, Endpoint>

Attributes:

 Tunnel Encaps Attribute (23)

 Tunnel Type (15): SR Policy

 SR Path Ingress Protection Sub-TLV

 Primary Ingress Sub-TLV

 Service Sub-TLV

 Traffic Description Sub-TLV

 Preference Sub-TLV

 Binding SID Sub-TLV

 Explicit NULL Label Policy (ENLP) Sub-TLV

 Priority Sub-TLV

 Policy Name Sub-TLV

 Segment List Sub-TLV

 Weight Sub-TLV

 Segment Sub-TLV

 Segment Sub-TLV

 ...

...

Where:

- o SR Policy SAFI NLRI is defined in [I-D.ietf-idr-segment-routing-te-policy].
- o Tunnel Encapsulation Attribute is defined in [I-D.ietf-idr-tunnel-encaps].
- o Tunnel Type of SR Policy is defined in [I-D.ietf-idr-segment-routing-te-policy].
- o SR Path Ingress Protection, Primary Ingress, Service and Traffic Description Sub-TLVs are defined in this document.
- o Preference, Binding SID, ENLP, Priority, Policy Name, Segment List, Weight and Segment Sub-TLVs are defined in [I-D.ietf-idr-segment-routing-te-policy].

After receiving a SR policy with a SR Path Ingress Protection Sub-TLV, the backup ingress node will install one or more candidate paths into its "BGP table". Another module such as SRPM will choose one or more paths and install the forwarding entries for them in the data plane.

The forwarding entries for the paths installed in the data plane will be set to be inactive if the flag A in the SR Path Ingress Protection Sub-TLV is zero. When the primary ingress node fails, these forwarding entries are set to be active. The failure of the primary ingress may be detected by the backup ingress node through using a mechanism such as BFD. The IP address of the primary ingress in the Primary Ingress Sub-TLV may be used for detecting the failure of the primary ingress node.

If the flag A in the SR Path Ingress Protection Sub-TLV is one, then the forwarding entries for the paths installed in the data plane will be set to be active.

When there is a Service Sub-TLV in the SR Path Ingress Protection Sub-TLV, the ID or Label in the Service Sub-TLV will be included in the forwarding entries. When a packet is imported into a backup SR path using the forwarding entries, the service ID or Label is pushed first and then the sequence of segments represented in the Segment List Sub-TLV.

7. Security Considerations

Protocol extensions defined in this document do not affect the BGP security other than those as discussed in the Security Considerations section of [RFC5575].

8. Acknowledgements

The authors of this document would like to thank Dhruv Dhody for the comments.

9. IANA Considerations

9.1. BGP Tunnel Encapsulation Attribute Sub-TLVs

Under Existing Registry Name: "BGP Tunnel Encapsulation Attribute Sub-TLVs", IANA is requested to assign a new Sub-TLV value for SR Path Ingress Protection as follows:

Value	sub-TLV Name	Reference
-----	-----	-----
TBD1	SR Path Ingress Protection Sub-TLV	This Document

9.2. Ingress Protection Information Sub-TLVs

A new registry called "Ingress Protection Information Sub-TLVs" is defined in this document. IANA is requested to create and maintain new registry:

o Ingress Protection Information Sub-TLVs

Initial values for the registry are given below. The future assignments are to be made through IETF Review [RFC5226].

Value	sub-TLV Name	Reference
-----	-----	-----
0	Reserved	
1	Primary Ingress IPv4 Address Sub-TLV	This Document
2	Primary Ingress IPv6 Address Sub-TLV	This Document
3	Service Label Sub-TLV	This Document
4	32 Bits Service ID Sub-TLV	This Document
5	128 Bits Service ID Sub-TLV	This Document
6	IPv4 FEC Sub-TLV	This Document
7	IPv6 FEC Sub-TLV	This Document
8	Interface Index Sub-TLV	This Document
9	Interface IPv4 Address Sub-TLV	This Document
10	Interface IPv6 Address Sub-TLV	This Document
11-255	Unassigned	

10. References

10.1. Normative References

- [I-D.ietf-idr-segment-routing-te-policy]
Previdi, S., Filsfils, C., Talaulikar, K., Mattes, P., Jain, D., and S. Lin, "Advertising Segment Routing Policies in BGP", Work in Progress, Internet-Draft, draft-ietf-idr-segment-routing-te-policy-17, 14 April 2022, <<https://www.ietf.org/archive/id/draft-ietf-idr-segment-routing-te-policy-17.txt>>.
- [I-D.ietf-idr-tunnel-encaps]
Patel, K., Velde, G. V. D., Sangli, S. R., and J. Scudder, "The BGP Tunnel Encapsulation Attribute", Work in Progress, Internet-Draft, draft-ietf-idr-tunnel-encaps-22, 7 January 2021, <<https://www.ietf.org/archive/id/draft-ietf-idr-tunnel-encaps-22.txt>>.
- [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>>.
- [RFC7356] Ginsberg, L., Previdi, S., and Y. Yang, "IS-IS Flooding Scope Link State PDUs (LSPs)", RFC 7356, DOI 10.17487/RFC7356, September 2014, <<https://www.rfc-editor.org/info/rfc7356>>.
- [RFC8424] Chen, H., Ed. and R. Torvi, Ed., "Extensions to RSVP-TE for Label Switched Path (LSP) Ingress Fast Reroute (FRR) Protection", RFC 8424, DOI 10.17487/RFC8424, August 2018, <<https://www.rfc-editor.org/info/rfc8424>>.

10.2. Informative References

- [I-D.bashandy-rtgwg-segment-routing-ti-lfa]
Bashandy, A., Filsfils, C., Decraene, B., Litkowski, S., Francois, P., Voyer, D., Clad, F., and P. Camarillo, "Topology Independent Fast Reroute using Segment Routing", Work in Progress, Internet-Draft, draft-bashandy-rtgwg-segment-routing-ti-lfa-05, 4 October 2018, <<https://www.ietf.org/archive/id/draft-bashandy-rtgwg-segment-routing-ti-lfa-05.txt>>.
- [I-D.hegde-spring-node-protection-for-sr-te-paths]
Hegde, S., Bowers, C., Litkowski, S., Xu, X., and F. Xu, "Node Protection for SR-TE Paths", Work in Progress, Internet-Draft, draft-hegde-spring-node-protection-for-sr-te-paths-07, 30 July 2020, <<https://www.ietf.org/archive/id/draft-hegde-spring-node-protection-for-sr-te-paths-07.txt>>.

- [I-D.hu-spring-segment-routing-proxy-forwarding]
Hu, Z., Chen, H., Yao, J., Bowers, C., Yongqing, and
Yisong, "SR-TE Path Midpoint Restoration", Work in
Progress, Internet-Draft, draft-hu-spring-segment-routing-
proxy-forwarding-19, 11 April 2022,
<[https://www.ietf.org/archive/id/draft-hu-spring-segment-
routing-proxy-forwarding-19.txt](https://www.ietf.org/archive/id/draft-hu-spring-segment-routing-proxy-forwarding-19.txt)>.
- [I-D.ietf-spring-segment-routing-policy]
Filsfils, C., Talaulikar, K., Voyer, D., Bogdanov, A., and
P. Mattes, "Segment Routing Policy Architecture", Work in
Progress, Internet-Draft, draft-ietf-spring-segment-
routing-policy-22, 22 March 2022,
<[https://www.ietf.org/archive/id/draft-ietf-spring-
segment-routing-policy-22.txt](https://www.ietf.org/archive/id/draft-ietf-spring-segment-routing-policy-22.txt)>.
- [I-D.sivabalan-pce-binding-label-sid]
Sivabalan, S., Filsfils, C., Tantsura, J., Hardwick, J.,
Previdi, S., and C. Li, "Carrying Binding Label/Segment-ID
in PCE-based Networks.", Work in Progress, Internet-Draft,
draft-sivabalan-pce-binding-label-sid-07, 8 July 2019,
<[https://www.ietf.org/archive/id/draft-sivabalan-pce-
binding-label-sid-07.txt](https://www.ietf.org/archive/id/draft-sivabalan-pce-binding-label-sid-07.txt)>.
- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an
IANA Considerations Section in RFCs", RFC 5226,
DOI 10.17487/RFC5226, May 2008,
<<https://www.rfc-editor.org/info/rfc5226>>.
- [RFC5462] Andersson, L. and R. Asati, "Multiprotocol Label Switching
(MPLS) Label Stack Entry: "EXP" Field Renamed to "Traffic
Class" Field", RFC 5462, DOI 10.17487/RFC5462, February
2009, <<https://www.rfc-editor.org/info/rfc5462>>.
- [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>>.

Authors' Addresses

Huaimo Chen
Futurewei
Boston, MA,
United States of America
Email: huaimo.chen@futurewei.com

Mehmet Toy
Verizon
United States of America
Email: mehmet.toy@verizon.com

Aijun Wang
China Telecom
Beiqijia Town, Changping District
Beijing
102209
China
Email: wangaj3@chinatelecom.cn

Zhenqiang Li
China Mobile
32 Xuanwumen West Ave, Xicheng District
Beijing
100053
China
Email: lizhengqiang@chinamobile.com

Lei Liu
Fujitsu
United States of America
Email: liulei.kddi@gmail.com

Xufeng Liu
Volta Networks
McLean, VA
United States of America
Email: xufeng.liu.ietf@gmail.com

Network Working Group
Internet Draft
Intended status: Informational
Expires: May 21, 2020
Consulting

L. Dunbar
Futurewei
S. Hares
Hickory Hill

November 21, 2019

SDWAN WAN Ports Property Advertisement in BGP UPDATE
draft-dunbar-idr-sdwan-port-safi-06

Abstract

The document describes how the SDWAN SAFI, which is assigned by IANA in the First Come First Server range, is used for SDWAN edge nodes to propagate its WAN port properties to its controller.

In the context of this document, BGP Route Reflectors (RR) is the component of the SDWAN Controller that receives the BGP UPDATE from SDWAN edges and in turns propagate the information to a group of authorized SDWAN edges reachable via overlay networks.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/ietf/lid-abstracts.txt>

The list of Internet-Draft Shadow Directories can be accessed at <http://www.ietf.org/shadow.html>

This Internet-Draft will expire on Dec 5, 2020.

Copyright Notice

Copyright (c) 2019 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction.....	3
2. Conventions used in this document.....	3
2.1. Information to be propagated for SDWAN UPDATE.....	4
2.2. SAFI under the MP-NLRI.....	6
2.3. How about a new Path Attribute under BGP UPDATE?.....	6
3. SDWAN WAN Port Identifier encoding in the MP-NLRI Path Attribute	6
4. WAN Port Properties encoding in the Tunnel Path Attribute.....	8
4.1. Port Ext SubTLV for NAT.....	9
4.2. IPsec Security Association Property.....	10
4.3. Remote Endpoint.....	11
5. Manageability Considerations.....	12
6. Security Considerations.....	12
7. IANA Considerations.....	12
8. References.....	12
8.1. Normative References.....	12
8.2. Informative References.....	13
9. Acknowledgments.....	14

1. Introduction

[Net2Cloud-Problem] introduces using SDWAN to reach dynamic workloads in multiple third-party data centers and aggregate multiple underlay paths, including public untrusted networks, provided by different service providers.

[SDWAN-BGP-USAGE] describes multiple SDWAN scenarios and illustrates how BGP is used as control plane for the SDWAN networks.

The document describes BGP UPDATE for SDWAN edge nodes to propagate its WAN port properties to RR.

2. Conventions used in this document

Cloud DC: Off-Premise Data Centers that usually host applications and workload owned by different organizations or tenants.

Controller: Used interchangeably with SDWAN controller to manage SDWAN overlay path creation/deletion and monitor the path conditions between sites.

CPE-Based VPN: Virtual Private Secure network formed among CPEs. This is to differentiate from most commonly used PE-based VPNs a la RFC 4364.

MP-NLRI: The MP_REACH_NLRI Path Attribute defined in RFC4760.

SDWAN End-point: An WAN port (logical or physical) of a SDWAN edge node. (If "endpoint" is used, it refers to a SDWAN End-point).

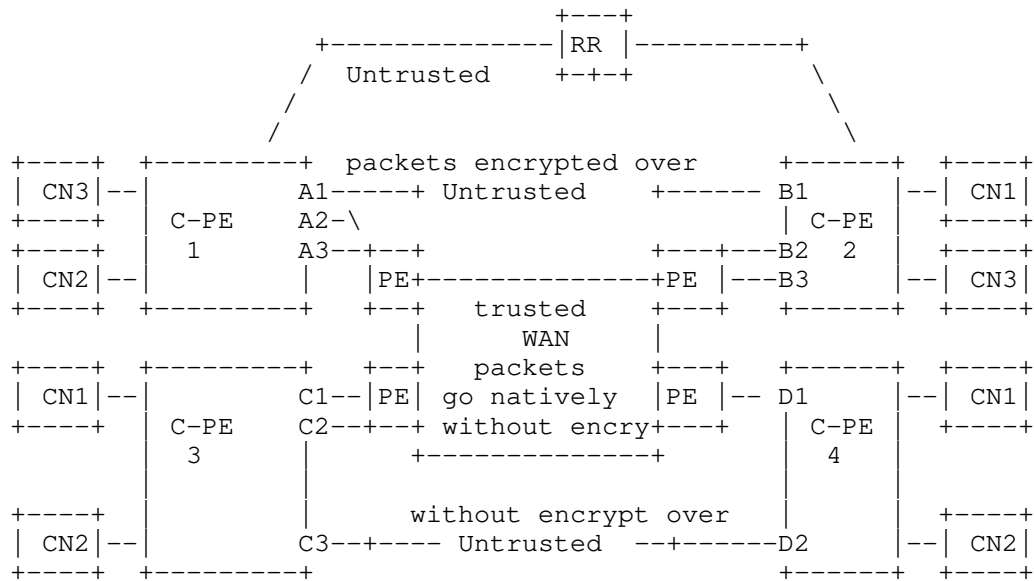
OnPrem: On Premises data centers and branch offices

SDWAN: Software Defined Wide Area Network. In this document, "SDWAN" refers to the solutions of pooling WAN bandwidth from multiple underlay networks to get better WAN bandwidth management, visibility & control. When the underlay networks are private networks, traffic can be

forwarded without additional encryption; when the underlay networks are public, such as Internet, some traffic needs to be encrypted when forwarding through those WAN ports (depending on user provided policies).

2.1. Information to be propagated for SDWAN UPDATE

Figure below shows the Hybrid SDWAN scenario:



CN: Client Network

Figure 1: Hybrid SDWAN

Using C-PE2 for illustration, C-PE2 needs to send out two separate BGP UPDATE messages.

BGP UPDATE #1 is to propagate C-PE2 attached routes, which are the regular VPN (L3VPN or EVPN) BGP Route UPDATE message,

MP-NLRI Path Attribute

```
Nexthop (C-PE2)
NLRI
  10.1.x.x.
  VLAN 15
  12.1.1x
Tunnel-Encap Path Attribute
  Details of any tunnels that applicable to the routes carried
  by the MP-NLRI Path Attribute
```

BGP UPDATE #2 is to propagate C-PE2's WAN port properties to RR, which should include:

- Identifier for the WAN Port
- The NAT property for the WAN Port
- The minimum IPsec information for establishing Port based IPsec.

Separating WAN port properties UPDATE from client routes UPDATE makes the implementation simpler, because the properties of a SDWAN node's WAN Port can change independent from the client routes attached to the C-PE2. WAN port properties change can be caused by many factors, such as ISP service agreement changes for the service connected to the WAN Port, the WAN port being disabled, or its IPsec property changes, etc. Since most SDWAN edges only have a small number of WAN ports, the disadvantage of multiple BGP UPDATE messages to advertise properties of those WAN ports is relatively small.

Following the same approach used by [idr-segment-routing-te-policy] where the SR Policy identifier is encoded in the MP-NLRI Path Attribute and the detailed SR Policies are encoded in the Tunnel Path attribute, the BGP UPDATE for SDWAN WAN port can have the WAN Port identifier encoded in the MP-NLRI Path Attribute and the associated WAN Port properties encoded in the Tunnel Path Attribute.

Receivers of the UPDATE can associate the SDWAN node identifier, site identifier with the node's WAN Port properties.

2.2. SAFI under the MP-NLRI

It is possible to continue using the same IP SAFI in the MP-NLRI [RFC4760] Path Attribute for advertising the SDWAN WAN port properties. If the same IP SAFI used, receiver needs extra logic to differentiate regular BGP MP-NLRI routes advertisement from the SDWAN WAN port properties advertisement and recognize the extra Site ID field added to the MP-NLRI. The benefit of using the same IP SAFI is that the UPDATE can traverse existing routers without being dropped. However, the SDWAN UPDATE is only between SDWAN edge and the RR, all the intermediate nodes treat the UPDATE message as regular IP data frame.

That is why it is simpler to follow the same approach used by [idr-segment-routing-te-policy] to have a unique SAFI (IANA assigned SDWAN SAFI = 74) mainly to differentiate the SDWAN UPDATE from regular route UPDATE.

This SDWAN SAFI is for a scenario where one SDWAN edge node has multiple WAN ports, some of which connected to private networks and others connected to public untrusted networks [Scenario #2 described in the [SDWAN-BGP-USAGE]]. The same routes attached to the SDWAN can be reached by the private networks without encryption (for better performance) or by the public networks with encryption.

2.3. How about a new Path Attribute under BGP UPDATE?

It is also possible to have a new Path Attribute, say SDWAN Path Attribute, combined with Tunnel Path Attribute to advertise SDWAN WAN Port properties. Besides having a different Path Attribute ID, everything else is same as using MP-NLRI & Tunnel Path Attributes.

3. SDWAN WAN Port Identifier encoding in the MP-NLRI Path Attribute

SDWAN WAN Port Identifier needs the following attributes

- locally significant port number,
- the location of the SDWAN device, and
- the globally routable address for the WAN Port.

Here is the encoding for those attributes in the NLRI field within the MP_REACH_NLRI Path Attribute of RFC4760, under a SDWAN SAFI (code = 74):

-----+		
	NLRI Length	1 octet
-----+		
	SDWAN-Type	2 Octets
-----+		
	Port-Local-ID	4 octets
-----+		
	SDWAN-Site-ID	4 octets
-----+		
	SDWAN-Node-ID	4 or 16 octets
-----+		

where:

- NLRI Length: 1 octet of length expressed in bits as defined in [RFC4760].
- SDWAN-Type: to define the encoding of the rest of the SDWAN NLRI. There could be different sub-TLVs for different SDWAN WAN ports and their associated policies.
- Port local ID: SDWAN edge node Port identifier, which can be locally significant. Each port can have unique properties. For example, some ports may get ISP or DHCP assigned IP addresses (IPv4 or IPv6), some may have private IP addresses that packets to/from those ports have to traverse NAT. The detailed properties about the port are further encoded in the subTLVs, e.g. Port-subTLV under the Tunnel Path Attribute.
- SDWAN-Site-ID: used to identify a common property shared by a set of SDWAN edge nodes, such as the property of a specific geographic location shared by a group of SDWAN edge nodes. The property is used to steer an overlay route to traverse specific geographic locations for various reasons, such as to comply

regulatory rules, to utilize specific value added services, or others.

- SDWAN EdgeNode ID: the SDWAN edge node identifier, which has to be a routable address (IPv4 or IPv6) within the WAN.

4. WAN Port Properties encoding in the Tunnel Path Attribute

The content of the SDWAN Port properties is encoded in the Tunnel Encapsulation Attribute defined in [Tunnel-Encap] using a new Tunnel-Type TLV (code point to be assigned by IANA from the "BGP Tunnel Encapsulation Attribute Tunnel Types" registry).

Tunnel Encaps Path Attribute (Code = 23)

Tunnel Type: SDWAN-WAN-Port

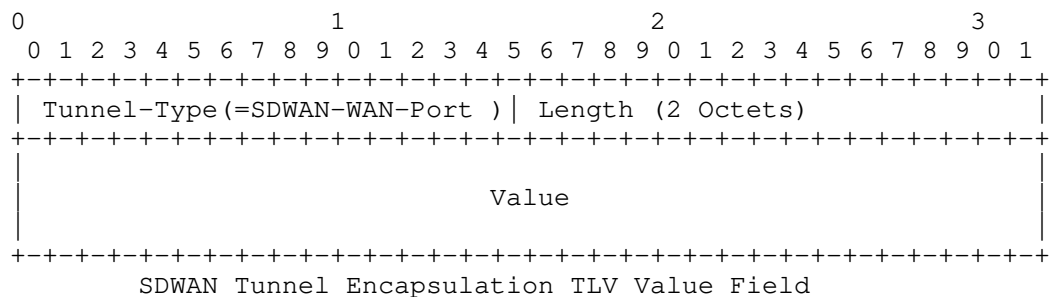
Followed by the detailed properties encoded as subTLV, such as

SubTLV for NAT

SubTLV for IPsec-SA Attribute

SubTLV for ISP connected to the WAN port

The Tunnel Encaps Attribute are defined as follows:



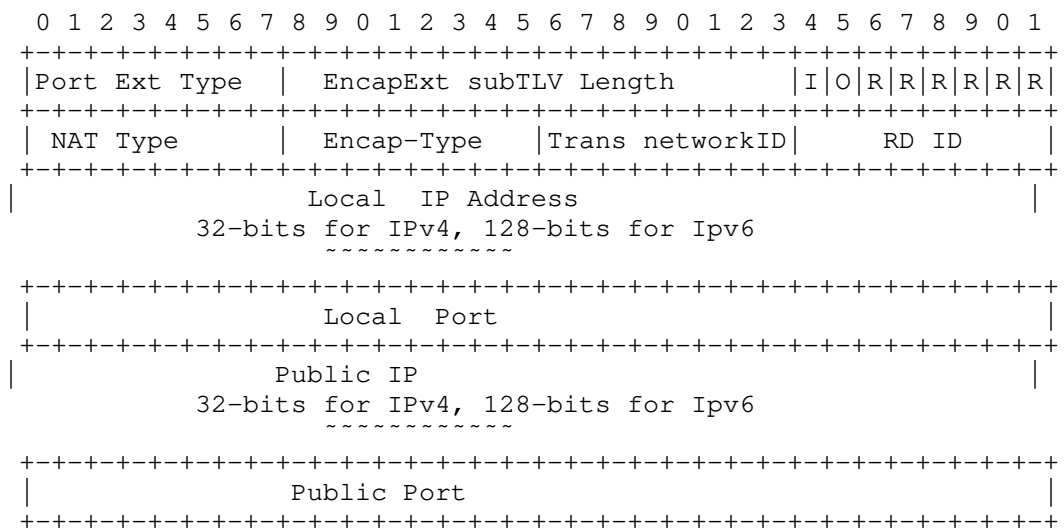
Where:

Tunnel Type is SDWAN-WAN-Port (to be assigned by IANA).

4.1. Port Ext SubTLV for NAT

NAT information is encoded is the Port Ext sub-TLV is for describing the NAT property if the port has private address and the network identifier to which the WAN port is connected, etc.

A SDWAN edge node can inquire STUN (Session Traversal of UDP Through Network Address Translation RFC 3489) Server to get the NAT property, the public IP address and the Public Port number to pass to peers.



Where:

- o Port Ext Type: indicate it is the Port Ext SubTLV.
- o PortExt subTLV Length: the length of the subTLV.
- o Flags:
 - I bit (CPE port address or Inner address scheme)
 - If set to 0, indicate the inner (private) address is IPv4.
 - If set to 1, it indicates the inner address is IPv6.
 - O bit (Outer address scheme):
 - If set to 0, indicate the public (outer) address is IPv4.

If set to 1, it indicates the public (outer) address is IPv6.

- R bits: reserved for future use. Must be set to 0 now.

- o NAT Type.without NAT; 1:1 static NAT; Full Cone; Restricted Cone; Port Restricted Cone; Symmetric; or Unknown (i.e. no response from the STUN server).
- o Encap Type.the supported encapsulation types for the port facing public network, such as IPsec+GRE, IPsec+VxLAN, IPsec without GRE, GRE (when packets don't need encryption)
- o Transport Network ID.Central Controller assign a global unique ID to each transport network.
- o RD ID.Routing Domain ID.Need to be global unique.
- o Local IP.The local (or private) IP address of the port.
- o Local Port.used by Remote SDWAN edge node for establishing IPsec to this specific port.
- o Public IP.The IP address after the NAT. If NAT is not used, this field is set to NULL.
- o Public Port.The Port after the NAT. If NAT is not used, this field is set to NULL.

4.2. IPsec Security Association Property

The IPsecSA sub-TLV is for the SDWAN edge node to establish IPsec security association with their peers via the port that face untrusted network. The minimum set of the IPsec information is from [CONTROLLER-IKE].

```

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
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|IPsec-SA Type |IPsecSA Length|Flag|
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
| Transform   | Transport   | AH   | ESP   |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|               Key Counter               |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
| key1 length | Public Key |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
| key2 length | Nonce     |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

```

+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| key3 length | key3 (for potential other keys |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|           Duration           |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+

```

Where:

- o IPsec-SA SubTLV Type: to be assigned by IANA. The type value has to be between 128~255 because IPsec-SA subTLV needs 2 bytes for length to carry the needed information.
- o IPsec-SA subTLV Length (2 Byte): 25 (or more)
- o Flags: 1 octet of flags. None are defined at this stage. Flags SHOULD be set to zero on transmission and MUST be ignored on receipt.
- o Transform (1 Byte): the value can be AH, ESP, or AH+ESP.
- o Transport (1 byte): the value can be Tunnel Mode or Transport mode
- o AH (1 byte): AH authentication algorithms supported, which can be md5 | sha1 | sha2-256 | sha2-384 | sha2-512 | sm3. Each SDWAN edge node can have multiple authentication algorithms; send to its peers to negotiate the strongest one.
- o ESP (1 byte): ESP authentication algorithms supported, which can be md5 | sha1 | sha2-256 | sha2-384 | sha2-512 | sm3. Each SDWAN edge node can have multiple authentication algorithms; send to its peers to negotiate the strongest one. Default algorithm is AES-256.
- o Rekey Counter: 4 bytes
- o Public Key: IPsec public key
- o Nonce: IPsec Nonce
- o Key3.other potential key
- o Duration: SA life span.

4.3. Remote Endpoint

The Remote Endpoint sub-TLV is not used for SDWAN NLRI because

- o The SDWAN Node ID and Site ID are already encoded in the SDWAN NLRI,
- o The network connected by the SDWAN WAN port might have identifier that is more than the AS number. SDWAN controller might use its own specific identifier for the network.

- o The Transport-Network-ID in the EncapExt sub-TLV represents the SDWAN unique network identifier.

If the Remote Endpoint Sub-TLV is present, it is ignored by other SDWAN edge nodes.

5. Manageability Considerations

TBD - this needs to be filled out before publishing

6. Security Considerations

The document describes the encoding for SDWAN edge nodes to advertise its SDWAN WAN ports properties to their peers via untrusted & unsecure networks.

The secure propagation is achieved by secure channels, such as TLS, SSL, or IPsec, between the SDWAN edge nodes and the local controller RR.

[More details need to be filled in here]

7. IANA Considerations

This document requires the following IANA actions.

- o SDWAN Overlay SAFI = 74 assigned by IANA
- o SDWAN Route Type

8. References

8.1. Normative References

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

8.2. Informative References

- [RFC8192] S. Hares, et al, "Interface to Network Security Functions (I2NSF) Problem Statement and Use Cases", July 2017
- [RFC5521] P. Mohapatra, E. Rosen, "The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", April 2009.
- [CONTROLLER-IKE] D. Carrel, et al, "IPsec Key Exchange using a Controller", draft-carrel-ipsecme-controller-ike-01, work-in-progress.
- [Tunnel-Encap] E. Rosen, et al, "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-09, Feb 2018.
- [VPN-over-Internet] E. Rosen, "Provide Secure Layer L3VPNs over Public Infrastructure", draft-rosen-bess-secure-l3vpn-00, work-in-progress, July 2018
- [DMVPN] Dynamic Multi-point VPN:
<https://www.cisco.com/c/en/us/products/security/dynamic-multipoint-vpn-dmvpn/index.html>
- [DSVPN] Dynamic Smart VPN:
<http://forum.huawei.com/enterprise/en/thread-390771-1-1.html>
- [ITU-T-X1036] ITU-T Recommendation X.1036, "Framework for creation, storage, distribution and enforcement of policies for network security", Nov 2007.
- [Net2Cloud-Problem] L. Dunbar and A. Malis, "Seamless Interconnect Underlay to Cloud Overlay Problem Statement", draft-dm-net2cloud-problem-statement-02, June 2018
- [Net2Cloud-gap] L. Dunbar, A. Malis, and C. Jacquenet, "Gap Analysis of Interconnecting Underlay with Cloud Overlay", draft-dm-net2cloud-gap-analysis-02, work-in-progress, Aug 2018.

[Tunnel-Encap] E. Rosen, et al "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-10, Aug 2018.

9. Acknowledgments

Acknowledgements to Wang Haibo, Hao Weiguo, and ShengCheng for implementation contribution; Many thanks to Jim Guichard, John Scudder, Darren Dukes, Andy Malis, Rachel Huang and Donald Eastlake for their review and contributions.

This document was prepared using 2-Word-v2.0.template.dot.

Authors' Addresses

Linda Dunbar
Futurewei
Email: ldunbar@futurewei.com

Sue Hares
Hickory Hill Consulting
Email: shares@ndzh.com

idr
Internet-Draft
Intended status: Standards Track
Expires: September 10, 2020

J. Hu
Nokia
March 9, 2020

BGP Provisioned IPsec Tunnel Configuration
draft-hujun-idr-bgp-ipsec-02

Abstract

This document defines a method of using BGP to provide IPsec tunnel configuration along with NLRI, it uses and extends tunnel encapsulation attribute as specified in [I-D.ietf-idr-tunnel-encaps] for IPsec tunnel.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

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

This Internet-Draft will expire on September 10, 2020.

Copyright Notice

Copyright (c) 2020 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	2
1.1. Terminology	3
2. Tunnel Encapsulation Attribute for IPsec	3
2.1. Local and Remote Prefix sub-TLV	4
2.2. Public Routing Instance sub-TLV	5
2.3. IPsec Configuration Tag sub-TLV	5
3. Operation	6
4. Semantics and Usage of IPsec Tunnel Encapsulation attribute .	10
4.1. Nested Tunnel	10
4.2. Other Operation Specifics	11
5. IANA Considerations	11
6. Security Considerations	12
7. Change Log	13
8. References	13
8.1. Normative References	13
8.2. Informative References	14
Author's Address	15

1. Introduction

IPsec is the standard for IP layer traffic protection, however in a big network where mesh connections are needed, configuring large number of IPsec tunnels is error prone and not scalable. So instead of pre-provision IPsec tunnels on each router, this document defines a method to allow router to advertise the IPsec tunnel configurations it requires to reach a given NLRI via BGP. This document does not intend to be one solution for all cases, the main use case is to simplify IPsec tunnel provision in networks under single administrative domain; it uses standard based components (IPsec/IKEv2[RFC7296] and BGP) with limited changes. There is no change to IPsec/IKEv2, and only limited changes to BGP.

IPsec tunnel in this document means IPsec tunnel mode as defined in [RFC4301].

IPsec tunnel configurations typically include following parts:

- o tunnel endpoint address (local and remote)
- o public routing instance, routing instance where IPsec packet is forwarded in
- o private routing instance, routing instance where payload packet is forwarded in
- o tunnel authentication method and credentials

- o IKE SA and CHILD SA transform (a.k.a crypto algorithms)
- o CHILD SA traffic selector
- o other: like lifetime, DPD timer, use of PFS ..etc

In order to minimize amount configurations signal via BGP, only following configurations are explicit advertised:

- o local tunnel endpoint address: BGP tunnel encapsulation attribute
- o public routing instance: sub-TLV in tunnel encapsulation attribute
- o CHILD SA traffic selector address range: NLRI and/or sub-TLV in tunnel encapsulation attribute

Other configurations are either derived or via tag mapping:

- o remote tunnel endpoint address: dynamic learned when received IKEv2 IKE_SA_INIT request
- o private routing instance: via route-target in same BGP UPDATE
- o tunnel authentication/credentials, traffic selector protocol/port range, IKE SA and CHILD SA transform, lifetime, DPD timer, PFS ..etc: all these configurations are implicitly signaled via IPsec configuration tag sub-TLV in tunnel encapsulation attribute

[I-D.ietf-idr-tunnel-encaps] defines a generic tunnel encapsulation attribute for BGP, however it needs to be extended to support IPsec tunnel.

1.1. Terminology

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

2. Tunnel Encapsulation Attribute for IPsec

This document extends tunnel encapsulation attribute specified in [I-D.ietf-idr-tunnel-encaps] by introducing following changes:

- o A tunnel type for IPsec tunnel: ESP tunnel mode (AH tunnel mode is not included in this document). Existing type 4 (IPsec in Tunnel-

mode) in IANA "BGP Tunnel Encapsulation Attribute Tunnel Types" registry could be reused

- o A new sub-TLV for public routing instance
- o A new sub-TLV for remote address prefix
- o A new sub-TLV for local address prefix
- o A new sub-TLV for IPsec configuration tag

Following existing sub-TLVs apply to IPsec tunnel encapsulation attribute:

- o Remote Endpoint: IPsec tunnel endpoint address
- o Embedded Label Handling: see Section 4 for detail

2.1. Local and Remote Prefix sub-TLV

Local prefix sub-TLV is an optional sub-TLV used to specify a list of address prefix that used as local traffic selector address ranges; if local prefix sub-TLV is not included, then prefixes in NLRI will be used; Remote prefix sub-TLV is a mandatory sub-TLV used to specify a list of address prefix that used as remote traffic selector address ranges; The IP version of local/remote prefix MUST be as same as IP version of prefix in NLRI. A single all zero prefix means any prefix is allowed. Local and remote prefix sub-TLV has same encoding as following:

```

+-----+
| list of prefixes (variable) |
+-----+
```

Figure 1: Source Prefix sub-TLV

Each prefix is encoded as following:

```

+-----+
| prefix Length (1 octet) |
+-----+
| Prefix (4 or 16 octets) |
+-----+
```

Figure 2: prefix

For a given IPsec tunnel TLV, local prefix sub-TLV MUST appear either zero or one time; remote prefix sub-TLV MUST appear only one time.

2.2. Public Routing Instance sub-TLV

Public routing instance sub-TLV is an optional sub-TLV used to specify the routing instance to which the remote point address belongs, if tunnel encapsulation attribute doesn't include this TLV, then the routing instance is the same to which BGP session belongs. the value field of the sub-TLV consist a route target community as defined in [RFC4360].

For a given IPsec tunnel TLV, public routing instance sub-TLV MUST appear either zero or one time.

2.3. IPsec Configuration Tag sub-TLV

This sub-TLV represents the IPsec configurations (like IPsec transform) that are not explicit advertised by other sub-TLVs specified in this documentation; the meaning of this sub-TLV is local to the administrative domain. Follow are some examples:

- o tag value T1 map to following configurations:
 - * Certificate trust-anchor: CA-1
 - * IKE_SA/CHILD_SA transform: AES-GCM-128
 - * Diffie-Hellman Group: 15
 - * Perfect Forward Secrecy: No
 - * local/remote Traffic selector protocol: any
 - * local/remote Traffic selector port range: any
 - * IKE_SA lifetime: 24 hours
 - * CHILD_SA lifetime: 1 hour
 - * DPD interval: 30 seconds
 - * ESP extended sequence number: no
- o tag value T2 map to following configurations:
 - * Certificate trust-anchor: CA-2
 - * IKE_SA/CHILD_SA transform: AES-GCM-256
 - * Diffie-Hellman Group: 20

- * Perfect Forward Secrecy: Yes with group 20
- * local/remote Traffic selector protocol: UDP
- * local/remote Traffic selector port range: any
- * IKE_SA lifetime: 48 hours
- * CHILD_SA lifetime: 2 hours
- * DPD interval: 10 seconds
- * ESP extended sequence number: yes

The value field of this sub-TLV is 4 octets long. each IPsec tunnel TLV SHOULD only contain one IPsec configuration tag sub-TLV;

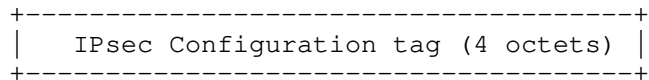


Figure 3: IPsec Configuration Tag

For a given IPsec tunnel TLV, IPsec configuration tag sub-TLV MUST appear only one time.

3. Operation

Following are the rules of operation:

1. All routers are in same administrative domain
2. All routers are pre-provisioned with Mapping between IPsec configuration tag value and IPsec configurations include authentication method/credentials
3. If a given NLRI need IPsec protection, then advertising router need to include an IPsec tunnel encapsulation attribute, along with the NLRI in BGP UPDATE U;
4. When a router need to forward a packet along a path is determined by a BGP UPDATE which has a tunnel encapsulation attribute that contains one or more IPsec tunnel TLV, and router decides use IPsec based on local policy, then the router use first feasible CHILD_SA, a CHILD SA is considered as feasible when it meets all following conditions:

- * its private routing instance is same as routing instance to which the packet to be forwarded belongs
 - * its public routing instance is same as indicated by the Public Routing Instance sub-TLV; if the sub-TLV doesn't exist, then it is same as routing instance to which BGP session belongs
 - * its peer tunnel address is same as indicated by Remote Endpoint sub-TLV
 - * the source and destination address of the packet to be forwarded falls in the range of CHILD SA's traffic selector
 - * its transform and other configuration maps to the tag indicated in the IPsec configuration tag sub-TLV
5. If router can't find such CHILD SA, then it will use IKEv2 to create one; if there are multiple IPsec tunnel TLVs in U, then it need to select one from feasible TLVs, a IPsec tunnel TLV is considered as feasible when it meets all following requirements:
- * the source address of the packet must fall in one of Remote Prefixes
 - * the destination address of the packet must fall one of Source Prefixes
 - * the Remote Endpoint, along with Public Routing Instance sub-TLV identifies an IP address that is reachable
6. If there are multiple feasible IPsec tunnel TLV exists, then select the TLV using following rules in order:
1. TLV with smallest local address range as indicated by Remote Prefix sub-TLV
 2. TLV with smallest remote address range as indicated by Local Prefix sub-TLV (NLRI prefix if local prefix sub-TLV is not included in TLV)
7. After an IPsec TLV is selected, router uses IKEv2 to create the CHILD_SA:
- * public/private routing instance, peer's tunnel address are chosen based on above rules
 - * Traffic Selector:

- ```
* For each TS in TSi:

+ address range: the prefix specified in Remote Prefix sub-
 TLV

+ protocol: tag mapped configuration

+ port range: tag mapped configuration

* for each TS in TSr:

+ address range: prefixes specified by Local Prefix sub-TLV
 if it exists; otherwise use the prefix specified by the
 NLRI

+ protocol: tag mapped configuration

+ port range: tag mapped configuration
```

The operation of BGP provisioned IPsec configuration is illustrated with following example:

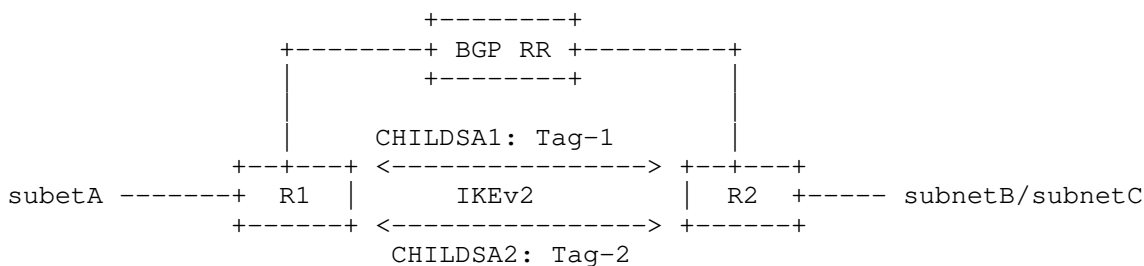


Figure 4: Operation Example

There are following traffic protection requirements:

- o subnetA - subnetB: ESP tunnel, CHACHA20\_POLY1305 , mapping to tag Tag-1
- o subnetA - subnetC: ESP tunnel, NULL-AES-GMAC-256 , mapping to tag Tag-2
- o note: other IPsec configurations, like IKE\_SA lifetime ..etc, are the same for both Tag-1 and Tag-2; not listed here for sake of



Both R1 and R2 are provisioned with IPsec authentication credentials and configurations corresponding to Tag-1 and Tag-2; both Tag-1 and Tag-2 map to traffic selector protocol any and port range any.

- o R1 advertise subnetA in BGP UPDATE, which has a tunnel encapsulation attribute that contains two IPsec tunnel TLVs:
  - \* TLV-1: endpoint R1TunnelAddr, tag sub-TLV Tag-1 and subnetB in Remote Prefix sub-TLV.
  - \* TLV-2: endpoint R1TunnelAddr, tag sub-TLV Tag-2 and subnetC in Remote Prefix sub-TLV.
- o R2 advertise subnetB in BGP UPDATE, which has a tunnel encapsulation attribute that contains one IPsec tunnel TLV: R2TunnelAddr, tag sub-TLV Tag-1 and subnetA in Remote Prefix sub-TLV.
- o R2 advertise subnetC in BGP UPDATE, which has a tunnel encapsulation attribute that contains one IPsec tunnel TLV: R2TunnelAddr, tag sub-TLV Tag-2 and subnetA in Remote Prefix sub-TLV.
- o R1 received a packet from subnetA destined to subnetB, since BGP UPDATE contain subnetB also contains an IPsec tunnel encapsulation attribute, there is no existing CHILD SA could be used, based on the rules described in this section, R1 select TLV-1 and uses IKEv2 to establish an IPsec tunnel to R2TunnelAddr, using certificate authentication, create 1st CHILD SA CHILDSA1:
  - \* ESP transform: CHACHA20\_POLY1305
  - \* Traffic Selector:
    - + TSi: address subnetA, protocol any, port any
    - + TSr: address subnetB, protocol any, port any
- o after tunnel is created, R1 and R2 could forward traffic between subnetA and subnetB over CHILDSA1
- o R1 received a packet from subnetA destined to subnetC, CHILDSA1 can't be used for this packet, R1 select TLV-2 to create 2nd CHILD SA, and given there is already an IKE SA between R1 and R2, R1 uses existing IKESA to create CHILDSA2:
  - \* ESP transform: NULL-AES-GMAC-256

\* Traffic Selector:

- + TSi: address subnetA, protocol any, port any
- + TSr: address subnetC, protocol any, port any
- o R1 and R2 could forward traffic between subnetA and subnetC over CHILDSA2

#### 4. Semantics and Usage of IPsec Tunnel Encapsulation attribute

IPsec tunnel encapsulation TLV has same usage and semantics as defined in [I-D.ietf-idr-tunnel-encaps] with following specific to IPsec tunnel:

- o Due to nature of IPsec, the payload packet could only be IPv4 or IPv6 packet, so it MAY be carried in any BGP UPDATE message whose AFI/SAFI is 1/1 (IPv4 Unicast), 2/1 (IPv6 Unicast).
- o For 1/128 (VPN-IPv4 Labeled Unicast), 2/128 (VPN-IPv6 Labeled Unicast), these NLRI has embedded label, which cause the payload packet can't be encapsulated in ESP packet, however with IPsec tunnel encapsulation, the label could be ignored during encapsulation since CHILD SA itself could be used to identify the private routing instance; so an UPDATE that include IPsec tunnel encapsulation attribute, which contains value 2 of Embedded Label Handling Sub-TLV, could be used to signal this type of setup.
- o For other types of AFI/SAFI, a nested tunnel setup could be used to get IPsec protection, for example, an 25/70 (EVPN) payload packet could be encapsulated in VXLAN over IPsec tunnel. See Section 4.1 for further detail.

##### 4.1. Nested Tunnel

A nested tunnel could be used for payload packet type that can't be encapsulated in IPsec tunnel directly, e.g. an Ethernet packet of EVPN service. Following is an example of using VXLAN over IPsec tunnel for EVPN service:

- o R1 need to forward an Ethernet packet P
- o the path along which P is to be forwarded is determined by BGP UPDATE U1, which has a VXLAN tunnel encapsulation attribute and the next-hop is router R2
- o the best path to R2 is a BGP route that was advertised in BGP UPDATE U2, which has an IPsec tunnel encapsulation TLV.

- o R1 will encapsulate P in a VXLAN tunnel as indicated in U1, then encapsulate VXLAN packet into IPsec tunnel as indicated in U2
- o if tag sub-TLV is used, then both U1 and U2 MUST have matching tag sub-TLV, otherwise the VXLAN packet will not be sent through IPsec tunnels identified in U2

#### 4.2. Other Operation Specifics

Following are some operation specific rules:

1. An IPsec dead peer detection mechanism, like IKEv2 DPD or BFD over IPsec, SHOULD be used to monitor liveness of IPsec tunnel;
2. If IPsec peer goes down, as described in section 5 of [I-D.ietf-idr-tunnel-encaps], packet forwarding router chooses another functional tunnel, specified by another tunnel TLV of same BGP route if there is any, to forward the packet; if there is no such tunnel, then router MAY drop the packet or MAY forward packet as it would had the Tunnel Encapsulation attribute not been present. this is matter of local policy.
3. After IPsec peer goes down, packet forwarding router SHOULD try to re-establish IPsec tunnel with certain hold-down timer and back-off mechanism. the detail is up to implementation. also IKEv2 session resumption [RFC5723] MAY be used to efficiently re-create tunnel;
4. When router receives a packet destined to a BGP route it advertised but does not have any of tunnel encapsulation in the BGP route, it MAY drop it or MAY accept it; this is matter of local policy. by default, the packet should be accepted.
5. As with all types of tunnel technology, IPsec tunnel adds overhead (crypto & encapsulation) to the packet, which often causes MTU issues, deployment SHOULD take tunnel overhead into MTU consideration.

#### 5. IANA Considerations

This document reuses "IPsec in Tunnel-mode"(4) as BGP Tunnel Encapsulation Attribute Tunnel Types.

This document will request new values in IANA "BGP Tunnel Encapsulation Attribute Sub-TLVs" registry for following sub-TLV:

- o public routing instance
- o remote address prefix
- o local address prefix
- o IPsec configuration tag

## 6. Security Considerations

IKEv2 is used to create IPsec tunnel, which ensures following:

- o Traffic protection keys are generated dynamically during IKEv2 negotiation, only known by participating peer of the IPsec tunnel; there is no central node to manage and distribute all keys.
- o IKEv2 rekey mechanism refresh keys regularly; PFS(Perfect Forward Secrecy) provides additional protection;
- o Secure authentication mechanism that only allow authenticated peer to create tunnel
- o Traffic Selector guarantee that only agreed traffic is allowed to be forwarded within the IPsec tunnel;
- o Using a separate, dedicate protocol(IKEv2) for key management/ authentication ensure they are not tied to BGP, all existing and future IKEv2 features could be used without changing BGP;

There is concern that malicious party might manipulate IPsec tunnel encapsulation attribute to divert traffic, however this risk could be mitigated by IKEv2 mutual authentication.

BGP route filter include outbound route filter [RFC5291], Origin Validation [RFC6811] and BGPSec [RFC8205] could be used to further secure BGP UPDATE message.

IKEv2 cookie [RFC7296] and varies mechanisms defined including client puzzle defined in [RFC8019] could be used to protect IKEv2 from Distributed Denial-of-Service Attacks.

Follow latest IETF ESP/IKEv2 implementation requirement and guidance ([RFC8221] and [RFC8247] at time of writing) to make sure always using secure and up-to-date cryptographic algorithms;

## 7. Change Log

- o v00 March 04, 2019: initial draft
- o v01 Sep 04, 2019:
  - \* replaces color sub-TLV with a new IPsec configuration tag sub-TLV
  - \* add rule on selecting TLV when there multiple feasible TLVs in section Section 3
  - \* change crypto used in example of section Section 3
  - \* change title from "BGP Signaled IPsec Tunnel Configuration" to "BGP Provisioned IPsec Tunnel Configuration"
  - \* Add a section Section 4.2 on some operation specifics
  - \* add more content in Section 6
  - \* add specification of number of time each new sub-TLV allowed in a given tunnel TLV
  - \* add clarification in section Section 1 to clarify IPsec tunnel means IPsec tunnel mode
  - \* traffic selector protocol and port range now come from tag mapped configuration
- o v02 March 09, 2020
  - \* increase version number to keep draft afloat

## 8. References

### 8.1. Normative References

- [I-D.ietf-idr-tunnel-encaps]  
Patel, K., Velde, G., and S. Ramachandra, "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-15 (work in progress), December 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>>.

- [RFC4301] Kent, S. and K. Seo, "Security Architecture for the Internet Protocol", RFC 4301, DOI 10.17487/RFC4301, December 2005, <<https://www.rfc-editor.org/info/rfc4301>>.
- [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>>.
- [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>>.

## 8.2. Informative References

- [RFC5291] Chen, E. and Y. Rekhter, "Outbound Route Filtering Capability for BGP-4", RFC 5291, DOI 10.17487/RFC5291, August 2008, <<https://www.rfc-editor.org/info/rfc5291>>.
- [RFC5723] Sheffer, Y. and H. Tschofenig, "Internet Key Exchange Protocol Version 2 (IKEv2) Session Resumption", RFC 5723, DOI 10.17487/RFC5723, January 2010, <<https://www.rfc-editor.org/info/rfc5723>>.
- [RFC6811] Mohapatra, P., Scudder, J., Ward, D., Bush, R., and R. Austein, "BGP Prefix Origin Validation", RFC 6811, DOI 10.17487/RFC6811, January 2013, <<https://www.rfc-editor.org/info/rfc6811>>.
- [RFC7296] Kaufman, C., Hoffman, P., Nir, Y., Eronen, P., and T. Kivinen, "Internet Key Exchange Protocol Version 2 (IKEv2)", STD 79, RFC 7296, DOI 10.17487/RFC7296, October 2014, <<https://www.rfc-editor.org/info/rfc7296>>.
- [RFC8019] Nir, Y. and V. Smyslov, "Protecting Internet Key Exchange Protocol Version 2 (IKEv2) Implementations from Distributed Denial-of-Service Attacks", RFC 8019, DOI 10.17487/RFC8019, November 2016, <<https://www.rfc-editor.org/info/rfc8019>>.
- [RFC8205] Lepinski, M., Ed. and K. Sriram, Ed., "BGPsec Protocol Specification", RFC 8205, DOI 10.17487/RFC8205, September 2017, <<https://www.rfc-editor.org/info/rfc8205>>.

- [RFC8221] Wouters, P., Migault, D., Mattsson, J., Nir, Y., and T. Kivinen, "Cryptographic Algorithm Implementation Requirements and Usage Guidance for Encapsulating Security Payload (ESP) and Authentication Header (AH)", RFC 8221, DOI 10.17487/RFC8221, October 2017, <<https://www.rfc-editor.org/info/rfc8221>>.
- [RFC8247] Nir, Y., Kivinen, T., Wouters, P., and D. Migault, "Algorithm Implementation Requirements and Usage Guidance for the Internet Key Exchange Protocol Version 2 (IKEv2)", RFC 8247, DOI 10.17487/RFC8247, September 2017, <<https://www.rfc-editor.org/info/rfc8247>>.

## Author's Address

Hu Jun  
Nokia  
777 East Middlefield Road  
Mountain View CA 95148  
United States

Email: [jun.hu@nokia.com](mailto:jun.hu@nokia.com)

Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: 3 October 2022

A. Azimov  
Qrator Labs & Yandex  
E. Bogomazov  
Qrator Labs  
R. Bush  
Internet Initiative Japan & Arrcus, Inc.  
K. Patel  
Arrcus  
K. Sriram  
USA NIST  
1 April 2022

Route Leak Prevention and Detection using Roles in UPDATE and OPEN  
Messages  
draft-ietf-idr-bgp-open-policy-24

Abstract

Route leaks are the propagation of BGP prefixes that violate assumptions of BGP topology relationships, e.g., announcing a route learned from one transit provider to another transit provider or a lateral (i.e., non-transit) peer or announcing a route learned from one lateral peer to another lateral peer or a transit provider. These are usually the result of misconfigured or absent BGP route filtering or lack of coordination between autonomous systems (ASes). Existing approaches to leak prevention rely on marking routes by operator configuration, with no check that the configuration corresponds to that of the eBGP neighbor, or enforcement that the two eBGP speakers agree on the peering relationship. This document enhances the BGP OPEN message to establish an agreement of the peering relationship on each eBGP session between autonomous systems in order to enforce appropriate configuration on both sides. Propagated routes are then marked according to the agreed relationship, allowing both prevention and detection of route leaks.

Requirements Language

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

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.



Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

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

This Internet-Draft will expire on 3 October 2022.

#### Copyright Notice

Copyright (c) 2022 IETF Trust and the persons identified as the document authors. All rights reserved.

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

#### Table of Contents

|                                                   |    |
|---------------------------------------------------|----|
| 1. Introduction . . . . .                         | 3  |
| 2. Terminology . . . . .                          | 3  |
| 2.1. Peering Relationships . . . . .              | 4  |
| 3. BGP Role . . . . .                             | 5  |
| 3.1. BGP Role Capability . . . . .                | 5  |
| 3.2. Role Correctness . . . . .                   | 6  |
| 4. BGP Only to Customer (OTC) Attribute . . . . . | 8  |
| 5. Additional Considerations . . . . .            | 10 |
| 6. IANA Considerations . . . . .                  | 10 |
| 7. Security Considerations . . . . .              | 11 |
| 8. References . . . . .                           | 12 |
| 8.1. Normative References . . . . .               | 12 |
| 8.2. Informative References . . . . .             | 13 |
| Acknowledgments . . . . .                         | 14 |
| Contributors . . . . .                            | 14 |
| Authors' Addresses . . . . .                      | 14 |

## 1. Introduction

Route leaks are the propagation of BGP prefixes that violate assumptions of BGP topology relationships, e.g., announcing a route learned from one transit provider to another transit provider or a lateral (i.e., non-transit) peer or announcing a route learned from one lateral peer to another lateral peer or a transit provider [RFC7908]. These are usually the result of misconfigured or absent BGP route filtering or lack of coordination between autonomous systems (ASes).

Existing approaches to leak prevention rely on marking routes by operator configuration, with no check that the configuration corresponds to that of the eBGP neighbor, or enforcement that the two eBGP speakers agree on the relationship. This document enhances the BGP OPEN message to establish an agreement of the relationship on each eBGP session between autonomous systems in order to enforce appropriate configuration on both sides. Propagated routes are then marked according to the agreed relationship, allowing both prevention and detection of route leaks.

This document specifies a means of replacing the operator-driven configuration-based method of route leak prevention, described above, with an in-band method for route leak prevention and detection.

This method uses a new configuration parameter, BGP Role, which is negotiated using a BGP Role Capability in the OPEN message [RFC5492]. An eBGP speaker may require the use of this capability and confirmation of BGP Role with a neighbor for the BGP OPEN to succeed.

An optional, transitive BGP Path Attribute, called Only to Customer (OTC), is specified in Section 4. It prevents ASes from creating leaks and detects leaks created by the ASes in the middle of an AS path. The main focus/applicability is the Internet (IPv4 and IPv6 unicast route advertisements).

## 2. Terminology

The terms "local AS" and "remote AS" are used to refer to the two ends of an eBGP session. The "local AS" is the AS where the protocol action being described is to be performed, and "remote AS" is the AS at the other end of the eBGP session in consideration.

The use of the term "route is ineligible" in this document has the same meaning as in [RFC4271], i.e., "route is ineligible to be installed in Loc-RIB and will be excluded from the next phase of route selection."

## 2.1. Peering Relationships

The terms for peering relationships defined and used in this document (see below) do not necessarily represent business relationships based on payment agreements. These terms are used to represent restrictions on BGP route propagation, sometimes known as the Gao-Rexford model [Gao]. The terms Provider, Customer, and Peer used here are synonymous to the terms "transit provider", "customer", and "lateral (i.e., non-transit) peer", respectively, used in [RFC7908].

The following is a list of BGP Roles for eBGP peering and the corresponding rules for route propagation:

Provider: MAY propagate any available route to a Customer.

Customer: MAY propagate any route learned from a Customer, or locally originated, to a Provider. All other routes MUST NOT be propagated.

Route Server (RS): MAY propagate any available route to a Route Server Client (RS-Client).

Route Server Client (RS-Client): MAY propagate any route learned from a Customer, or locally originated, to an RS. All other routes MUST NOT be propagated.

Peer: MAY propagate any route learned from a Customer, or locally originated, to a Peer. All other routes MUST NOT be propagated.

If the local AS has one of the above Roles (in the order shown), then the corresponding peering relationship with the remote AS is Provider-to-Customer, Customer-to-Provider, RS-to-RS-Client, RS-Client-to-RS, or Peer-to-Peer (i.e., lateral peers), respectively. These are called normal peering relationships.

If the local AS has more than one peering role with the remote AS such peering relation is called Complex. An example is when the peering relationship is Provider-to-Customer for some prefixes while it is Peer-to-Peer for other prefixes [Gao].

A BGP speaker may apply policy to reduce what is announced, and a recipient may apply policy to reduce the set of routes they accept.

Violation of the route propagation rules listed above may result in route leaks [RFC7908]. Automatic enforcement of these rules should significantly reduce route leaks that may otherwise occur due to manual configuration mistakes.

As specified in Section 4, the Only to Customer (OTC) Attribute is used to identify all the routes in the AS that have been received from a Peer, Provider, or RS.

### 3. BGP Role

The BGP Role characterizes the relationship between the eBGP speakers forming a session. One of the Roles described below SHOULD be configured at the local AS for each eBGP session (see definitions in Section 2) based on the local AS's knowledge of its Role. The only exception is when the eBGP connection is Complex (see Section 5). BGP Roles are mutually confirmed using the BGP Role Capability (described in Section 3.1) on each eBGP session.

Allowed Roles for eBGP sessions are:

- \* Provider - the local AS is a transit Provider of the remote AS;
- \* Customer - the local AS is a transit Customer of the remote AS;
- \* RS - the local AS is a Route Server (usually at an Internet exchange point) and the remote AS is its RS-Client;
- \* RS-Client - the local AS is a client of an RS and the RS is the remote AS;
- \* Peer - the local and remote ASes are Peers (i.e., have a lateral peering relationship).

#### 3.1. BGP Role Capability

The BGP Role Capability is defined as follows:

- \* Code - 9
- \* Length - 1 (octet)
- \* Value - integer corresponding to speaker's BGP Role (see Table 1).

| Value | Role name (for the local AS) |
|-------|------------------------------|
| 0     | Provider                     |
| 1     | RS                           |
| 2     | RS-Client                    |
| 3     | Customer                     |
| 4     | Peer (i.e., Lateral Peer)    |
| 5-255 | Unassigned                   |

Table 1: Predefined BGP Role Values

If BGP Role is locally configured, the eBGP speaker MUST advertise BGP Role Capability in the BGP OPEN message. An eBGP speaker MUST NOT advertise multiple versions of the BGP Role Capability. The error handling when multiple BGP Role Capabilities are received is described in Section 3.2.

### 3.2. Role Correctness

Section 3.1 described how BGP Role encodes the relationship on each eBGP session between autonomous systems (ASes).

The mere receipt of BGP Role Capability does not automatically guarantee the Role agreement between two eBGP neighbors. If the BGP Role Capability is advertised, and one is also received from the peer, the Roles MUST correspond to the relationships in Table 2. If the Roles do not correspond, the BGP speaker MUST reject the connection using the Role Mismatch Notification (code 2, subcode TBD).

| Local AS Role | Remote AS Role |
|---------------|----------------|
| Provider      | Customer       |
| Customer      | Provider       |
| RS            | RS-Client      |
| RS-Client     | RS             |
| Peer          | Peer           |

Table 2: Allowed Pairs of Role Capabilities

For backward compatibility, if the BGP Role Capability is sent but one is not received, the BGP Speaker SHOULD ignore the absence of the BGP Role Capability and proceed with session establishment. The locally configured BGP Role is used for the procedures described in Section 4.

An operator may choose to apply a "strict mode" in which the receipt of a BGP Role Capability from the remote AS is required. When operating in the "strict mode", if the BGP Role Capability is sent, but one is not received, then the connection is rejected using the Role Mismatch Notification (code 2, subcode TBD). See comments in Section 7.

If an eBGP speaker receives multiple but identical BGP Role Capabilities with the same value in each, then the speaker considers them to be a single BGP Role Capability and proceeds [RFC5492]. If multiple BGP Role Capabilities are received and not all of them have the same value, then the BGP speaker MUST reject the connection using the Role Mismatch Notification (code 2, subcode TBD).

The BGP Role value for the local AS (in conjunction with the OTC Attribute in the received UPDATE message) is used in the route leak prevention and detection procedures described in Section 4.

#### 4. BGP Only to Customer (OTC) Attribute

The Only to Customer (OTC) Attribute is an optional transitive path attribute of the UPDATE message with Attribute Type Code 35 and a length of 4 octets. The purpose of this attribute is to enforce that once a route is sent to a Customer, Peer, or RS-Client (see definitions in Section 2.1), it will subsequently go only to Customers. The attribute value is an AS number (ASN) determined by the procedures described below.

The following ingress procedure applies to the processing of the OTC Attribute on route receipt:

1. If a route with the OTC Attribute is received from a Customer or RS-Client, then it is a route leak and MUST be considered ineligible (see Section 2).
2. If a route with the OTC Attribute is received from a Peer (i.e., remote AS with a Peer Role) and the Attribute has a value that is not equal to the remote (i.e., Peer's) AS number, then it is a route leak and MUST be considered ineligible.
3. If a route is received from a Provider, Peer, or RS, and the OTC Attribute is not present, then it MUST be added with a value equal to the AS number of the remote AS.

The following egress procedure applies to the processing of the OTC Attribute on route advertisement:

1. If a route is to be advertised to a Customer, Peer, or RS-Client (when the sender is an RS), and the OTC Attribute is not present, then when advertising the route, an OTC Attribute MUST be added with a value equal to the AS number of the local AS.
2. If a route already contains the OTC Attribute, it MUST NOT be propagated to Providers, Peers, or RS(s).

The above-described procedures provide both leak prevention for the local AS and leak detection and mitigation multiple hops away. In the case of prevention at the local AS, the presence of an OTC Attribute indicates to the egress router that the route was learned from a Peer, Provider, or RS, and it can be advertised only to the customers. The same OTC Attribute which is set locally also provides a way to detect route leaks by an AS multiple hops away if a route is received from a Customer, Peer, or RS-Client. For example, if an AS sets the OTC Attribute on a route sent to a Peer and the route is subsequently received by a compliant AS from a Customer, then the receiving AS detects (based on the presence of the OTC Attribute) that the route is a leak.

The OTC Attribute might be set at the egress of the remote AS or at the ingress of the local AS, i.e., if the remote AS is non-compliant with this specification, then the local AS will have to set the OTC Attribute if it is absent. In both scenarios, the OTC value will be the same. This makes the scheme more robust and benefits early adopters.

The OTC Attribute is considered malformed if the length value is not 4. An UPDATE message with a malformed OTC Attribute SHALL be handled using the approach of "treat-as-withdraw" [RFC7606].

The BGP Role negotiation and OTC Attribute based procedures specified in this document are NOT RECOMMENDED to be used between autonomous systems in an AS Confederation [RFC5065]. If an OTC Attribute is added on egress from the AS Confederation, its value MUST equal the AS Confederation Identifier. Also, on egress from the AS Confederation, an UPDATE MUST NOT contain an OTC Attribute with a value corresponding to any Member-AS Number other than the AS Confederation Identifier.

The procedures specified in this document in scenarios that use private AS numbers behind an Internet-facing ASN (e.g., a data center network [RFC7938] or stub customer) may be used, but any details are outside the scope of this document. On egress from the Internet-facing AS, the OTC Attribute MUST NOT contain a value other than the Internet-facing ASN.

Once the OTC Attribute has been set, it MUST be preserved unchanged (this also applies to an AS Confederation).

The described ingress and egress procedures are applicable only for the address families AFI 1 (IPv4) and AFI 2 (IPv6) with SAFI 1 (unicast) in both cases and MUST NOT be applied to other address families by default. The operator MUST NOT have the ability to modify the procedures defined in this section.



## 5. Additional Considerations

Roles MUST NOT be configured on an eBGP session with a Complex peering relationship. If multiple eBGP sessions can segregate the Complex peering relationship into eBGP sessions with normal peering relationships, BGP Roles SHOULD be used on each of the resulting eBGP sessions.

An operator may want to achieve an equivalent outcome by configuring policies on a per-prefix basis to follow the definitions of peering relations as described in Section 2.1. However, in this case, there are no in-band measures to check the correctness of the per-prefix peering configuration.

The incorrect setting of BGP Roles and/or OTC Attributes may affect prefix propagation. Further, this document does not specify any special handling of an incorrect AS number in the OTC Attribute.

In AS migration scenarios [RFC7705], a given router may represent itself as any one of several different ASes. This should not be a problem since the egress procedures in Section 4 specify that the OTC Attribute is to be attached as part of route transmission. Therefore, a router is expected to set the OTC value equal to the ASN it is currently representing itself as.

Section 6 of [RFC7606] documents possible negative impacts of "treat-as-withdraw" behavior. Such negative impacts may include forwarding loops or blackholes. It also discusses debugging considerations related to this behavior.

## 6. IANA Considerations

IANA has registered a new BGP Capability (Section 3.1) in the "Capability Codes" registry's "IETF Review" range [RFC5492]. The description for the new capability is "BGP Role". IANA has assigned the value 9 [to be removed upon publication: <https://www.iana.org/assignments/capability-codes/capability-codes.xhtml>]. This document is the reference for the new capability.

The BGP Role capability includes a Value field, for which IANA is requested to create and maintain a new sub-registry called "BGP Role Value" in the Capability Codes registry. Assignments consist of a Value and a corresponding Role name. Initially, this registry is to be populated with the data contained in Table 1 found in Section 3.1. Future assignments may be made by the "IETF Review" policy as defined in [RFC8126]. The registry is as shown in Table 3.

| Value | Role name (for the local AS)  | Reference     |
|-------|-------------------------------|---------------|
| 0     | Provider                      | This document |
| 1     | RS                            | This document |
| 2     | RS-Client                     | This document |
| 3     | Customer                      | This document |
| 4     | Peer (i.e., Lateral Peer)     | This document |
| 5-255 | To be assigned by IETF Review |               |

Table 3: IANA Registry for BGP Role

IANA has registered a new OPEN Message Error subcode named the "Role Mismatch" (see Section 3.2) in the OPEN Message Error subcodes registry. IANA has assigned the value 11 [to be removed upon publication: <https://www.iana.org/assignments/bgp-parameters/bgp-parameters.xhtml#bgp-parameters-6>]. This document is the reference for the new subcode.

Due to improper use of the values 8, 9, and 10 in the OPEN Message Error subcodes registry, this document requested IANA to mark these values as "Deprecated". IANA has marked values 8-10 as "Deprecated" in the OPEN Message Error subcodes registry. This document is listed as the reference.

IANA has also registered a new path attribute named "Only to Customer (OTC)" (see Section 4) in the "BGP Path Attributes" registry. IANA has assigned code value 35 [To be removed upon publication: <http://www.iana.org/assignments/bgp-parameters/bgp-parameters.xhtml#bgp-parameters-2>]. This document is the reference for the new attribute.

## 7. Security Considerations

The security considerations of BGP (as specified in [RFC4271] and [RFC4272]) apply.

This document proposes a mechanism using BGP Role for the prevention and detection of route leaks that are the result of BGP policy misconfiguration. A misconfiguration of the BGP Role may affect prefix propagation. For example, if a downstream (i.e., towards a Customer) peering link were misconfigured with a Provider or Peer

Role, this will limit the number of prefixes that can be advertised in this direction. On the other hand, if an upstream provider were misconfigured (by a local AS) with the Customer Role, this may result in propagating routes that are received from other Providers or Peers. But the BGP Role negotiation and the resulting confirmation of Roles make such misconfigurations unlikely.

Setting the strict mode of operation for BGP Role negotiation as the default may result in a situation where the eBGP session will not come up after a software update. Implementations with such default behavior are strongly discouraged.

Removing the OTC Attribute or changing its value can limit the opportunity for route leak detection. Such activity can be done on purpose as part of an on-path attack. For example, an AS can remove the OTC Attribute on a received route and then leak the route to its transit provider. This kind of threat is not new in BGP and it may affect any Attribute (Note: BGPsec [RFC8205] offers protection only for the AS\_PATH Attribute).

Adding an OTC Attribute when the route is advertised from Customer to Provider will limit the propagation of the route. Such a route may be considered as ineligible by the immediate Provider or its Peers or upper layer Providers. This kind of OTC Attribute addition is unlikely to happen on the Provider side because it will limit the traffic volume towards its Customer. On the Customer side, adding an OTC Attribute for traffic engineering purposes is also discouraged because it will limit route propagation in an unpredictable way.

## 8. References

### 8.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [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>>.
- [RFC5065] Traina, P., McPherson, D., and J. Scudder, "Autonomous System Confederations for BGP", RFC 5065, DOI 10.17487/RFC5065, August 2007, <<https://www.rfc-editor.org/info/rfc5065>>.

- [RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, DOI 10.17487/RFC5492, February 2009, <<https://www.rfc-editor.org/info/rfc5492>>.
- [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>>.
- [RFC7908] Sriram, K., Montgomery, D., McPherson, D., Osterweil, E., and B. Dickson, "Problem Definition and Classification of BGP Route Leaks", RFC 7908, DOI 10.17487/RFC7908, June 2016, <<https://www.rfc-editor.org/info/rfc7908>>.
- [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>>.

## 8.2. Informative References

- [Gao] Gao, L. and J. Rexford, "Stable Internet routing without global coordination", IEEE/ACM Transactions on Networking, Volume 9, Issue 6, pp 689-692, DOI 10.1109/90.974523, December 2001, <<https://ieeexplore.ieee.org/document/974523>>.
- [RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis", RFC 4272, DOI 10.17487/RFC4272, January 2006, <<https://www.rfc-editor.org/info/rfc4272>>.
- [RFC7705] George, W. and S. Amante, "Autonomous System Migration Mechanisms and Their Effects on the BGP AS\_PATH Attribute", RFC 7705, DOI 10.17487/RFC7705, November 2015, <<https://www.rfc-editor.org/info/rfc7705>>.
- [RFC7938] Lapukhov, P., Premji, A., and J. Mitchell, Ed., "Use of BGP for Routing in Large-Scale Data Centers", RFC 7938, DOI 10.17487/RFC7938, August 2016, <<https://www.rfc-editor.org/info/rfc7938>>.
- [RFC8205] Lepinski, M., Ed. and K. Sriram, Ed., "BGPsec Protocol Specification", RFC 8205, DOI 10.17487/RFC8205, September 2017, <<https://www.rfc-editor.org/info/rfc8205>>.

## Acknowledgments

The authors wish to thank Alvaro Retana, Bruno Decraene, Jeff Haas, John Scudder, Sue Hares, Ben Maddison, Andrei Robachevsky, Daniel Ginsburg, Ruediger Volk, Pavel Lunin, Gyan Mishra, and Ignas Bagdonas for review, comments, and suggestions during the course of this work. Thanks are also due to many IESG reviewers whose comments greatly helped improve the clarity, accuracy, and presentation in the document.

## Contributors

Brian Dickson  
Independent  
Email: brian.peter.dickson@gmail.com

Doug Montgomery  
USA National Institute of Standards and Technology  
Email: dougm@nist.gov

## Authors' Addresses

Alexander Azimov  
Qrator Labs & Yandex  
Ulitsa Iva Tolstogo 16  
Moscow  
119021  
Russian Federation  
Email: a.e.azimov@gmail.com

Eugene Bogomazov  
Qrator Labs  
1-y Magistralnyy tupik 5A  
Moscow  
123290  
Russian Federation  
Email: eb@qrator.net

Randy Bush  
Internet Initiative Japan & Arrcus, Inc.  
5147 Crystal Springs  
Bainbridge Island, Washington 98110  
United States of America  
Email: randy@psg.com

Keyur Patel  
Arrcus  
2077 Gateway Place, Suite #400  
San Jose, CA 95119  
United States of America  
Email: keyur@arrcus.com

Kotikalapudi Sriram  
USA National Institute of Standards and Technology  
100 Bureau Drive  
Gaithersburg, MD 20899  
United States of America  
Email: ksriram@nist.gov

Inter-Domain Routing  
Internet-Draft  
Intended status: Standards Track  
Expires: January 9, 2020

K. Talaulikar, Ed.  
Cisco Systems  
H. Gredler  
Rtbrick  
J. Medved  
Cisco Systems, Inc.  
S. Previdi  
Individual Contributor  
A. Farrel  
Old Dog Consulting  
S. Ray  
Individual Contributor  
July 8, 2019

Distribution of Link-State and Traffic Engineering Information Using BGP  
draft-ketant-idr-rfc7752bis-01

## Abstract

In a number of environments, a component external to a network is called upon to perform computations based on the network topology and current state of the connections within the network, including Traffic Engineering (TE) information. This is information typically distributed by IGP routing protocols within the network.

This document describes a mechanism by which link-state and TE information can be collected from networks and shared with external components using the BGP routing protocol. This is achieved using a new BGP Network Layer Reachability Information (NLRI) encoding format. The mechanism is applicable to physical and virtual IGP links. The mechanism described is subject to policy control.

Applications of this technique include Application-Layer Traffic Optimization (ALTO) servers and Path Computation Elements (PCEs).

## Requirements Language

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

## Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

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

This Internet-Draft will expire on January 9, 2020.

## Copyright Notice

Copyright (c) 2019 IETF Trust and the persons identified as the document authors. All rights reserved.

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

## Table of Contents

|                                                     |    |
|-----------------------------------------------------|----|
| 1. Introduction . . . . .                           | 3  |
| 2. Motivation and Applicability . . . . .           | 5  |
| 2.1. MPLS-TE with PCE . . . . .                     | 5  |
| 2.2. ALTO Server Network API . . . . .              | 7  |
| 3. BGP Speaker Roles for BGP-LS . . . . .           | 8  |
| 4. Carrying Link-State Information in BGP . . . . . | 9  |
| 4.1. TLV Format . . . . .                           | 9  |
| 4.2. The Link-State NLRI . . . . .                  | 10 |
| 4.2.1. Node Descriptors . . . . .                   | 15 |
| 4.2.2. Link Descriptors . . . . .                   | 19 |
| 4.2.3. Prefix Descriptors . . . . .                 | 21 |
| 4.3. The BGP-LS Attribute . . . . .                 | 23 |
| 4.3.1. Node Attribute TLVs . . . . .                | 24 |
| 4.3.2. Link Attribute TLVs . . . . .                | 27 |



|                                                                               |    |
|-------------------------------------------------------------------------------|----|
| 4.3.3. Prefix Attribute TLVs . . . . .                                        | 32 |
| 4.4. Private Use . . . . .                                                    | 35 |
| 4.5. BGP Next-Hop Information . . . . .                                       | 36 |
| 4.6. Inter-AS Links . . . . .                                                 | 36 |
| 4.7. Handling of Unreachable IGP Nodes . . . . .                              | 36 |
| 4.8. Router-ID Anchoring Example: ISO Pseudonode . . . . .                    | 38 |
| 4.9. Router-ID Anchoring Example: OSPF Pseudonode . . . . .                   | 39 |
| 4.10. Router-ID Anchoring Example: OSPFv2 to IS-IS Migration . . . . .        | 40 |
| 5. Link to Path Aggregation . . . . .                                         | 40 |
| 5.1. Example: No Link Aggregation . . . . .                                   | 41 |
| 5.2. Example: ASBR to ASBR Path Aggregation . . . . .                         | 41 |
| 5.3. Example: Multi-AS Path Aggregation . . . . .                             | 42 |
| 6. IANA Considerations . . . . .                                              | 42 |
| 6.1. Guidance for Designated Experts . . . . .                                | 43 |
| 7. Manageability Considerations . . . . .                                     | 44 |
| 7.1. Operational Considerations . . . . .                                     | 44 |
| 7.1.1. Operations . . . . .                                                   | 44 |
| 7.1.2. Installation and Initial Setup . . . . .                               | 44 |
| 7.1.3. Migration Path . . . . .                                               | 44 |
| 7.1.4. Requirements on Other Protocols and Functional<br>Components . . . . . | 44 |
| 7.1.5. Impact on Network Operation . . . . .                                  | 45 |
| 7.1.6. Verifying Correct Operation . . . . .                                  | 45 |
| 7.2. Management Considerations . . . . .                                      | 45 |
| 7.2.1. Management Information . . . . .                                       | 45 |
| 7.2.2. Fault Management . . . . .                                             | 45 |
| 7.2.3. Configuration Management . . . . .                                     | 48 |
| 7.2.4. Accounting Management . . . . .                                        | 48 |
| 7.2.5. Performance Management . . . . .                                       | 48 |
| 7.2.6. Security Management . . . . .                                          | 49 |
| 8. TLV/Sub-TLV Code Points Summary . . . . .                                  | 49 |
| 9. Security Considerations . . . . .                                          | 50 |
| 10. Contributors . . . . .                                                    | 51 |
| 11. Acknowledgements . . . . .                                                | 51 |
| 12. References . . . . .                                                      | 52 |
| 12.1. Normative References . . . . .                                          | 52 |
| 12.2. Informative References . . . . .                                        | 54 |
| Appendix A. Changes from RFC 7752 . . . . .                                   | 56 |
| Authors' Addresses . . . . .                                                  | 57 |

## 1. Introduction

The contents of a Link-State Database (LSDB) or of an IGP's Traffic Engineering Database (TED) describe only the links and nodes within an IGP area. Some applications, such as end-to-end Traffic Engineering (TE), would benefit from visibility outside one area or Autonomous System (AS) in order to make better decisions.

The IETF has defined the Path Computation Element (PCE) [RFC4655] as a mechanism for achieving the computation of end-to-end TE paths that cross the visibility of more than one TED or that require CPU-intensive or coordinated computations. The IETF has also defined the ALTO server [RFC5693] as an entity that generates an abstracted network topology and provides it to network-aware applications.

Both a PCE and an ALTO server need to gather information about the topologies and capabilities of the network in order to be able to fulfill their function.

This document describes a mechanism by which link-state and TE information can be collected from networks and shared with external components using the BGP routing protocol [RFC4271]. This is achieved using a new BGP Network Layer Reachability Information (NLRI) encoding format. The mechanism is applicable to physical and virtual links. The mechanism described is subject to policy control.

A router maintains one or more databases for storing link-state information about nodes and links in any given area. Link attributes stored in these databases include: local/remote IP addresses, local/remote interface identifiers, link metric and TE metric, link bandwidth, reservable bandwidth, per Class-of-Service (CoS) class reservation state, preemption, and Shared Risk Link Groups (SRLGs). The router's BGP process can retrieve topology from these LSDBs and distribute it to a consumer, either directly or via a peer BGP speaker (typically a dedicated Route Reflector), using the encoding specified in this document.

An illustration of the collection of link-state and TE information and its distribution to consumers is shown in the Figure 1 below.

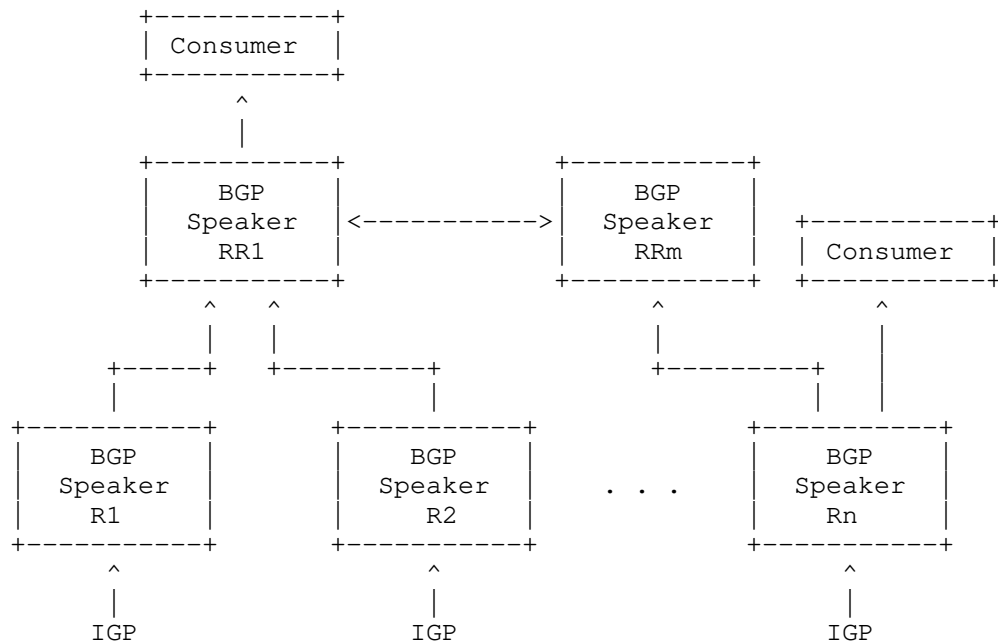


Figure 1: Collection of Link-State and TE Information

A BGP speaker may apply configurable policy to the information that it distributes. Thus, it may distribute the real physical topology from the LSDB or the TED. Alternatively, it may create an abstracted topology, where virtual, aggregated nodes are connected by virtual paths. Aggregated nodes can be created, for example, out of multiple routers in a Point of Presence (POP). Abstracted topology can also be a mix of physical and virtual nodes and physical and virtual links. Furthermore, the BGP speaker can apply policy to determine when information is updated to the consumer so that there is a reduction of information flow from the network to the consumers. Mechanisms through which topologies can be aggregated or virtualized are outside the scope of this document

## 2. Motivation and Applicability

This section describes use cases from which the requirements can be derived.

### 2.1. MPLS-TE with PCE

As described in [RFC4655], a PCE can be used to compute MPLS-TE paths within a "domain" (such as an IGP area) or across multiple domains (such as a multi-area AS or multiple ASes).

- o Within a single area, the PCE offers enhanced computational power that may not be available on individual routers, sophisticated policy control and algorithms, and coordination of computation across the whole area.
- o If a router wants to compute a MPLS-TE path across IGP areas, then its own TED lacks visibility of the complete topology. That means that the router cannot determine the end-to-end path and cannot even select the right exit router (Area Border Router (ABR)) for an optimal path. This is an issue for large-scale networks that need to segment their core networks into distinct areas but still want to take advantage of MPLS-TE.

Previous solutions used per-domain path computation [RFC5152]. The source router could only compute the path for the first area because the router only has full topological visibility for the first area along the path, but not for subsequent areas. Per-domain path computation uses a technique called "loose-hop-expansion" [RFC3209] and selects the exit ABR and other ABRs or AS Border Routers (ASBRs) using the IGP-computed shortest path topology for the remainder of the path. This may lead to sub-optimal paths, makes alternate/back-up path computation hard, and might result in no TE path being found when one really does exist.

The PCE presents a computation server that may have visibility into more than one IGP area or AS, or may cooperate with other PCEs to perform distributed path computation. The PCE obviously needs access to the TED for the area(s) it serves, but [RFC4655] does not describe how this is achieved. Many implementations make the PCE a passive participant in the IGP so that it can learn the latest state of the network, but this may be sub-optimal when the network is subject to a high degree of churn or when the PCE is responsible for multiple areas.

The following figure shows how a PCE can get its TED information using the mechanism described in this document.

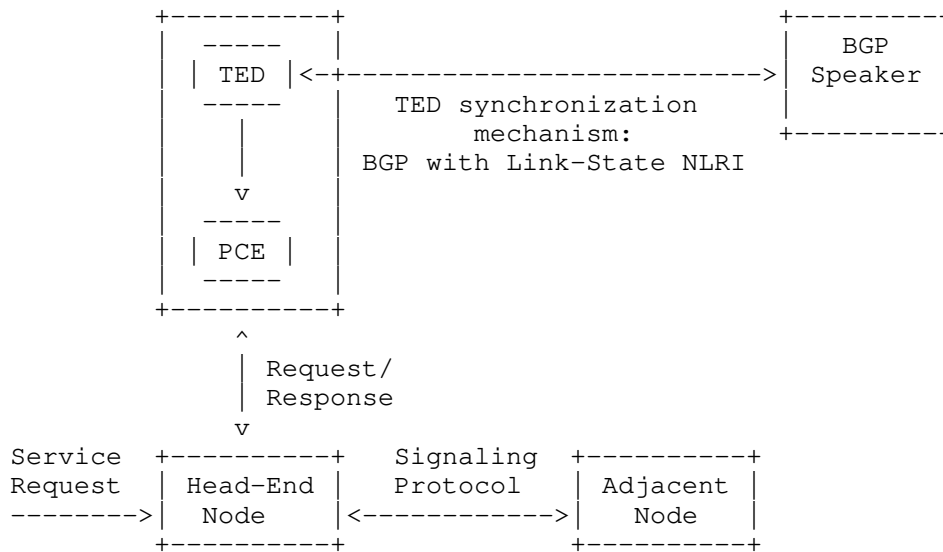


Figure 2: External PCE Node Using a TED Synchronization Mechanism

The mechanism in this document allows the necessary TED information to be collected from the IGP within the network, filtered according to configurable policy, and distributed to the PCE as necessary.

## 2.2. ALTO Server Network API

An ALTO server [RFC5693] is an entity that generates an abstracted network topology and provides it to network-aware applications over a web-service-based API. Example applications are peer-to-peer (P2P) clients or trackers, or Content Distribution Networks (CDNs). The abstracted network topology comes in the form of two maps: a Network Map that specifies allocation of prefixes to Partition Identifiers (PIDs), and a Cost Map that specifies the cost between PIDs listed in the Network Map. For more details, see [RFC7285].

ALTO abstract network topologies can be auto-generated from the physical topology of the underlying network. The generation would typically be based on policies and rules set by the operator. Both prefix and TE data are required: prefix data is required to generate ALTO Network Maps, and TE (topology) data is required to generate ALTO Cost Maps. Prefix data is carried and originated in BGP, and TE data is originated and carried in an IGP. The mechanism defined in this document provides a single interface through which an ALTO server can retrieve all the necessary prefix and network topology data from the underlying network. Note that an ALTO server can use

other mechanisms to get network data, for example, peering with multiple IGP and BGP speakers.

The following figure shows how an ALTO server can get network topology information from the underlying network using the mechanism described in this document.

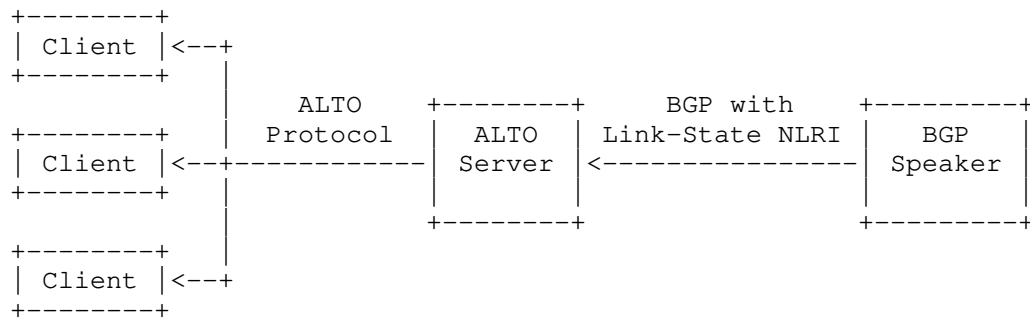


Figure 3: ALTO Server Using Network Topology Information

### 3. BGP Speaker Roles for BGP-LS

In the illustration shown in Figure 1, the BGP Speakers can be seen playing different roles in the distribution of information using BGP-LS. This section introduces terms that explain the different roles of the BGP Speakers which are then used through the rest of this document.

- o **BGP-LS Producer:** The BGP Speakers R1, R2, ... Rn, originate link-state information from their underlying link-state IGP protocols into BGP-LS. If R1 and R2 are in the same IGP area, then likely they are originating the same link-state information into BGP-LS. R1 may also source information from sources other than IGP, e.g. its local node information. The term BGP-LS Producer refers to the BGP Speaker that is originating link-state information into BGP.
- o **BGP-LS Consumer:** The BGP Speakers RR1 and Rn are handing off the BGP-LS information that they have collected to a consumer application. The BGP protocol implementation and the consumer application may be on the same or different nodes. The term BGP-LS Consumer refers to the consumer application/process and not the BGP Speaker. This document only covers the BGP implementation. The consumer application and the design of interface between BGP and consumer application may be implementation specific and outside the scope of this document.

- o BGP-LS Propagator: The BGP Speaker RRm propagates the BGP-LS information between the BGP Speaker Rn and the BGP Speaker RR1. The BGP implementation on RRm is doing the propagation of BGP-LS updates and performing BGP best path calculations. Similarly, the BGP Speaker RR1 is receiving BGP-LS information from R1, R2 and RRm and propagating the information to the BGP-LS Consumer after performing BGP best path calculations. The term BGP-LS Propagator refers to the BGP Speaker that is performing BGP protocol processing on the link-state information.

The above roles are not mutually exclusive. The same BGP Speaker may be the producer for some link-state information and propagator for some other link-state information while also providing this information to a consumer application. Nothing precludes a BGP implementation performing some of the validation and processing on behalf of the BGP-LS Consumer as long as it does not impact the semantics of its role as BGP-LS Propagator as described in this document.

The rest of this document refers to the role when describing procedures that are specific to that role. When the role is not specified, then the said procedure applies to all BGP Speakers.

#### 4. Carrying Link-State Information in BGP

This specification contains two parts: definition of a new BGP NLRI that describes links, nodes, and prefixes comprising IGP link-state information and definition of a new BGP path attribute (BGP-LS Attribute) that carries link, node, and prefix properties and attributes, such as the link and prefix metric or auxiliary Router-IDs of nodes, etc.

It is desirable to keep the dependencies on the protocol source of this attribute to a minimum and represent any content in an IGP-neutral way, such that applications that want to learn about a link-state topology do not need to know about any OSPF or IS-IS protocol specifics.

This section mainly describes the procedures at a BGP-LS Producer that originate link-state information into BGP-LS.

##### 4.1. TLV Format

Information in the new Link-State NLRIs and the BGP-LS Attribute is encoded in Type/Length/Value triplets. The TLV format is shown in Figure 4 and applies to both the NLRI and the BGP-LS Attribute encodings.

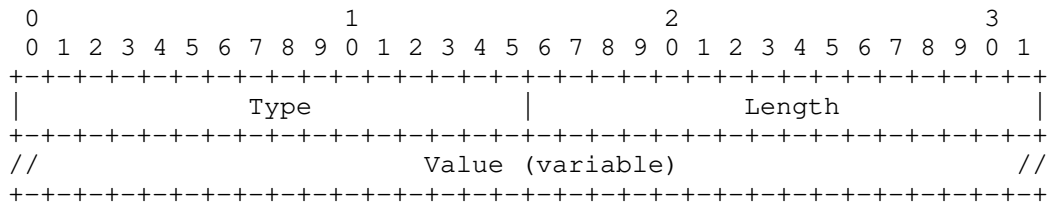


Figure 4: TLV Format

The Length field defines the length of the value portion in octets (thus, a TLV with no value portion would have a length of zero). The TLV is not padded to 4-octet alignment. Unknown and unsupported types MUST be preserved and propagated within both the NLRI and the BGP-LS Attribute. The presence of unrecognized or unexpected TLVs MUST NOT result in the NLRI or the BGP-LS Attribute being considered as malformed.

In order to compare NLRIs with unknown TLVs, all TLVs within the NLRI MUST be ordered in ascending order by TLV Type. If there are multiple TLVs of the same type within a single NLRI, then the TLVs sharing the same type MUST be in ascending order based on the value field. Comparison of the value fields is performed by treating the entire field as an opaque hexadecimal string. Standard string comparison rules apply. NLRIs having TLVs which do not follow the above ordering rules MUST be considered as malformed by a BGP-LS Propagator. This ensures that multiple copies of the same NLRI from multiple BGP-LS Producers and the ambiguity arising there from is prevented.

All TLVs within the NLRI that are not specified as mandatory are considered optional. All TLVs within the BGP-LS Attribute are considered optional unless specified otherwise.

The TLVs within the BGP-LS Attribute need not be ordered in any specific order.

#### 4.2. The Link-State NLRI

The MP\_REACH\_NLRI and MP\_UNREACH\_NLRI attributes are BGP's containers for carrying opaque information. This specification defines three Link-State NLRI types that describes either a node, a link, and a prefix.

All non-VPN link, node, and prefix information SHALL be encoded using AFI 16388 / SAFI 71. VPN link, node, and prefix information SHALL be encoded using AFI 16388 / SAFI 72.



In order for two BGP speakers to exchange Link-State NLRI, they MUST use BGP Capabilities Advertisement to ensure that they are both capable of properly processing such NLRI. This is done as specified in [RFC4760], by using capability code 1 (multi-protocol BGP), with AFI 16388 / SAFI 71 for BGP-LS, and AFI 16388 / SAFI 72 for BGP-LS-VPN.

New Link-State NLRI Types may be introduced in the future. Since supported NLRI type values within the address family are not expressed in the Multiprotocol BGP (MP-BGP) capability [RFC4760], it is possible that a BGP speaker has advertised support for Link-State but does not support a particular Link-State NLRI type. In order to allow introduction of new Link-State NLRI types seamlessly in the future, without the need for upgrading all BGP speakers in the propagation path (e.g. a route reflector), this document deviates from the default handling behavior specified by [RFC7606] for Link-State address-family. An implementation MUST handle unrecognized Link-State NLRI types as opaque objects and MUST preserve and propagate them.

The format of the Link-State NLRI is shown in the following figures.

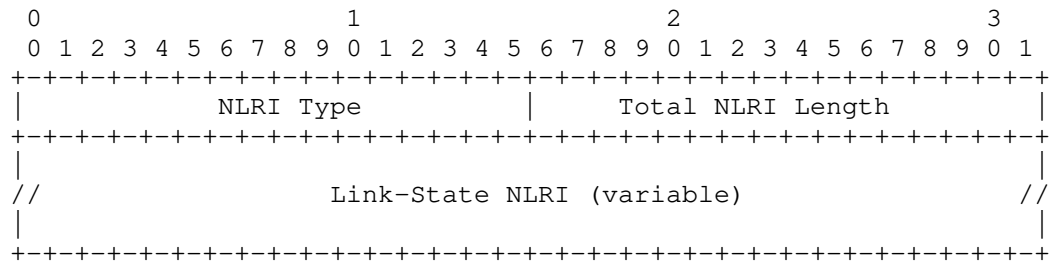


Figure 5: Link-State AFI 16388 / SAFI 71 NLRI Format

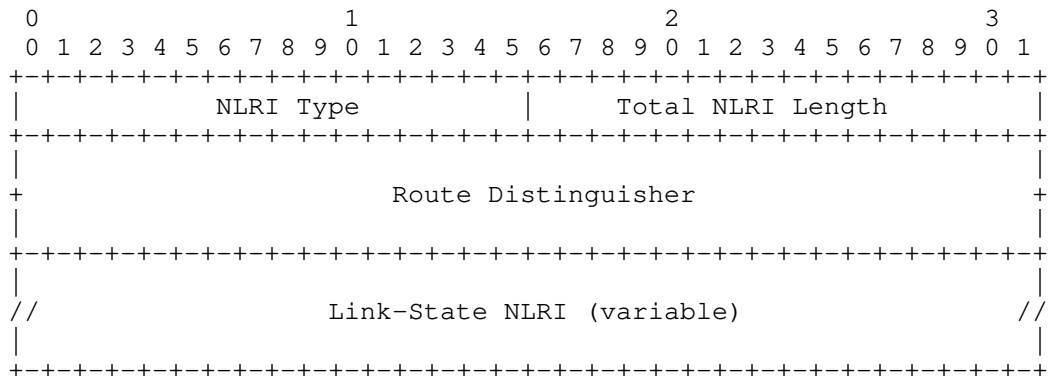


Figure 6: Link-State VPN AFI 16388 / SAFI 72 NLRI Format

The Total NLRI Length field contains the cumulative length, in octets, of the rest of the NLRI, not including the NLRI Type field or itself. For VPN applications, it also includes the length of the Route Distinguisher.

| Type        | NLRI Type                 |
|-------------|---------------------------|
| 1           | Node NLRI                 |
| 2           | Link NLRI                 |
| 3           | IPv4 Topology Prefix NLRI |
| 4           | IPv6 Topology Prefix NLRI |
| 65000-65535 | Private Use               |

Table 1: NLRI Types

Route Distinguishers are defined and discussed in [RFC4364].

The Node NLRI (NLRI Type = 1) is shown in the following figure.

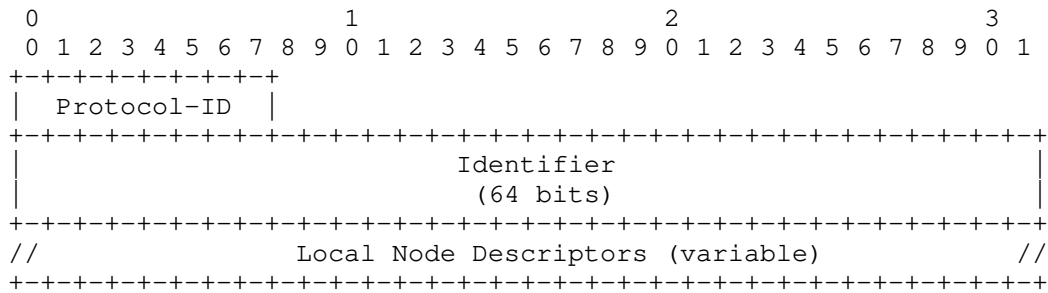


Figure 7: The Node NLRI Format

The Link NLRI (NLRI Type = 2) is shown in the following figure.

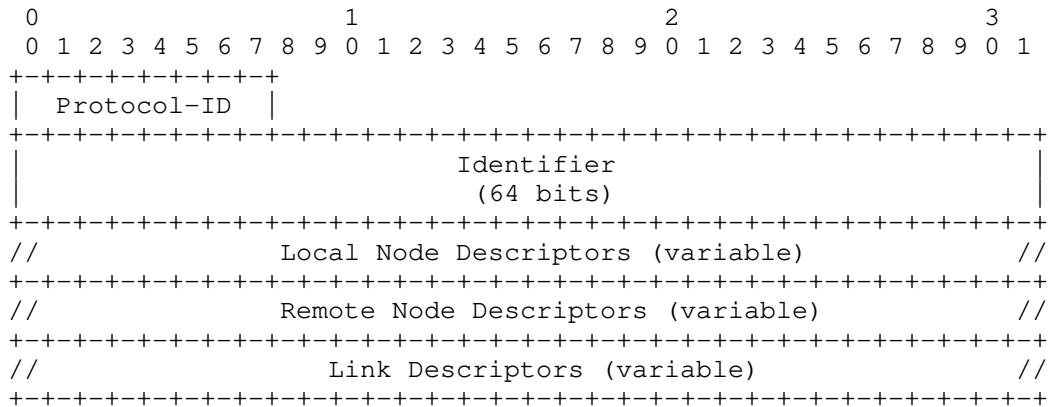


Figure 8: The Link NLRI Format

The IPv4 and IPv6 Prefix NLRIs (NLRI Type = 3 and Type = 4) use the same format, as shown in the following figure.

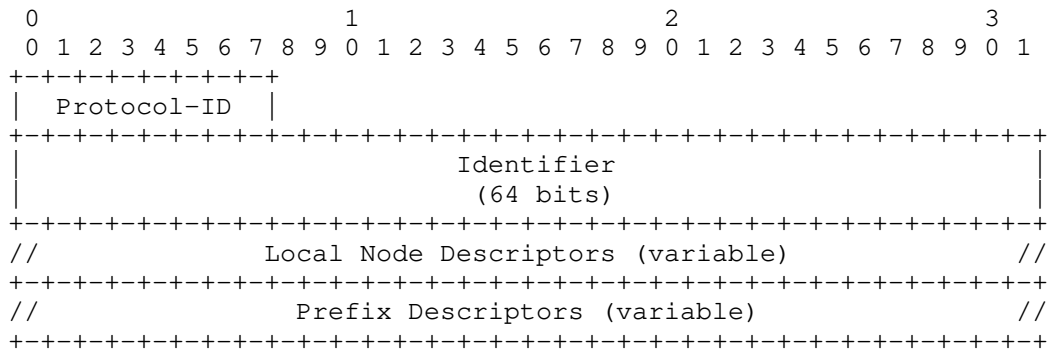


Figure 9: The IPv4/IPv6 Topology Prefix NLRI Format

The Protocol-ID field can contain one of the following values:

| Protocol-ID | NLRI information source protocol |
|-------------|----------------------------------|
| 1           | IS-IS Level 1                    |
| 2           | IS-IS Level 2                    |
| 3           | OSPFv2                           |
| 4           | Direct                           |
| 5           | Static configuration             |
| 6           | OSPFv3                           |
| 200-255     | Private Use                      |

Table 2: Protocol Identifiers

The 'Direct' and 'Static configuration' protocol types SHOULD be used when BGP-LS is sourcing local information. For all information derived from other protocols, the corresponding Protocol-ID MUST be used. If BGP-LS has direct access to interface information and wants to advertise a local link, then the Protocol-ID 'Direct' SHOULD be used. For modeling virtual links, such as described in Section 5, the Protocol-ID 'Static configuration' SHOULD be used.

A router MAY run multiple protocol instances of OSPF or ISIS where by it becomes a border router between multiple IGP domains. Both OSPF and IS-IS MAY also run multiple routing protocol instances over the same link. See [RFC8202] and [RFC6549]. These instances define independent IGP routing domains. The 64-bit Identifier field carries a BGP-LS Instance Identifier (Instance-ID) that is used to identify the IGP routing domain where the NLRI belongs. The NLRIs representing link-state objects (nodes, links, or prefixes) from the same IGP routing instance MUST have the same Identifier field value.

NLRIs with different Identifier field values MUST be considered to be from different IGP routing instances. The Identifier field value 0 is RECOMMENDED to be used when there is only a single protocol instance in the network where BGP-LS is operational.

An implementation which supports multiple IGP instances MUST support the configuration of unique BGP-LS Instance-IDs at the routing protocol instance level. The network operator MUST assign consistent BGP-LS Instance-ID values on all BGP-LS Producers within a given IGP domain. Unique BGP-LS Instance-ID values MUST be assigned to routing protocol instances operating in different IGP domains. This allows the BGP-LS Consumer to build an accurate segregated multi-domain topology based on the Identifier field even when the topology is advertised via BGP-LS by multiple BGP-LS Producers in the network.

When the above described semantics and recommendations are not followed, a BGP-LS Consumer may see duplicate link-state objects for the same node, link or prefix when there are multiple BGP-LS Producers deployed. This may also result in the BGP-LS Consumers getting an inaccurate network-wide topology.

When adding, removing or modifying a TLV/sub-TLV from a Link-State NLRI, the BGP-LS Producer MUST withdraw the old NLRI by including it in the MP\_UNREACH\_NLRI. Not doing so can result in duplicate and inconsistent link-state objects hanging around in the BGP-LS table.

Each Node Descriptor and Link Descriptor consists of one or more TLVs, as described in the following sections.

#### 4.2.1. Node Descriptors

Each link is anchored by a pair of Router-IDs that are used by the underlying IGP, namely, a 48-bit ISO System-ID for IS-IS and a 32-bit Router-ID for OSPFv2 and OSPFv3. An IGP may use one or more additional auxiliary Router-IDs, mainly for Traffic Engineering purposes. For example, IS-IS may have one or more IPv4 and IPv6 TE Router-IDs [RFC5305] [RFC6119]. These auxiliary Router-IDs MUST be included in the node attribute described in Section 4.3.1 and MAY be included in link attribute described in Section 4.3.2. The advertisement of the TE Router-IDs help a BGP-LS Consumer to correlate multiple link-state objects (e.g. in different IGP instances or areas/levels) to the same node in the network.

It is desirable that the Router-ID assignments inside the Node Descriptor are globally unique. However, there may be Router-ID spaces (e.g., ISO) where no global registry exists, or worse, Router-IDs have been allocated following the private-IP allocation described

in RFC 1918 [RFC1918]. BGP-LS uses the Autonomous System (AS) Number to disambiguate the Router-IDs, as described in Section 4.2.1.1.

#### 4.2.1.1. Globally Unique Node/Link/Prefix Identifiers

One problem that needs to be addressed is the ability to identify an IGP node globally (by "globally", we mean within the BGP-LS database collected by all BGP-LS speakers that talk to each other). This can be expressed through the following two requirements:

- (A)    The same node MUST NOT be represented by two keys (otherwise, one node will look like two nodes).
- (B)    Two different nodes MUST NOT be represented by the same key (otherwise, two nodes will look like one node).

We define an "IGP domain" to be the set of nodes (hence, by extension links and prefixes) within which each node has a unique IGP representation by using the combination of Area-ID, Router-ID, Protocol-ID, Multi-Topology ID, and Instance-ID. The problem is that BGP may receive node/link/prefix information from multiple independent "IGP domains", and we need to distinguish between them. Moreover, we can't assume there is always one and only one IGP domain per AS. During IGP transitions, it may happen that two redundant IGPs are in place.

The mapping of the Instance-ID to the Identifier field as described earlier along with a set of sub-TLVs described in Section 4.2.1.4, allows specification of a flexible key for any given node/link information such that global uniqueness of the NLRI is ensured.

#### 4.2.1.2. Local Node Descriptors

The Local Node Descriptors TLV contains Node Descriptors for the node anchoring the local end of the link. This is a mandatory TLV in all three types of NLRIs (node, link, and prefix). The Type is 256. The length of this TLV is variable. The value contains one or more Node Descriptor Sub-TLVs defined in Section 4.2.1.4.

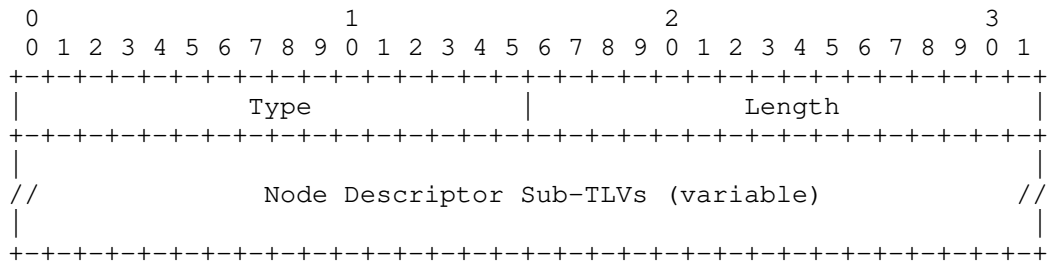


Figure 10: Local Node Descriptors TLV Format

## 4.2.1.3. Remote Node Descriptors

The Remote Node Descriptors TLV contains Node Descriptors for the node anchoring the remote end of the link. This is a mandatory TLV for Link NLRIs. The type is 257. The length of this TLV is variable. The value contains one or more Node Descriptor Sub-TLVs defined in Section 4.2.1.4.

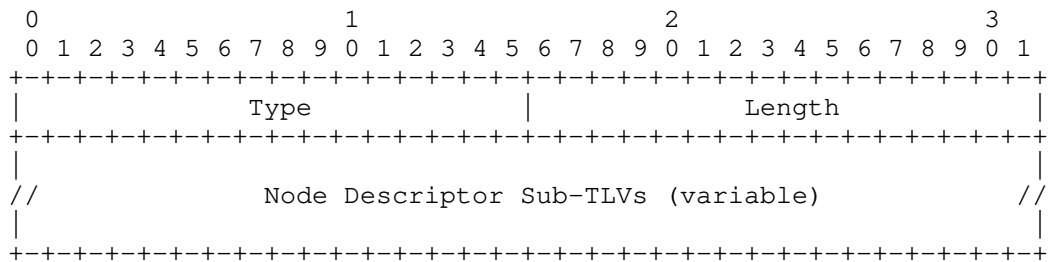


Figure 11: Remote Node Descriptors TLV Format

## 4.2.1.4. Node Descriptor Sub-TLVs

The Node Descriptor Sub-TLV type code points and lengths are listed in the following table:

| Sub-TLV Code Point | Description                    | Length   |
|--------------------|--------------------------------|----------|
| 512                | Autonomous System              | 4        |
| 513                | BGP-LS Identifier (deprecated) | 4        |
| 514                | OSPF Area-ID                   | 4        |
| 515                | IGP Router-ID                  | Variable |

Table 3: Node Descriptor Sub-TLVs

The sub-TLV values in Node Descriptor TLVs are defined as follows:

**Autonomous System:** Opaque value (32-bit AS Number). This is an optional TLV. The value SHOULD be set to the AS Number associated with the BGP process originating the link-state information. An implementation MAY provide a configuration option on the BGP-LS Producer to use a value different.

**BGP-LS Identifier:** Opaque value (32-bit ID). This is an optional TLV. In conjunction with Autonomous System Number (ASN), uniquely identifies the BGP-LS domain. The combination of ASN and BGP-LS ID MUST be globally unique. All BGP-LS speakers within an IGP flooding-set (set of IGP nodes within which an LSP/LSA is flooded) MUST use the same ASN, BGP-LS ID tuple. If an IGP domain consists of multiple flooding-sets, then all BGP-LS speakers within the IGP domain SHOULD use the same ASN, BGP-LS ID tuple.

**Area-ID:** Used to identify the 32-bit area to which the NLRI belongs. This is a mandatory TLV when originating information from OSPF. The Area Identifier allows different NLRIs of the same router to be discriminated.

**IGP Router-ID:** Opaque value. This is a mandatory TLV when originating information from IS-IS, OSPF, direct or static. For an IS-IS non-pseudonode, this contains a 6-octet ISO Node-ID (ISO system-ID). For an IS-IS pseudonode corresponding to a LAN, this contains the 6-octet ISO Node-ID of the Designated Intermediate System (DIS) followed by a 1-octet, nonzero PSN identifier (7 octets in total). For an OSPFv2 or OSPFv3 non-pseudonode, this contains the 4-octet Router-ID. For an OSPFv2 pseudonode representing a LAN, this contains the 4-octet Router-ID of the Designated Router (DR) followed by the 4-octet IPv4 address of the DR's interface to the LAN (8 octets in total). Similarly, for an OSPFv3 pseudonode, this contains the 4-octet Router-ID of the DR followed by the 4-octet interface identifier of the DR's interface to the LAN (8 octets in total). The TLV size in combination with the protocol identifier enables the decoder to determine the type of the node. For Direct or Static configuration, the value SHOULD be taken from an IPv4 or IPv6 address (e.g. loopback interface) configured on the node.

There can be at most one instance of each sub-TLV type present in any Node Descriptor. The sub-TLVs within a Node Descriptor MUST be arranged in ascending order by sub-TLV type. This needs to be done in order to compare NLRIs, even when an implementation encounters an unknown sub-TLV. Using stable sorting, an implementation can do binary comparison of NLRIs and hence allow incremental deployment of new key sub-TLVs.



The BGP-LS Identifier was introduced by [RFC7752] and it's use is being deprecated by this document. Implementations MUST continue to support this sub-TLV for backward compatibility. The default value of 0 is RECOMMENDED to be use when a BGP-LS Producer includes this sub-TLV when originating information into BGP-LS. Implementations MAY provide an option to configure this value for backward compatibility reasons. The use of the Instance-ID in the Identifier field is the RECOMMENDED way of segregation of different IGP domains in BGP-LS.

#### 4.2.2. Link Descriptors

The Link Descriptor field is a set of Type/Length/Value (TLV) triplets. The format of each TLV is shown in Section 4.1. The Link Descriptor TLVs uniquely identify a link among multiple parallel links between a pair of anchor routers. A link described by the Link Descriptor TLVs actually is a "half-link", a unidirectional representation of a logical link. In order to fully describe a single logical link, two originating routers advertise a half-link each, i.e., two Link NLRIs are advertised for a given point-to-point link.

A BGP-LS Consumer should not consider a link between two nodes as being available unless it has received the two Link NLRIs corresponding to the half-link representation of that link from both the nodes. This check is similar to the 'two way connectivity check' that is performed by link-state IGP and is also required to be done by BGP-LS Consumers of link-state topology.

A BGP-LS Producer MAY suppress the advertisement of a Link NLRI, corresponding to a half link, from a link-state IGP unless it has verified that the link is being reported in the IS-IS LSP or OSPF Router LSA by both the nodes connected by that link. This 'two way connectivity check' is performed by link-state IGP during their computation and may be leveraged before passing information for any half-link that is reported from these IGP to BGP-LS. This ensures that only those Link State IGP adjacencies which are established get reported via Link NLRIs. Such a 'two way connectivity check' may be also required in certain cases (e.g. with OSPF) to obtain the proper link identifiers of the remote node.

The format and semantics of the Value fields in most Link Descriptor TLVs correspond to the format and semantics of Value fields in IS-IS Extended IS Reachability sub-TLVs, defined in [RFC5305], [RFC5307], and [RFC6119]. Although the encodings for Link Descriptor TLVs were originally defined for IS-IS, the TLVs can carry data sourced by either IS-IS or OSPF.

The following TLVs are defined as Link Descriptors in the Link NLRI:

| TLV Code Point | Description                   | IS-IS TLV /Sub-TLV | Reference (RFC/Section) |
|----------------|-------------------------------|--------------------|-------------------------|
| 258            | Link Local/Remote Identifiers | 22/4               | [RFC5307]/1.1           |
| 259            | IPv4 interface address        | 22/6               | [RFC5305]/3.2           |
| 260            | IPv4 neighbor address         | 22/8               | [RFC5305]/3.3           |
| 261            | IPv6 interface address        | 22/12              | [RFC6119]/4.2           |
| 262            | IPv6 neighbor address         | 22/13              | [RFC6119]/4.3           |
| 263            | Multi-Topology Identifier     | ---                | Section 4.2.2.1         |

Table 4: Link Descriptor TLVs

The information about a link present in the LSA/LSP originated by the local node of the link determines the set of TLVs in the Link Descriptor of the link.

If interface and neighbor addresses, either IPv4 or IPv6, are present, then the IP address TLVs MUST be included and the Link Local/Remote Identifiers TLV MUST NOT be included in the Link Descriptor. The Link Local/Remote Identifiers TLV MAY be included in the link attribute when available.

If interface and neighbor addresses are not present and the link local/remote identifiers are present, then the Link Local/Remote Identifiers TLV MUST be included in the Link Descriptor.

The Multi-Topology Identifier TLV MUST be included in Link Descriptor if the underlying IGP link object is associated with a non-default topology.

#### 4.2.2.1. Multi-Topology ID

The Multi-Topology ID (MT-ID) TLV carries one or more IS-IS or OSPF Multi-Topology IDs for a link, node, or prefix.

Semantics of the IS-IS MT-ID are defined in Section 7.2 of RFC 5120 [RFC5120]. Semantics of the OSPF MT-ID are defined in Section 3.7 of RFC 4915 [RFC4915]. Bits R are reserved and SHOULD be set to 0 when

originated and ignored on receipt. If the value in the MT-ID TLV is derived from OSPF, then the upper 5 bits of the MT-ID field MUST be set to 0.

The format of the MT-ID TLV is shown in the following figure.

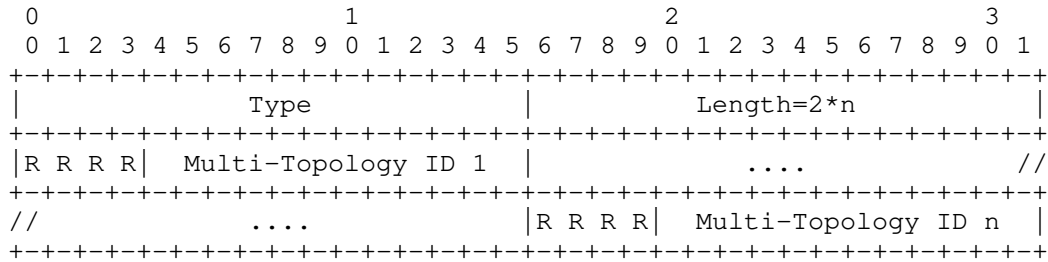


Figure 12: Multi-Topology ID TLV Format

where Type is 263, Length is  $2*n$ , and  $n$  is the number of MT-IDs carried in the TLV.

The MT-ID TLV MAY be present in a Link Descriptor, a Prefix Descriptor, or the BGP-LS attribute of a Node NLRI. In a Link or Prefix Descriptor, only a single MT-ID TLV containing the MT-ID of the topology where the link or the prefix is reachable is allowed. In case one wants to advertise multiple topologies for a given Link Descriptor or Prefix Descriptor, multiple NLRIs MUST be generated where each NLRI contains a single unique MT-ID. In the BGP-LS attribute of a Node NLRI, one MT-ID TLV containing the array of MT-IDs of all topologies where the node is reachable is allowed.

#### 4.2.3. Prefix Descriptors

The Prefix Descriptor field is a set of Type/Length/Value (TLV) triplets. Prefix Descriptor TLVs uniquely identify an IPv4 or IPv6 prefix originated by a node. The following TLVs are defined as Prefix Descriptors in the IPv4/IPv6 Prefix NLRI:

| TLV Code Point | Description                 | Length   | Reference (RFC/Section) |
|----------------|-----------------------------|----------|-------------------------|
| 263            | Multi-Topology Identifier   | variable | Section 4.2.2.1         |
| 264            | OSPF Route Type             | 1        | Section 4.2.3.1         |
| 265            | IP Reachability Information | variable | Section 4.2.3.2         |

Table 5: Prefix Descriptor TLVs

The Multi-Topology Identifier TLV MUST be included in Prefix Descriptor if the underlying IGP prefix object is associated with a non-default topology.

#### 4.2.3.1. OSPF Route Type

The OSPF Route Type TLV is a mandatory TLV corresponding to Prefix NLRIs originated from OSPF. It is used to identify the OSPF route type of the prefix. An OSPF prefix MAY be advertised in the OSPF domain with multiple route types. The Route Type TLV allows the discrimination of these advertisements. The format of the OSPF Route Type TLV is shown in the following figure.

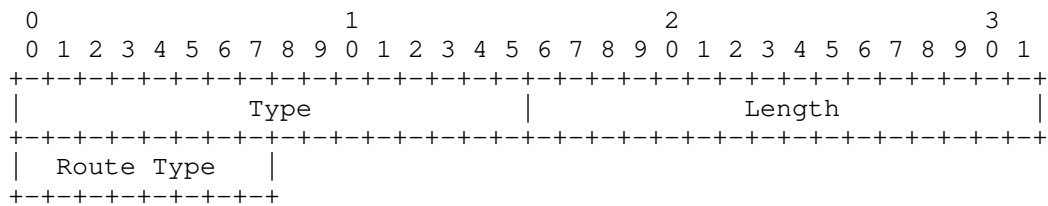


Figure 13: OSPF Route Type TLV Format

where the Type and Length fields of the TLV are defined in Table 5. The OSPF Route Type field values are defined in the OSPF protocol and can be one of the following:

- o Intra-Area (0x1)
- o Inter-Area (0x2)
- o External 1 (0x3)
- o External 2 (0x4)

- o NSSA 1 (0x5)
- o NSSA 2 (0x6)

#### 4.2.3.2. IP Reachability Information

The IP Reachability Information TLV is a mandatory TLV for IPv4 & IPv6 Prefix NLRI types. The TLV contains one IP address prefix (IPv4 or IPv6) originally advertised in the IGP topology. Its purpose is to glue a particular BGP service NLRI by virtue of its BGP next hop to a given node in the LSDB. A router SHOULD advertise an IP Prefix NLRI for each of its BGP next hops. The format of the IP Reachability Information TLV is shown in the following figure:

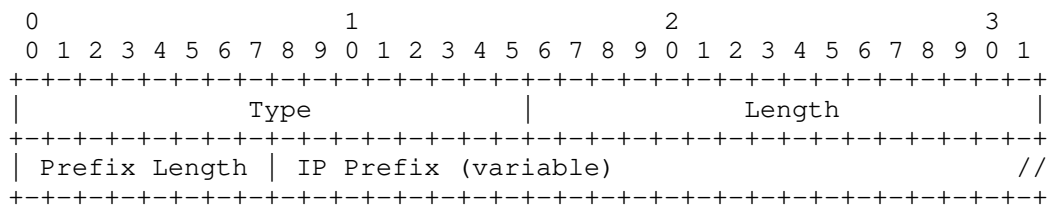


Figure 14: IP Reachability Information TLV Format

The Type and Length fields of the TLV are defined in Table 5. The following two fields determine the reachability information of the address family. The Prefix Length field contains the length of the prefix in bits. The IP Prefix field contains the most significant octets of the prefix, i.e., 1 octet for prefix length 1 up to 8, 2 octets for prefix length 9 to 16, 3 octets for prefix length 17 up to 24, 4 octets for prefix length 25 up to 32, etc.

#### 4.3. The BGP-LS Attribute

The BGP-LS Attribute is an optional, non-transitive BGP attribute that is used to carry link, node, and prefix parameters and attributes. It is defined as a set of Type/Length/Value (TLV) triplets, described in the following section. This attribute SHOULD only be included with Link-State NLRIs. This attribute MUST be ignored for all other address families.

The Node Attribute TLVs, Link Attribute TLVs and Prefix Attribute TLVs are sets of TLVs that may be encoded in the BGP-LS Attribute associated with a Node NLRI, Link NLRI and Prefix NLRI respectively.

The BGP-LS Attribute may potentially grow large in size depending on the amount of link-state information associated with a single Link-State NLRI. The BGP specification [RFC4271] mandates a maximum BGP

message size of 4096 octets. It is RECOMMENDED that an implementation support [I-D.ietf-idr-bgp-extended-messages] in order to accommodate larger size of information within the BGP-LS Attribute. BGP-LS Producers MUST ensure that they limit the TLVs included in the BGP-LS Attribute to ensure that a BGP update message for a single Link-State NLRI does not cross the maximum limit for a BGP message. The determination of the types of TLVs to be included MAY be made by the BGP-LS Producer based on the BGP-LS Consumer applications requirement and is outside the scope of this document. When a BGP-LS Propagator finds that it is exceeding the maximum BGP message size due to addition or update of some other BGP Attribute (e.g. AS\_PATH), it MUST consider the BGP-LS Attribute to be malformed and handle the propagation as described in Section 7.2.2.

#### 4.3.1. Node Attribute TLVs

The following Node Attribute TLVs are defined for the BGP-LS Attribute associated with a Node NLRI:

| TLV Code Point | Description                  | Length   | Reference (RFC/Section) |
|----------------|------------------------------|----------|-------------------------|
| 263            | Multi-Topology Identifier    | variable | Section 4.2.2.1         |
| 1024           | Node Flag Bits               | 1        | Section 4.3.1.1         |
| 1025           | Opaque Node Attribute        | variable | Section 4.3.1.5         |
| 1026           | Node Name                    | variable | Section 4.3.1.3         |
| 1027           | IS-IS Area Identifier        | variable | Section 4.3.1.2         |
| 1028           | IPv4 Router-ID of Local Node | 4        | [RFC5305]/4.3           |
| 1029           | IPv6 Router-ID of Local Node | 16       | [RFC6119]/4.1           |

Table 6: Node Attribute TLVs

##### 4.3.1.1. Node Flag Bits TLV

The Node Flag Bits TLV carries a bit mask describing node attributes. The value is a variable-length bit array of flags, where each bit represents a node capability.

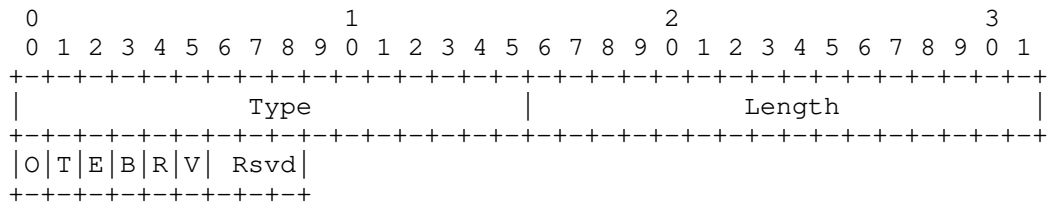


Figure 15: Node Flag Bits TLV Format

The bits are defined as follows:

| Bit             | Description             | Reference  |
|-----------------|-------------------------|------------|
| 'O'             | Overload Bit            | [ISO10589] |
| 'T'             | Attached Bit            | [ISO10589] |
| 'E'             | External Bit            | [RFC2328]  |
| 'B'             | ABR Bit                 | [RFC2328]  |
| 'R'             | Router Bit              | [RFC5340]  |
| 'V'             | V6 Bit                  | [RFC5340]  |
| Reserved (Rsvd) | Reserved for future use |            |

Table 7: Node Flag Bits Definitions

#### 4.3.1.2. IS-IS Area Identifier TLV

An IS-IS node can be part of one or more IS-IS areas. Each of these area addresses is carried in the IS-IS Area Identifier TLV. If multiple area addresses are present, multiple TLVs are used to encode them. The IS-IS Area Identifier TLV may be present in the BGP-LS attribute only when advertised in the Link-State Node NLRI.

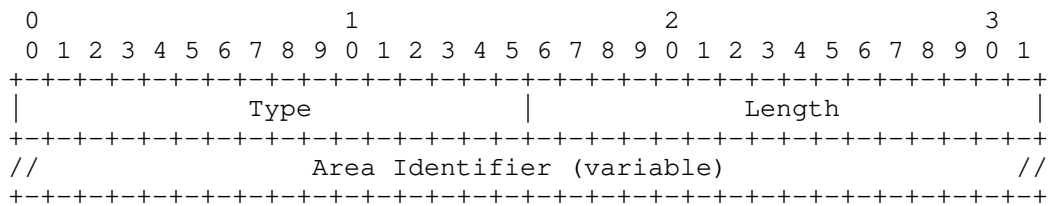


Figure 16: IS-IS Area Identifier TLV Format

## 4.3.1.3. Node Name TLV

The Node Name TLV is optional. Its structure and encoding has been borrowed from [RFC5301]. The Value field identifies the symbolic name of the router node. This symbolic name can be the Fully Qualified Domain Name (FQDN) for the router, it can be a subset of the FQDN (e.g., a hostname), or it can be any string operators want to use for the router. The use of FQDN or a subset of it is strongly RECOMMENDED. The maximum length of the Node Name TLV is 255 octets.

The Value field is encoded in 7-bit ASCII. If a user interface for configuring or displaying this field permits Unicode characters, that user interface is responsible for applying the ToASCII and/or ToUnicode algorithm as described in [RFC5890] to achieve the correct format for transmission or display.

[RFC5301] describes an IS-IS-specific extension and [RFC5642] describes an OSPF extension for advertisement of Node Name which MAY encoded in the Node Name TLV.

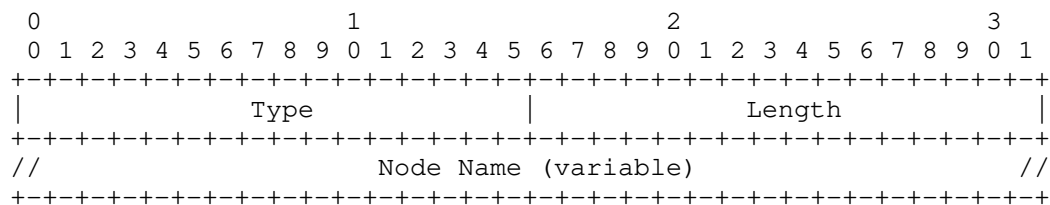


Figure 17: Node Name Format

## 4.3.1.4. Local IPv4/IPv6 Router-ID TLVs

The local IPv4/IPv6 Router-ID TLVs are used to describe auxiliary Router-IDs that the IGP might be using, e.g., for TE and migration purposes such as correlating a Node-ID between different protocols. If there is more than one auxiliary Router-ID of a given type, then each one is encoded in its own TLV.

## 4.3.1.5. Opaque Node Attribute TLV

The Opaque Node Attribute TLV is an envelope that transparently carries optional Node Attribute TLVs advertised by a router. An originating router shall use this TLV for encoding information specific to the protocol advertised in the NLRI header Protocol-ID field or new protocol extensions to the protocol as advertised in the NLRI header Protocol-ID field for which there is no protocol-neutral representation in the BGP Link-State NLRI. The primary use of the Opaque Node Attribute TLV is to bridge the document lag between,



e.g., a new IGP link-state attribute being defined and the protocol-neutral BGP-LS extensions being published. A router, for example, could use this extension in order to advertise the native protocol's Node Attribute TLVs, such as the OSPF Router Informational Capabilities TLV defined in [RFC7770] or the IGP TE Node Capability Descriptor TLV described in [RFC5073].

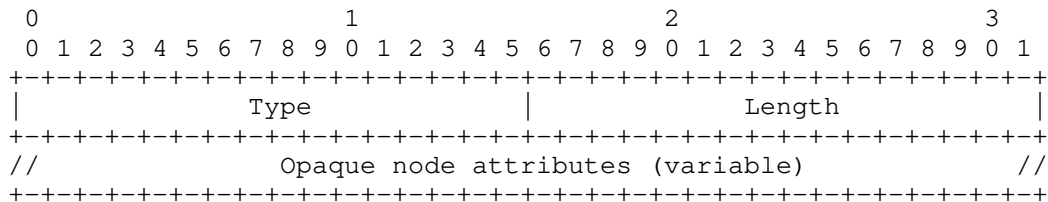


Figure 18: Opaque Node Attribute Format

#### 4.3.2. Link Attribute TLVs

Link Attribute TLVs are TLVs that may be encoded in the BGP-LS attribute with a Link NLRI. Each 'Link Attribute' is a Type/Length/Value (TLV) triplet formatted as defined in Section 4.1. The format and semantics of the Value fields in some Link Attribute TLVs correspond to the format and semantics of the Value fields in IS-IS Extended IS Reachability sub-TLVs, defined in [RFC5305] and [RFC5307]. Other Link Attribute TLVs are defined in this document. Although the encodings for Link Attribute TLVs were originally defined for IS-IS, the TLVs can carry data sourced by either IS-IS or OSPF.

The following Link Attribute TLVs are defined for the BGP-LS Attribute associated with a Link NLRI:

| TLV Code Point | Description                    | IS-IS TLV /Sub-TLV | Reference (RFC/Section) |
|----------------|--------------------------------|--------------------|-------------------------|
| 1028           | IPv4 Router-ID of Local Node   | 134/---            | [RFC5305]/4.3           |
| 1029           | IPv6 Router-ID of Local Node   | 140/---            | [RFC6119]/4.1           |
| 1030           | IPv4 Router-ID of Remote Node  | 134/---            | [RFC5305]/4.3           |
| 1031           | IPv6 Router-ID of Remote Node  | 140/---            | [RFC6119]/4.1           |
| 1088           | Administrative group (color)   | 22/3               | [RFC5305]/3.1           |
| 1089           | Maximum link bandwidth         | 22/9               | [RFC5305]/3.4           |
| 1090           | Max. reservable link bandwidth | 22/10              | [RFC5305]/3.5           |
| 1091           | Unreserved bandwidth           | 22/11              | [RFC5305]/3.6           |
| 1092           | TE Default Metric              | 22/18              | Section 4.3.2.3         |
| 1093           | Link Protection Type           | 22/20              | [RFC5307]/1.2           |
| 1094           | MPLS Protocol Mask             | ---                | Section 4.3.2.2         |
| 1095           | IGP Metric                     | ---                | Section 4.3.2.4         |
| 1096           | Shared Risk Link Group         | ---                | Section 4.3.2.5         |
| 1097           | Opaque Link Attribute          | ---                | Section 4.3.2.6         |
| 1098           | Link Name                      | ---                | Section 4.3.2.7         |

Table 8: Link Attribute TLVs

## 4.3.2.1. IPv4/IPv6 Router-ID TLVs

The local/remote IPv4/IPv6 Router-ID TLVs are used to describe auxiliary Router-IDs that the IGP might be using, e.g., for TE purposes. All auxiliary Router-IDs of both the local and the remote node MUST be included in the link attribute of each Link NLRI. If there is more than one auxiliary Router-ID of a given type, then multiple TLVs are used to encode them.

## 4.3.2.2. MPLS Protocol Mask TLV

The MPLS Protocol Mask TLV carries a bit mask describing which MPLS signaling protocols are enabled. The length of this TLV is 1. The

value is a bit array of 8 flags, where each bit represents an MPLS Protocol capability.

Generation of the MPLS Protocol Mask TLV is only valid for and SHOULD only be used with originators that have local link insight, for example, the Protocol-IDs 'Static configuration' or 'Direct' as per Table 2. The MPLS Protocol Mask TLV MUST NOT be included in NLRIs with the other Protocol-IDs listed in Table 2.

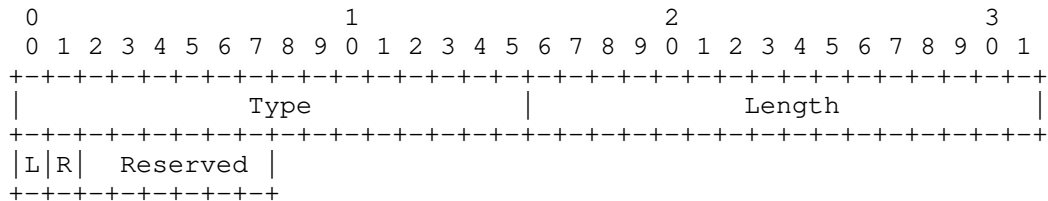


Figure 19: MPLS Protocol Mask TLV

The following bits are defined:

| Bit        | Description                                 | Reference |
|------------|---------------------------------------------|-----------|
| 'L'        | Label Distribution Protocol (LDP)           | [RFC5036] |
| 'R'        | Extension to RSVP for LSP Tunnels (RSVP-TE) | [RFC3209] |
| 'Reserved' | Reserved for future use                     |           |

Table 9: MPLS Protocol Mask TLV Codes

#### 4.3.2.3. TE Default Metric TLV

The TE Default Metric TLV carries the Traffic Engineering metric for this link. The length of this TLV is fixed at 4 octets. If a source protocol uses a metric width of less than 32 bits, then the high-order bits of this field MUST be padded with zero.

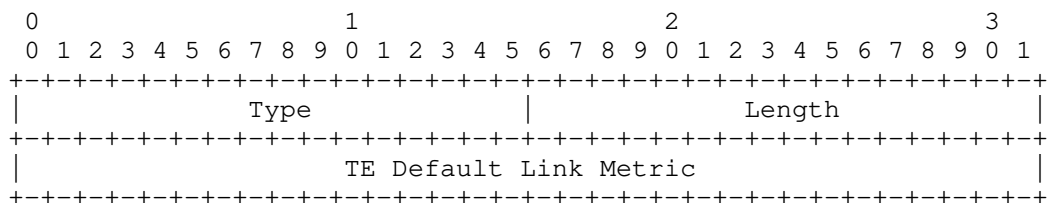


Figure 20: TE Default Metric TLV Format

## 4.3.2.4. IGP Metric TLV

The IGP Metric TLV carries the metric for this link. The length of this TLV is variable, depending on the metric width of the underlying protocol. IS-IS small metrics have a length of 1 octet (the two most significant bits are ignored). OSPF link metrics have a length of 2 octets. IS-IS wide metrics have a length of 3 octets.

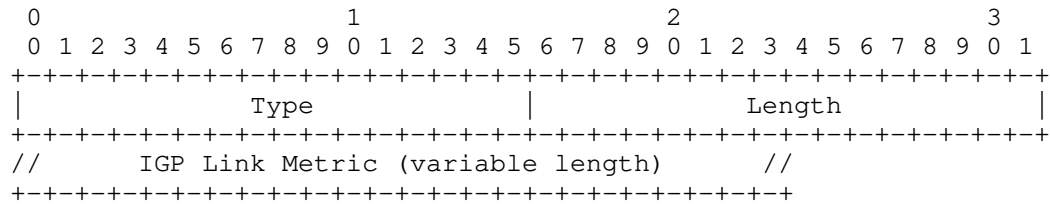


Figure 21: IGP Metric TLV Format

## 4.3.2.5. Shared Risk Link Group TLV

The Shared Risk Link Group (SRLG) TLV carries the Shared Risk Link Group information (see Section 2.3 ("Shared Risk Link Group Information") of [RFC4202]). It contains a data structure consisting of a (variable) list of SRLG values, where each element in the list has 4 octets, as shown in Figure 22. The length of this TLV is 4 \* (number of SRLG values).

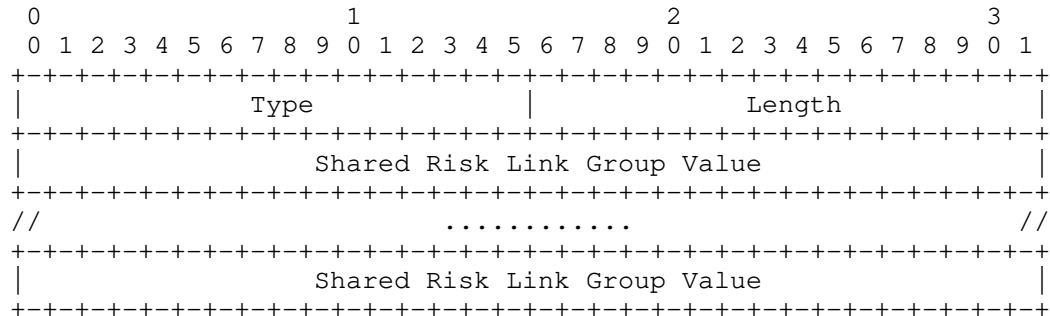


Figure 22: Shared Risk Link Group TLV Format

The SRLG TLV for OSPF-TE is defined in [RFC4203]. In IS-IS, the SRLG information is carried in two different TLVs: the IPv4 (SRLG) TLV (Type 138) defined in [RFC5307] and the IPv6 SRLG TLV (Type 139) defined in [RFC6119]. In Link-State NLRI, both IPv4 and IPv6 SRLG information are carried in a single TLV.

## 4.3.2.6. Opaque Link Attribute TLV

The Opaque Link Attribute TLV is an envelope that transparently carries optional Link Attribute TLVs advertised by a router. An originating router shall use this TLV for encoding information specific to the protocol advertised in the NLRI header Protocol-ID field or new protocol extensions to the protocol as advertised in the NLRI header Protocol-ID field for which there is no protocol-neutral representation in the BGP Link-State NLRI. The primary use of the Opaque Link Attribute TLV is to bridge the document lag between, e.g., a new IGP link-state attribute being defined and the 'protocol-neutral' BGP-LS extensions being published.

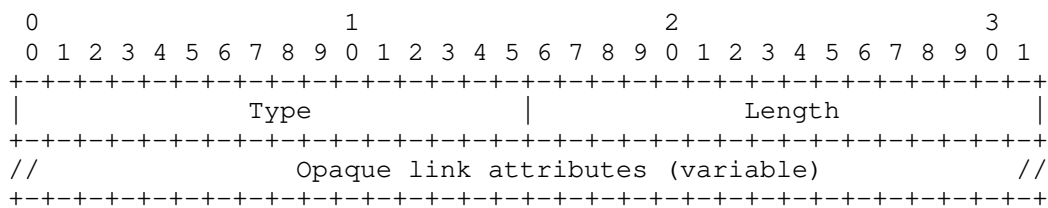


Figure 23: Opaque Link Attribute TLV Format

## 4.3.2.7. Link Name TLV

The Link Name TLV is optional. The Value field identifies the symbolic name of the router link. This symbolic name can be the FQDN for the link, it can be a subset of the FQDN, or it can be any string operators want to use for the link. The use of FQDN or a subset of it is strongly RECOMMENDED. The maximum length of the Link Name TLV is 255 octets.

The Value field is encoded in 7-bit ASCII. If a user interface for configuring or displaying this field permits Unicode characters, that user interface is responsible for applying the ToASCII and/or ToUnicode algorithm as described in [RFC5890] to achieve the correct format for transmission or display.

How a router derives and injects link names is outside of the scope of this document.

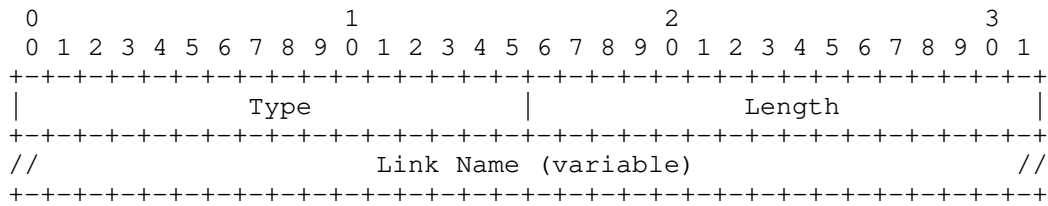


Figure 24: Link Name TLV Format

#### 4.3.3. Prefix Attribute TLVs

Prefixes are learned from the IGP topology (IS-IS or OSPF) with a set of IGP attributes (such as metric, route tags, etc.) that are advertised in the BGP-LS Attribute with Prefix NLRI types 3 and 4.

The following Prefix Attribute TLVs are defined for the BGP-LS Attribute associated with a Prefix NLRI:

| TLV Code Point | Description             | Length   | Reference       |
|----------------|-------------------------|----------|-----------------|
| 1152           | IGP Flags               | 1        | Section 4.3.3.1 |
| 1153           | IGP Route Tag           | 4*n      | [RFC5130]       |
| 1154           | IGP Extended Route Tag  | 8*n      | [RFC5130]       |
| 1155           | Prefix Metric           | 4        | [RFC5305]       |
| 1156           | OSPF Forwarding Address | 4        | [RFC2328]       |
| 1157           | Opaque Prefix Attribute | variable | Section 4.3.3.6 |

Table 10: Prefix Attribute TLVs

##### 4.3.3.1. IGP Flags TLV

The IGP Flags TLV contains IS-IS and OSPF flags and bits originally assigned to the prefix. The IGP Flags TLV is encoded as follows:

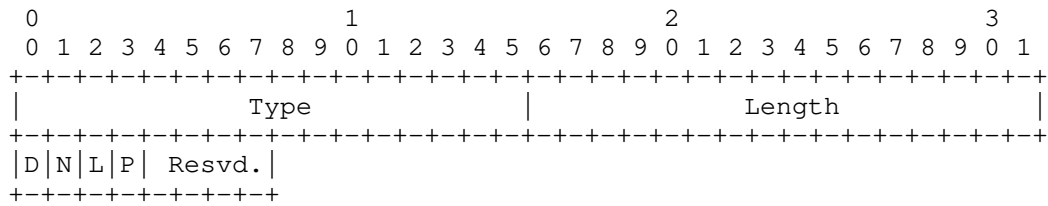


Figure 25: IGP Flag TLV Format

The Value field contains bits defined according to the table below:

| Bit      | Description               | Reference |
|----------|---------------------------|-----------|
| 'D'      | IS-IS Up/Down Bit         | [RFC5305] |
| 'N'      | OSPF "no unicast" Bit     | [RFC5340] |
| 'L'      | OSPF "local address" Bit  | [RFC5340] |
| 'P'      | OSPF "propagate NSSA" Bit | [RFC5340] |
| Reserved | Reserved for future use.  |           |

Table 11: IGP Flag Bits Definitions

#### 4.3.3.2. IGP Route Tag TLV

The IGP Route Tag TLV carries original IGP Tags (IS-IS [RFC5130] or OSPF) of the prefix and is encoded as follows:

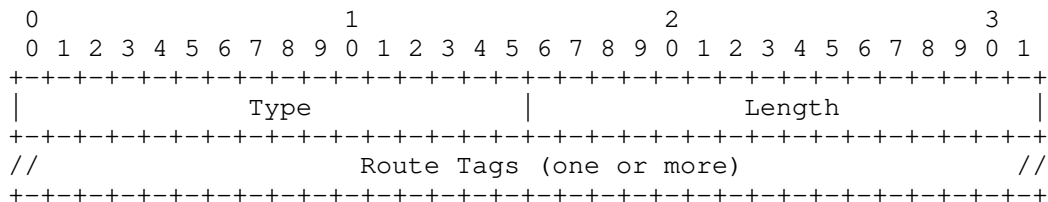


Figure 26: IGP Route Tag TLV Format

Length is a multiple of 4.

The Value field contains one or more Route Tags as learned in the IGP topology.

## 4.3.3.3. Extended IGP Route Tag TLV

The Extended IGP Route Tag TLV carries IS-IS Extended Route Tags of the prefix [RFC5130] and is encoded as follows:

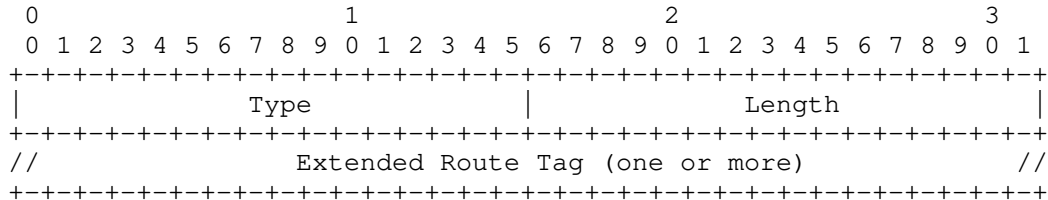


Figure 27: Extended IGP Route Tag TLV Format

Length is a multiple of 8.

The Extended Route Tag field contains one or more Extended Route Tags as learned in the IGP topology.

## 4.3.3.4. Prefix Metric TLV

The Prefix Metric TLV is an optional attribute and may only appear once. If present, it carries the metric of the prefix as known in the IGP topology as described in Section 4 of [RFC5305] (and therefore represents the reachability cost to the prefix). If not present, it means that the prefix is advertised without any reachability.

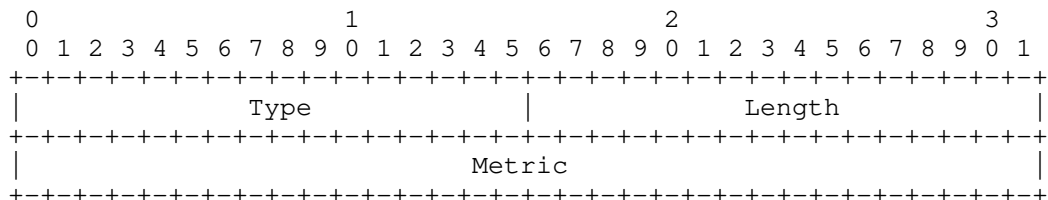


Figure 28: Prefix Metric TLV Format

Length is 4.

## 4.3.3.5. OSPF Forwarding Address TLV

The OSPF Forwarding Address TLV [RFC2328] [RFC5340] carries the OSPF forwarding address as known in the original OSPF advertisement. Forwarding address can be either IPv4 or IPv6.



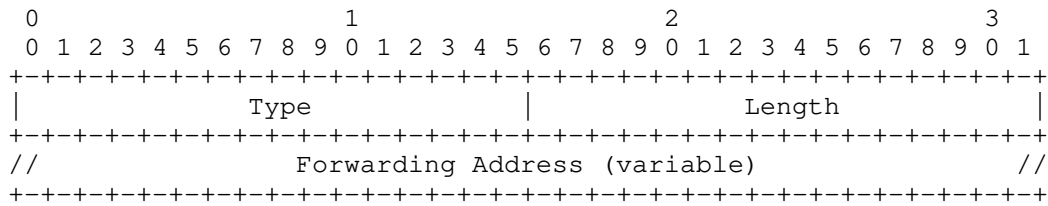


Figure 29: OSPF Forwarding Address TLV Format

Length is 4 for an IPv4 forwarding address, and 16 for an IPv6 forwarding address.

#### 4.3.3.6. Opaque Prefix Attribute TLV

The Opaque Prefix Attribute TLV is an envelope that transparently carries optional Prefix Attribute TLVs advertised by a router. An originating router shall use this TLV for encoding information specific to the protocol advertised in the NLRI header Protocol-ID field or new protocol extensions to the protocol as advertised in the NLRI header Protocol-ID field for which there is no protocol-neutral representation in the BGP Link-State NLRI. The primary use of the Opaque Prefix Attribute TLV is to bridge the document lag between, e.g., a new IGP link-state attribute being defined and the protocol-neutral BGP-LS extensions being published.

The format of the TLV is as follows:

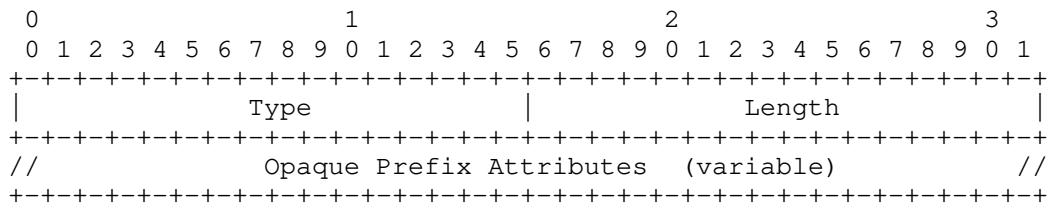


Figure 30: Opaque Prefix Attribute TLV Format

Type is as specified in Table 10. Length is variable.

#### 4.4. Private Use

TLVs for Vendor Private use are supported using the code point range reserved as indicated in Section 6. For such TLV use in the NLRI or BGP-LS Attribute, the format as described in Section 4.1 is to be used and a 4 octet field MUST be included as the first field in the value to carry the Enterprise Code. For a private use NLRI Type, a 4 octet field MUST be included as the first field in the NLRI

immediately following the Total NLRI Length field of the Link-State NLRI format as described in Section 4.2 to carry the Enterprise Code. The Enterprise Codes are listed at <http://www.iana.org/assignments/enterprise-numbers>. This enables use vendor specific extensions without conflicts.

#### 4.5. BGP Next-Hop Information

BGP link-state information for both IPv4 and IPv6 networks can be carried over either an IPv4 BGP session or an IPv6 BGP session. If an IPv4 BGP session is used, then the next hop in the MP\_REACH\_NLRI SHOULD be an IPv4 address. Similarly, if an IPv6 BGP session is used, then the next hop in the MP\_REACH\_NLRI SHOULD be an IPv6 address. Usually, the next hop will be set to the local endpoint address of the BGP session. The next-hop address MUST be encoded as described in [RFC4760]. The Length field of the next-hop address will specify the next-hop address family. If the next-hop length is 4, then the next hop is an IPv4 address; if the next-hop length is 16, then it is a global IPv6 address; and if the next-hop length is 32, then there is one global IPv6 address followed by a link-local IPv6 address. The link-local IPv6 address should be used as described in [RFC2545]. For VPN Subsequent Address Family Identifier (SAFI), as per custom, an 8-byte Route Distinguisher set to all zero is prepended to the next hop.

The BGP Next Hop attribute is used by each BGP-LS speaker to validate the NLRI it receives. In case identical NLRIs are sourced by multiple BGP-LS Producers, the BGP Next Hop attribute is used to tiebreak as per the standard BGP path decision process. This specification doesn't mandate any rule regarding the rewrite of the BGP Next Hop attribute.

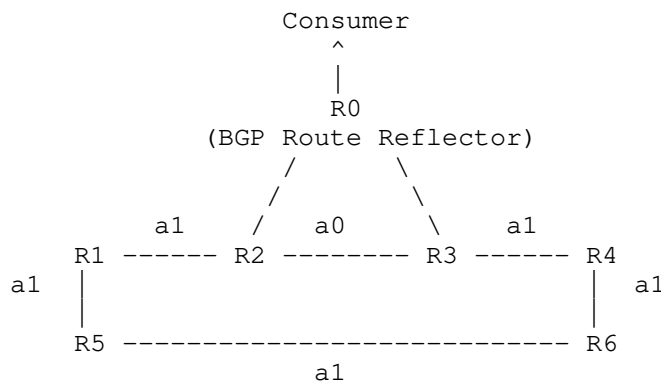
#### 4.6. Inter-AS Links

The main source of TE information is the IGP, which is not active on inter-AS links. In some cases, the IGP may have information of inter-AS links [RFC5392] [RFC5316]. In other cases, an implementation SHOULD provide a means to inject inter-AS links into BGP-LS. The exact mechanism used to provision the inter-AS links is outside the scope of this document

#### 4.7. Handling of Unreachable IGP Nodes

The origination and propagation of IGP link-state information via BGP needs to provide a consistent and true view of the topology of the IGP domain. BGP-LS provides an abstraction of the protocol specifics and BGP-LS Consumers may be varied types of applications.

A BGP-LS Consumer talks to a BGP route-reflector (RR) R0 which is aggregating the BGP-LS feed from the BGP-LS Producers R2 and R3. Here R2 and R3 provide a redundant topology feed via BGP-LS to R0. Normally, R0 would receive two identical copies of all the Link-State NLRIs from both R2 and R3 and it would pick one of them (say R2) based on the standard BGP best path decision process.



Consider a scenario where the link between R5 and R6 is lost (thereby partitioning the area 1) and its impact on the OSPF LSDB at R2 and R3.

At the same time, R6 has removed the link 6-5 from its Router LSA and this updated LSA is available at R3. Similarly, R3 also has a stale copy of R5's Router LSA having the link 5-6 in it. Based on it's LSDB, R3 will advertise only the half-link 5-6 that it has derived from R5's stale Router LSA.

[Page 37]

Also if R2 continues to report Link-State NLRIs corresponding to the stale copy of Router LSA of R4 and R6 nodes then R0 would prefer them over the valid Link-State NLRIs for R4 and R6 that it is receiving from R3 based on its BGP decision process. This would result in the BGP-LS Consumer getting stale and inaccurate topology information. This problems scenario is avoided if R2 were to not advertise the link-state information corresponding to R4 and R6 and if R3 were to not advertise similarly for R1 and R5.

A BGP-LS Producer MUST withdraw all link-state objects advertised by it in BGP when the node that originated its corresponding LSP/LSAs is determined to have become unreachable in the IGP and it MUST re-advertise those link-state objects only after that node becomes reachable again in the IGP domain.

#### 4.8. Router-ID Anchoring Example: ISO Pseudonode

Encoding of a broadcast LAN in IS-IS provides a good example of how Router-IDs are encoded. Consider Figure 32. This represents a Broadcast LAN between a pair of routers. The "real" (non-pseudonode) routers have both an IPv4 Router-ID and IS-IS Node-ID. The pseudonode does not have an IPv4 Router-ID. Node1 is the DIS for the LAN. Two unidirectional links (Node1, Pseudonode1) and (Pseudonode1, Node2) are being generated.

The Link NLRI of (Node1, Pseudonode1) is encoded as follows. The IGP Router-ID TLV of the local Node Descriptor is 6 octets long and contains the ISO-ID of Node1, 1920.0000.2001. The IGP Router-ID TLV of the remote Node Descriptor is 7 octets long and contains the ISO-ID of Pseudonode1, 1920.0000.2001.02. The BGP-LS attribute of this link contains one local IPv4 Router-ID TLV (TLV type 1028) containing 192.0.2.1, the IPv4 Router-ID of Node1.

The Link NLRI of (Pseudonode1, Node2) is encoded as follows. The IGP Router-ID TLV of the local Node Descriptor is 7 octets long and contains the ISO-ID of Pseudonode1, 1920.0000.2001.02. The IGP Router-ID TLV of the remote Node Descriptor is 6 octets long and contains the ISO-ID of Node2, 1920.0000.2002. The BGP-LS attribute of this link contains one remote IPv4 Router-ID TLV (TLV type 1030) containing 192.0.2.2, the IPv4 Router-ID of Node2.

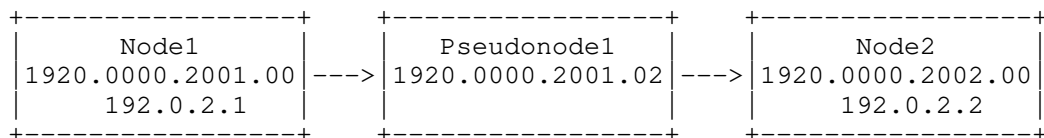


Figure 32: IS-IS Pseudonodes

#### 4.9. Router-ID Anchoring Example: OSPF Pseudonode

Encoding of a broadcast LAN in OSPF provides a good example of how Router-IDs and local Interface IPs are encoded. Consider Figure 33. This represents a Broadcast LAN between a pair of routers. The "real" (non-pseudonode) routers have both an IPv4 Router-ID and an Area Identifier. The pseudonode does have an IPv4 Router-ID, an IPv4 Interface Address (for disambiguation), and an OSPF Area. Node1 is the DR for the LAN; hence, its local IP address 10.1.1.1 is used as both the Router-ID and Interface IP for the pseudonode keys. Two unidirectional links, (Node1, Pseudonode1) and (Pseudonode1, Node2), are being generated.

The Link NLRI of (Node1, Pseudonode1) is encoded as follows:

- o Local Node Descriptor
  - TLV #515: IGP Router-ID: 11.11.11.11
  - TLV #514: OSPF Area-ID: ID:0.0.0.0
- o Remote Node Descriptor
  - TLV #515: IGP Router-ID: 11.11.11.11:10.1.1.1
  - TLV #514: OSPF Area-ID: ID:0.0.0.0

The Link NLRI of (Pseudonode1, Node2) is encoded as follows:

- o Local Node Descriptor
  - TLV #515: IGP Router-ID: 11.11.11.11:10.1.1.1
  - TLV #514: OSPF Area-ID: ID:0.0.0.0
- o Remote Node Descriptor
  - TLV #515: IGP Router-ID: 33.33.33.34
  - TLV #514: OSPF Area-ID: ID:0.0.0.0

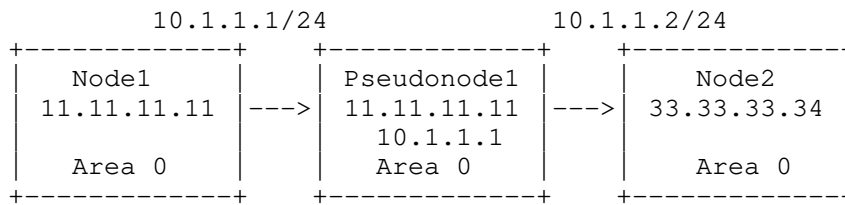


Figure 33: OSPF Pseudonodes

The LAN subnet 10.1.1.0/24 is not included in the Router LSA of Node1 or Node2. The Network LSA for this LAN advertised by the DR Node1 contains the subnet mask for the LAN along with the DR address. A Prefix NLRI corresponding to the LAN subnet is advertised with the Pseudonode1 used as the Local node using the DR address and the subnet mask from the Network LSA.

#### 4.10. Router-ID Anchoring Example: OSPFv2 to IS-IS Migration

Graceful migration from one IGP to another requires coordinated operation of both protocols during the migration period. Such a coordination requires identifying a given physical link in both IGPs. The IPv4 Router-ID provides that "glue", which is present in the Node Descriptors of the OSPF Link NLRI and in the link attribute of the IS-IS Link NLRI.

Consider a point-to-point link between two routers, A and B, that initially were OSPFv2-only routers and then IS-IS is enabled on them. Node A has IPv4 Router-ID and ISO-ID; node B has IPv4 Router-ID, IPv6 Router-ID, and ISO-ID. Each protocol generates one Link NLRI for the link (A, B), both of which are carried by BGP-LS. The OSPFv2 Link NLRI for the link is encoded with the IPv4 Router-ID of nodes A and B in the local and remote Node Descriptors, respectively. The IS-IS Link NLRI for the link is encoded with the ISO-ID of nodes A and B in the local and remote Node Descriptors, respectively. In addition, the BGP-LS attribute of the IS-IS Link NLRI contains the TLV type 1028 containing the IPv4 Router-ID of node A, TLV type 1030 containing the IPv4 Router-ID of node B, and TLV type 1031 containing the IPv6 Router-ID of node B. In this case, by using IPv4 Router-ID, the link (A, B) can be identified in both the IS-IS and OSPF protocol.

#### 5. Link to Path Aggregation

Distribution of all links available in the global Internet is certainly possible; however, it is not desirable from a scaling and privacy point of view. Therefore, an implementation may support a link to path aggregation. Rather than advertising all specific links

of a domain, an ASBR may advertise an "aggregate link" between a non-adjacent pair of nodes. The "aggregate link" represents the aggregated set of link properties between a pair of non-adjacent nodes. The actual methods to compute the path properties (of bandwidth, metric, etc.) are outside the scope of this document. The decision whether to advertise all specific links or aggregated links is an operator's policy choice. To highlight the varying levels of exposure, the following deployment examples are discussed.

#### 5.1. Example: No Link Aggregation

Consider Figure 34. Both AS1 and AS2 operators want to protect their inter-AS {R1, R3}, {R2, R4} links using RSVP-FRR LSPs. If R1 wants to compute its link-protection LSP to R3, it needs to "see" an alternate path to R3. Therefore, the AS2 operator exposes its topology. All BGP-TE-enabled routers in AS1 "see" the full topology of AS2 and therefore can compute a backup path. Note that the computing router decides if the direct link between {R3, R4} or the {R4, R5, R3} path is used.

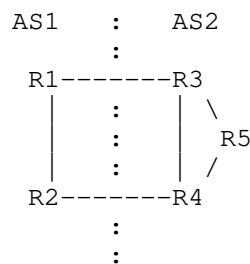


Figure 34: No Link Aggregation

#### 5.2. Example: ASBR to ASBR Path Aggregation

The brief difference between the "no-link aggregation" example and this example is that no specific link gets exposed. Consider Figure 35. The only link that gets advertised by AS2 is an "aggregate" link between R3 and R4. This is enough to tell AS1 that there is a backup path. However, the actual links being used are hidden from the topology.

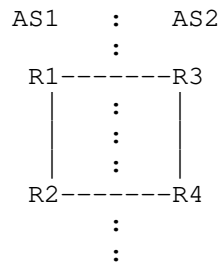


Figure 35: ASBR Link Aggregation

### 5.3. Example: Multi-AS Path Aggregation

Service providers in control of multiple ASes may even decide to not expose their internal inter-AS links. Consider Figure 36. AS3 is modeled as a single node that connects to the border routers of the aggregated domain.

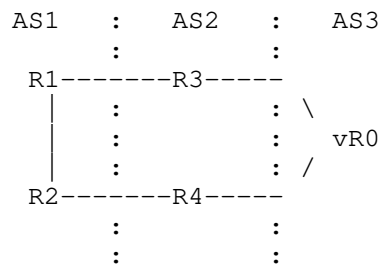


Figure 36: Multi-AS Aggregation

## 6. IANA Considerations

IANA has assigned address family number 16388 (BGP-LS) in the "Address Family Numbers" registry with [RFC7752] as a reference.

IANA has assigned SAFI values 71 (BGP-LS) and 72 (BGP-LS-VPN) in the "SAFI Values" sub-registry under the "Subsequent Address Family Identifiers (SAFI) Parameters" registry.

IANA has assigned value 29 (BGP-LS Attribute) in the "BGP Path Attributes" sub-registry under the "Border Gateway Protocol (BGP) Parameters" registry.

IANA has created a new "Border Gateway Protocol - Link State (BGP-LS) Parameters" registry at <http://www.iana.org/assignments/bgp-ls-parameters>. All of the following registries are BGP-LS specific and are accessible under this registry:



- o "BGP-LS NLRI-Types" registry

Value 0 is reserved. The maximum value is 65535. The range 65000-65535 is for Private Use. The registry has been populated with the values shown in Table 1. Allocations within the registry require documentation of the proposed use of the allocated value (Specification Required) and approval by the Designated Expert assigned by the IESG (see [RFC8126]).

- o "BGP-LS Protocol-IDs" registry

Value 0 is reserved. The maximum value is 255. The range 200-255 is for Private Use. The registry has been populated with the values shown in Table 2. Allocations within the registry require documentation of the proposed use of the allocated value (Specification Required) and approval by the Designated Expert assigned by the IESG (see [RFC8126]).

- o "BGP-LS Well-Known Instance-IDs" registry

This registry was setup via [RFC7752] and is no longer required. It may be retained as deprecated.

- o "BGP-LS Node Descriptor, Link Descriptor, Prefix Descriptor, and Attribute TLVs" registry

Values 0-255 are reserved. Values 256-65535 will be used for code points. The range 65000-65535 is for Private Use. The registry has been populated with the values shown in Table 12. Allocations within the registry require documentation of the proposed use of the allocated value (Specification Required) and approval by the Designated Expert assigned by the IESG (see [RFC8126]).

#### 6.1. Guidance for Designated Experts

In all cases of review by the Designated Expert (DE) described here, the DE is expected to ascertain the existence of suitable documentation (a specification) as described in [RFC8126] and to verify that the document is permanently and publicly available. The DE is also expected to check the clarity of purpose and use of the requested code points. Last, the DE must verify that any specification produced in the IETF that requests one of these code points has been made available for review by the IDR working group and that any specification produced outside the IETF does not conflict with work that is active or already published within the IETF.

## 7. Manageability Considerations

This section is structured as recommended in [RFC5706].

### 7.1. Operational Considerations

#### 7.1.1. Operations

Existing BGP operational procedures apply. No new operation procedures are defined in this document. It is noted that the NLRI information present in this document carries purely application-level data that has no immediate impact on the corresponding forwarding state computed by BGP. As such, any churn in reachability information has a different impact than regular BGP updates, which need to change the forwarding state for an entire router. It is expected that the distribution of this NLRI SHOULD be handled by dedicated route reflectors in most deployments providing a level of isolation and fault containment between different NLRI types. In the event of dedicated route reflectors not being available, other alternate mechanisms like separation of BGP instances or separate BGP sessions (e.g. using different addresses for peering) for Link-State information distribution SHOULD be used.

#### 7.1.2. Installation and Initial Setup

Configuration parameters defined in Section 7.2.3 SHOULD be initialized to the following default values:

- o The Link-State NLRI capability is turned off for all neighbors.
- o The maximum rate at which Link-State NLRIs will be advertised/withdrawn from neighbors is set to 200 updates per second.

#### 7.1.3. Migration Path

The proposed extension is only activated between BGP peers after capability negotiation. Moreover, the extensions can be turned on/off on an individual peer basis (see Section 7.2.3), so the extension can be gradually rolled out in the network.

#### 7.1.4. Requirements on Other Protocols and Functional Components

The protocol extension defined in this document does not put new requirements on other protocols or functional components.

#### 7.1.5. Impact on Network Operation

Frequency of Link-State NLRI updates could interfere with regular BGP prefix distribution. A network operator MAY use a dedicated Route-Reflector infrastructure to distribute Link-State NLRIs.

Distribution of Link-State NLRIs SHOULD be limited to a single admin domain, which can consist of multiple areas within an AS or multiple ASes.

#### 7.1.6. Verifying Correct Operation

Existing BGP procedures apply. In addition, an implementation SHOULD allow an operator to:

- o List neighbors with whom the speaker is exchanging Link-State NLRIs.

### 7.2. Management Considerations

#### 7.2.1. Management Information

The IDR working group has documented and continues to document parts of the Management Information Base and YANG models for managing and monitoring BGP speakers and the sessions between them. It is currently believed that the BGP session running BGP-LS is not substantially different from any other BGP session and can be managed using the same data models.

#### 7.2.2. Fault Management

This section describes the fault management actions, as described in [RFC7606], that are to be performed for handling of BGP update messages for BGP-LS.

A Link-State NLRI MUST NOT be considered as malformed or invalid based on the inclusion/exclusion of TLVs or contents of the TLV fields (i.e. semantic errors), as described in Section 4.1 and Section 4.2.

A BGP-LS Speaker MUST perform the following syntactic validation of the Link-State NLRI to determine if it is malformed.

- o Does the sum of all TLVs found in the BGP MP\_REACH\_NLRI attribute correspond to the BGP MP\_REACH\_NLRI length?
- o Does the sum of all TLVs found in the BGP MP\_UNREACH\_NLRI attribute correspond to the BGP MP\_UNREACH\_NLRI length?

- o Does the sum of all TLVs found in a Link-State NLRI correspond to the Total NLRI Length field of all its Descriptors?
- o Is the length of the TLVs and, when the TLV is recognized then, its sub-TLVs in the NLRI valid?
- o Has the syntactic correctness of the NLRI fields been verified as per [RFC7606]?
- o Has the rule regarding ordering of TLVs been followed as described in Section 4.1?

When the error determined allows for the router to skip the malformed NLRI(s) and continue processing of the rest of the update message (e.g. when the TLV ordering rule is violated), then it MUST handle such malformed NLRIs as 'Treat-as-withdraw'. In other cases, where the error in the NLRI encoding results in the inability to process the BGP update message (e.g. length related encoding errors), then the router SHOULD handle such malformed NLRIs as 'AFI/SAFI disable' when other AFI/SAFI besides BGP-LS are being advertised over the same session. Alternately, the router MUST perform 'session reset' when the session is only being used for BGP-LS or when it 'AFI/SAFI disable' action is not possible.

A BGP-LS Attribute MUST NOT be considered as malformed or invalid based on the inclusion/exclusion of TLVs or contents of the TLV fields (i.e. semantic errors), as described in Section 4.1 and Section 4.3.

A BGP-LS Speaker MUST perform the following syntactic validation of the BGP-LS Attribute to determine if it is malformed.

- o Does the sum of all TLVs found in the BGP-LS Attribute correspond to the BGP-LS Attribute length?
- o Has the syntactic correctness of the Attributes (including BGP-LS Attribute) been verified as per [RFC7606]?
- o Is the length of each TLV and, when the TLV is recognized then, its sub-TLVs in the BGP-LS Attribute valid?

When the error determined allows for the router to skip the malformed BGP-LS Attribute and continue processing of the rest of the update message (e.g. when the BGP-LS Attribute length and the total Path Attribute Length are correct but some TLV/sub-TLV length within the BGP-LS Attribute is invalid), then it MUST handle such malformed BGP-LS Attribute as 'Attribute Discard'. In other cases, where the error in the BGP-LS Attribute encoding results in the inability to process

the BGP update message then the handling is the same as described above for the malformed NLRI.

Note that the 'Attribute Discard' action results in the loss of all TLVs in the BGP-LS Attribute and not the removal of a specific malformed TLV. The removal of specific malformed TLVs may give a wrong indication to a BGP-LS Consumer of that specific information being deleted or not available.

When a BGP Speaker receives an update message with Link-State NLRI(s) in the MP\_REACH\_NLRI but without the BGP-LS Attribute, it is most likely an indication that a BGP Speaker preceding it has performed the 'Attribute Discard' fault handling. An implementation SHOULD preserve and propagate the Link-State NLRIs in such an update message so that the BGP-LS Consumers can detect the loss of link-state information for that object and not assume its deletion/withdraw. This also makes it possible for a network operator to trace back to the BGP-LS Propagator which actually detected a fault with the BGP-LS Attribute.

An implementation SHOULD log an error for any errors found during syntax validation for further analysis.

A BGP-LS Propagator SHOULD NOT perform semantic validation of the Link-State NLRI or the BGP-LS Attribute to determine if it is malformed or invalid. Some types of semantic validation that are not to be performed by a BGP-LS Propagator are as follows (and this is not to be considered as an exhaustive list):

- o is a mandatory TLV present or not?
- o is the length of a fixed length TLV correct or the length of a variable length TLV a valid/missible?
- o are the values of TLV fields valid or permissible?
- o are the inclusion and use of TLVs/sub-TLVs with specific Link-State NLRI types valid?

Each TLV MAY indicate the valid and permissible values and their semantics that can to be used only by a BGP-LS Consumer for its semantic validation. However, the handling of any errors may be specific to the particular application and outside the scope of this document. A BGP-LS Consumer should ignore unrecognized and unexpected TLV types in both the NLRI and BGP-LS Attribute portions and not consider their presence as an error.

### 7.2.3. Configuration Management

An implementation SHOULD allow the operator to specify neighbors to which Link-State NLRIs will be advertised and from which Link-State NLRIs will be accepted.

An implementation SHOULD allow the operator to specify the maximum rate at which Link-State NLRIs will be advertised/withdrawn from neighbors.

An implementation SHOULD allow the operator to specify the maximum number of Link-State NLRIs stored in a router's Routing Information Base (RIB).

An implementation SHOULD allow the operator to create abstracted topologies that are advertised to neighbors and create different abstractions for different neighbors.

An implementation SHOULD allow the operator to configure a 64-bit Instance-ID.

An implementation SHOULD allow the operator to configure ASN and BGP-LS identifiers (refer Section 4.2.1.4).

An implementation SHOULD allow the operator to configure the maximum size of the BGP-LS Attribute that may be used on a BGP-LS Producer.

### 7.2.4. Accounting Management

Not Applicable.

### 7.2.5. Performance Management

An implementation SHOULD provide the following statistics:

- o Total number of Link-State NLRI updates sent/received
- o Number of Link-State NLRI updates sent/received, per neighbor
- o Number of errored received Link-State NLRI updates, per neighbor
- o Total number of locally originated Link-State NLRIs

These statistics should be recorded as absolute counts since system or session start time. An implementation MAY also enhance this information by recording peak per-second counts in each case.

## 7.2.6. Security Management

An operator SHOULD define an import policy to limit inbound updates as follows:

- o Drop all updates from peers that are only serving BGP-LS Consumers.

An implementation MUST have the means to limit inbound updates.

## 8. TLV/Sub-TLV Code Points Summary

This section contains the global table of all TLVs/sub-TLVs defined in this document.

| TLV Code Point | Description                    | IS-IS TLV/<br>Sub-TLV | Reference<br>(RFC/Section) |
|----------------|--------------------------------|-----------------------|----------------------------|
| 256            | Local Node Descriptors         | ---                   | Section 4.2.1.2            |
| 257            | Remote Node Descriptors        | ---                   | Section 4.2.1.3            |
| 258            | Link Local/Remote Identifiers  | 22/4                  | [RFC5307]/1.1              |
| 259            | IPv4 interface address         | 22/6                  | [RFC5305]/3.2              |
| 260            | IPv4 neighbor address          | 22/8                  | [RFC5305]/3.3              |
| 261            | IPv6 interface address         | 22/12                 | [RFC6119]/4.2              |
| 262            | IPv6 neighbor address          | 22/13                 | [RFC6119]/4.3              |
| 263            | Multi-Topology ID              | ---                   | Section 4.2.2.1            |
| 264            | OSPF Route Type                | ---                   | Section 4.2.3              |
| 265            | IP Reachability Information    | ---                   | Section 4.2.3              |
| 512            | Autonomous System              | ---                   | Section 4.2.1.4            |
| 513            | BGP-LS Identifier (deprecated) | ---                   | Section 4.2.1.4            |
| 514            | OSPF Area-ID                   | ---                   | Section 4.2.1.4            |
| 515            | IGP Router-ID                  | ---                   | Section 4.2.1.4            |
| 1024           | Node Flag Bits                 | ---                   | Section 4.3.1.1            |
| 1025           | Opaque Node Attribute          | ---                   | Section 4.3.1.5            |
| 1026           | Node Name                      | variable              | Section 4.3.1.3            |
| 1027           | IS-IS Area Identifier          | variable              | Section 4.3.1.2            |

|      |                                |         |                 |
|------|--------------------------------|---------|-----------------|
| 1028 | IPv4 Router-ID of Local Node   | 134/--- | [RFC5305]/4.3   |
| 1029 | IPv6 Router-ID of Local Node   | 140/--- | [RFC6119]/4.1   |
| 1030 | IPv4 Router-ID of Remote Node  | 134/--- | [RFC5305]/4.3   |
| 1031 | IPv6 Router-ID of Remote Node  | 140/--- | [RFC6119]/4.1   |
| 1088 | Administrative group (color)   | 22/3    | [RFC5305]/3.1   |
| 1089 | Maximum link bandwidth         | 22/9    | [RFC5305]/3.4   |
| 1090 | Max. reservable link bandwidth | 22/10   | [RFC5305]/3.5   |
| 1091 | Unreserved bandwidth           | 22/11   | [RFC5305]/3.6   |
| 1092 | TE Default Metric              | 22/18   | Section 4.3.2.3 |
| 1093 | Link Protection Type           | 22/20   | [RFC5307]/1.2   |
| 1094 | MPLS Protocol Mask             | ---     | Section 4.3.2.2 |
| 1095 | IGP Metric                     | ---     | Section 4.3.2.4 |
| 1096 | Shared Risk Link Group         | ---     | Section 4.3.2.5 |
| 1097 | Opaque Link Attribute          | ---     | Section 4.3.2.6 |
| 1098 | Link Name                      | ---     | Section 4.3.2.7 |
| 1152 | IGP Flags                      | ---     | Section 4.3.3.1 |
| 1153 | IGP Route Tag                  | ---     | [RFC5130]       |
| 1154 | IGP Extended Route Tag         | ---     | [RFC5130]       |
| 1155 | Prefix Metric                  | ---     | [RFC5305]       |
| 1156 | OSPF Forwarding Address        | ---     | [RFC2328]       |
| 1157 | Opaque Prefix Attribute        | ---     | Section 4.3.3.6 |

Table 12: Summary Table of TLV/Sub-TLV Code Points

## 9. Security Considerations

Procedures and protocol extensions defined in this document do not affect the BGP security model. See the Security Considerations section of [RFC4271] for a discussion of BGP security. Also refer to [RFC4272] and [RFC6952] for analysis of security issues for BGP.

In the context of the BGP peerings associated with this document, a BGP speaker MUST NOT accept updates from a peer that is only



providing information to a BGP-LS Consumer. That is, a participating BGP speaker should be aware of the nature of its relationships for link-state relationships and should protect itself from peers sending updates that either represent erroneous information feedback loops or are false input. Such protection can be achieved by manual configuration of consumer peers at the BGP speaker.

An operator SHOULD employ a mechanism to protect a BGP speaker against DDoS attacks from BGP-LS Consumers. The principal attack a consumer may apply is to attempt to start multiple sessions either sequentially or simultaneously. Protection can be applied by imposing rate limits.

Additionally, it may be considered that the export of link-state and TE information as described in this document constitutes a risk to confidentiality of mission-critical or commercially sensitive information about the network. BGP peerings are not automatic and require configuration; thus, it is the responsibility of the network operator to ensure that only trusted consumers are configured to receive such information.

#### 10. Contributors

We would like to thank Robert Varga for the significant contribution he gave to RFC7752.

#### 11. Acknowledgements

This document update to the BGP-LS specification [RFC7752] is a result of feedback and inputs from the discussions in the IDR working group. It also incorporates certain details and clarifications based on implementation and deployment experience with BGP-LS.

Cengiz Alaettinoglu and Parag Amritkar brought forward the need to clarify the advertisement of LAN subnet for OSPF.

We would like to thank Balaji Rajagopalan, Srihari Sangli and Shraddha Hegde for their review and feedback on this document.

We would like to thank Nischal Sheth, Alia Atlas, David Ward, Derek Yeung, Murtuza Lightwala, John Scudder, Kaliraj Vairavakkalai, Les Ginsberg, Liem Nguyen, Manish Bhardwaj, Matt Miller, Mike Shand, Peter Psenak, Rex Fernando, Richard Woundy, Steven Luong, Tamas Mondal, Waqas Alam, Vipin Kumar, Naiming Shen, Carlos Pignataro, Balaji Rajagopalan, Yakov Rekhter, Alvaro Retana, Barry Leiba, and Ben Campbell for their comments on RFC7752.

## 12. References

### 12.1. Normative References

- [I-D.ietf-idr-bgp-extended-messages]  
Bush, R., Patel, K., and D. Ward, "Extended Message support for BGP", draft-ietf-idr-bgp-extended-messages-33 (work in progress), July 2019.
- [ISO10589]  
International Organization for Standardization, "Intermediate System to Intermediate System intra-domain routing information exchange protocol for use in conjunction with the protocol for providing the connectionless-mode network service (ISO 8473)", ISO/IEC 10589, November 2002.
- [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>>.
- [RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, DOI 10.17487/RFC2328, April 1998, <<https://www.rfc-editor.org/info/rfc2328>>.
- [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>>.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, DOI 10.17487/RFC3209, December 2001, <<https://www.rfc-editor.org/info/rfc3209>>.
- [RFC4202] Kompella, K., Ed. and Y. Rekhter, Ed., "Routing Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4202, DOI 10.17487/RFC4202, October 2005, <<https://www.rfc-editor.org/info/rfc4202>>.
- [RFC4203] Kompella, K., Ed. and Y. Rekhter, Ed., "OSPF Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4203, DOI 10.17487/RFC4203, October 2005, <<https://www.rfc-editor.org/info/rfc4203>>.

- [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>>.
- [RFC4915] Psenak, P., Mirtorabi, S., Roy, A., Nguyen, L., and P. Pillay-Esnault, "Multi-Topology (MT) Routing in OSPF", RFC 4915, DOI 10.17487/RFC4915, June 2007, <<https://www.rfc-editor.org/info/rfc4915>>.
- [RFC5036] Andersson, L., Ed., Minei, I., Ed., and B. Thomas, Ed., "LDP Specification", RFC 5036, DOI 10.17487/RFC5036, October 2007, <<https://www.rfc-editor.org/info/rfc5036>>.
- [RFC5120] Przygienda, T., Shen, N., and N. Sheth, "M-ISIS: Multi Topology (MT) Routing in Intermediate System to Intermediate Systems (IS-ISs)", RFC 5120, DOI 10.17487/RFC5120, February 2008, <<https://www.rfc-editor.org/info/rfc5120>>.
- [RFC5130] Previdi, S., Shand, M., Ed., and C. Martin, "A Policy Control Mechanism in IS-IS Using Administrative Tags", RFC 5130, DOI 10.17487/RFC5130, February 2008, <<https://www.rfc-editor.org/info/rfc5130>>.
- [RFC5301] McPherson, D. and N. Shen, "Dynamic Hostname Exchange Mechanism for IS-IS", RFC 5301, DOI 10.17487/RFC5301, October 2008, <<https://www.rfc-editor.org/info/rfc5301>>.
- [RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, DOI 10.17487/RFC5305, October 2008, <<https://www.rfc-editor.org/info/rfc5305>>.
- [RFC5307] Kompella, K., Ed. and Y. Rekhter, Ed., "IS-IS Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 5307, DOI 10.17487/RFC5307, October 2008, <<https://www.rfc-editor.org/info/rfc5307>>.
- [RFC5340] Coltun, R., Ferguson, D., Moy, J., and A. Lindem, "OSPF for IPv6", RFC 5340, DOI 10.17487/RFC5340, July 2008, <<https://www.rfc-editor.org/info/rfc5340>>.

- [RFC5642] Venkata, S., Harwani, S., Pignataro, C., and D. McPherson, "Dynamic Hostname Exchange Mechanism for OSPF", RFC 5642, DOI 10.17487/RFC5642, August 2009, <<https://www.rfc-editor.org/info/rfc5642>>.
- [RFC5890] Klensin, J., "Internationalized Domain Names for Applications (IDNA): Definitions and Document Framework", RFC 5890, DOI 10.17487/RFC5890, August 2010, <<https://www.rfc-editor.org/info/rfc5890>>.
- [RFC6119] Harrison, J., Berger, J., and M. Bartlett, "IPv6 Traffic Engineering in IS-IS", RFC 6119, DOI 10.17487/RFC6119, February 2011, <<https://www.rfc-editor.org/info/rfc6119>>.
- [RFC6549] Lindem, A., Roy, A., and S. Mirtorabi, "OSPFv2 Multi-Instance Extensions", RFC 6549, DOI 10.17487/RFC6549, March 2012, <<https://www.rfc-editor.org/info/rfc6549>>.
- [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>>.
- [RFC7752] Gredler, H., Ed., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and Traffic Engineering (TE) Information Using BGP", RFC 7752, DOI 10.17487/RFC7752, March 2016, <<https://www.rfc-editor.org/info/rfc7752>>.
- [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>>.
- [RFC8202] Ginsberg, L., Previdi, S., and W. Henderickx, "IS-IS Multi-Instance", RFC 8202, DOI 10.17487/RFC8202, June 2017, <<https://www.rfc-editor.org/info/rfc8202>>.

## 12.2. Informative References

- [RFC1918] Rekhter, Y., Moskowitz, B., Karrenberg, D., de Groot, G., and E. Lear, "Address Allocation for Private Internets", BCP 5, RFC 1918, DOI 10.17487/RFC1918, February 1996, <<https://www.rfc-editor.org/info/rfc1918>>.

- [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>>.
- [RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, DOI 10.17487/RFC4655, August 2006, <<https://www.rfc-editor.org/info/rfc4655>>.
- [RFC5073] Vasseur, J., Ed. and J. Le Roux, Ed., "IGP Routing Protocol Extensions for Discovery of Traffic Engineering Node Capabilities", RFC 5073, DOI 10.17487/RFC5073, December 2007, <<https://www.rfc-editor.org/info/rfc5073>>.
- [RFC5152] Vasseur, JP., Ed., Ayyangar, A., Ed., and R. Zhang, "A Per-Domain Path Computation Method for Establishing Inter-Domain Traffic Engineering (TE) Label Switched Paths (LSPs)", RFC 5152, DOI 10.17487/RFC5152, February 2008, <<https://www.rfc-editor.org/info/rfc5152>>.
- [RFC5316] Chen, M., Zhang, R., and X. Duan, "ISIS Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", RFC 5316, DOI 10.17487/RFC5316, December 2008, <<https://www.rfc-editor.org/info/rfc5316>>.
- [RFC5392] Chen, M., Zhang, R., and X. Duan, "OSPF Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", RFC 5392, DOI 10.17487/RFC5392, January 2009, <<https://www.rfc-editor.org/info/rfc5392>>.
- [RFC5693] Seedorf, J. and E. Burger, "Application-Layer Traffic Optimization (ALTO) Problem Statement", RFC 5693, DOI 10.17487/RFC5693, October 2009, <<https://www.rfc-editor.org/info/rfc5693>>.
- [RFC5706] Harrington, D., "Guidelines for Considering Operations and Management of New Protocols and Protocol Extensions", RFC 5706, DOI 10.17487/RFC5706, November 2009, <<https://www.rfc-editor.org/info/rfc5706>>.

- [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>>.
- [RFC7285] Alimi, R., Ed., Penno, R., Ed., Yang, Y., Ed., Kiesel, S., Previdi, S., Roome, W., Shalunov, S., and R. Woundy, "Application-Layer Traffic Optimization (ALTO) Protocol", RFC 7285, DOI 10.17487/RFC7285, September 2014, <<https://www.rfc-editor.org/info/rfc7285>>.
- [RFC7770] Lindem, A., Ed., Shen, N., Vasseur, JP., Aggarwal, R., and S. Shaffer, "Extensions to OSPF for Advertising Optional Router Capabilities", RFC 7770, DOI 10.17487/RFC7770, February 2016, <<https://www.rfc-editor.org/info/rfc7770>>.

#### Appendix A. Changes from RFC 7752

This section lists the high-level changes from RFC 7752 and provides reference to the document sections wherein those have been introduced.

1. Update the Figure 1 in Section 1 and added Section 3 to illustrate the different roles of a BGP implementation in conveying link-state information.
2. In Section 4.1, clarification about the TLV handling aspects that are applicable to both the NLRI and BGP-LS Attribute parts and those that are applicable only for the NLRI portion. An implementation may have missed the part about handling of unrecognized TLV and so, based on [RFC7606] guidelines, might discard the unknown NLRI types. This aspect is now unambiguously clarified in Section 4.2. Also, the ascending order of TLVs in the BGP-LS Attribute is not necessary.
3. Clarification of mandatory and optional TLVs in both NLRI and BGP-LS Attribute portions all through the document.
4. Handling of the growth of the BGP-LS Attribute is covered in Section 4.3.
5. Clarification on the use of Identifier field in the Link-State NLRI in Section 4.2 is provided. It was defined ambiguously to refer to only multi-instance IGP on a single link while it can also be used for multiple IGP protocol instances on a router. The IANA registry is accordingly being removed.

6.    The BGP-LS Identifier TLV in the Node Descriptors has been deprecated. Its use was not well specified by [RFC7752] and there has been some amount of confusion between implementators on its usage for identification of IGP domains as against the use of the Identifier doing the same functionality as the Instance-ID when running multiple instances of IGP routing protocols.
7.    Moved MT-ID TLV from the Node Descriptor section to under the Link Descriptor section since it is not a Node Descriptor sub-TLV. Also fixed the ambiguity in the encoding of OSPF MT-ID in this TLV. MT-ID TLV use is now elevated to SHOULD when it is enabled in the underlying IGP.
8.    Update the usage of OSPF Route Type TLV to mandate its use for OSPF prefixes in Section 4.2.3.1 since this is required for segregation of intra-area prefixes that are used to reach a node (e.g. a loopback) from other types of inter-area and external prefixes.
9.    Updated the Node Name TLV in Section 4.3.1.3 with the OSPF specification.
10.   Clarified the advertisement of the prefix corresponding to the LAN segment in an OSPF network in Section 4.9.
11.   Introduced Private Use TLV code point space and specified their encoding in Section 4.4.
12.   Introduced Section 4.7 where issues related to consistency of reporting IGP link-state along with their solutions are covered.
13.   Handling of large size of BGP-LS Attribute with growth in BGP-LS information is explained in Section 4.3 along with mitigation of errors arising out of it.
14.   Added recommendation for isolation of BGP-LS sessions from other BGP route exchange to avoid errors and faults in BGP-LS affecting the normal BGP routing.
15.   Updated the Fault Management section with detailed rules based on the role in the BGP-LS information propagation flow.

Authors' Addresses

Ketan Talaulikar (editor)  
Cisco Systems  
India

Email: ketant@cisco.com

Hannes Gredler  
Rtbrick

Email: hannes@rtbrick.com

Jan Medved  
Cisco Systems, Inc.  
170, West Tasman Drive  
San Jose, CA 95134  
US

Email: jmedved@cisco.com

Stefano Previdi  
Individual Contributor  
Rome  
Italy

Email: stefano@previdi.net

Adrian Farrel  
Old Dog Consulting

Email: adrian@olddog.co.uk

Saikat Ray  
Individual Contributor

Email: raysaikat@gmail.com



Interdomain Routing Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: February 9, 2020

C. Li  
Huawei Technologies  
H. Chen  
China Telecom  
M. Chen  
J. Dong  
Z. Li  
Huawei Technologies  
August 8, 2019

SR Policies Extensions for Path Segment and Bidirectional Path in BGP-LS  
draft-li-idr-bgp-ls-sr-policy-path-segment-03

## Abstract

This document specifies the way of collecting configuration and states of SR policies carrying Path Segment and bidirectional path information by using BGP-LS. Such information can be used by external components for many use cases such as performance measurement, path re-optimization and end-to-end protection.

## 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].

## Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

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

This Internet-Draft will expire on February 9, 2020.

## Copyright Notice

Copyright (c) 2019 IETF Trust and the persons identified as the document authors. All rights reserved.

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

## Table of Contents

|                                                       |    |
|-------------------------------------------------------|----|
| 1. Introduction . . . . .                             | 2  |
| 2. Terminology . . . . .                              | 3  |
| 3. Carrying SR Path Sub-TLVs in BGP-LS . . . . .      | 3  |
| 3.1. SR Path Segment Sub-TLV . . . . .                | 5  |
| 3.2. Sub-TLVs for Bidirectional Path . . . . .        | 6  |
| 3.2.1. SR Bidirectional Path Sub-TLV . . . . .        | 6  |
| 3.2.2. SR Reverse Path Segment List Sub-TLV . . . . . | 7  |
| 4. Operations . . . . .                               | 7  |
| 5. IANA Considerations . . . . .                      | 7  |
| 5.1. BGP-LS TLVs . . . . .                            | 7  |
| 5.2. BGP-LS SR Segment Descriptors . . . . .          | 8  |
| 6. Security Considerations . . . . .                  | 8  |
| 7. Acknowledgements . . . . .                         | 8  |
| 8. References . . . . .                               | 8  |
| 8.1. Normative References . . . . .                   | 8  |
| 8.2. Informative References . . . . .                 | 9  |
| Authors' Addresses . . . . .                          | 10 |

## 1. Introduction

Segment routing (SR) [RFC8402] is a source routing paradigm that allows the ingress node steers packets into a specific path according to the Segment Routing Policy [I-D.ietf-spring-segment-routing-policy].

However, the SR Policies defined in [I-D.ietf-spring-segment-routing-policy] only supports unidirectional SR paths and there is no path ID in a Segment List to identify an SR path. For identifying an SR path and supporting bidirectional path [I-D.ietf-spring-mpls-path-segment], new policies carrying Path Segment and bidirectional path information are defined in

[I-D.li-idr-sr-policy-path-segment-distribution], as well as the extensions to BGP to distribute new SR policies. The Path Segment can be a Path Segment in SR-MPLS [I-D.ietf-spring-mpls-path-segment], or other IDs that can identify a path.

In many network scenarios, the configuration and state of each TE Policy is required by a controller which allows the network operator to optimize several functions and operations through the use of a controller aware of both topology and state information [I-D.ietf-idr-te-lsp-distribution].

To collect the TE Policy information that is locally available in a router, [I-D.ietf-idr-te-lsp-distribution] describes a new mechanism by using BGP-LS update messages.

Based on the mechanism defined in [I-D.ietf-idr-te-lsp-distribution], this document describes a mechanism to distribute configuration and states of the new SR policies defined in [I-D.li-idr-sr-policy-path-segment-distribution] to external components using BGP-LS.

## 2. Terminology

This memo makes use of the terms defined in [RFC8402] and [I-D.ietf-idr-te-lsp-distribution].

## 3. Carrying SR Path Sub-TLVs in BGP-LS

A mechanism to collect states of SR Policies via BGP-LS is proposed by [I-D.ietf-idr-te-lsp-distribution]. The characteristics of an SR policy can be described by a TE Policy State TLV, which is carried in the optional non-transitive BGP Attribute "LINK\_STATE Attribute" defined in [RFC7752]. The TE Policy State TLV contains several sub-TLVs such as SR TE Policy sub-TLVs. Rather than replicating SR TE Policy sub-TLVs, [I-D.ietf-idr-te-lsp-distribution] reuses the equivalent sub-TLVs as defined in [I-D.ietf-idr-segment-routing-te-policy].

[I-D.li-idr-sr-policy-path-segment-distribution] defines the BGP extensions for Path Segment. The Path Segment can appear at both segment-list level and candidate path level upon the use case. The encoding is shown below.

```
SR Policy SAFI NLRI: <Distinguisher, Policy-Color, Endpoint>
Attributes:
 Tunnel Encaps Attribute (23)
 Tunnel Type: SR Policy
 Binding SID
 Preference
 Priority
 Policy Name
 Explicit NULL Label Policy (ENLP)
 Path Segment
 Segment List
 Weight
 Path Segment
 Segment
 Segment
 ...
 Segment List
 Weight
 Path Segment
 Segment
 Segment
 ...
 ...
```

Figure 1. Path Segment in SR policy

Also, [I-D.li-idr-sr-policy-path-segment-distribution] defines SR policy extensions for bidirectional SR path, the encoding is shown below:

```

SR Policy SAFI NLRI: <Distinguisher, Policy-Color, Endpoint>
 Attributes: Tunnel Encaps Attribute (23)
 Tunnel Type: SR Policy
 Binding SID
 Preference
 Priority
 Policy Name
 Explicit NULL Label Policy (ENLP)
 Bidirectional Path
 Segment List
 Weight
 Path Segment
 Segment
 Segment
 ...
 Reverse Segment List
 Weight
 Path Segment
 Segment
 Segment
 ...

```

Figure 2. SR policy for Bidirectional path

In order to collect configuration and states of unidirectional and bidirectional SR policies defined in [I-D.li-idr-sr-policy-path-segment-distribution], new sub-TLVs in SR TE Policy sub-TLVs should be defined. Likewise, rather than replicating SR Policy sub-TLVs, this document can reuse the equivalent sub-TLVs as defined in [I-D.li-idr-sr-policy-path-segment-distribution].

### 3.1. SR Path Segment Sub-TLV

This section reuses the SR Path Segment sub-TLV defined in [I-D.li-idr-sr-policy-path-segment-distribution] to describe a Path Segment , and it can be included in the Segment List sub-TLV as defined in [I-D.ietf-idr-te-lsp-distribution] . An SR Path Segment sub-TLV can be associated with an SR path specified by a Segment List sub-TLV, and it MUST appear only once within a Segment List sub-TLV. Also, it can be used for identifying an SR candidate path or an SR Policy defined in [I-D.ietf-spring-segment-routing-policy].

The format of Path Segment TLV is included below for reference.

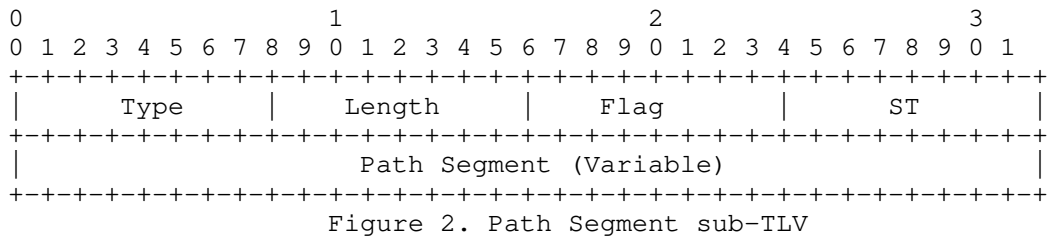


Figure 2. Path Segment sub-TLV

All fields, including type and length, are defined in [I-D.li-idr-sr-policy-path-segment-distribution].

### 3.2. Sub-TLVs for Bidirectional Path

In some scenarios like mobile backhaul transport network, there are requirements to support bidirectional path. In SR, a bidirectional path can be represented as a binding of two unidirectional SR paths [I-D.ietf-spring-mpls-path-segment].

[I-D.li-idr-sr-policy-path-segment-distribution] defines new sub-TLVs to describe an SR bidirectional path. An SRpolicy carrying SR bidirectional path information is expressed in Figure 1.

#### 3.2.1. SR Bidirectional Path Sub-TLV

This section reuses the SR bidirectional path sub-TLV defined in [I-D.li-idr-sr-policy-path-segment-distribution] to specify a bidirectional path, which contains a Segment List sub-TLV [I-D.ietf-idr-segment-routing-te-policy] and an associated Reverse Path Segment List as defined in [I-D.li-idr-sr-policy-path-segment-distribution]. The SR bidirectional path sub-TLV has the following format:

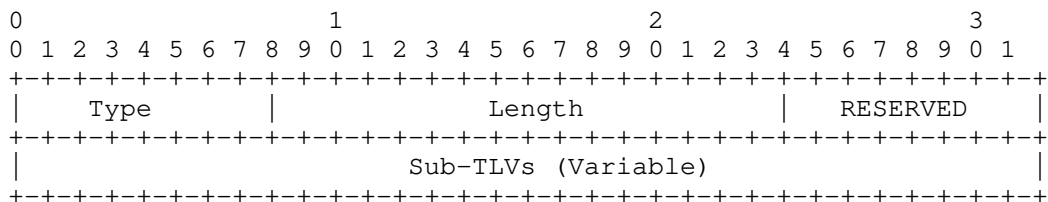


Figure 3. SR Bidirectional path sub-TLV

All fields, including type and length, are defined in [I-D.li-idr-sr-policy-path-segment-distribution].

### 3.2.2. SR Reverse Path Segment List Sub-TLV

This section reuses the SR Reverse Path Segment List sub-TLV defined in [I-D.li-idr-sr-policy-path-segment-distribution] to specify an reverse SR path associated with the path specified by the Segment List in the same SR Bidirectional Path Sub-TLV, and it has the following format:

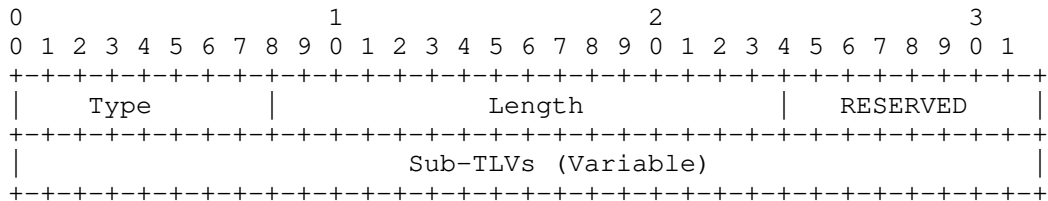


Figure 4. SR Reverse Path Segment List Sub-TLV

All fields, including type and length, are defined in [I-D.li-idr-sr-policy-path-segment-distribution].

## 4. Operations

No new operation procedures are defined in this document, the operations procedures of [RFC7752] can apply to this document.

Typically but not limited to, the uni/bidirectional SR policies carrying path identification information can be distributed by the ingress node.

Generally, BGP-LS is used for collecting link states and synchronizing with the external component. The consumer of the uni/bidirectional SR policies carrying path identification information is not BGP LS process by itself, and it can be any applications such as performance measurement [I-D.gandhi-spring-udp-pm] and path re-computation or re-optimization, etc. The operation of sending information to other precesses is out of scope of this document.

## 5. IANA Considerations

### 5.1. BGP-LS TLVs

IANA maintains a registry called "Border Gateway Protocol - Link State (BGP-LS) Parameters" with a sub-registry called "Node Anchor, Link Descriptor and Link Attribute TLVs". The following TLV codepoints are suggested (for early allocation by IANA):

| Codepoint | Description                   | Reference     |
|-----------|-------------------------------|---------------|
| 1212      | Path Segment sub-TLV          | This document |
| 1213      | SR Bidirectional Path sub-TLV | This document |
| 1214      | Reverse Segment List sub-TLV  | This document |

## 5.2. BGP-LS SR Segment Descriptors

This document defines new sub-TLVs in the registry "SR Segment Descriptor Types" [I-D.ietf-idr-te-lsp-distribution] to be assigned by IANA:

| Codepoint | Description          | Reference     |
|-----------|----------------------|---------------|
| 14        | Path Segment sub-TLV | This document |

## 6. Security Considerations

TBA

## 7. Acknowledgements

Many thanks to Shraddha Hedge for her detailed review and professional comments.

## 8. References

### 8.1. Normative References

- [I-D.ietf-idr-segment-routing-te-policy]  
 Previdi, S., Filsfils, C., Mattes, P., Rosen, E., Jain, D., and S. Lin, "Advertising Segment Routing Policies in BGP", draft-ietf-idr-segment-routing-te-policy-07 (work in progress), July 2019.
- [I-D.ietf-idr-te-lsp-distribution]  
 Previdi, S., Talaulikar, K., Dong, J., Chen, M., Gredler, H., and J. Tantsura, "Distribution of Traffic Engineering (TE) Policies and State using BGP-LS", draft-ietf-idr-te-lsp-distribution-11 (work in progress), May 2019.



- [I-D.ietf-spring-mpls-path-segment]  
Cheng, W., Li, H., Chen, M., Gandhi, R., and R. Zigler,  
"Path Segment in MPLS Based Segment Routing Network",  
draft-ietf-spring-mpls-path-segment-00 (work in progress),  
March 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.
- [I-D.li-idr-sr-policy-path-segment-distribution]  
Li, C., Chen, M., Dong, J., and Z. Li, "Segment Routing  
Policies for Path Segment and Bidirectional Path", draft-  
li-idr-sr-policy-path-segment-distribution-01 (work in  
progress), October 2018.
- [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>>.
- [RFC7752] Gredler, H., Ed., Medved, J., Previdi, S., Farrel, A., and  
S. Ray, "North-Bound Distribution of Link-State and  
Traffic Engineering (TE) Information Using BGP", RFC 7752,  
DOI 10.17487/RFC7752, March 2016,  
<<https://www.rfc-editor.org/info/rfc7752>>.
- [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>>.

## 8.2. Informative References

- [I-D.gandhi-spring-udp-pm]  
Gandhi, R., Filsfils, C., daniel.voyer@bell.ca, d.,  
Salsano, S., Ventre, P., and M. Chen, "UDP Path for In-  
band Performance Measurement for Segment Routing  
Networks", draft-gandhi-spring-udp-pm-02 (work in  
progress), September 2018.
- [I-D.ietf-mpls-bfd-directed]  
Mirsky, G., Tantsura, J., Varlashkin, I., and M. Chen,  
"Bidirectional Forwarding Detection (BFD) Directed Return  
Path", draft-ietf-mpls-bfd-directed-11 (work in progress),  
April 2019.

Authors' Addresses

Cheng Li  
Huawei Technologies  
Huawei Campus, No. 156 Beiqing Rd.  
Beijing 100095  
China

Email: chengli13@huawei.com

Huanan Chen  
China Telecom  
109 West Zhongshan Ave  
Guangzhou  
China

Email: chenhn8.gd@chinatelecom.cn

Mach(Guoyi) Chen  
Huawei Technologies  
Huawei Campus, No. 156 Beiqing Rd.  
Beijing 100095  
China

Email: Mach.chen@huawei.com

Jie Dong  
Huawei Technologies  
Huawei Campus, No. 156 Beiqing Rd.  
Beijing 100095  
China

Email: jie.dong@huawei.com

Zhenbin Li  
Huawei Technologies  
Huawei Campus, No. 156 Beiqing Rd.  
Beijing 100095  
China

Email: lizhenbin@huawei.com

Interdomain Routing Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: May 6, 2020

C. Li  
Huawei Technologies  
Y. Zhu  
China Telecom  
A. Sawaf  
Saudi Telecom Company  
Z. Li  
Huawei Technologies  
November 03, 2019

Segment Routing Path MTU in BGP  
draft-li-idr-sr-policy-path-mtu-03

Abstract

Segment Routing is a source routing paradigm that explicitly indicates the forwarding path for packets at the ingress node. An SR policy is a set of candidate SR paths consisting of one or more segment lists with necessary path attributes. However, the path maximum transmission unit (MTU) information for SR path is not available in the SR policy since the SR does not require signaling. This document defines extensions to BGP to distribute path MTU information within SR policies.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

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

This Internet-Draft will expire on May 6, 2020.

Copyright Notice

Copyright (c) 2019 IETF Trust and the persons identified as the document authors. All rights reserved.

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

## Table of Contents

|                                       |   |
|---------------------------------------|---|
| 1. Introduction . . . . .             | 2 |
| 2. Terminology . . . . .              | 3 |
| 2.1. Requirements Language . . . . .  | 3 |
| 3. SR Policy for Path MTU . . . . .   | 3 |
| 3.1. SR Path MTU Sub-TLV . . . . .    | 4 |
| 4. Operations . . . . .               | 5 |
| 5. IANA Considerations . . . . .      | 6 |
| 6. Security Considerations . . . . .  | 6 |
| 7. Contributors . . . . .             | 6 |
| 8. Acknowledgements . . . . .         | 6 |
| 9. References . . . . .               | 6 |
| 9.1. Normative References . . . . .   | 6 |
| 9.2. Informative References . . . . . | 7 |
| Authors' Addresses . . . . .          | 8 |

## 1. Introduction

Segment routing (SR) [RFC8402] is a source routing paradigm that explicitly indicates the forwarding path for packets at the ingress node. The ingress node steers packets into a specific path according to the Segment Routing Policy (SR Policy) as defined in [I-D.ietf-spring-segment-routing-policy]. In order to distribute SR policies to the headend, [I-D.ietf-idr-segment-routing-te-policy] specifies a mechanism by using BGP.

The maximum transmission unit (MTU) is the largest size packet or frame, in bytes, that can be sent in a network. An MTU that is too large might cause retransmissions. Too small an MTU might cause the router to send and handle relatively more header overhead and acknowledgments.

When an LSP is created across a set of links with different MTU sizes, the ingress router needs to know what the smallest MTU is on the LSP path. If this MTU is larger than the MTU of one of the intermediate links, traffic might be dropped, because MPLS packets cannot be fragmented. Also, the ingress router may not be aware of

this type of traffic loss, because the control plane for the LSP would still function normally. [RFC3209] specify the mechanism of MTU signaling in RSVP. Likewise, SRv6 packets will be dropped if the packet size is larger than path MTU, since IPv6 packet can not be fragmented on transmission [RFC8200] .

The host may discover the PMTU by Path MTU Discovery (PMTUD) [RFC8201] or other mechanisms. But the ingress still needs to examine the packet size for dropping too large packets to avoid malicious traffic or error traffic. Also, the packet size may exceeds the PMTU because of the new encapsulation of SR-MPLS or SRv6 packet at the ingress.

In order to check whether the Packet size exceeds the PMTU or not, the ingress node needs to know the Path MTU associated to the forwarding path. However, the path maximum transmission unit (MTU) information for SR path is not available since the SR does not require signaling.

This document defines extensions to BGP to distribute path MTU information within SR policies. The Link MTU information can be obtained via BGP-LS [I-D.zhu-idr-bgp-ls-path-mtu] or some other means. With the Link MTU, the controller can compute the PMTU and convey the information via the BGP SR policy.

## 2. Terminology

This memo makes use of the terms defined in [RFC8402] and [RFC3209].

### 2.1. Requirements Language

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

## 3. SR Policy for Path MTU

As defined in [I-D.ietf-idr-segment-routing-te-policy] , the SR policy encoding structure is as follows:

SR Policy SAFI NLRI: <Distinguisher, Policy-Color, Endpoint>  
Attributes:

- Tunnel Encaps Attribute (23)
  - Tunnel Type: SR Policy
  - Binding SID
  - Preference
  - Priority
  - Policy Name
  - Explicit NULL Label Policy (ENLP)
  - Segment List
    - Weight
    - Segment
    - Segment
    - ...
  - ...
- ...

As introduced in Section 1, each SR path has it's path MTU. SR policy with SR path MTU information is expressed as below:

SR Policy SAFI NLRI: <Distinguisher, Policy-Color, Endpoint>  
Attributes:

- Tunnel Encaps Attribute (23)
  - Tunnel Type: SR Policy
  - Binding SID
  - Preference
  - Priority
  - Policy Name
  - Explicit NULL Label Policy (ENLP)
  - Segment List
    - Weight
    - Path MTU
    - Segment
    - Segment
    - ...
  - ...
- ...

### 3.1. SR Path MTU Sub-TLV

An SR Path MTU sub-TLV is an Optional sub-TLV. When it appears, it must appear only once at most within a Segment List sub-TLV. If multiple Path MTU sub-TLVs appear within a Segment List sub-TLV, the first one will be processed, and the rest will be ignored. An SR Path MTU sub-TLV is associated with an SR path specified by a segment list sub-TLV or path segment as defined in [I-D.ietf-spring-mpls-path-segment] and [I-D.li-spring-srv6-path-segment]. It has the following format:

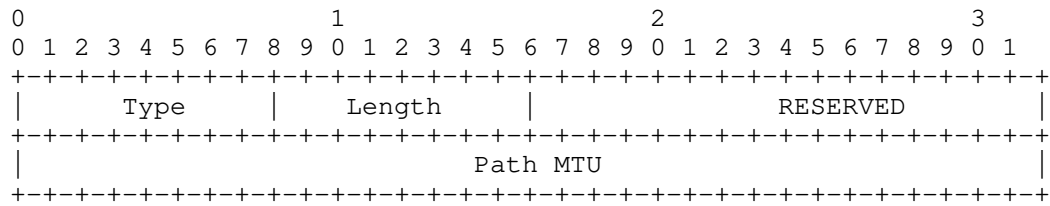


Figure 1. Path MTU sub-TLV

Where:

Type: to be assigned by IANA.

Length: the total length of the value field not including Type and Length fields.

Reserved: 16 bits reserved and MUST be set to 0 on transmission and MUST be ignored on receipt.

Path MTU: 4 bytes value of path MTU in octets. The value can be calculated by a central controller or other devices based on the information that learned via IGP of BGP-LS or other means.

Whenever the path MTU of a physical or logical interface is changed, a new SR policy with new path MTU information should be updated accordingly by BGP.

#### 4. Operations

The document does not bring new operation beyond the description of operations defined in [I-D.ietf-idr-segment-routing-te-policy]. The existing operations defined in [I-D.ietf-idr-segment-routing-te-policy] can apply to this document directly.

Typically but not limit to, the SR policies carrying path MTU information are configured by a controller.

After configuration, the SR policies carrying path MTU information will be advertised by BGP update messages. The operation of advertisement is the same as defined in [I-D.ietf-idr-segment-routing-te-policy], as well as the reception.

The consumer of the SR policies is not the BGP process. The operation of sending information to consumers is out of scope of this document.

## 5. IANA Considerations

This document defines a new Sub-TLV in registries "SR Policy List Sub- TLVs" [I-D.ietf-idr-segment-routing-te-policy]:

| Value | Description      | Reference     |
|-------|------------------|---------------|
| TBA   | Path MTU sub-TLV | This document |

## 6. Security Considerations

TBA

## 7. Contributors

Jun Qiu

Huawei Technologies

China

Email: qiu jun8@huawei.com

## 8. Acknowledgements

TBA

## 9. References

### 9.1. Normative References

[I-D.ietf-idr-segment-routing-te-policy]  
Previdi, S., Filsfils, C., Mattes, P., Rosen, E., Jain, D., and S. Lin, "Advertising Segment Routing Policies in BGP", draft-ietf-idr-segment-routing-te-policy-07 (work in progress), July 2019.

[I-D.ietf-spring-segment-routing-policy]  
Filsfils, C., Sivabalan, S., Voyer, D., Bogdanov, A., 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>>.
- [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>>.
- [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>>.

## 9.2. Informative References

- [I-D.ietf-spring-mpls-path-segment]  
Cheng, W., Li, H., Chen, M., Gandhi, R., and R. Zigler, "Path Segment in MPLS Based Segment Routing Network", draft-ietf-spring-mpls-path-segment-01 (work in progress), September 2019.
- [I-D.li-spring-srv6-path-segment]  
Li, C., Cheng, W., Chen, M., Dhody, D., Li, Z., Dong, J., and R. Gandhi, "Path Segment for SRv6 (Segment Routing in IPv6)", draft-li-spring-srv6-path-segment-03 (work in progress), August 2019.
- [I-D.zhu-idr-bgp-ls-path-mtu]  
Zhu, Y., Hu, Z., Yan, G., and J. Yao, "BGP-LS Extensions for Advertising Path MTU", draft-zhu-idr-bgp-ls-path-mtu-01 (work in progress), July 2019.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, DOI 10.17487/RFC3209, December 2001, <<https://www.rfc-editor.org/info/rfc3209>>.
- [RFC8200] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", STD 86, RFC 8200, DOI 10.17487/RFC8200, July 2017, <<https://www.rfc-editor.org/info/rfc8200>>.
- [RFC8201] McCann, J., Deering, S., Mogul, J., and R. Hinden, Ed., "Path MTU Discovery for IP version 6", STD 87, RFC 8201, DOI 10.17487/RFC8201, July 2017, <<https://www.rfc-editor.org/info/rfc8201>>.

Authors' Addresses

Cheng Li  
Huawei Technologies  
Huawei Campus, No. 156 Beiqing Rd.  
Beijing 100095  
China

Email: chenglil3@huawei.com

YongQing Zhu  
China Telecom  
109, West Zhongshan Road, Tianhe District.  
Guangzhou  
China

Email: zhuyq.gd@chinatelecom.cn

Ahmed El Sawaf  
Saudi Telecom Company  
Riyadh  
Saudi Arabia

Email: aelsawaf.c@stc.com.sa

Zhenbin Li  
Huawei Technologies  
Huawei Campus, No. 156 Beiqing Rd.  
Beijing 100095  
China

Email: lizhenbin@huawei.com

Interdomain Routing Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: February 9, 2020

C. Li  
Huawei Technologies  
H. Chen  
China Telecom  
M. Chen  
J. Dong  
Z. Li  
Huawei Technologies  
August 8, 2019

SR Policy Extensions for Path Segment and Bidirectional Path  
draft-li-idr-sr-policy-path-segment-01

Abstract

A Segment Routing (SR) policy is a set of candidate SR paths consisting of one or more segment lists with necessary path attributes. For each SR path, it may also have its own path attributes, and Path Segment is one of them. A Path Segment is defined to identify an SR path, which can be used for performance measurement, path correlation, and end-2-end path protection. Path Segment can be also used to correlate two unidirectional SR paths into a bidirectional SR path which is required in some scenarios, for example, mobile backhaul transport network.

This document defines extensions to BGP to distribute SR policies carrying Path Segment and bidirectional path information.

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].

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any

time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on February 9, 2020.

#### Copyright Notice

Copyright (c) 2019 IETF Trust and the persons identified as the document authors. All rights reserved.

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

#### Table of Contents

|                                                                                  |    |
|----------------------------------------------------------------------------------|----|
| 1. Introduction . . . . .                                                        | 2  |
| 2. Terminology . . . . .                                                         | 3  |
| 3. Path Segment in SR Policy . . . . .                                           | 3  |
| 3.1. SR Path Segment Sub-TLV . . . . .                                           | 4  |
| 4. SR Policy for Bidirectional Path . . . . .                                    | 5  |
| 4.1. SR Bidirectional Path Sub-TLV . . . . .                                     | 6  |
| 4.2. SR Reverse Path Segment List Sub-TLV . . . . .                              | 7  |
| 5. Operations . . . . .                                                          | 8  |
| 6. IANA Considerations . . . . .                                                 | 8  |
| 6.1. Existing Registry: BGP Tunnel Encapsulation Attribute<br>sub-TLVs . . . . . | 8  |
| 7. Security Considerations . . . . .                                             | 9  |
| 8. Acknowledgements . . . . .                                                    | 9  |
| 9. References . . . . .                                                          | 9  |
| 9.1. Normative References . . . . .                                              | 9  |
| 9.2. Informative References . . . . .                                            | 10 |
| Authors' Addresses . . . . .                                                     | 10 |

#### 1. Introduction

Segment routing (SR) [RFC8402] is a source routing paradigm that explicitly indicates the forwarding path for packets at the ingress node. The ingress node steers packets into a specific path according to the Segment Routing Policy (SR Policy) as defined in [I-D.ietf-spring-segment-routing-policy]. For distributing SR policies to the headend, [I-D.ietf-idr-segment-routing-te-policy]

specifies a mechanism by using BGP, and new sub-TLVs are defined for SR Policies in BGP UPDATE message.

In many use cases such as performance measurement, the path to which the packets belong is required to be identified. Furthermore, in some scenarios, for example, mobile backhaul transport network, there are requirements to support bidirectional path. However, there is no path identification information for each Segment List in the SR Policies defined in [I-D.ietf-spring-segment-routing-policy]. Also, the SR Policies defined in [I-D.ietf-spring-segment-routing-policy] only supports unidirectional SR paths.

Therefore, this document defines the extension to SR policies that carry Path Segment in the Segment List and support bidirectional path. The Path Segment can be a Path Segment in SR-MPLS [I-D.ietf-spring-mpls-path-segment], or other IDs that can identify a path. Also, this document defines extensions to BGP to distribute SR policies carrying Path Segment and bidirectional path information.

## 2. Terminology

This memo makes use of the terms defined in [RFC8402] and [I-D.ietf-idr-segment-routing-te-policy].

## 3. Path Segment in SR Policy

As defined in [I-D.ietf-idr-segment-routing-te-policy], the SR Policy encoding structure is as follows:

SR Policy SAFI NLRI: <Distinguisher, Policy-Color, Endpoint>  
Attributes:

- Tunnel Encaps Attribute (23)
  - Tunnel Type: SR Policy
    - Binding SID
    - Preference
    - Priority
    - Policy Name
    - Explicit NULL Label Policy (ENLP)
    - Segment List
      - Weight
      - Segment
      - Segment
      - ...
    - ...

An SR path can be specified by an Segment List sub-TLV that contains a set of segment sub-TLVs and other sub-TLVs as shown above. As

defined in [I-D.ietf-spring-segment-routing-policy], a candidate path includes multiple SR paths specified by SID list. The Path Segment can be used for identifying an SR path(specified by SID list). Also, it can be used for identifying an SR candidate path or an SR Policy in some use cases if needed. New SR Policy encoding structure is expressed as below:

SR Policy SAFI NLRI: <Distinguisher, Policy-Color, Endpoint>

Attributes:

Tunnel Encaps Attribute (23)

Tunnel Type: SR Policy

Binding SID

Preference

Priority

Policy Name

Explicit NULL Label Policy (ENLP)

Path Segment

Segment List

Weight

Path Segment

Segment

Segment

...

Segment List

Weight

Path Segment

Segment

Segment

...

...

The Path Segment can appear at both segment-list level and candidate path level, and generally it SHOULD also appear only at one level depending upon use case. Path segment at segment list level and at candidate path level may be same or may be different based on usecase and the ID allocation scope. When multiple Path Segments appear in both levels, it means the the Path Segment associated with candidate path and segment list SHOULD both be inserted into the SID list.

### 3.1. SR Path Segment Sub-TLV

This section defines an SR Path Segment sub-TLV.

An SR Path Segment sub-TLV can be included in the segment list sub-TLV to identify an SID list, and it MUST appear only once within a Segment List sub-TLV. It has the following format:

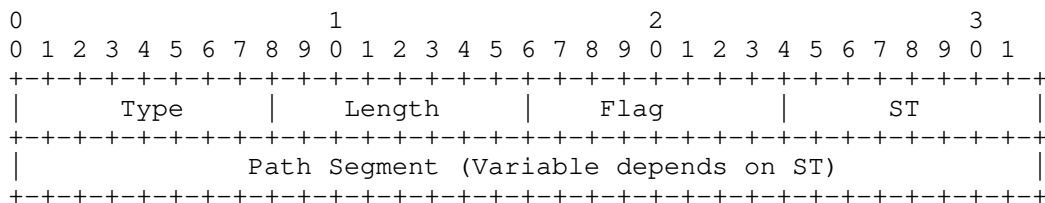


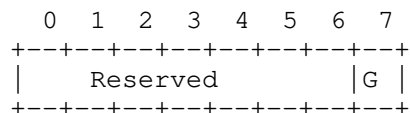
Figure 1. Path Segment sub-TLV

Where:

Type: to be assigned by IANA (suggested value 10).

Length: the total length of the value field not including Type and Length fields.

Flag: 8 bits of flags. Following flags are defined:



G-Flag: Global flag. Set when the Path Segment is global within an SR domain.

Reserved: 5 bits reserved and MUST be set to 0 on transmission and MUST be ignored on receipt.

ST: Segment type, specifies the type of the Path Segment, and it has following types:

- o 0: SR-MPLS Path Segment
- o 1-255:Reserved

**Path Segment:** The Path Segment of an SR path. The Path Segment type is indicated by the Segment Type(ST) field. It can be a Path Segment in SR-MPLS [I-D.ietf-spring-mpls-path-segment], which is 32-bits value, which is a 128-bits value, or other IDs that can identify a path.

#### 4. SR Policy for Bidirectional Path

In some scenarios, for example, mobile backhaul transport network, there are requirements to support bidirectional path. In SR, a bidirectional path can be represented as a binding of two unidirectional SR paths. This document also defines new sub-TLVs to

describe an SR bidirectional path. An SR policy carrying SR bidirectional path information is expressed as below:

```

SR Policy SAFI NLRI: <Distinguisher, Policy-Color, Endpoint>
 Attributes: Tunnel Encaps Attribute (23)
 Tunnel Type: SR Policy
 Binding SID
 Preference
 Priority
 Policy Name
 Explicit NULL Label Policy (ENLP)
 Bidirectionanl Path
 Segment List
 Weight
 Path Segment
 Segment
 Segment
 ...
 Reverse Segment List
 Weight
 Path Segment
 Segment
 Segment
 ...

```

#### 4.1. SR Bidirectional Path Sub-TLV

This section defines an SR bidirectional path sub-TLV to specify a bidirectional path, which contains a Segment List sub-TLV [I-D.ietf-idr-segment-routing-te-policy] and an associated Reverse Path Segment List as defined at section 4.2. The SR bidirectional path sub-TLV has the following format:

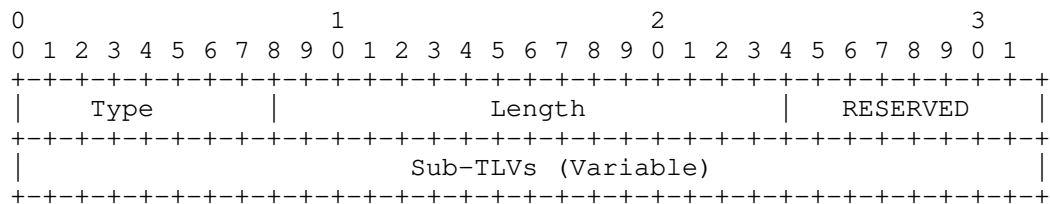


Figure 2. SR Bidirectional path sub-TLV

Where:

Type: TBA, and the suggest value is 14.



Length: the total length of the sub-TLVs encoded within the SR Bidirectional Path Sub-TLV not including Type and Length fields.

RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.

Sub-TLVs:

- o An Segment List sub-TLV
- o An associated Reverse Path Segment List sub-TLV

#### 4.2. SR Reverse Path Segment List Sub-TLV

An SR Reverse Path Segment List sub-TLV is defined to specify an SR reverse path associated with the path specified by the Segment List in the same SR Bidirectional Path Sub-TLV, and it has the following format:

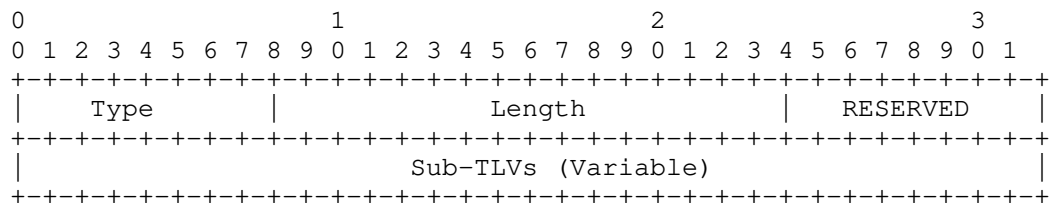


Figure 2. SR Reverse Path Segment List Sub-TLV

where:

Type: TBA, and suggest value is 127.

Length: the total length of the sub-TLVs encoded within the SR Reverse Path Segment List Sub-TLV not including the Type and Length fields.

RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.

sub-TLVs, reuse the sub-TLVs in Segment List defined in [I-D.ietf-idr-segment-routing-te-policy].

- o An optional single Weight sub-TLV.
- o An mandatory SR Path Segment sub-TLV that contains the Path Segment of the reverse SR path.
- o Zero or more Segment sub-TLVs to specify the reverse SR path.

The Segment sub-TLVs in the Reverse Path Segment List sub-TLV provides the information of the reverse SR path, which can be used for directing egress BFD peer to use specific path for the reverse direction of the BFD session [I-D.ietf-mppls-bfd-directed] or other applications.

## 5. Operations

The document does not bring new operation beyond the description of operations defined in [I-D.ietf-idr-segment-routing-te-policy]. The existing operations defined in [I-D.ietf-idr-segment-routing-te-policy] can apply to this document directly.

Typically but not limit to, the unidirectional or bidirectional SR policies carrying path identification information are configured by a controller.

After configuration, the unidirectional or bidirectional SR policies carrying path identification information will be advertised by BGP update messages. The operation of advertisement is the same as defined in [I-D.ietf-idr-segment-routing-te-policy], as well as the reception.

The consumer of the unidirectional or bidirectional SR policies is not the BGP process, it can be any applications, such as performance measurement [I-D.gandhi-spring-udp-pm]. The operation of sending information to consumers is out of scope of this document.

## 6. IANA Considerations

This document defines new Sub-TLVs in following registries:

### 6.1. Existing Registry: BGP Tunnel Encapsulation Attribute sub-TLVs

This document defines new sub-TLVs in the registry "BGP Tunnel Encapsulation Attribute sub-TLVs" to be assigned by IANA:

| Codepoint | Description                   | Reference     |
|-----------|-------------------------------|---------------|
| 14        | Path Segment sub-TLV          | This document |
| 15        | SR Bidirectional Path sub-TLV | This document |
| 127       | Reverse Segment List sub-TLV  | This document |

This document defines new sub-TLVs in the registry "SR Policy List Sub-TLVs" [I-D.ietf-idr-segment-routing-te-policy] to be assigned by IANA:

| Codepoint | Description          | Reference     |
|-----------|----------------------|---------------|
| 14        | Path Segment sub-TLV | This document |

## 7. Security Considerations

TBA

## 8. Acknowledgements

Many thanks to Shraddha Hedge for her detailed review and professional comments.

## 9. References

### 9.1. Normative References

- [I-D.ietf-idr-segment-routing-te-policy]  
Previdi, S., Filsfils, C., Mattes, P., Rosen, E., Jain, D., and S. Lin, "Advertising Segment Routing Policies in BGP", draft-ietf-idr-segment-routing-te-policy-07 (work in progress), July 2019.
- [I-D.ietf-spring-mpls-path-segment]  
Cheng, W., Li, H., Chen, M., Gandhi, R., and R. Zigler, "Path Segment in MPLS Based Segment Routing Network", draft-ietf-spring-mpls-path-segment-00 (work in progress), March 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>>.
- [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>>.

## 9.2. Informative References

[I-D.gandhi-spring-udp-pm]

Gandhi, R., Filsfils, C., daniel.voyer@bell.ca, d., Salsano, S., Ventre, P., and M. Chen, "UDP Path for In-band Performance Measurement for Segment Routing Networks", draft-gandhi-spring-udp-pm-02 (work in progress), September 2018.

[I-D.ietf-mpls-bfd-directed]

Mirsky, G., Tantsura, J., Varlashkin, I., and M. Chen, "Bidirectional Forwarding Detection (BFD) Directed Return Path", draft-ietf-mpls-bfd-directed-11 (work in progress), April 2019.

## Authors' Addresses

Cheng Li  
Huawei Technologies  
Huawei Campus, No. 156 Beiqing Rd.  
Beijing 100095  
China

Email: chengli13@huawei.com

Huanan Chen  
China Telecom  
109 West Zhongshan Ave  
Guangzhou  
China

Email: chenhn8.gd@chinatelecom.cn

Mach(Guoyi) Chen  
Huawei Technologies  
Huawei Campus, No. 156 Beiqing Rd.  
Beijing 100095  
China

Email: Mach.chen@huawei.com

Jie Dong  
Huawei Technologies  
Huawei Campus, No. 156 Beiqing Rd.  
Beijing 100095  
China

Email: jie.dong@huawei.com

Zhenbin Li  
Huawei Technologies  
Huawei Campus, No. 156 Beiqing Rd.  
Beijing 100095  
China

Email: lizhenbin@huawei.com

Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: April 6, 2021

Z. Li  
L. Li  
Huawei  
H. Chen  
Futurewei  
Y. Fan  
Casa Systems  
X. Liu  
Volta Networks  
L. Liu  
Fujitsu  
October 3, 2020

BGP Request for Candidate Paths of SR TE Policies  
draft-li-ldr-bgp-request-cp-sr-te-policy-02

Abstract

An SR Policy is a set of candidate paths. The headend of an SR Policy may learn multiple candidate paths for an SR Policy via a number of different mechanisms, e.g., CLI, NetConf, PCEP, or BGP. This document defines extensions to BGP for the headend to request BGP speaker (controller) for advertising the candidate paths.

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].

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

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

This Internet-Draft will expire on April 6, 2021.

## Copyright Notice

Copyright (c) 2020 IETF Trust and the persons identified as the document authors. All rights reserved.

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

## Table of Contents

|                                                      |    |
|------------------------------------------------------|----|
| 1. Introduction . . . . .                            | 2  |
| 2. Terminology . . . . .                             | 3  |
| 3. Overview of BGP Request for SR-TE Paths . . . . . | 3  |
| 4. BGP Request UPDATE Message . . . . .              | 4  |
| 4.1. Extension of SR Policy NLRI . . . . .           | 4  |
| 4.2. New SR Policy Sub-TLVs . . . . .                | 5  |
| 4.2.1. SR Path Attributes Sub-TLV . . . . .          | 5  |
| 4.2.2. Synchronization Sub-TLV . . . . .             | 6  |
| 4.2.3. Metric Sub-TLV . . . . .                      | 8  |
| 4.2.4. Include Route Sub-TLV . . . . .               | 9  |
| 4.2.5. Load Balance Sub-TLV . . . . .                | 10 |
| 4.2.6. Request Parameter Sub-TLV . . . . .           | 10 |
| 5. IANA . . . . .                                    | 11 |
| 6. Contributors . . . . .                            | 12 |
| 7. Acknowledgments . . . . .                         | 12 |
| 8. References . . . . .                              | 12 |
| 8.1. Normative References . . . . .                  | 12 |
| 8.2. Informative References . . . . .                | 13 |
| Authors' Addresses . . . . .                         | 14 |

## 1. Introduction

An SR Policy defined in [I-D.ietf-spring-segment-routing-policy] is a set of candidate paths. The headend of an SR Policy may be informed by various means including: Configuration, PCEP [RFC8281] or BGP [I-D.ietf-idr-segment-routing-te-policy]. All these mechanisms are Controller initiated, but in some situations the headend may want to pull a set of candidate paths from Controller rather than receive all information passively. Actually PCEP can use request and reply messages defined in [RFC5440] to match this requirement, but the

mechanism is not clear when controller advertises candidate paths via BGP.

This document defines a way to request controller (BGP speaker) to advertise candidate paths via BGP update messages. This makes BGP have the mechanism with request and reply similar to PCEP.

## 2. Terminology

RP: Request Parameters

LSPA: LSP Attributes

SVEC: Synchronization VEctor

IRO: Include Route Object

ERO: Explicit Route Object

MSD: Base MPLS Imposition Maximum SID Depth, as defined in [RFC8491]

NAI: Node or Adjacency Identifier

PCC: Path Computation Client

PCE: Path Computation Element

PCEP: Path Computation Element Communication Protocol

SID: Segment Identifier

SR: Segment Routing

SR-TE: Segment Routing Traffic Engineering

## 3. Overview of BGP Request for SR-TE Paths

[I-D.ietf-idr-segment-routing-te-policy] defines the extensions to BGP for a headend to receive candidate paths in a BGP UPDATE message from a controller (BGP speaker). In some situations a headend just wants to get these candidate paths when some special event occurs (for example, when it receives a customer route (VPN route) with a special color or special BGP attribute). This document defines the mechanism in which the headend requests the controller to advertise the expected SR policy with the candidate paths when this special situation occurs.



At first, the headend decides to get a new candidate path from the controller based on some trigger event. This trigger mechanism is out of scope of this document.

Then, the headend creates a new BGP request UPDATE message (defined below in this document) and sends it to the controller. The message contains the constraints/attributes of SR-TE paths such as affinity, metric, SRLG, and so on. This special request UPDATE message is called request message or request for short. It SHOULD NOT be used for BGP best path selection.

After receiving the request message, the controller will calculate one or a set of paths (i.e., segment lists) according to the request from the headend and advertise the SR Policy with the paths computed to the headend using [I-D.ietf-idr-segment-routing-te-policy]. How to calculate the paths is out of scope of this document.

#### 4. BGP Request UPDATE Message

A BGP request UPDATE message is based on the update message defined in [I-D.ietf-idr-segment-routing-te-policy] with some extensions described below.

##### 4.1. Extension of SR Policy NLRI

The SR Policy NLRI defined in [I-D.ietf-idr-segment-routing-te-policy] has the following format:

|        |               |                |
|--------|---------------|----------------|
| -----+ |               |                |
|        | NLRI Length   | 1 octet        |
| -----+ |               |                |
|        | Distinguisher | 4 octets       |
| -----+ |               |                |
|        | Policy Color  | 4 octets       |
| -----+ |               |                |
|        | Endpoint      | 4 or 16 octets |
| -----+ |               |                |

where:

- o NLRI Length: 1 octet of length expressed in bits as defined in [RFC4760].
- o Distinguisher: 4-octet value uniquely identifying the policy in the context of <color, endpoint> tuple. The distinguisher has no semantic value and is solely used by the SR Policy originator to make unique (from an NLRI perspective) multiple occurrences of the same SR Policy.

- o Policy Color: 4-octet value identifying (with the endpoint) the policy. The color is used to match the color of the destination prefixes to steer traffic into the SR Policy [I-D.ietf-spring-segment-routing-policy]
- o Endpoint: identifies the endpoint of a policy. The Endpoint may represent a single node or a set of nodes (e.g., an anycast address). The Endpoint is an IPv4 (4-octet) address or an IPv6 (16-octet) address according to the AFI of the NLRI.

NLRI Length, Policy Color, Endpoint field remains unchanged, while the Distinguisher field will be set to FF:FF:FF:FF to indicate that the UPDATE message with this NLRI is a request message to the controller.

#### 4.2. New SR Policy Sub-TLVs

The content of the SR Policy is encoded in the Tunnel Encapsulation Attribute TLV of type 23 defined in [I-D.ietf-idr-tunnel-encaps] containing a new Tunnel Type TLV of type 15. The SR Policy Encoding structure is as follows:

SR Policy SAFI NLRI: <Distinguisher, Policy-Color, Endpoint>

Attributes:

Tunnel Encaps Attribute (23)  
Tunnel Type (15): SR Policy  
    <Sub-TLVs>

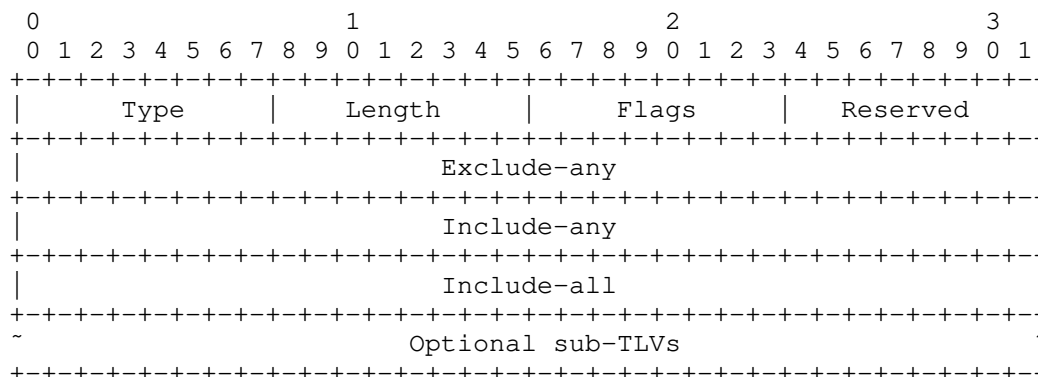
Preference, Binding SID, Priority, Policy Name, ENLP, Segment List, Weight and Segment Sub-TLVs are all defined in [I-D.ietf-idr-segment-routing-te-policy] for a SR Policy to be advertised to a headend.

Additional 6 new Sub-TLVs are defined below for the request mechanism. They are SR Path Attributes, Synchronization, Metric, Include Route, Load Balance, and Request Parameters Sub-TLVs.

##### 4.2.1. SR Path Attributes Sub-TLV

A SR Path Attributes Sub-TLV contains the attributes of the SR paths requested, which are similar to an LSP Attributes Object defined in [RFC5440] and [RFC3209].

It is optional and specifies various attributes or constraints of the paths requested. Its format is shown below.



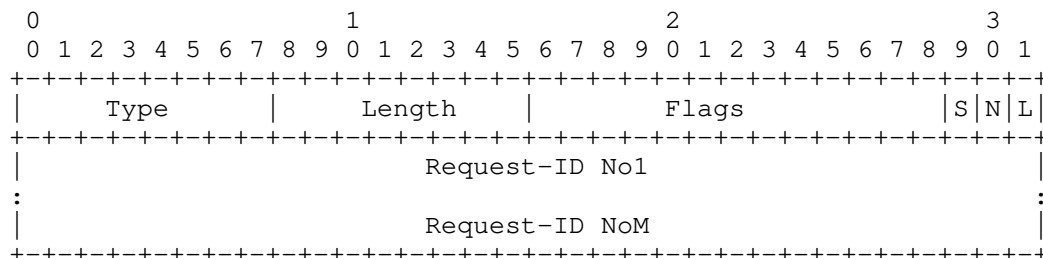
where:

- o Type: TBD1
- o Length: specifies the length of the value field not including Type and Length fields.
- o Flags (8 bits): No flag is currently defined. Undefined flags MUST be set to zero on transmission and be ignored on receipt.
- o Reserved (8 bits): This field MUST be set to zero on transmission and be ignored on receipt.
- o Exclude-any: A 32-bit vector representing a set of attribute filters associated with a path any of which renders a link unacceptable.
- o Include-any: A 32-bit vector representing a set of attribute filters associated with a path any of which renders a link acceptable (with respect to this test). A null set (all bits set to zero) automatically passes.
- o Include-all: A 32-bit vector representing a set of attribute filters associated with a path all of which must be present for a link to be acceptable (with respect to this test). A null set (all bits set to zero) automatically passes.
- o Optional sub-TLVs: No optional sub-TLV is currently defined.

#### 4.2.2. Synchronization Sub-TLV

A Synchronization Sub-TLV allows a headend to request the synchronization of a set of M dependent or independent SR path

requests. This TLV is similar to the SVEC Object defined in [RFC5440]. It is optional and has the following format.



where:

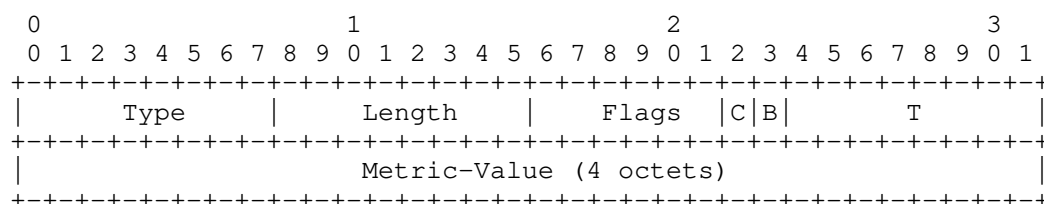
- o Type: TBD2
- o Length: specifies the length of the value field not including Type and Length fields.
- o Flags (16 bits): Defines the potential dependency among a set of SR paths (i.e., segment lists). Three flags are defined as follows:
  - \* L (Link diverse) bit: when set, it indicates that the computed SR paths (i.e., segment lists) MUST NOT have any link in common.
  - \* N (Node diverse) bit: when set, it indicates that the computed SR paths (i.e., segment lists) MUST NOT have any node in common.
  - \* S (SRLG diverse) bit: when set, it indicates that the computed SR paths (i.e., segment lists) MUST NOT share any SRLG (Shared Risk Link Group).
- o Request-ID No1, ..., NoM: each of which uniquely identifies one of M SR path requests.

In case of M synchronized independent path requests, the bits L, N, and S are set to zero.

Unassigned flags MUST be set to zero on transmission and MUST be ignored on receipt.

## 4.2.3. Metric Sub-TLV

A Metric Sub-TLV carries the same content as a Metric Object defined in [RFC5440] and [I-D.ietf-pce-segment-routing]. It has following format:



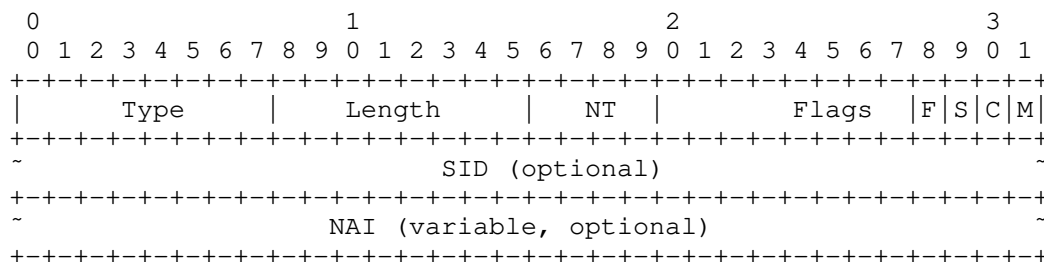
- o Type: TBD3.
- o Length: specifies the length of the value field not including Type and Length fields.
- o Flags (8 bits): Two flags are currently defined:
  - \* B (Bound - 1 bit): When set, the metric-value indicates a bound (a maximum) for the path metric that must not be exceeded for the headend to consider the computed path as acceptable. The path metric must be less than or equal to the value specified in the metric-value field. When the B flag is cleared, the metric-value field is not used to reflect a bound constraint.
  - \* C (Computed Metric - 1 bit): When set, it indicates that the controller MUST provide the computed path metric value (should a path satisfying the constraints be found) in the advertisement message for the corresponding metric.
  - \* Unassigned flags MUST be set to zero on transmission and MUST be ignored on receipt.
- o T (Type - 8 bits): Specifies the metric type. Four metric types are currently defined:
  - \* T=1: IGP metric
  - \* T=2: TE metric
  - \* T=3: Hop Counts
  - \* T=11: Maximum SID Depth of the requested path

- o Metric-Value (32 bits): It is a metric value encoded in 32 bits IEEE floating point format (see [IEEE.754.1985]).

#### 4.2.4. Include Route Sub-TLV

The Include Route Sub-TLV is optional and can be used to specify that the computed candidate path MUST traverse a set of specified network elements. The Include Route Sub-TLV carries the same content as IRO Object defined in [RFC5440], [RFC3209] and SR-ERO defined in [I-D.ietf-pce-segment-routing]

The Include Route Sub-TLV has following format:



Where:

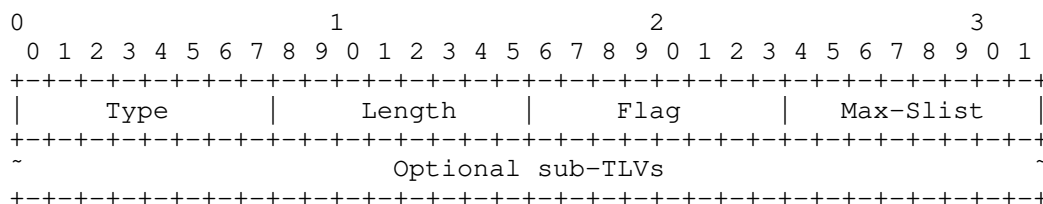
- o Type: TBD4.
- o Length: It specifies the length of the value field not including Type and Length fields.
- o NAI Type (NT): It indicates the type and format of the NAI contained, if any is present. If the F bit is set to zero, then the NT field has no meaning and MUST be ignored by the receiver. This document describes the following NT values:
  - \* NT=0: The NAI is absent.
  - \* NT=1: The NAI is an IPv4 node ID.
  - \* NT=2: The NAI is an IPv6 node ID.
  - \* NT=3: The NAI is an IPv4 adjacency.
  - \* NT=4: The NAI is an IPv6 adjacency with global IPv6 addresses.
  - \* NT=5: The NAI is an unnumbered adjacency with IPv4 node IDs.

- \* NT=6: The NAI is an IPv6 adjacency with link-local IPv6 addresses.

- o SID and NAI are the same as SR-ERO defined in [I-D.ietf-pce-segment-routing]

#### 4.2.5. Load Balance Sub-TLV

A Load Balance Sub-TLV specifies how many SR paths (i.e., segment lists) should be computed for a path request. It has following format:

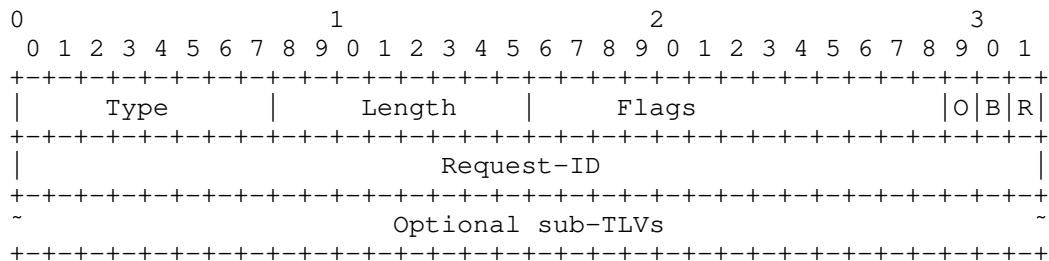


Where:

- o Type: TBD5.
- o Length: It specifies the length of the value field not including Type and Length fields.
- o Flags (8 bits): No flag is currently defined. The Flags field MUST be set to zero on transmission and MUST be ignored on receipt.
- o Max-Slist (8 bits): It indicates the maximum number of SR paths (i.e., segment lists) to be computed for the request. The load is distributed among these SR paths.
- o Optional sub-TLVs: No Optional sub-TLV is currently defined.

#### 4.2.6. Request Parameter Sub-TLV

A Request Parameter (RP) Sub-TLV specifies the request identifier and other parameters for a path request. It has the format below.



Where:

- o Type: TBD6.
- o Length: It specifies the length of the value field not including Type and Length fields.
- o Flags (16 bits): Three flag bits are currently defined as follows:
  - \* R (Reoptimization - 1 bit): when set, it indicates that the SR path request message is for the reoptimization of an existing SR path, which is represented by a segment list Sub-TLV in the message.
  - \* B (Bi-directional - 1 bit): when set, it indicates that the SR path request relates to bi-directional paths that has the same traffic engineering requirements including fate sharing, TE links, and other requirements (such as latency and jitter) in each direction.
  - \* O (strict/loose - 1 bit): when set, it indicates that a loose path is acceptable. Otherwise (i.e., when cleared), it indicates that a path exclusively made of strict hops is required.

## 5. IANA

Under Existing Registry Name: "BGP Tunnel Encapsulation Attribute Sub-TLVs", IANA is requested to assign new Sub-TLV values for SR Path Request as follows:



| Type | Value | Sub-TLV Name               | Reference     |
|------|-------|----------------------------|---------------|
| TBD1 |       | SR Path Attributes Sub-TLV | This document |
| TBD2 |       | Synchronization Sub-TLV    | This document |
| TBD3 |       | Metric Sub-TLV             | This document |
| TBD4 |       | Include Route Sub-TLV      | This document |
| TBD5 |       | Load Balance Sub-TLV       | This document |
| TBD6 |       | Request Parameters Sub-TLV | This document |

## 6. Contributors

TBD

## 7. Acknowledgments

TBD

## 8. References

### 8.1. Normative References

- [I-D.ietf-idr-segment-routing-te-policy]  
 Previdi, S., Filsfils, C., Talaulikar, K., Mattes, P.,  
 Rosen, E., Jain, D., and S. Lin, "Advertising Segment  
 Routing Policies in BGP", draft-ietf-idr-segment-routing-  
 te-policy-09 (work in progress), May 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-08 (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>>.

- [RFC5440] Vasseur, JP., Ed. and JL. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, DOI 10.17487/RFC5440, March 2009, <<https://www.rfc-editor.org/info/rfc5440>>.
- [RFC8281] Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "Path Computation Element Communication Protocol (PCEP) Extensions for PCE-Initiated LSP Setup in a Stateful PCE Model", RFC 8281, DOI 10.17487/RFC8281, December 2017, <<https://www.rfc-editor.org/info/rfc8281>>.

## 8.2. Informative References

- [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-19 (work in progress), September 2020.
- [I-D.ietf-pce-segment-routing]  
Sivabalan, S., Filsfils, C., Tantsura, J., Henderickx, W., and J. Hardwick, "PCEP Extensions for Segment Routing", draft-ietf-pce-segment-routing-16 (work in progress), March 2019.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, DOI 10.17487/RFC3209, December 2001, <<https://www.rfc-editor.org/info/rfc3209>>.
- [RFC4090] Pan, P., Ed., Swallow, G., Ed., and A. Atlas, Ed., "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, DOI 10.17487/RFC4090, May 2005, <<https://www.rfc-editor.org/info/rfc4090>>.
- [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>>.
- [RFC5420] Farrel, A., Ed., Papadimitriou, D., Vasseur, JP., and A. Ayyangar, "Encoding of Attributes for MPLS LSP Establishment Using Resource Reservation Protocol Traffic Engineering (RSVP-TE)", RFC 5420, DOI 10.17487/RFC5420, February 2009, <<https://www.rfc-editor.org/info/rfc5420>>.

Authors' Addresses

Zhenbin Li  
Huawei  
156 Beiqing Road  
Beijing, 100095  
P.R. China

Email: lizhenbin@huawei.com

Lei Li  
Huawei  
156 Beiqing Road  
Beijing, 100095  
P.R. China

Email: lily.lilei@huawei.com

Huaimo Chen  
Futurewei  
Boston, MA  
USA

Email: Huaimo.chen@futurewei.com

Yanhe Fan  
Casa Systems  
USA

Email: yfan@casa-systems.com

Xufeng Liu  
Volta Networks

McLean, VA  
USA

Email: xufeng.liu.ietf@gmail.com

Lei Liu  
Fujitsu

USA

Email: [liulei.kddi@gmail.com](mailto:liulei.kddi@gmail.com)

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

M. Zheng  
A. Lindem  
Cisco Systems  
J. Haas  
Juniper Networks, Inc.  
July 7, 2019

BGP BFD Strict-Mode  
draft-merciaz-idr-bgp-bfd-strict-mode-02

Abstract

This document specifies extensions to RFC4271 BGP-4 that enable a BGP speaker to negotiate additional Bidirectional Forwarding Detection (BFD) extensions using a BGP capability. This BFD capability enables a BGP speaker to prevent a BGP session from being established until a BFD session is established. It is referred to as BGP BFD "strict-mode". BGP BFD strict-mode will be supported when both the local speaker and its remote peer are BFD strict-mode capable.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

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

This Internet-Draft will expire on January 8, 2020.

Copyright Notice

Copyright (c) 2019 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect

to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

|                                           |   |
|-------------------------------------------|---|
| 1. Introduction . . . . .                 | 2 |
| 2. Requirements Language . . . . .        | 3 |
| 3. BFD Capability . . . . .               | 3 |
| 4. Operation . . . . .                    | 4 |
| 5. Manageability Considerations . . . . . | 5 |
| 6. Security Considerations . . . . .      | 5 |
| 7. IANA Considerations . . . . .          | 5 |
| 8. Acknowledgement . . . . .              | 5 |
| 9. Normative References . . . . .         | 6 |
| Authors' Addresses . . . . .              | 6 |

## 1. Introduction

Bidirectional Forwarding Detection BFD [RFC5882] enables routers to monitor data plane connectivity and to detect faults in the bidirectional forwarding path between them. This capability is leveraged by routing protocols such as BGP [RFC4271] to rapidly react to topology changes in the face of path failures.

The BFD interaction with BGP is specified in Section 10.2 of [RFC5882]. When BFD is enabled for a BGP neighbor, faults in the bidirectional forwarding detected by BFD result in session termination. It is possible in some failure scenarios for the network to be in a state such that a BGP session may be established but a BFD session cannot be established. In some other scenarios, it may be possible to establish a BGP session, but a degraded or poor-quality link may result in the corresponding BFD session going up and down frequently.

To avoid situations which result in routing churn and to minimize the impact of network interruptions, it will be beneficial to disallow BGP to establish a session until BFD session is successfully established and has stabilized. We refer to this mode of operation as BGP BFD "strict-mode". However, always using "strict-mode" would preclude BGP operation in an environment where not all routers support BFD strict-mode or have BFD enabled. This document defines BGP "strict-mode" operation as preventing BGP session establishment until both the local and remote speakers have a stable BFD session. The document also specifies the BGP protocol extensions for BGP capability [RFC5492] for announcing BFD parameters including a BGP

speaker's support for "strict-mode", i.e., requiring a BFD session for BGP session establishment.

## 2. Requirements Language

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

## 3. BFD Capability

The BGP Capability [RFC5492] for BFD parameters will allow a BGP speaker's BFD capabilities including its support for BFD strict-mode. This capability is defined as follows:

Capability code: TBD

Capability length: 1 octet

Capability value: Consists of 1 octet BFD flags as follows:

```
+-----+
| BFD Flags (8 bits) |
+-----+
```

The use and meaning of the fields are as follows:

BFD Flags: This field contains bit flags relating to BFD.

```
 0 1 2 3 4 5 6 7
+---+---+---+---+---+---+
| S | Reserved |
+---+---+---+---+---+---+
```

The most significant bit is defined as state of Strict-Mode ("Strict-Mode", or "S") bit, which can be used by a BGP speaker to signal its support for BFD Strict-mode. When set (value 1), this bit indicates that the BGP speaker has the BFD "Strict-mode" enabled. If both local BGP speaker and its peer have BFD strict-mode enabled, then BGP session establishment will be prevented until a BFD session is established between the peering addresses. A BGP speaker with BFD

strict-mode enabled MUST advertise the BFD capability with "S" bit set.

The remaining bits are reserved and SHOULD be set to zero by the sender and MUST be ignored by the receiver.

#### 4. Operation

A BGP speaker which supports capabilities advertisement and has BFD strict-mode enabled MUST include the BGP BFD capability with the "S" Bit set in the BGP capabilities it advertises.

A BGP speaker which supports BFD capability, examines the list of capabilities present in the Capabilities BFD Parameter that the speaker receives from its peer. If both the local and remote BGP speakers have BFD strict-mode enabled, the BGP finite state machine does not transition to the Established state from OpenSent or OpenConfirm state [RFC4271] until the BFD session is in the Up state (see below for AdminDown state). This means that a KEEPALIVE message is not sent nor is the KeepaliveTimer set.

If the BFD session does not transition to the Up state, and the HoldTimer has been negotiated to a non-zero value, the BGP FSM will close the session appropriately. If the HoldTimer has been negotiated to a zero value, the session should be closed after a time of X. This time X is referred as "BGP BFD Hold time". The proposed default BGP BFD Hold time value is 30 seconds. The BGP BFD Hold time value is configurable.

If BFD session is in the AdminDown state, then the BGP finite state machine will proceed normally without input from BFD. This means that BFD session "AdminDown" state WILL NOT prevent the BGP state transition to Established state from OpenConfirm.

Once the BFD session has transitioned to the Up state, the BGP FSM may proceed to transition to the Established state from the OpenSent or OpenConfirm state appropriately. I.e. a KEEPALIVE message is sent, and the KeepaliveTimer is started.

If either BGP peer has not advertised the BFD Capability with strict-mode enabled, then a BFD session WILL NOT be required for the BGP session to reach Established state. This does not preclude usage of BFD after BGP session establishment [RFC5882].



## 5. Manageability Considerations

Auto-configuration is possible for the enabling BGP BFD restrict-mode. However, the configuration automation is out of the scope of this document.

A BGP NOTIFICATION message subcode indicating BFD Hold timer expiration may be required for network management. (To be discussed in the next revision of this document.)

## 6. Security Considerations

The mechanism defined in this document interacts with the BGP finite state machine when so configured. The security considerations of BFD thus become considerations for BGP-4 [RFC4271] so used. The use of the BFD Authentication mechanism defined in [RFC5880] is thus RECOMMENDED when used to protect BGP-4 [RFC4271].

## 7. IANA Considerations

This document defines a new BGP capability - BFD Capability. The Capability Code for BFD Capability is TBD.

IANA is requested to establish a "BGP BFD Capability Flags" registry within the "Border Gateway Protocol (BGP) Parameters" grouping. The Registration Procedure should be Standards Action, the initial values as follows:

| Bit Position | Name        | Short Name | Reference     |
|--------------|-------------|------------|---------------|
| 0            | Strict-Mode | S          | this document |
| 1-7          | Unassigned  |            | this document |

## 8. Acknowledgement

The authors would like to acknowledge the review and inputs from Shyam Sethuram, Mohammed Mirza, Bruno Decraene, and Carlos Pignataro.

## 9. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [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>>.
- [RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, DOI 10.17487/RFC5492, February 2009, <<https://www.rfc-editor.org/info/rfc5492>>.
- [RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, DOI 10.17487/RFC5880, June 2010, <<https://www.rfc-editor.org/info/rfc5880>>.
- [RFC5882] Katz, D. and D. Ward, "Generic Application of Bidirectional Forwarding Detection (BFD)", RFC 5882, DOI 10.17487/RFC5882, June 2010, <<https://www.rfc-editor.org/info/rfc5882>>.
- [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>>.

## Authors' Addresses

Mercia Zheng  
Cisco Systems  
821 Alder Drive  
MILPITAS, CALIFORNIA 95035  
UNITED STATES

Email: [merciarz@cisco.com](mailto:merciarz@cisco.com)

Acee Lindem  
Cisco Systems  
301 Midenhall Way  
GARY, NC 27513  
UNITED STATES

Email: [acee@cisco.com](mailto:acee@cisco.com)

Jeffrey Haas  
Juniper Networks, Inc.  
1133 Innovation Way  
SUNNYVALE, CALIFORNIA 94089  
UNITED STATES

Email: [jhaas@juniper.net](mailto:jhaas@juniper.net)

BESS Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: April 28, 2022

A. Sajassi, Ed.  
A. Banerjee  
S. Thoria  
Cisco  
D. Carrel  
Graphiant  
B. Weis  
Independent  
J. Drake  
Juniper Networks  
October 25, 2021

Secure EVPN  
draft-sajassi-bess-secure-evpn-05

Abstract

The applications of EVPN-based solutions ([RFC7432] and [RFC8365]) have become pervasive in Data Center, Service Provider, and Enterprise segments. It is being used for fabric overlays and inter-site connectivity in the Data Center market segment, for Layer-2, Layer-3, and IRB VPN services in the Service Provider market segment, and for fabric overlay and WAN connectivity in Enterprise networks. For Data Center and Enterprise applications, there is a need to provide inter-site and WAN connectivity over public Internet in a secured manner with same level of privacy, integrity, and authentication for tenant's traffic as IPsec tunneling using IKEv2. This document presents a solution where BGP point-to-multipoint signaling is leveraged for key and policy exchange among PE devices to create private pair-wise IPsec Security Associations without IKEv2 point-to-point signaling or any other direct peer-to-peer session establishment messages.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

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

This Internet-Draft will expire on April 28, 2022.

## Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

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

## Table of Contents

|                                                                |    |
|----------------------------------------------------------------|----|
| 1. Introduction . . . . .                                      | 3  |
| 1.1. Requirements Language . . . . .                           | 5  |
| 2. Terminology . . . . .                                       | 5  |
| 3. Requirements . . . . .                                      | 7  |
| 3.1. Tenant's Layer-2 and Layer-3 data and control traffic . . | 7  |
| 3.2. Tenant's Unicast and Multicast Data Protection . . . . .  | 7  |
| 3.3. P2MP Signaling for SA setup and Maintenance . . . . .     | 7  |
| 3.4. Granularity of Security Association Tunnels . . . . .     | 7  |
| 3.5. Support for Policy and DH-Group List . . . . .            | 8  |
| 4. SA and Key Management . . . . .                             | 8  |
| 4.1. Generating Initial IPsec SAs . . . . .                    | 8  |
| 4.2. Rekey of IPsec SAs . . . . .                              | 10 |
| 4.2.1. Single IPsec Device Rekey . . . . .                     | 11 |
| 4.2.2. Multiple IPsec Device Rekey . . . . .                   | 13 |
| 5. IPsec Database Generation . . . . .                         | 16 |
| 5.1. The Security Policy Database (SPD) . . . . .              | 16 |
| 5.2. Security Association Database (SAD) . . . . .             | 16 |
| 5.2.1. Generating Keying Material for IPsec SAs . . . . .      | 16 |
| 5.2.1.1. g^ir . . . . .                                        | 16 |
| 5.2.1.2. Nonces . . . . .                                      | 17 |
| 5.2.1.3. SPIs . . . . .                                        | 17 |
| 5.2.1.4. IPsec key generation . . . . .                        | 18 |
| 5.3. Peer Authorization Database (PAD) . . . . .               | 19 |
| 6. Policy distributed through the BGP RR . . . . .             | 19 |
| 6.1. IPsec policy negotiation . . . . .                        | 20 |
| 7. BGP Component . . . . .                                     | 21 |
| 7.1. Zero Touch Bring-up (ZTB) . . . . .                       | 21 |
| 7.2. Configuration Management . . . . .                        | 21 |
| 7.3. Orchestration . . . . .                                   | 22 |

|                                                         |    |
|---------------------------------------------------------|----|
| 7.4. Signaling . . . . .                                | 22 |
| 8. Solution Description . . . . .                       | 22 |
| 8.1. Inheritance of Security Policies . . . . .         | 23 |
| 8.2. Distribution of Public Keys and Policies . . . . . | 24 |
| 8.2.1. Minimal DIM . . . . .                            | 24 |
| 8.2.2. Multiple Policies . . . . .                      | 25 |
| 8.2.3. Multiple DH-groups . . . . .                     | 25 |
| 8.2.4. Multiple or Single ESP SA policies . . . . .     | 25 |
| 8.3. Initial IPsec SAs Generation . . . . .             | 25 |
| 8.4. Re-Keying . . . . .                                | 26 |
| 8.5. IPsec Databases . . . . .                          | 26 |
| 9. Encapsulation . . . . .                              | 26 |
| 9.1. Standard ESP Encapsulation . . . . .               | 26 |
| 9.2. ESP Encapsulation within UDP packet . . . . .      | 27 |
| 10. BGP Encoding . . . . .                              | 28 |
| 10.1. The Base (Minimal Set) DIM Sub-TLV . . . . .      | 29 |
| 10.2. The Key Exchange Sub-TLV . . . . .                | 29 |
| 10.3. ESP SA Proposals Sub-TLV . . . . .                | 30 |
| 10.3.1. Transform Substructure . . . . .                | 30 |
| 11. Applicability . . . . .                             | 31 |
| 12. Acknowledgements . . . . .                          | 32 |
| 13. IANA Considerations . . . . .                       | 32 |
| 14. Security Considerations . . . . .                   | 32 |
| 15. References . . . . .                                | 33 |
| 15.1. Normative References . . . . .                    | 33 |
| 15.2. Informative References . . . . .                  | 34 |
| Appendix A. Additional Stuff . . . . .                  | 35 |
| Authors' Addresses . . . . .                            | 35 |

## 1. Introduction

The applications of EVPN-based solutions have become pervasive in Data Center, Service Provider, and Enterprise segments. It is being used for fabric overlays and inter-site connectivity in the Data Center market segment, for Layer-2, Layer-3, and IRB VPN services in the Service Provider market segment, and for fabric overlay and WAN connectivity in the Enterprise networks. For Data Center and Enterprise applications, there is a need to provide inter-site and WAN connectivity over public Internet in a secured manner with the same level of privacy, integrity, and authentication for tenant's traffic as used in IPsec tunneling using IKEv2. This document presents a solution where BGP point-to-multipoint signaling is leveraged for key and policy exchange among PE devices to create private pair-wise IPsec Security Associations without IKEv2 point-to-point signaling or any other direct peer-to-peer session establishment messages. This method is specially recommended for large scale deployment where large meshes of IKEv2 sessions among PE devices are not appropriate.

EVPN uses BGP as control-plane protocol for distribution of information needed for discovery of PEs participating in a VPN, discovery of PEs participating in a redundancy group, customer MAC addresses and IP prefixes/addresses, aliasing information, tunnel encapsulation types, multicast tunnel types, multicast group memberships, and other information. The advantages of using BGP control plane in EVPN are well understood including the following:

1. A full mesh of BGP sessions among PE devices can be avoided by using Route Reflector (RR) where a PE only needs to setup a single BGP session between itself and the RR as opposed to setting up N BGP sessions to N other remote PEs; therefore, reducing number of BGP sessions from  $O(N^2)$  to  $O(N)$  in the network. Furthermore, RR hierarchy can be leveraged to scale the number of BGP routes on the RR.
2. MP-BGP route filtering and constrained route distribution can be leveraged to ensure that the control-plane traffic for a given VPN is only distributed to the PEs participating in that VPN.

For setting up point-to-point security association (i.e., IPsec tunnel) between a pair of EVPN PEs, it is important to leverage BGP point-to-multipoint singling architecture using the RR along with its route filtering and constrain mechanisms to achieve the performance and the scale needed for large number of security associations (IPsec tunnels) along with their frequent re-keying requirements. Using BGP signaling along with the RR (instead of peer-to-peer protocol such as IKEv2) reduces number of message exchanges needed for SAs establishment and maintenance from  $O(N^2)$  to  $O(N)$  in the network.

Many key exchange methods (such as IKEv2) use a Diffie-Hellman (DH) algorithm to derive keys. When combined with an authentication method, the key exchange method allows two network devices to generate private pair-wise keys with each other. This document presents a key exchange method making use of the PE-to-RR trust model, where an RR is used to distribute keying material and policy between PE devices, also resulting in the PEs generating private pair-wise keys with each other. DH public values are provided to controllers from IPsec devices, where the controller relays the DH public values to authorized peers of that IPsec device as defined by a centralized policy. PE devices then create and install private pair-wise IPsec session keys to be used to secure communications with their peers.

Although IKEv2 is not used in this approach, the key management interfaces between IKEv2 and IPsec defined in RFC 7296 are maintained as much as possible.

### 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] RFC 8174 [RFC8174] when, and only when, they appear in all capitals, as shown here.

## 2. Terminology

- o AC: Attachment Circuit.
- o ARP: Address Resolution Protocol.
- o BD: Broadcast Domain. As per RFC7432 [RFC7432], an EVI consists of a single or multiple BDs. In case of VLAN-bundle and VLAN-based service models (see RFC7432 [RFC7432]), a BD is equivalent to an EVI. In case of VLAN-aware bundle service model, an EVI contains multiple BDs. Also, in this document, BD and subnet are equivalent terms.
- o BD Route Target: refers to the Broadcast Domain assigned Route Target RFC4364 [RFC4364]. In case of VLAN-aware bundle service model, all the BD instances in the MAC-VRF share the same Route Target.
- o BT: Bridge Table. The instantiation of a BD in a MAC-VRF, as per RFC7432 [RFC7432].
- o DGW: Data Center Gateway.
- o Ethernet A-D route: Ethernet Auto-Discovery (A-D) route, as per [RFC7432].
- o Ethernet NVO tunnel: refers to Network Virtualization Overlay tunnels with Ethernet payload. Examples of this type of tunnels are VXLAN or GENEVE [GENEVE].
- o EVI: EVPN Instance spanning the NVE/PE devices that are participating on that EVPN, as per [RFC7432].
- o EVPN: Ethernet Virtual Private Networks, as per [RFC7432].
- o GRE: Generic Routing Encapsulation.
- o GW IP: Gateway IP Address.
- o IPL: IP Prefix Length.



- o IP NVO tunnel: it refers to Network Virtualization Overlay tunnels with IP payload (no MAC header in the payload).
- o IP-VRF: A VPN Routing and Forwarding table for IP routes on an NVE/PE. The IP routes could be populated by EVPN and IP-VPN address families. An IP-VRF is also an instantiation of a layer 3 VPN in an NVE/PE.
- o IRB: Integrated Routing and Bridging interface. It connects an IP-VRF to a BD (or subnet).
- o MAC-VRF: A Virtual Routing and Forwarding table for Media Access Control (MAC) addresses on an NVE/PE, as per [RFC7432]. A MAC-VRF is also an instantiation of an EVI in an NVE/PE.
- o ML: MAC address length.
- o ND: Neighbor Discovery Protocol.
- o NVE: Network Virtualization Edge.
- o GENEVE: Generic Network Virtualization Encapsulation, [GENEVE].
- o NVO: Network Virtualization Overlays.
- o RT-2: EVPN route type 2, i.e., MAC/IP advertisement route, as defined in [RFC7432].
- o RT-5: EVPN route type 5, i.e., IP Prefix route. As defined in Section 3 of [EVPN-PREFIX].
- o SBD: Supplementary Broadcast Domain. A BD that does not have any ACs, only IRB interfaces, and it is used to provide connectivity among all the IP-VRFs of the tenant. The SBD is only required in IP-VRF- to-IP- VRF use-cases (see Section 4.4.).
- o SN: Subnet.
- o TS: Tenant System.
- o VA: Virtual Appliance.
- o VNI: Virtual Network Identifier. As in [RFC8365], the term is used as a representation of a 24-bit NVO instance identifier, with the understanding that VNI will refer to a VXLAN Network Identifier in VXLAN, or Virtual Network Identifier in GENEVE, etc. unless it is stated otherwise.

- o VTEP: VXLAN Termination End Point, as in RFC 7348 [RFC7348].
- o VXLAN: Virtual Extensible LAN, as in RFC 7348 [RFC7348].

This document also assumes familiarity with the terminology of [RFC7432], [RFC8365], and [RFC7365].

### 3. Requirements

The requirements for secured EVPN are captured in the following subsections.

#### 3.1. Tenant's Layer-2 and Layer-3 data and control traffic

Tenant's layer-2 and layer-3 data and control traffic must be protected by IPsec cryptographic methods. This implies not only tenant's data traffic must be protected by IPsec but also tenant's control and routing information that are advertised in BGP must also be protected by IPsec. This in turn implies that BGP session must be protected by IPsec.

#### 3.2. Tenant's Unicast and Multicast Data Protection

Tenant's layer-2 and layer-3 unicast traffic must be protected by IPsec. In addition to that, tenant's layer-2 broadcast, unknown unicast, and multicast traffic as well as tenant's layer-3 multicast traffic must be protected by IPsec when ingress replication or assisted replication are used. The use of BGP P2MP signaling for setting up P2MP SAs in P2MP multicast tunnels is for future study.

#### 3.3. P2MP Signaling for SA setup and Maintenance

BGP P2MP signaling must be used for IPsec SAs setup and maintenance. This reduces the number of message exchanges from  $O(N^2)$  to  $O(N)$  among the participating PE devices.

#### 3.4. Granularity of Security Association Tunnels

The solution must support the setup and maintenance of IPsec SAs at the following level of granularities:

- o Per PE: A single IPsec tunnel between a pair of PEs to be used for all tenants' traffic supported by the pair of PEs.
- o Per tenant: A single IPsec tunnel per tenant per pair of PEs. For example, if there are 1000 tenants supported on a pair of PEs, then 1000 IPsec tunnels are required between that pair of PEs.

- o Per subnet: A single IPsec tunnel per subnet (e.g., per VLAN/EVI) of a tenant on a pair of PEs.
- o Per L3 flow: A single IPsec tunnel per pair of IP addresses of a tenant on a pair of PEs.
- o Per L2 flow: A single IPsec tunnel per pair of MAC addresses of a tenant on a pair of PEs.
- o Per AC pair: A single IPsec tunnel per pair of Attachment Circuits between a pair of PEs.

### 3.5. Support for Policy and DH-Group List

The solution must support a single policy and DH group for all SAs as well as supporting multiple policies and DH groups among the SAs.

## 4. SA and Key Management

The BGP Route Reflector (RR) acts as a trusted third party, which relays policy and keying material between PE devices. Communications between the RR and the PEs MUST be authenticated, encrypted, and integrity-protected. All algorithms are selected by the management station associated with the RR. The combination of the RR and a set of PE devices comprises of a cooperating group of devices that make up a VPN, where each PE device is authorized to communicate with other PE devices in the group. Policies can allow a PE device to communicate with all other PE devices in the group, or may restrict it to a subset of those devices.

DH public values from each PE are distributed to other authorized peer PEs via the RR. Each PE device creates and maintains a DH pair, which it uses to communicate with other members of the VPN. This distribution of DH public values (and other related values) is intended to be embedded into the BGP protocol as described later. In particular, the RR provides a mechanism for secure key management. However, it does not provide policy information or configuration as that is assumed to be provided by the management station.

### 4.1. Generating Initial IPsec SAs

When an PE device (PE) begins operation, it generates a private/public DH pair, using an algorithm defined in the IKEv2 Diffie-Hellman Group Transform IDs [IKEV2-IANA]. If the device does not have any active peers it simply distributes its DH public value to the BGP RR, along with a nonce to be used during SA creation. Whenever a private/public DH pair is created, a new nonce MUST also be created. Whenever DH public values are transmitted, they are

transmitted with the corresponding nonce. Whenever a DH private or DH public value is used, it is used along with the corresponding nonce. However, in the diagrams and descriptions below, the nonces are often left out for the sake of clarity.

Upon receiving a peer's DH public value and nonce, the receiver creates IPsec SAs (as described in Section 5.2). For each peer, a pair of IPsec SAs are created by combining the PE device's own DH private value with the DH public number received from the Controller.

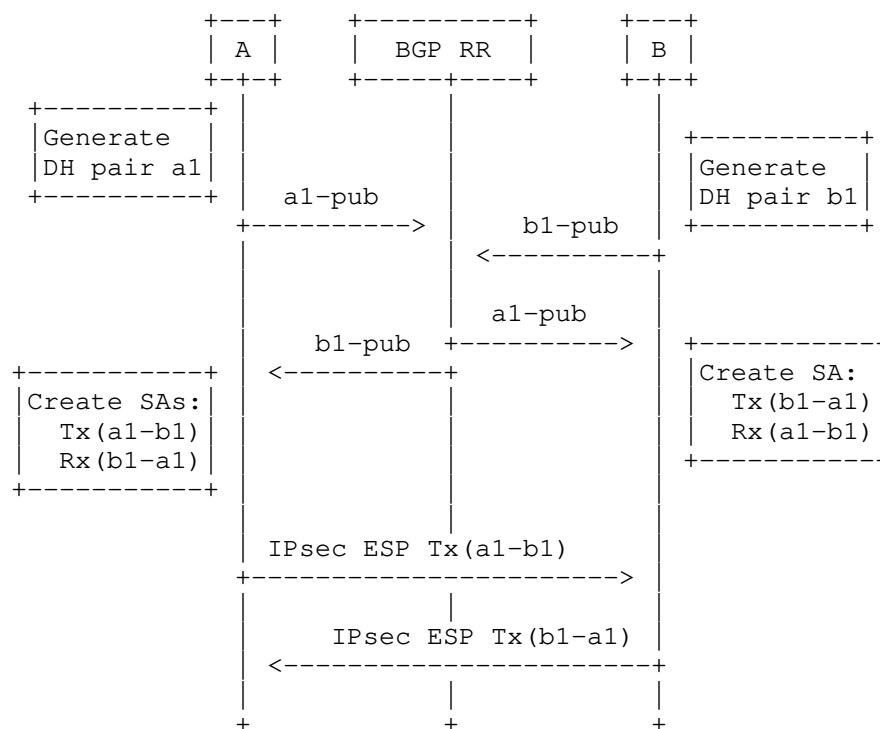


Figure 1: Generation of Initial IPsec SAs between two peers

Figure 1 shows IPsec SA generation between a pair of PE devices. Two PE devices (A and B shown in Figure 1) join the network. Each creates its own DH pair (labelled "a1" on A and "b1" on B), and distributes the DH public value (labelled a1-pub and b1-pub) to the BGP RR. The BGP RR forwards the DH public value to all authorized peers, although for simplicity of exposition the figure only shows the two IPsec devices.

When each device receives the peer's DH public value, a pair of IPsec SAs are generated: one outbound and one inbound. As shown in the

figure, A generates an outbound SA labeled Tx(a1-b1), representing that it has been generated using A's DH pair labeled a1 and B's DH pair labeled b1. B generates the same IPsec SA as an inbound SA, which is labeled Rx(a1-b1). Similarly, A generates an inbound IPsec SA labelled Rx(b1-a1), which is the same IPsec SA on B which is labelled Tx(b1-a1).

This process repeats on both A and B as they discover other PE devices with which they are authorized to communicate.

#### 4.2. Rekey of IPsec SAs

Any IPsec device may initiate a rekey at any time. Common reasons to perform a rekey include a local time or volume based policy, or may be the result of a cipher counter mode Initialization Vector (IV) counter nearing its final value. The rekey process is performed individually for each remote peer. If rekeying is performed with multiple peers simultaneously, then the decision process and rules described in this rekey are performed independently for each peer.

A decision process choosing an outbound IPsec SA is followed when certain events occur, as described in the rules below. The same decision process is followed regardless of whether the device is performing a rekey or responding to a peer's rekey. The decision process is:

1. Determine the outbound SAs with the remote peer's most recently distributed DH public value.
2. Determine which of those outbound SAs are "live". A "live" outbound SA is one built from a DH value from the local peer for which it has observed inbound traffic using any SA based on the same local DH pair. This proves that the remote peer is prepared to receive traffic protected by that DH pair.
3. Choose the "live" outbound SA built from the local peer's most recent DH public value.

A rekey operation follows these four basic rules.

Rule 1: When an IPsec device needs to perform a rekey with a remote peer, it creates a new pair of IPsec SAs by combining the new DH private value with the peer's DH public values. If the remote peer is also in the midst of a rollover and its DH public value has already been received, then this may result in creating two sets of SAs: one pair with the remote peer's old DH public value, and one pair with the remote peer's new DH public value.

- Rule 2: When an IPsec device receives a new remote peer's DH public value from the controller it creates and installs a new pair of IPsec SAs by combining the remote peer's new DH public value with its own current local DH private values. If both devices are in the midst of a rollover, this may result in creating two sets of SAs with the remote peer's new DH public value: one with the local old DH private value, and one with the local new DH private value. The outbound SA decision process is performed.
- Rule 3: The first IPsec packet received by a rekeying IPsec device on an inbound SA using its new DH pair causes it to perform the outbound SA decision process. It may also shorten the lifetime of IPsec SAs using its own old DH pair that are shared with this peer, as they are no longer in use (other than the inbound SA might receive packets in transit).
- Rule 4: The first IPsec packet received from a remote rekeying IPsec device using the remote peer's new DH pair allows the IPsec device to shorten the lifetime of IPsec SAs shared with this peer using unused remote DH pairs.

Two examples follow: a single IPsec device performing a rekey with its peers, and two IPsec devices performing a simultaneous rekey. The same rekey operations described above are exhibited in both cases.

#### 4.2.1. Single IPSec Device Rekey

When a single IPsec device begins a rekey, it first generates a new DH pair and generates new IPsec SA pairs for each peer with which it is communicating. It does this by combining the new DH private value with each peer's existing DH public value. Only when the new IPsec SAs have been installed and the device is prepared to receive on those new SAs does it then distribute the new DH public value to the Controller, which forwards the new DH public value to its authorized peers. The rekeying IPsec device continues to transmit on the old SAs for each peer until it observes that peer begin to transmit on the new SAs.

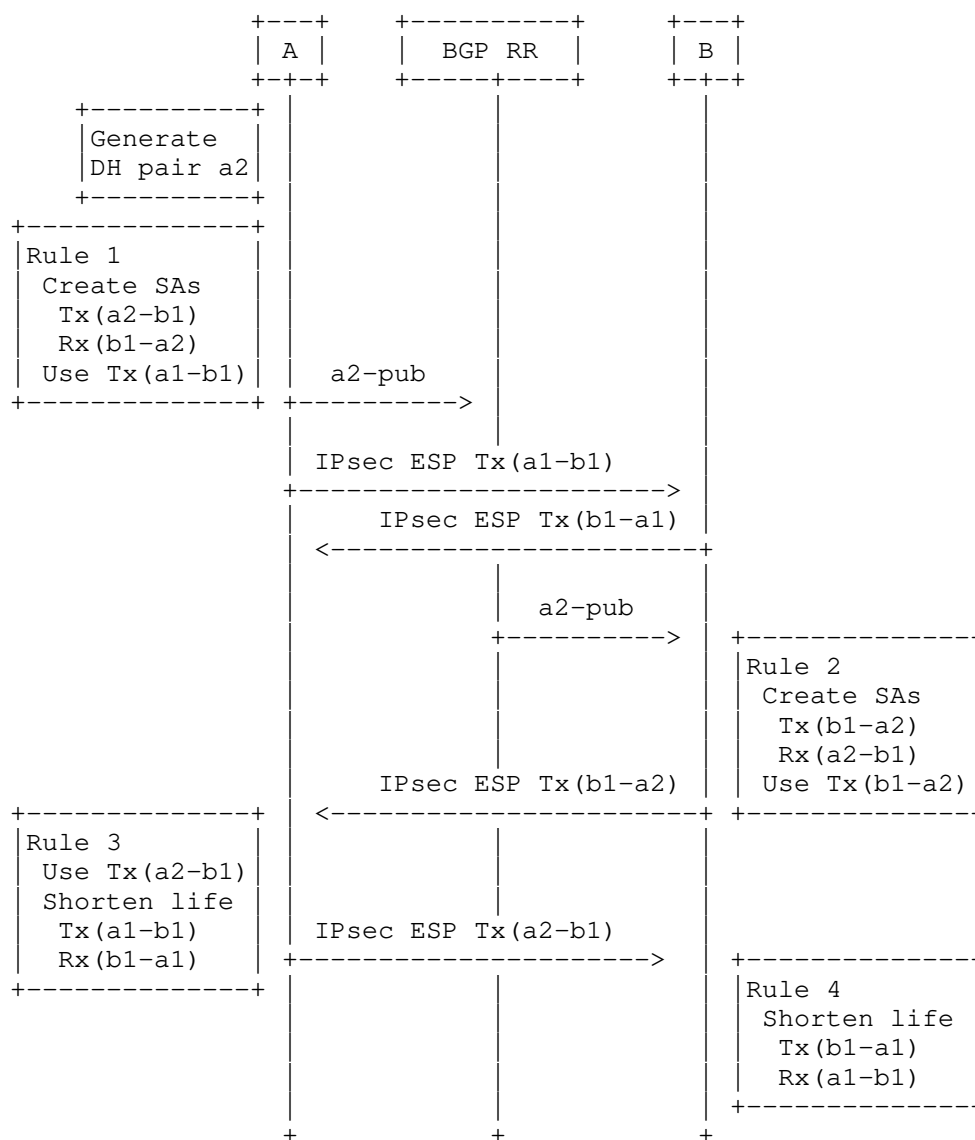


Figure 2: Single IPsec Device Rekey between two peers

In Figure 3, device A is shown as performing a rekey, and it creates a DH pair labelled "a2". The following steps are followed.

1. Rule 1 requires creating new IPsec SAs for each peer. In this example, A creates a new outbound IPsec SA to communicate with B labelled Tx(a2-b1), and a new inbound IPsec SA labelled

Rx(b1-a2). A continues to transmit on Tx(a1-b1) (generated as shown in Figure 2).

2. A distributes the new public value (a2-pub) to the Controller who forwards it to A's authorized peers, which includes B. During this time, both A and B continue to use the initial IPsec SAs setup between them using a1 and b1.
3. When B receives a2 from the controller, B follows Rule 2 by creating Tx(b1-a2), Rx(a2-b1). B also follows the outbound SA decision process, which causes it to change its outbound IPsec SA to A to Tx(b1-a2).
4. When A receives a packet protected by Rx(b1-a2), it follows Rule 3 and performs the outbound SA decision process. This causes it to change its outbound IPsec SA to Use Tx(a2-b1). It also optionally shortens the lifetime of the old IPsec SAs shared with this peer.
5. When B receives a packet protected by Tx(a2-b1), it follows Rule 4, in which it may shorten the lifetime of the old IPsec SAs shared with this peer using DH pairs that are no longer in use.

At the end of the rekey, both A and B retain a single DH pair, and a single set of IPsec SAs between them.

#### 4.2.2. Multiple IPsec Device Rekey

When two or more IPsec device simultaneously begin a rekey, they each follow the rekeying method described in the previous section. Every rekeying IPsec device generates a new DH pair and generates new IPsec SA pairs for each peer with which it is communicating by combining their new DH private value with each peer's existing DH public value. When this completes on a particular IPsec device, it distributes the new DH public value to the Controller, which forwards it to its authorized peers. Each continues to transmit on the existing SAs for each peer until it observes that peer transmitting on the new SAs. During a simultaneous rekey up to four pairs of IPsec SAs may be temporarily created, but the four rules ensure that they converge on a single new set of IPsec SAs.



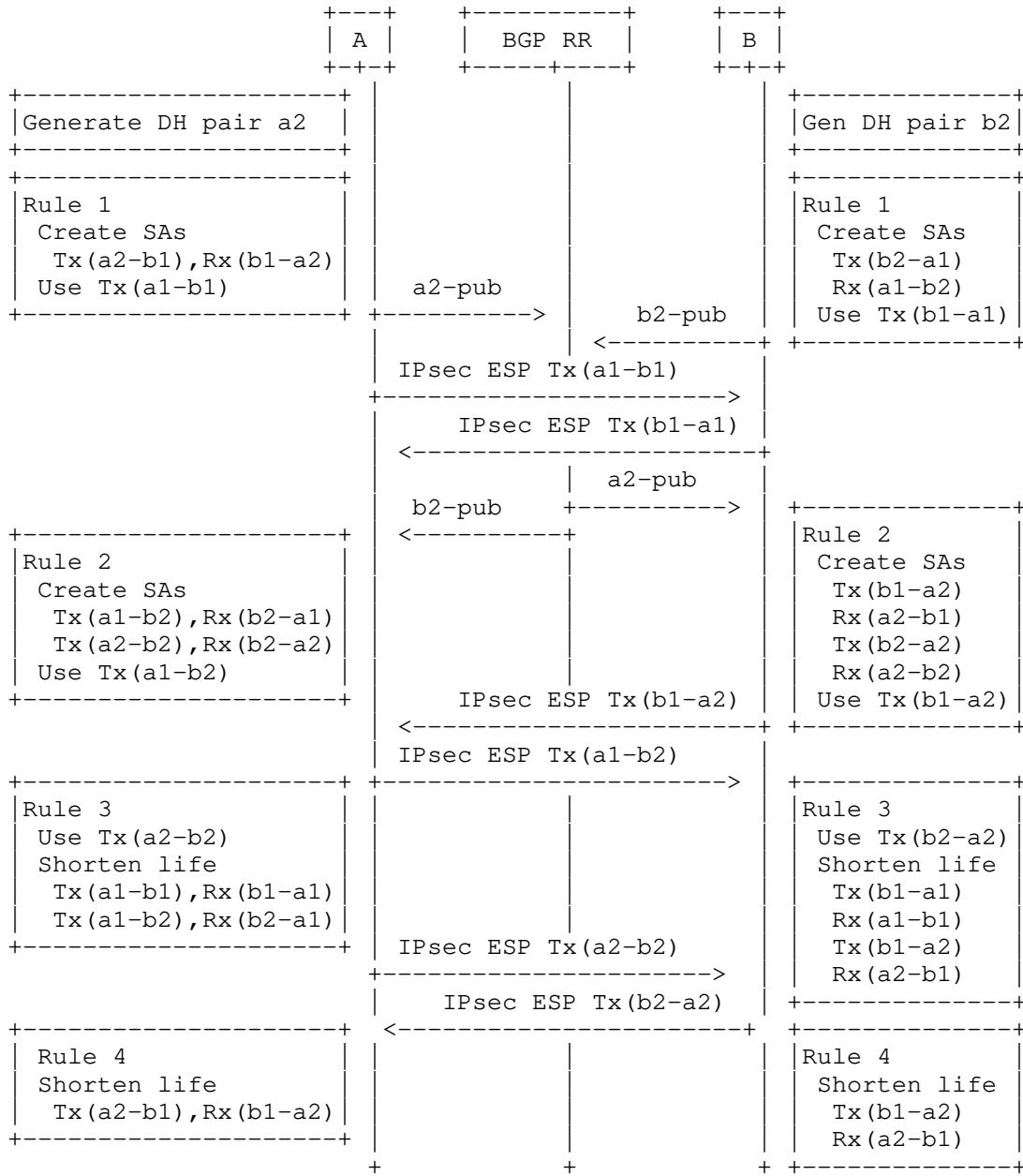


Figure 3: Simultaneous IPsec Device Rekey between two peers

In Figure 4, device A and device B are both shown as performing a rekey. Their initial state corresponds to the final state shown in

Figure 2 (i.e., they are communicating using a single pair of IPsec SAs created from DH pairs "a1" and "b1").

1. A and B follow Rule 1, which includes creating new IPsec SAs for each peer. In this example, A creates a new outbound IPsec SA to communicate with B labelled Tx(a2-b1), and a new inbound IPsec SA labelled Rx(b1-a2). B creates a new outbound IPsec SA to communicate with A labelled Tx(a1-b2), and a new inbound IPsec SA labelled Rx(b2-a1). A and B continue to transmit on IPsec SAs previously created from DH pairs "a1" and "b1".
2. A distributes the new public value (a2-pub) to the Controller who forwards it to A's authorized peers, which includes B. B also distributes the new public value (b2-pub) to the Controller who forwards it to B's authorized peers, which includes A.
3. When A and B receive each other's new peer DH public value from the controller they follow Rule 2. But because now there are four DH values that could be in use between A and B, they must be prepared to use IPsec SAs using each permutation of DH values: a1-b1, a1-b2, a2-b1, a2-b2. Prior to implementing Rule 2, each has already created sets of IPsec SAs matching two of the permutations, so just two more sets must be generated during Rule 2.
  - \* One pair is created using the IPsec device's old DH pair with the peer's new DH pair. This is necessary, because the peer may transmit on this pair.
  - \* One pair is created using the IPsec device's new DH pair with the peer's new DH pair. This is the set of IPsec SAs that will be used at the end of the rekey process.

Each peer begins transmitting on an IPsec SA that combines the remote peer's new DH pair and its own old DH pair, which is the most recent "live" SA on which it can transmit. I.e., A begins transmitting on Tx(a1-b2) and B begins transmitting on Tx(b1-a2).

4. When A receives a packet protected by Rx(b1-a2), it understands that the remote peer has received its new DH public value. A also understands that because of Rule 2 that B must have created IPsec SAs using a2-b2. This allows A to follow Rule 3 and change its outbound IPsec SA to Use Tx(a2-b2). Similarly, when B receives a packet protected by Rx(a1-b2), B recognizes that it can also begin to transmit using Tx(b2-a2). Note that it is also possible that A will receive a packet protected by Rx(b2-a2) or B will receive a packet protected by Rx(a2-b2), and then knows it can transmit on an IPsec SA using both of the new DH pairs.

5. Also in Rule 3, Both A and B optionally shorten the lifetime of older IPsec SAs shared with this peer derived from unused DH pairs to be cleaned up. A shortens the lifetime of SAs based on a1. B shortens the lifetime of SAs based on b1.
6. When A and B receive a packet protected by the remote peer's latest DH pair, they shortens the lifetime of SAs based on the remote peer's unused DH pair.

## 5. IPsec Database Generation

The PAD, SPD, and SAD all need to be setup as defined in the IPsec Security Architecture [RFC4301].

### 5.1. The Security Policy Database (SPD)

The SPD is implemented using methods outside the scope of this document. The SPD describes the type of traffic that will be protected between IPsec devices and the policy (e.g., ciphers) used to create SAs.

### 5.2. Security Association Database (SAD)

The SAD is constructed from IPsec policy (e.g., ciphers) obtained (depending on the controller protocol method) either from the controller or distributed by a peer (see Section 6).

Keying Material is generated following the method defined in IKEv2, and depends on SPIs, nonces, and the Diffie-Hellman shared secret.

The following sections describe how the necessary values are determined.

#### 5.2.1. Generating Keying Material for IPsec SAs

##### 5.2.1.1. $g^{ir}$

A DH public value is distributed from the peer.

A DH shared secret ( $g^{ir}$ ) is computed using the peer's public value, and the device's private value. The DH group to be used must be known by the device. Options include distribution by an SDN controller, or distribution by the peer with the DH public value (see Section 6).

#### 5.2.1.2. Nonces

Nonces are distributed with a DH public value, and are used only with that value. It is RECOMMENDED that nonces are generated as described in Section 2.10 of [RFC7296].

IKEv2 Key derivation specifies an initiator's nonce ( $N_i$ ) and a responder's nonce ( $N_r$ ). While neither peer is truly initiating a session, in order to fit the IKE key material models the roles must be assigned. The initiator is chosen as the peer with the larger nonce and the responder is the peer with the smaller. This does mean that the roles can change for each rekey and for each SA within a rekey.

#### 5.2.1.3. SPIs

SPI values that are unique to each generation of keying material need to be determined. While each peer could distribute its own inbound SA value, the SPI value would be used by many peers. Although this is not a problem for an SA lookup (lookup can include the source and destination IP addresses), experience has shown that this is sub-optimal for some hardware SA lookup algorithms. Instead, this specification proposes generating values that are unpredictable and indistinguishable from randomly-generated SPI values.

SPI values are generated using the IKEv2 prf+ function, where nonces are used as the input to the prf. This produces a statistically random SPI value that should be unique. However, with a 32 bit value there is still a very small, but non-zero, chance of SPIs repeating for a given pair of peers. To prevent this and ensure uniqueness in the operational window, we also use the lower 2 bits from each peer's rekey counter.

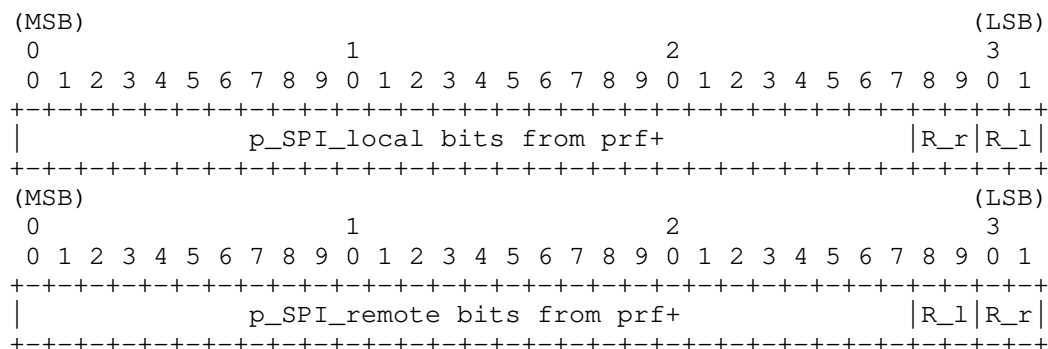
First the SPIs are taken from the prf+ function as 32 bit values and assigned based on which peer is taking the role of initiator and which is taking the role of responder. The  $p\_SPI\_i$  is taken by the device providing  $N_i$ , where  $p\_SPI\_r$  is taken by the other device.

$$\{p\_SPI\_i \mid p\_SPI\_r\} = \text{prf+}(N_i \mid N_r, \text{"SPI generation"})$$

Next  $p\_SPI\_i$  and  $p\_SPI\_r$  are mapped from initiator and responder roles to local and remote roles based on the choice of  $N_i$  and  $N_r$  in 5.2.1.2 and are renamed to  $p\_SPI\_local$  and  $p\_SPI\_remote$ .

Then, 2 2-bit Rotation Numbers (RN) are generated from the 2 least significant bits (LSB) of the 2 rekey counter values (see Section 6). These 4 bits replace the least significant bits of  $p\_SPI\_local$  and  $p\_SPI\_remote$  with the local RN bits taking the least significant

position in `p_SPI_local` and the remote RN bits taking the least significant position in `p_SPI_remote`. This shown in the following two diagrams with `RN_local` shown as `R_l` and `RN_remote` shown as `R_r`.



The reason for changing terminology from initiator/responder to local/remote is because the roles of initiator/responder can change in every rekey. The order of `RN_local` and `RN_remote` needs to remain constant. If that order was based on initiator/responder, there's a risk that if the initiator and responder roles changed and the two peers re-keyed on different frequencies, they could end up with identical RN values.

In some circumstances additional values may also need to be added to the `prf` for peers to ensure that they have implemented the same policy. Appendix A.3.1 includes a discussion of when this might be needed. In these cases, only the `prf+` inputs are modified and the Rotation Numbers MUST still be added as above.

Because a device is not choosing its inbound SPI, its SA lookup process needs to be aware that duplicates could occur across different peers. In that case, the inbound SA Lookup SHOULD include a source IP address in addition to the SPI value (see Section 4.1 of [RFC4301]).

#### 5.2.1.4. IPsec key generation

As described in previous sections, a DH public value and a nonce are distributed by peers. These are used to generate IPsec keys following the method defined in the IKEv2. SKEYSEED is generated following Section 2.14 of [RFC7296]:

$$\text{SKEYSEED} = \text{prf}(\text{Ni} \parallel \text{Nr}, g^{\text{air}})$$

KEYMAT can be similarly derived as defined by IKEv2 (Section 2.17 of [RFC7296]), although only SK\_d is required to be generated (shown in Section 2.14 of [RFC7296]).

$$\text{SK\_d} = \text{prf+} (\text{SKEYSEED}, \text{Ni} \mid \text{Nr} \mid \text{SPIi} \mid \text{SPIr})$$
$$\text{KEYMAT} = \text{prf+}(\text{SK\_d}, \text{Ni} \mid \text{Nr})$$

However, with the simplification where only SK\_d is generated, it can be observed that the derivation of SK\_d could be skipped entirely, and an optimized derivation of KEYMAT could be as follows:

$$\text{KEYMAT} = \text{prf+} (\text{SKEYSEED}, \text{Ni} \mid \text{Nr} \mid \text{SPIi} \mid \text{SPIr})$$

Note: A single specification for generating KEYMAT will be determined in a future version of this document.

### 5.3. Peer Authorization Database (PAD)

The PAD identifies authorized peers. PAD entries are either statically configured, or may be dynamically updated by the controller.

The PAD omits authentication data for each peer, because it has delegated authentication and authorization to the controller.

The controller protocol MUST be able to describe an identity that a receiver can match against its local PAD database, to ensure that the peer is an authorized peer.

## 6. Policy distributed through the BGP RR

An IPsec device distributes to a controller a DH public value and the associated information and policy needed to create IPsec SAs in a Device Information Message (DIM). The controller then distributes the DIM to all authorized peers of that device. The following data elements MUST be embedded in a DIM message:

- o DH public number (used for key computation)
- o Nonce (used for key computation and SPI generation)
- o Peer identity (used to identify a peer in the PAD)
- o An Indication whether this is the initial distributed policy
- o A rekey counter, which increases for each unique DIM sent

In cases where a single fixed IPsec policy has been pre-distributed, it is not necessary for the peer to send or receive that policy in a DIM. However, in cases where an IPsec device needs to indicate the policy it is willing to use, the following data elements SHOULD be included in a DIM:

- o An IPsec policy or policies
- o A lifetime bounding the use of the DH public number. When this DH public number is used to create an IPsec SA, the shortest lifetime is used as an SA lifetime for the pair of generated IPsec SAs. When the lifetime expires, the local version of the DIM and IPsec SAs generated from it MUST be deleted.

Appendix A suggests different ways that this policy may be included in a controller protocol. This document does not define a normative protocol format, because the DIM very likely needs to be integrated into an existing controller protocol rather than be an independent key management protocol. However, the controller protocol MUST provide a strong authentication between the device and the controller, and integrity of the messages MUST be provided. Confidentiality of the messages SHOULD also be provided. It is important that the controller protocol be protected with algorithms that are at least as strong as the algorithms used to protect the IPsec packets.

#### 6.1. IPsec policy negotiation

In many controller based networks, there is a single IPsec policy used by all devices and there is no need to redistribute or select policy details. However, in some circumstances, there may be a need to have multiple policy options. This could happen when a controller changes the policy and wants to smoothly migrate all devices to the new policy. Or it could happen if a network supports devices with different capabilities. In these cases, devices need to be able to choose the correct policy to use for each other device, and must do this without sending additional messages and without sending individual messages to each peer. When a device supports multiple policies, it MUST include those policies within the DIM. This is done by sending multiple distinct policies, in order of preference, where the first policy is the most preferred. The policy to use is selected by taking the receiver's list of policies (i.e., the list advertised by the device that generates  $N_r$ ), starting with the first policy, compare against the initiator's (device that generates  $N_i$ ) list, and choosing the first one found in common. The method conforms to the IKEv2 Cryptographic Algorithm Negotiation described in Section 2.7 of [RFC7296]. (However, see additional discussion when IKEv2 payloads are used in Appendix A.3.1).

If there is no match, this indicates a controller configuration error. These devices MUST NOT establish new SAs until a DIM is received that does produce a match.

When a device supports more than one DH group, then a unique DH public number MUST be specified for each in order of preference. The selection of which DH group to use follows the same logic as Policy selection, using the receiver's list order until a match is found in the initiator's list.

## 7. BGP Component

The architecture that encompasses device-to-controller trust model, has several components among which is the signaling component. Secure EVPN Signaling, as defined in this document, is the BGP signaling component of the overall Architecture. We will briefly describe this Architecture here to further facilitate understanding how Secure EVPN fits into the overall architecture. The Architecture describes the components needed to create BGP based SD-WANs and how these components work together. Our intention is to list these components here along with their brief description and to describe this Architecture in details in a separate document where to specify the details for other parts of this architecture besides the BGP signaling component which is described in this document.

The Architecture consists of four components. These components are Zero Touch Bring-up, Configuration Management, Orchestration, and Signaling. In addition to these components, secure communications must be provided between the edge nodes and all servers/devices providing the architecture components.

### 7.1. Zero Touch Bring-up (ZTB)

The first component is a zero touch capability that allows an edge device to find and join its SD-WAN with little to no assistance other than power and network connectivity. The goal is to use existing work in this area. The requirements are that an edge device can locate its ZTB server/component of its SD-WAN controller in a secure manner and to proceed to receive its configuration.

### 7.2. Configuration Management

After an edge device joins its SD-WAN, it needs to be configured. Configuration covers all device configuration, not just the configuration related to Secure EVPN. The previous Zero Touch Bring-up component will have directed the edge device, either directly or indirectly, to its configuration server/component. One example of a configuration server is the I2NSF Controller. After a device has



been configured, it can engage in the next two components. Configuration may include updates over time and is not a one time only component.

### 7.3. Orchestration

This component is optional. It allows for more dynamic updates of configuration and statistics information. Orchestration can be more dynamic than configuration.

### 7.4. Signaling

Signaling is the component described in this document. The functionality of a Route Reflector is well understood. Here we describe the signaling component of BGP SD-WAN Architecture and the BGP extension/signaling for IPsec key management and policy.

## 8. Solution Description

This solution uses BGP P2MP signaling where an originating PE only send a message to the Route Reflector (RR) and then the RR reflects that message to the interested recipient PEs. The framework for such signaling is described in section 4 and it is referred to as device-to-controller trust model. This trust model is significantly different than the traditional peer-to-peer trust model where a P2P signaling protocol such as IKEv2 [RFC7296] is used in which the PE devices directly authenticate each other and agree upon security policy and keying material to protect communications between themselves. The device-to-controller trust model leverages P2MP signaling via the controller (e.g., the RR) to achieve much better scale and performance for establishment and maintenance of large number of pair-wise Security Associations (SAs) among the PEs.

This device-to-controller trust model first secures the control channel between each device and the controller using peer-to-peer protocol such as IKEv2 [RFC7296] to establish P2P SAs between each PE and the RR. It then uses this secured control channel for P2MP signaling in establishment of P2P SAs between each pair of PE devices.

Each PE advertises to other PEs via the RR the information needed in establishment of pair-wise SAs between itself and every other remote PEs. These pieces of information are sent as Sub-TLVs of IPsec tunnel type in BGP Tunnel Encapsulation attribute. These Sub-TLVs are detailed in section 10 and are based on the DIM message components in section 5 and the IKEv2 specification [RFC7296]. The IPsec tunnel TLVs along with its Sub-TLVs are sent along with the BGP route (NLRI) for a given level of granularity.

If only a single SA is required per pair of PE devices to multiplex user traffic for all tenants, then IPsec tunnel TLV is advertised along with IPv4 or IPv6 NLRI representing loopback address of the originating PE. It should be noted that this is not a VPN route but rather an IPv4 or IPv6 route.

If a SA is required per tenant between a pair of PE devices, then IPsec tunnel TLV can be advertised along with EVPN IMET route representing the tenant or can be advertised along with a new EVPN route representing the tenant.

If a SA is required per tenant's subnet (e.g., per VLAN) between a pair of PE devices, then IPsec tunnel TLV is advertised along with EVPN IMET route.

If a SA is required between a pair of tenant's devices represented by a pair of IP addresses, then IPsec tunnel TLV is advertised along with EVPN IP Prefix Advertisement Route or EVPN MAC/IP Advertisement route.

If a SA is required between a pair of tenant's devices represented by a pair of MAC addresses, then IPsec tunnel TLV is advertised along with EVPN MAC/IP Advertisement route.

If a SA is required between a pair of Attachment Circuits (ACs) on two PE devices (where an AC can be represented by {VLAN, port}), then IPsec tunnel TLV is advertised along with EVPN Ethernet AD route.

### 8.1. Inheritance of Security Policies

Operationally, it is easy to configure a security association between a pair of PEs using BGP signaling. This is the default security association that is used for traffic that flows between peers. However, in the event more finer granularity of security association is desired on the traffic flows, it is possible to set up SAs between a pair of tenants, a pair of subnets within a tenant, a pair of IPs between a subnet, and a pair of MACs between a subnet using the appropriate EVPN routes as described above. In the event, there are no security TLVs associated with an EVPN route, there is a strict order in the manner security associations are inherited for such a route. This results in an EVPN route inheriting the security associations of the parent in a hierarchical fashion. For example, traffic between an IP pair is protected using security TLVs announced along with the EVPN IP Prefix Advertisement Route or EVPN MAC/IP Advertisement route as a first choice. If such TLVs are missing with the associated route, then one checks to see if the subnets the IPs are associated with has security TLVs with the EVPN IMET route. If they are present, those associations are used in securing the

traffic. In the absence of them, the peer security associations are used. The order in which security associations are inherited are from the granular to the coarser, namely, IP/MAC associated TLVs with the EVPN route being the first preference, and the subnet, the tenant, and the peer associations preferred in that fashion.

It should be noted that when a security association is made it is possible for it to be re-used by a large number of traffic flows. For example, a tenant security association may be associated with a number of child subnet routes. Clearly it is mandatory to keep a tenant security association alive, if there are one or more subnet routes that want to use that association. Logically, the security associations between a pair of entities creates a single secure tunnel. It is thus possible to classify the incoming traffic in the most granular sense {IP/MAC, subnet, tenant, peer} to a particular secure tunnel that falls within its route hierarchy. The policy that is applied to such traffic is independent from its use of an existing or a new secure tunnel. It is clear that since any number of classified traffic flows can use a security association, such a security association will not be torn down, if at least there is one policy using such a secure tunnel.

## 8.2. Distribution of Public Keys and Policies

One of the requirements for this solution is to support a single DH group and a single policy for all SAs as well as to support multiple DH groups and policies among the SAs. The following subsections describe what pieces of information (what Sub-TLVs) are needed to be exchanged to support a single DH group and a single policy versus multiple DH groups and multiple policies.

### 8.2.1. Minimal DIM

For SA establishment, at the minimum, a PE needs to advertise to other PEs, its DIM values as specified in section 5. These include:

|    |                                           |
|----|-------------------------------------------|
| ID | Tunnel ID                                 |
| N  | Nonce                                     |
| RC | Rekey Counter                             |
| I  | Indication of initial policy distribution |
| KE | DH public value.                          |

When this minimal set of DIM values is sent, then it is assumed that all peer PEs share the same policy for which DH group to use, as well as which IPsec SA policy to employ. Section 5.1 defines the Minimal DIM sub-TLV as part of IPsec tunnel TLV in BGP Tunnel Encapsulation Attribute.

### 8.2.2. Multiple Policies

There can be scenarios for which there is a need to have multiple policy options. This can happen when there is a need for policy change and smooth migration among all PE devices to the new policy is required. It can also happen if different PE devices have different capabilities within the network. In these scenarios, PE devices need to be able to choose the correct policy to use for each other. This multi-policy scheme is described in section 6. In order to support this multi-policy feature, a PE device MUST distribute a policy list. This list consists of multiple distinct policies in order of preference, where the first policy is the most preferred one. The receiving PE selects the policy by taking the received list (starting with the first policy) and comparing that against its own list and choosing the first one found in common. If there is no match, this indicates a configuration error and the PEs MUST NOT establish new SAs until a message is received that does produce a match.

### 8.2.3. Multiple DH-groups

It can be the case that not all peers use the same DH group. When multiple DH groups are supported, the peer may include multiple KE Sub-TLVs. The order of the KE Sub-TLVs determines the preference. The preference and selection methods are specified in section 6.

### 8.2.4. Multiple or Single ESP SA policies

In order to specify an ESP SA Policy, a DIM may include one or more SA Sub-TLVs. When all peers are configured by a controller with the same ESP SA policy, they MAY leave the SA out of the DIM. This minimizes messaging when group configuration is static and known. However, it may also be desirable to include the SA. If a single SA is included, the peer is indicating what ESP SA policy it uses, but is not willing to negotiate. If multiple SA Sub-TLVs are included, the peer is indicating that it is willing to negotiate. The order of the SA Sub-TLVs determines the preference. The preference and selection methods are specified in section 6.

### 8.3. Initial IPsec SAs Generation

The procedure for generation of initial IPsec SAs is described in section 4. This section gives a summary of it in context of BGP signaling. When a PE device first comes up and wants to setup an IPsec SA between itself and each of the interested remote PEs, it generates a DH pair along for each [what word here? "tenant"?] using an algorithm defined in the IKEv2 Diffie-Hellman Group Transform IDs [IKEv2-IANA]. The originating PE distributes the DH public value along with the other values in the DIM (using IPsec Tunnel TLV in

Tunnel Encapsulation Attribute) to other remote PEs via the RR. Each receiving PE uses this DH public number and the corresponding nonce in creation of IPsec SA pair to the originating PE - i.e., an outbound SA and an inbound SA. The detail procedures are described in Section 4.1.

#### 8.4. Re-Keying

A PE can initiate re-keying at any time due to local time or volume based policy or due to the result of cipher counter nearing its final value. The rekey process is performed individually for each remote PE. If rekeying is performed with multiple PEs simultaneously, then the decision process and rules described in this rekey are performed independently for each PE. Section 4.2 describes this rekeying process in details and gives examples for a single IPsec device (e.g., a single PE) rekey versus multiple PE devices rekey simultaneously.

#### 8.5. IPsec Databases

The Peer Authorization Database (PAD), the Security Policy Database (SPD), and the Security Association Database (SAD) all need to be setup as defined in the IPsec Security Architecture RFC 4301 [RFC4301]. Section 5 of this document gives a summary description of how these databases are setup where key is exchanged via P2MP signaling through the RR and the policy can be either signaled via the RR (in case of multiple policies) or configured by the management station (in case of single policy).

### 9. Encapsulation

Vast majority of Encapsulation for Network Virtualization Overlay (NVO) networks in deployment are based on UDP/IP with UDP destination port ID indicating the type of NVO encapsulation (e.g., VxLAN, GPE, GENEVE, GUE) and UDP source port ID representing flow entropy for load-balancing of the traffic within the fabric based on n-tuple that includes UDP header. When encrypting NVO encapsulated packets using IP Encapsulating Security Payload (ESP), the following two options can be used: a) adding a UDP header before ESP header (e.g., UDP header in clear) and b) no UDP header before ESP header (e.g., standard ESP encapsulation). The following subsection describe these encapsulation in further details.

#### 9.1. Standard ESP Encapsulation

When standard IP Encapsulating Security Payload (ESP) is used (without outer UDP header) for encryption of NVO packets, it is used in transport mode as depicted below. When such encapsulation is

used, for BGP signaling, the Tunnel Type of Tunnel Encapsulation TLV is set to ESP-Transport and the Tunnel Type of Encapsulation Extended Community is set to NVO encapsulation type (e.g., VxLAN, GENEVE, GPE, etc.). This implies that the customer packets are first encapsulated using NVO encapsulation type and then it is further encapsulated and encrypted using ESP-Transport mode.

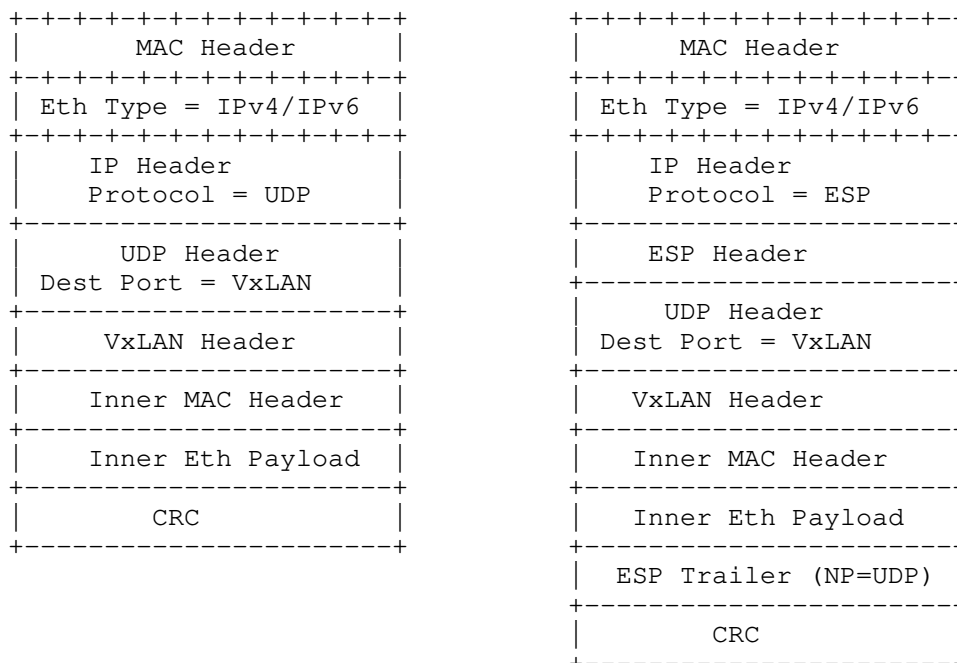


Figure 4

## 9.2. ESP Encapsulation within UDP packet

In scenarios where NAT traversal is required (RFC 3948 [RFC3948]) or where load balancing using UDP header is required, then ESP encapsulation within UDP packet as depicted in the following figure is used. The ESP for NVO applications is in transport mode. The outer UDP header (before the ESP header) has its source port set to flow entropy and its destination port set to 4500 (indicating ESP header follows). A non-zero SPI value in ESP header implies that this is a data packet (i.e., it is not an IKE packet). The Next Protocol field in the ESP trailer indicates what follows the ESP header, is a UDP header. This inner UDP header has a destination port ID that identifies NVO encapsulation type (e.g., VxLAN). Optimization of this packet format where only a single UDP header is used (only the outer UDP header) is for future study.

When such encapsulation is used, for BGP signaling, the Tunnel Type of Tunnel Encapsulation TLV is set to ESP-in-UDP-Transport and the Tunnel Type of Encapsulation Extended Community is set to NVO encapsulation type (e.g., VxLAN, GENEVE, GPE, etc.). This implies that the customer packets are first encapsulated using NVO encapsulation type and then it is further encapsulated and encrypted using ESP-in-UDP with Transport mode.

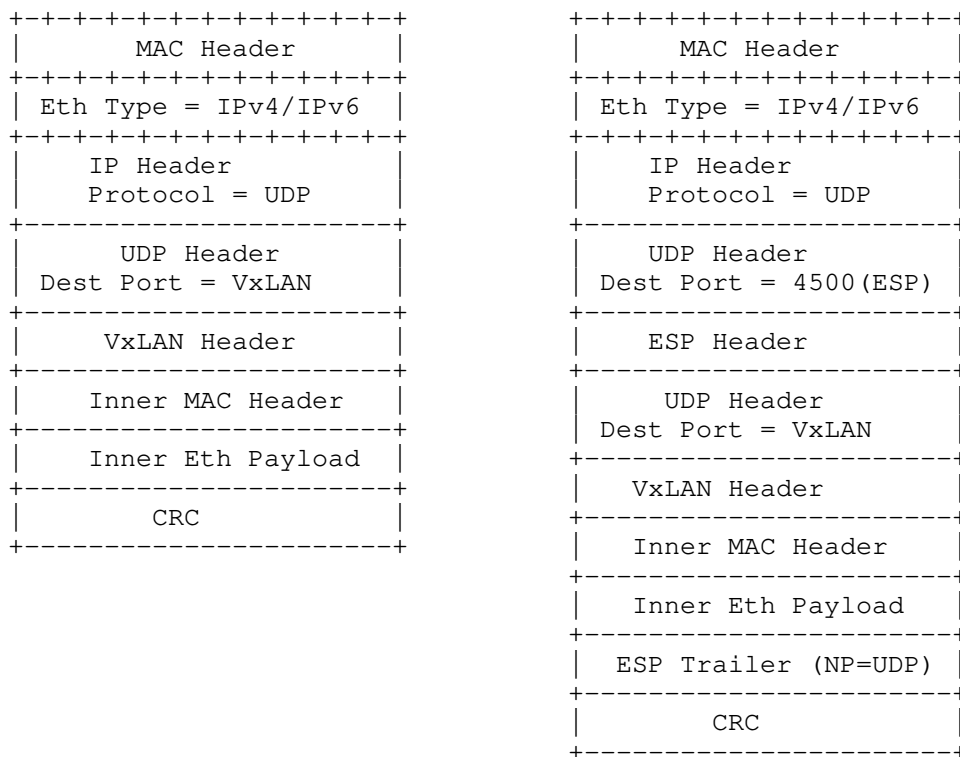


Figure 5

## 10. BGP Encoding

This document defines two new Tunnel Types along with its associated sub-TLVs for The Tunnel Encapsulation Attribute [TUNNEL-ENCAP]. These tunnel types correspond to ESP-Transport and ESP-in-UDP-Transport as described in section 4. The following sub-TLVs apply to both tunnel types unless stated otherwise.

### 10.1. The Base (Minimal Set) DIM Sub-TLV

The Base DIM is described in 3.2.1. One and only one Base DIM may be sent in the IPSec Tunnel TLV.

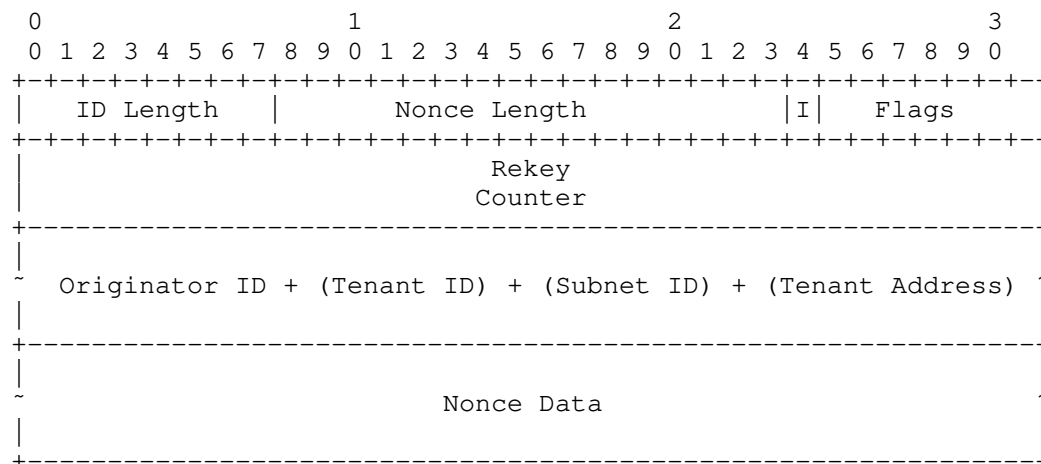


Figure 6

ID Length (16 bits) is the length of the Originator ID + (Tenant ID) + (Subnet ID) + (Tenant Address) in bytes. Nonce Length (8 bits) is the length of the Nonce Data in bytes I (1 bit) is the initial contact flag Flags (7 bits) are reserved and MUST be set to zero on transmit and ignored on receipt. The Rekey Counter is a 64 bit rekey counter The Originator ID + (Tenant ID) + (Subnet ID) + (Tenant Address) is the tunnel identifier and uniquely identifies the tunnel. Depending on the granularity of the tunnel, the fields in () may not be used - i.e., for a tunnel at the PE level of granularity, only Originator ID is required. The Nonce Data is the nonce. Its length is a multiple of 32 bits. Nonce lengths should be chosen to meet minimum requirements described in IKEv2 [RFC7296].

### 10.2. The Key Exchange Sub-TLV

The KE Sub-TLV is described in 3.2.1 and 3.2.2.1. A KE is always required. One or more KE Sub-TLVs may be included in the IPSec Tunnel TLV.



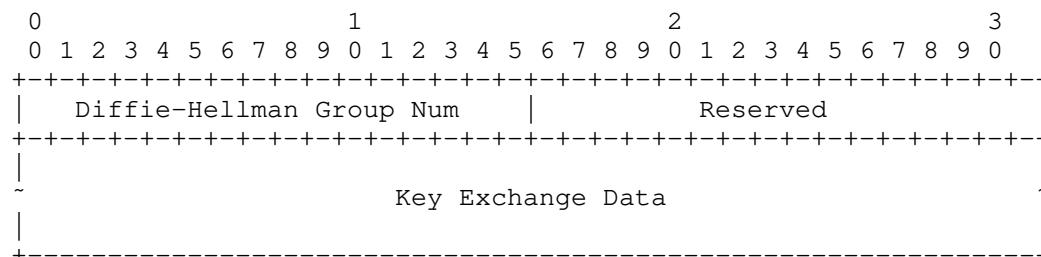


Figure 7

Diffie-Hellman Group Num 916 bits) identifies the Diffie-Hellman group in the Key Exchange Data was computed. Diffie-Hellman group numbers are discussed in IKEv2 [RFC7296] Appendix B and [RFC5114].

The Key Exchange payload is constructed by copying one's Diffie-Hellman public value into the "Key Exchange Data" portion of the payload. The length of the Diffie-Hellman public value is described for MOPT groups in [RFC7296] and for ECP groups in [RFC4753].

### 10.3. ESP SA Proposals Sub-TLV

The SA Sub-TLV is described in 3.2.2.2. Zero or more SA Sub-TLVs may be included in the IPSec Tunnel TLV.

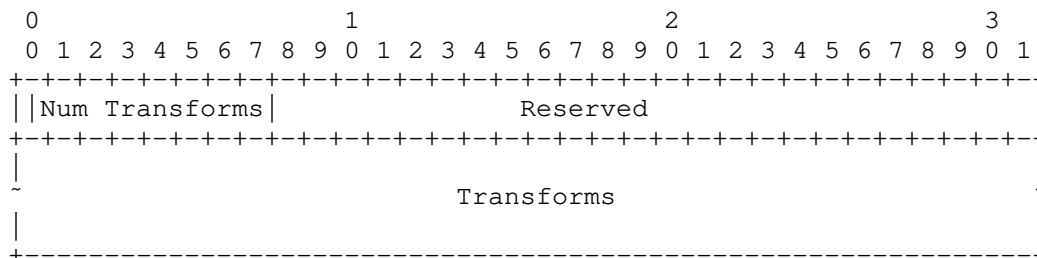


Figure 8

Num Transforms is the number of transforms included. Reserved is not used and MUST be set to zero on transmit and MUST be ignored on receipt.

#### 10.3.1. Transform Substructure

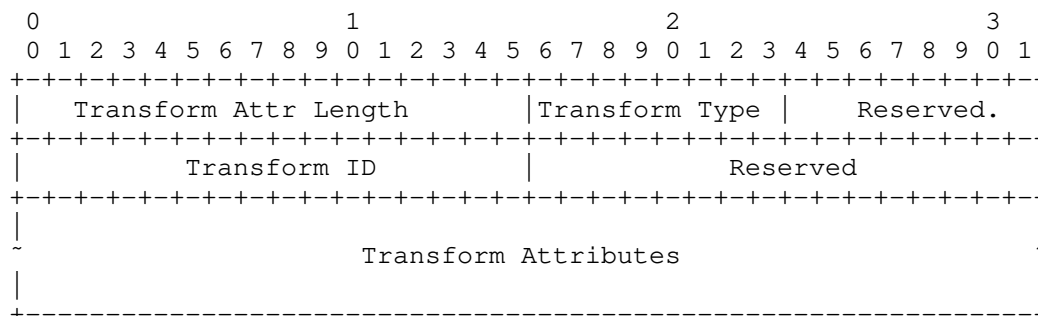


Figure 9

The Transform Attr Length is the length of the Transform Attributes field. The Transform Type is from Section 3.3.2 of [RFC7296] and [IKEV2IANA]. Only the values ENCR, INTEG, and ESN are allowed. The Transform ID specifies the transform identification value from [IKEV2IANA]. Reserved is unused and MUST be zero on transmit and MUST be ignored on receipt. The Transform Attributes are taken directly from 3.3.5 of [RFC7296].

## 11. Applicability

Although P2MP BGP signaling for establishment and maintenance of SAs among PE devices is described in this document in context of EVPN, there is no reason why it cannot be extended to other VPN technologies such as IP-VPN RFC 4364 [RFC4364], VPLS RFC 4761 [RFC4761] and RFC 4762 [RFC4762], and MVPN RFC 6513 [RFC6513] and RFC 6514 [RFC6514] with ingress replication. The reason EVPN has been chosen is because of its pervasiveness in DC, SP, and Enterprise applications and because of its ability to support SA establishment at different granularity levels such as: per PE, Per tenant, per subnet, per Ethernet Segment, per IP address, and per MAC. For other VPN technology types, a much smaller granularity levels can be supported. For example for VPLS, only the granularity of per PE and per subnet can be supported. For per-PE granularity level, the mechanism is the same among all the VPN technologies as IPsec tunnel type (and its associated TLV and sub-TLVs) are sent along with the PE's loopback IPv4 (or IPv6) address. For VPLS, if per-subnet (per bridge domain) granularity level needs to be supported, then the IPsec tunnel type and TLV are sent along with VPLS AD route.

The following table lists what level of granularity can be supported by a given VPN technology and with what BGP route.

| Functionality | EVPN          | IP-VPN        | MVPN        | VPLS    |
|---------------|---------------|---------------|-------------|---------|
| per PE        | IPv4/v6 route | IPv4/v6 route | IPv4/v6 rte | IPv4/v6 |
| per tenant    | IMET (or new) | lpbk (or new) | I-PMSI      | N/A     |
| per subnet    | IMET          | N/A           | N/A         | VPLS AD |
| per IP        | EVPN RT2/RT5  | VPN IP rt     | *,G or S,G  | N/A     |
| per MAC       | EVPN RT2      | N/A           | N/A         | N/A     |

Figure 10

## 12. Acknowledgements

TBD.

## 13. IANA Considerations

A new transitive extended community Type of 0x06 and Sub-Type of TBD for EVPN Attachment Circuit Extended Community needs to be allocated by IANA.

## 14. Security Considerations

This document proposes that a device re-use an ephemeral Diffie-Hellman exponential with multiple peers. There are some known potential vulnerabilities to this approach, which can be mitigated by the device first validating a peer's public value to be a safe public value before combining its own private value with it. The tests which MUST be performed are described in [RFC6989]. See [REUSE] for additional security considerations when reusing ephemeral Diffie-Hellman keys.

A controller acts as a "trusted third party", which asserts that a particular Diffie-Hellman public value is associated with a particular entity. A device receiving the public key is not required to validate the assertion.

A subverted controller can act as a "man-in-the-middle" between a pair of devices. The easiest attack would be for the attacker to adjust the routing for the desired traffic through a compromised gateway and directly observe the cleartext. It is also possible that a subverted controller could provide a device with a Diffie-Hellman public value that actually belongs to a compromised gateway rather

than the intended gateway, but doing so does not seem to be necessary. Nonetheless, the attack of a subverted controller can be mitigated by having a device sign its Diffie-Hellman public value (e.g, as a CMS Signed data object), where the receiver validates the digital signature on the object. However, this adds significant processing cost to a rekey and does not fit the controller-based network architecture model.

A subverted IPsec device whose DH pair has been compromised would be vulnerable to all of its IPsec traffic using that DH pair being compromised. Assuming the use of strong DH algorithms (including quantum resistant algorithms as they become available), the compromise would most likely be due to the device itself being compromised. Such a compromised device is also vulnerable to a direct plaintext compromise.

PFS is achieved between rekey periods, as DH pairs are required to be generated independently. However, because a device uses the same long-term key to generate session key with multiple peers, there is no PFS between sessions within the same rekey period. To reduce key exposure outside of a rekey period, when a connection is closed each endpoint MUST forget not only the keys used by the connection but also any information that could be used to recompute those keys. However, the DH private key value and the nonce distributed with it may be forgotten only once the last IPsec SA that uses the private key value is removed from the SAD and there is no chance that a new IPsec SA could be setup that requires the private key value.

If quantum resistance is considered to be an issue, the controller can distribute a PSK, which could be used to create the SK\_d in the manner shown in [I-D.ietf-ipsecme-qr-ikev2].

## 15. References

### 15.1. Normative References

- [GENEVE] Gross, J., et al., "Geneve: Generic Network Virtualization Encapsulation", 2018,  
<<https://tools.ietf.org/html/draft-ietf-nvo3-geneve-06>>.
- [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>>.

- [RFC3948] Huttunen, A., Swander, B., Volpe, V., DiBurro, L., and M. Stenberg, "UDP Encapsulation of IPsec ESP Packets", RFC 3948, DOI 10.17487/RFC3948, January 2005, <<https://www.rfc-editor.org/info/rfc3948>>.
- [RFC4301] Kent, S. and K. Seo, "Security Architecture for the Internet Protocol", RFC 4301, DOI 10.17487/RFC4301, December 2005, <<https://www.rfc-editor.org/info/rfc4301>>.
- [RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <<https://www.rfc-editor.org/info/rfc7432>>.
- [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>>.
- [RFC8365] Sajassi, A., Ed., Drake, J., Ed., Bitar, N., Shekhar, R., Uttaro, J., and W. Henderickx, "A Network Virtualization Overlay Solution Using Ethernet VPN (EVPN)", RFC 8365, DOI 10.17487/RFC8365, March 2018, <<https://www.rfc-editor.org/info/rfc8365>>.

## 15.2. Informative References

- [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>>.
- [RFC4761] Kompella, K., Ed. and Y. Rekhter, Ed., "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, DOI 10.17487/RFC4761, January 2007, <<https://www.rfc-editor.org/info/rfc4761>>.
- [RFC4762] Lasserre, M., Ed. and V. Kompella, Ed., "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, DOI 10.17487/RFC4762, January 2007, <<https://www.rfc-editor.org/info/rfc4762>>.
- [RFC6513] Rosen, E., Ed. and R. Aggarwal, Ed., "Multicast in MPLS/BGP IP VPNs", RFC 6513, DOI 10.17487/RFC6513, February 2012, <<https://www.rfc-editor.org/info/rfc6513>>.
- [RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, DOI 10.17487/RFC6514, February 2012, <<https://www.rfc-editor.org/info/rfc6514>>.

[RFC7348] Mahalingam, M., Dutt, D., Duda, K., Agarwal, P., Kreeger, L., Sridhar, T., Bursell, M., and C. Wright, "Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", RFC 7348, DOI 10.17487/RFC7348, August 2014, <<https://www.rfc-editor.org/info/rfc7348>>.

#### Appendix A. Additional Stuff

TBD.

#### Authors' Addresses

Ali Sajassi (editor)  
Cisco  
170 W Tasman Drive  
San Jose, CA  
USA

Email: [sajassi@cisco.com](mailto:sajassi@cisco.com)

Ayan Banerjee  
Cisco  
170 W Tasman Drive  
San Jose, CA  
USA

Email: [ayabaner@cisco.com](mailto:ayabaner@cisco.com)

Sameer Thoria  
Cisco  
170 W Tasman Drive  
San Jose, CA  
USA

Email: [sthoria@cisco.com](mailto:sthoria@cisco.com)

David Carrel  
Graphiant  
CA  
USA

Email: [carrel@graphiant.com](mailto:carrel@graphiant.com)

Brian Weis  
Independent  
CA  
USA

Email: bew.stds@gmail.com

John Drake  
Juniper Networks  
CA  
USA

Email: jdrake@juniper.net

IDR  
Internet-Draft  
Intended status: Standards Track  
Expires: January 23, 2020

S. Sangli  
R. Bonica  
Juniper Networks Inc.  
July 22, 2019

BGP based Virtual Private Network (VPN) Services over SRv6+ enabled IPv6  
networks  
draft-ssangli-idr-bgp-vpn-srv6-plus-02

## Abstract

This document defines BGP protocol extensions for encoding and carrying SRv6+ Per-Path Service Instruction information to support Virtual Private Network services. This is applicable when the VPN services are offered in a SRv6+ enabled IPv6 network such that the VPN payload is transported over IPv6. The Per-Path Service Instruction information is encoded in the IPv6 Destination Option Header in the IPv6 data packets.

## Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

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

This Internet-Draft will expire on January 23, 2020.

## Copyright Notice

Copyright (c) 2019 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must



include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

|                                                              |    |
|--------------------------------------------------------------|----|
| 1. Introduction . . . . .                                    | 2  |
| 2. Requirements Language . . . . .                           | 3  |
| 3. Per-Path Service Instruction Information . . . . .        | 3  |
| 4. Usage of Tunnel Encapsulation Attribute . . . . .         | 4  |
| 5. Procedures for Egress BGP Speaker . . . . .               | 6  |
| 6. Procedures for Ingress BGP Speaker . . . . .              | 7  |
| 7. BGP based L3 VPN services over IPv6 . . . . .             | 7  |
| 7.1. IPv4 VPN on SRv6+ enabled IPv6 Core . . . . .           | 7  |
| 7.2. IPv6 VPN on SRv6+ enabled IPv6 Core . . . . .           | 8  |
| 7.3. IPv4 Global Routes on SRv6+ enabled IPv6 Core . . . . . | 8  |
| 8. BGP based Ethernet VPN services over IPv6 . . . . .       | 9  |
| 8.1. Ethernet Per ES Auto-Discovery (A-D) route . . . . .    | 9  |
| 8.2. Ethernet per EVI Auto-Discovery (A-D) route . . . . .   | 10 |
| 8.3. MAC/IP Advertisement route . . . . .                    | 11 |
| 8.4. Inclusive Multicast Ethernet Route . . . . .            | 11 |
| 8.5. IP Prefix Route . . . . .                               | 11 |
| 9. Deployment Considerations . . . . .                       | 12 |
| 10. Backward Compatibility . . . . .                         | 13 |
| 11. Security Considerations . . . . .                        | 13 |
| 12. IANA Considerations . . . . .                            | 13 |
| 13. Acknowledgements . . . . .                               | 13 |
| 14. References . . . . .                                     | 13 |
| 14.1. Normative References . . . . .                         | 13 |
| 14.2. Informative References . . . . .                       | 15 |
| Authors' Addresses . . . . .                                 | 16 |

## 1. Introduction

Virtual Private Network (VPN) technologies allow network providers to emulate private networks with shared infrastructure. For example, assume that a set of red sites, set of blue sites and a set of green sites connect to a provider network. Furthermore, assume that red sites and blue sites wish to interconnect, exchange packets. However, the green sites wish to communicate with green sites only. The provider should allow its infrastructure network to scale to both the requirements without having to create multiple parallel network infrastructures. The IETF has standardized many VPN technologies viz. Layer 3 VPN (L3VPN) [RFC4364], Layer 2 VPN (L2VPN) [RFC6624], Virtual Private LAN Service (VPLS) [RFC4761], [RFC4762], Ethernet VPN (EVPN) [RFC7432], Pseudowires [RFC8077] to enable Layer 3 and Layer 2 VPN services.

The aforementioned technologies leverage MPLS network architecture :

- o to establish a MPLS tunnel from ingress PE to egress PE, thus making all P routers agnostic of VPN state.
- o to provide demultiplexing abstraction in the tunnelled packet so the payload packet can be forwarded at the egress router based on Routing table and/or interface.

In pure IPv6 deployments where there may be non-MPLS capable routers, it would be desirable to have alternate mechanism to provide VPN connectivity. This document describes BGP extensions and procedures applicable for SRv6+ enabled IPv6 networks, to provide VPN services over BGP.

## 2. Requirements Language

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

## 3. Per-Path Service Instruction Information

A SRv6+ [I-D.bonica-spring-srv6-plus] segment provides unidirectional connectivity from an ingress node to an egress node. A SRv6+ path contains one or more such segments. SRv6+ introduces the concept of Per-Segment Service Instruction and Per-Path Service Instruction. These instructions describe the additional packet processing performed on a node. The Per-Segment Service Instruction is executed on the segment egress node while the Per-Path Service Instruction is executed on the path egress node. The SR Path egress node advertises the service prefix reachability information to SR Path ingress node via Multi Protocol extensions in BGP [RFC4760].

For providing VPN services, aforementioned BGP extensions rely on MPLS architecture [RFC3031]. The BGP extensions specify the new encoding for Network Layer Reachability Information (NLRI) to include the MPLS VPN labels [RFC8277]. Such a MPLS VPN label is associated with a forwarding decision in the VPN Routing Instance on the egress BGP Router. The ingress BGP router will push the VPN label on the data packet destined to the egress BGP router. The transport tunnel from ingress router to egress router can be MPLS or GRE or L2TPv3, but inner payload is a MPLS packet as described in [RFC4023], [RFC4817], [RFC7510]. The intermediate routers do not process the VPN label [a.k.a.] embedded label as described in [I-D.ietf-idr-tunnel-encaps].

To provide BGP based VPN services on a non-MPLS IPv6 networks, it would be beneficial to retain the benefits of BGP protocol extensions while leveraging the benefits of IPv6 [RFC8200].

[I-D.bonica-6man-vpn-dest-opt] describes SRv6+ paths as programmable with Per-Path Service Instructions (PPSI) that determine how egress nodes process SRv6+ payloads. The PPSIs are carried in the PPSI Option encoded in the IPv6 Destination Option Header [RFC8200].

The Per-Path Service Instruction (PPSI) Identifier is defined as follows:

- o 32 bit quantity.

The PPSI Identifier have node-local significance and is assigned by the egress BGP router. The value of zero is reserved. The PPSI Identifier will serve 2 purposes.

- o It MUST uniquely identify the VPN Routing Instance for L3VPN or identify an Ethernet Segment for EVPN or identify a leaf property for EVPN TREE upon which forwarding decision can be taken.
- o It MAY provide information for special processing before the packet is forwarded.

The structure of 3 octet PPSI Identifier will be updated in the next version of this document.

The encoding of the Per-Path Service Instruction Identifier for VPNs is described in Section 7 and Section 8.

#### 4. Usage of Tunnel Encapsulation Attribute

This document defines a new Tunnel type : SRv6+. The format is as per below.

- o Tunnel Type (2 Octets) : To be assigned
- o Tunnel Length (2 Octets) : 1
- o Value : List of Sub-TLVs

[I-D.ietf-idr-tunnel-encaps] defines many sub-TLVs for the tunnels. The encoding for them are as follows:

- o Tunnel Endpoint Sub-TLV : As per [I-D.ietf-idr-tunnel-encaps]
- o Encapsulation Sub-TLV : Not needed.

- o IPv4 DS Field Sub-TLV : Not needed.
- o UDP Destination Port Sub-TLV : Not needed.
- o Protocol Type Sub-TLV : As per [I-D.ietf-idr-tunnel-encaps].
- o Color Sub-TLV : As per [I-D.ietf-idr-tunnel-encaps].
- o Embedded Label Handling Sub-TLV : 3.
- o MPLS Label Stack Sub-TLV : Not needed.
- o Prefix SID Sub-TLV : Not Needed.

The Tunnel Encapsulation Attribute is an Optional Transitive attribute as described in [I-D.ietf-idr-tunnel-encaps]. This attribute with SRv6+ tunnel type MUST be present in the BGP update carrying the Network Layer Reachability Information encoded with the PPSI Information. This document refers to the NLRI that is associated with SRv6+ Tunnel Encapsulation attribute as SRv6+\_NLRI. The document [I-D.ietf-idr-tunnel-encaps] defines the encoding for sub-TLV as follows.

- o Sub-TLV Type : 1 octet
- o Sub-TLV Length : 1 or 2 octets
- o Sub-TLV Value : defined per Sub-TLV as per below.

The Tunnel Endpoint Sub-TLV can specify the IPv6 address of the egress router as the final destination address of SRv6+ packet which is also referred to as SR Path destination address. The sub-fields on this sub-TLV is encoded as below.

- o Autonomous System Number : AS number of the IPv6 SR domain.
- o Address Family : 2 (refers to IPv6).
- o Address : IPv6 address of the egress interface present in SRv6+ domain.

The Value field may be set to 0 which indicates that next hop value in the NLRI should be chosen for the SRv6+ Path destination address.

The Embedded Label Handling Sub-TLV describes how the label field in the NLRI should be interpreted.

- o Value : MUST be set to 3.

The [I-D.ietf-idr-tunnel-encaps] specifies only 2 values. While the value 1 refers to label field as MPLS embedded label that is carried at the top of the label stack of the MPLS payload packet, the value 2 refers to label field to be either ignored or carried in the virtual network field of the encapsulation header.

This document defines another behavior for the label field. The value 3 will indicate that value in the label field MUST be inserted in the Destination Options Header of the IPv6 Tunnel header.

The Tunnel Encapsulation attribute can carry one or more Tunnel types. The local policy on the ingress router can determine which Tunnel type to be used for the NLRI.

## 5. Procedures for Egress BGP Speaker

The PPSI Information instructs the egress router to de-encapsulate the packet and forward the newly exposed payload inner packet through the specified interface or forward using the specified Routing Instance. The PPSI Identifier described in Section 3 will be assigned by the egress BGP Router except in the case of EVPN per ES AD route when P2MP tunnel is used for delivering BUM traffic in EVPN. If P2MP tunnel is used to deliver BUM traffic for EVPN, the PPSI Identifier used to identify an Ethernet Segment is assigned by the upstream ingress BGP Router. Otherwise, it is downstream assigned by the egress BGP router.

When the egress BGP Speaker advertises the NLRI, it will include the PPSI Information in the encoding described in Section 7 and Section 8. The egress BGP Speaker MUST include the Tunnel Encapsulation Attribute with Route type SRv6+ as described in Section 4 in such BGP updates.

By tagging the BGP update with Tunnel Encapsulation attribute of SRv6+ type, the BGP Speaker informs how the SRv6+\_NLRI should be decoded and processed by the receiving BGP Speaker.

Via the Remote Tunnel Endpoint Sub-TLV encoding, the egress BGP router may specify the SRv6+ Path Destination Address. The Protocol type Sub-TLV and the Color Sub-TLV may be used by the egress BGP router to influence the payload packets to be put on SRv6+ path. The Embedded Label Handling Sub-TLV MUST be set to 3 to inform that the label field MUST be inserted in the Destination Options Header at the ingress router as described in [I-D.bonica-6man-vpn-dest-opt].

A single PPSI Identifier may be associated with all the prefixes in a Routing Instance or a unique PPSI Identifier may be associated for each prefix in the Routing Instance. Similarly, a PPSI Identifier

may be assigned to identify an Ethernet segment or leaf AC property by EVPN. The choice is left to the Network Operator and is outside the scope of this document.

## 6. Procedures for Ingress BGP Speaker

Upon receiving a BGP update, the receiving BGP Speaker will look for Tunnel Encapsulation attribute. If the tunnel type carried in the Tunnel Encapsulation attribute is SRv6+, the BGP updates is said to be carrying the SRv6+\_NLRI and the Label field in the Network Layer Reachability Information is treated as Per-Path Service Instruction (PPSI) Identifier.

The tuple (PPSI Identifier, Prefix) is programmed in the forwarding infrastructure of the router. The manner in which this tuple is stored in the router is outside the scope of this document. If SRv6+ has been enabled on the router, such a tuple SHOULD be used for encoding the Destination Options Header as described in [I-D.bonica-6man-vpn-dest-opt].

The [I-D.ietf-idr-tunnel-encaps] describes how Tunnel Endpoint Sub-TLV has to be processed. It also describes the usage of the Protocol type Sub-TLV and the Color Sub-TLV. This may be used by the ingress BGP router to select the payload packets that should be put on SRv6+ path.

The Embedded Label Handling Sub-TLV value that is set to 3 indicates that ingress BGP router to insert value of label field in the Destination Options Header of the Tunnel IPv6 packet.

## 7. BGP based L3 VPN services over IPv6

The Egress and Ingress BGP speakers form a BGP peering session to exchange a set of prefixes described in [RFC4271] and Multi protocol extensions [RFC4760]. The BGP Router capable of SRv6+ that is enabled to carry L3 VPN services over IPv6 networks should follow the procedures mentioned in Section 5 and Section 6. The manner in which a BGP Router is configured for SRv6+ underlay and L3 VPN overlay is outside the scope of this document.

### 7.1. IPv4 VPN on SRv6+ enabled IPv6 Core

The IPv4 L3 VPN over IPv6 is defined in [RFC5549]. The MP\_REACH NLRI and Tunnel Encapsulation attribute encoding is as per below:

- o AFI : 1; SAFI : 128
- o Length of the Next Hop : 16 (or 32 if Link Local)

- o Network address of the Next Hop : IPv6 address of the egress BGP Router
- o NLRI : IPv4-VPN routes
- o Label : Per-Path Service Instruction Identifier
- o Tunnel Encapsulation Path Attribute : SRv6+ Type as described in Section 4

The PPSI Identifier is associated with VPN Routing Instance on the Egress PE. The Tunnel Encapsulation attribute with SRv6+ type MUST be appended to the Path attributes associated with the NLRI.

#### 7.2. IPv6 VPN on SRv6+ enabled IPv6 Core

The IPv6 L3 VPN over IPv6 is defined in [RFC4659]. The MP\_REACH NLRI and Tunnel Encapsulation attribute encoding is as per below:

- o AFI : 2; SAFI : 128
- o Length of the Next Hop : 16 (or 32 if Link Local)
- o Network address of the Next Hop : IPv6 address of the egress BGP Router
- o NLRI : IPv6-VPN routes
- o Label : Per-Path Service Instruction Identifier
- o Tunnel Encapsulation Path Attribute : SRv6+ Type as described in Section 4

The PPSI Identifier is associated with VPN Routing Instance on the Egress PE. The Tunnel Encapsulation attribute with SRv6+ type MUST be appended to the Path attribute associated with the NLRI.

#### 7.3. IPv4 Global Routes on SRv6+ enabled IPv6 Core

The IPv4 L3 VPN over IPv6 is defined in [RFC5549]. The MP\_REACH NLRI and Tunnel Encapsulation attribute encoding is per below:

- o AFI : 1; SAFI : 1
- o Length of the Next Hop : 16 (or 32 if Link Local)
- o Network address of the Next Hop : IPv6 address of the egress BGP Router

- o NLRI : IPv4 routes
- o Label : Per-Path Service Instruction Identifier
- o Tunnel Encapsulation Path Attribute : SRv6+ Type as described in Section 4

The PPSI Identifier is associated with VPN Routing Instance on the Egress PE. The Tunnel Encapsulation attribute with SRv6+ type MUST be appended to the Path attribute associated with the NLRI.

## 8. BGP based Ethernet VPN services over IPv6

The [RFC7432] describes the BGP extensions for carrying the Ethernet Virtual Private Network Overlay on MPLS network. It defines 4 types of EVPN NLRI. This document specifies changes to certain fields for those NLRIs.

- o Ethernet Auto-Discovery (A-D) route
- o MAC/IP Advertisement route
- o Inclusive Multicast Ethernet Tag route
- o IP Prefix route

### 8.1. Ethernet Per ES Auto-Discovery (A-D) route

The MP\_REACH and MP\_UNREACH attributes will carry this route in the NLRI encoding described in [RFC7432]. In addition to Tunnel Encapsulation attribute encoding, this document recommends to follow the [RFC7432] encoding except the following. For MPLS label carried in the Ethernet A-D per ESI route:

- o MPLS label : Per [RFC7432], it is set to zero.
- o Tunnel Encapsulation Path Attribute : SRv6+ Type as described in Section 4

The MPLS label field is not part of the route but treated as route attribute. For procedures and usage of this route, refer to [RFC7432]. The Tunnel Encapsulation attribute with SRv6+ type MUST be appended to the Path attribute associated with the NLRI.

An EVPN Ethernet per ES A-D route is usually signaled together with an ESI label extended community. For ESI Label carried in the ESI label extended community:



- o ESI Label: Per-Path Service Instruction Identifier

The Per-Path Service Instruction Identifier is used to identify an Ethernet segment attached to the BGP PE for EVPN.

If P2MP tunnel is used to deliver BUM traffic, then this PPSI Identifier is upstream assigned by the ingress router, otherwise it is downstream assigned by the egress router.

## 8.2. Ethernet per EVI Auto-Discovery (A-D) route

The MP\_REACH and MP\_UNREACH attributes will carry this route in the NLRI encoding described in [RFC7432]. In addition to Tunnel Encapsulation attribute encoding, this document recommends to follow the [RFC7432] encoding except the following:

- o MPLS label : Per-Path Service Instruction Identifier
- o Tunnel Encapsulation Path Attribute : SRv6+ Type as described in Section 4

The MPLS label field is not part of the route but treated as route attribute. For procedures and usage of this route, refer to [RFC7432]. The Tunnel Encapsulation attribute with SRv6+ type MUST be appended to the Path attribute associated with the NLRI.

In addition, for EVPN E-tree service, this route may be signaled together with an E-Tree Extended Community as it is specified in [RFC8317]. For the leaf label carried in the E-Tree Extended Community:

- o Leaf Label: Per-Path Service Instruction Identifier

In case of EVPN E-tree service, the per-path service identifier carried in the E-Tree extended community is used to signal a leaf AC property.

In the data plane, this PPSI identifier specified in the Destination Option header is used by an egress router to identify that a data packet is ingressed from a leaf AC such that appropriate forwarding decision can be made.

If P2MP tunnel is used to deliver BUM traffic, then this PPSI Identifier is upstream assigned by the ingress router. Otherwise it is downstream assigned by the egress router.

### 8.3. MAC/IP Advertisement route

The MP\_REACH and MP\_UNREACH attributes will carry this route in the NLRI encoding described in [RFC7432]. In addition to Tunnel Encapsulation attribute encoding, this document recommends to follow the [RFC7432] encoding except the following.

- o MPLS label1 : Per-Path Service Instruction Identifier1
- o MPLS label2 : Per-Path Service Instruction Identifier2
- o Tunnel Encapsulation Path Attribute : SRv6+ Type as described in Section 4

The MPLS label field is not part of the route but treated as route attribute. For procedures and usage of this route, refer to [RFC7432]. The Tunnel Encapsulation attribute with SRv6+ type MUST be appended to the Path attribute associated with the NLRI.

### 8.4. Inclusive Multicast Ethernet Route

The MP\_REACH and MP\_UNREACH attributes will carry this route in the NLRI encoding described in [RFC7432]. In addition to Tunnel Encapsulation attribute encoding, this document recommends to follow the [RFC7432] encoding except the following.

- o If MPLS label field in the PMSI Tunnel Attribute is non-zero, it is set to Per-Path Service Instruction Identifier.
- o Tunnel Encapsulation Path Attribute : SRv6+ Type as described in Section 4

The Tunnel Encapsulation attribute with SRv6+ type MUST be appended to the Path attribute associated with the NLRI.

### 8.5. IP Prefix Route

The MP\_REACH and MP\_UNREACH attributes will carry this route in the NLRI encoding described in [I-D.ietf-bess-evpn-prefix-advertisement]. In addition to Tunnel Encapsulation attribute encoding, this document recommends the following change:

- o MPLS label: if it is non-zero, it is set to Per-Path Service Instruction Identifier.
- o Tunnel Encapsulation Path Attribute : SRv6+ Type as described in Section 4

The MPLS label field is not part of the route but treated as route attribute. For procedures and usage of this route, refer to [I-D.ietf-bess-evpn-prefix-advertisement]. The Tunnel Encapsulation attribute with SRv6+ type MUST be appended to the Path attribute associated with the NLRI.

## 9. Deployment Considerations

This document proposes to reuse the NLRI encoding for BGP L3VPN and EVPN Network Layer Routing Information. However, care should be taken when BGP VPN overlay services are enabled on SRv6+ underlay such that Tunnel Encapsulation Path attribute with SRv6+ type MUST be appended. When a BGP router advertises SRv6+\_NLRI, it MUST NOT remove the Tunnel Encapsulation Path attribute.

The SRv6+ underlay is similar to other "tunnel" technologies viz MPLS, GRE, IP-in-IP, L2TPv3. The egress and ingress BGP routers can be connected via one or more such underlay technologies. A BGP speaker can advertise the VPN NLRI with the nexthop reachable via one or more such underlay paths. Each such mechanism can co-exist together as ships-in-night. However, when SRv6+\_NLRI is advertised by a egress BGP speaker and received by an ingress BGP speaker, they MUST follow the procedures mentioned in this document.

For migrating a BGP router to SRv6+ the following procedures can be followed.

- o Operator will enable SRv6+ underlay on the ingress and egress routers identifying the SRv6+ path from ingress router's interface to egress router's interface. The way to configure the ingress and egress routers are outside the scope of this document.
- o SRv6+ enabled ingress BGP router will setup the additional information in the forwarding table such that it can append an IPv6 tunnel header and encode the PPSI Option in the Destination Options Header.
- o SRv6+ enabled egress BGP router will setup the additional information in the forwarding table such that PPSI Identifier can be used to lookup to find the Routing Instance and make the forwarding decision.
- o Operator will enable BGP VPN overlay over SRv6+ underlay on ingress router. This means that ingress router will start looking for SRv6+\_NLRI in the BGP updates. The way to enable the BGP VPN overlay over SRv6+ underlay is outside the scope of this document.

- o The operator will enable BGP VPN overlay over SRv6+ underlay on egress router. With this, the egress router will create PPSI Identifier and associate it with Routing Instances. It then advertises the SRv6+\_NLRI to the ingress BGP router.
- o The ingress router will interpret the SRv6+\_NLRI and use PPSI identifier and follow the procedures in [I-D.bonica-spring-srv6-plus] to encode the Destination Options Header to forward the data packet.
- o Now that SRv6+ path is setup between ingress and egress BGP routers, on the egress BGP router the Operator can migrate the Routing Instances from MPLS VPN set of Instances to SRv6+ enabled set of Instances. The way to configure Routing Instances to achieve the above is outside the scope of this document.

#### 10. Backward Compatibility

The extension proposed in this document is backward compatible with procedures described for BGP enabled services.

#### 11. Security Considerations

This document does not introduce any new security considerations beyond those already specified in [RFC4271], [RFC8277] and [I-D.ietf-idr-tunnel-encaps].

#### 12. IANA Considerations

IANA is requested to assign a code point for SRv6+ Route Type for BGP Tunnel Encapsulation Path Attribute from BGP Tunnel Encapsulation Attribute Tunnel Types Registry.

#### 13. Acknowledgements

The authors would like to thank Jeff Haas and Wen Lin for careful review and suggestions.

#### 14. References

##### 14.1. Normative References

- [I-D.bonica-6man-vpn-dest-opt]  
Bonica, R., Kamite, Y., Lenart, C., So, N., Xu, F.,  
Presbury, G., Chen, G., Zhu, Y., Yang, G., and Y. Zhou,  
"The Per-Path Service Instruction (PPSI) Option", draft-  
bonica-6man-vpn-dest-opt-06 (work in progress), July 2019.

- [I-D.bonica-spring-srv6-plus]  
Bonica, R., Hegde, S., Kamite, Y., Alston, A., Henriques, D., Halpern, J., Linkova, J., and G. Chen, "IPv6 Support for Segment Routing: SRv6+", draft-bonica-spring-srv6-plus-04 (work in progress), July 2019.
- [I-D.ietf-bess-evpn-prefix-advertisement]  
Rabadan, J., Henderickx, W., Drake, J., Lin, W., and A. Sajassi, "IP Prefix Advertisement in EVPN", draft-ietf-bess-evpn-prefix-advertisement-11 (work in progress), May 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-13 (work in progress), July 2019.
- [RFC0791] Postel, J., "Internet Protocol", STD 5, RFC 791, DOI 10.17487/RFC0791, September 1981, <<https://www.rfc-editor.org/info/rfc791>>.
- [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>>.
- [RFC4303] Kent, S., "IP Encapsulating Security Payload (ESP)", RFC 4303, DOI 10.17487/RFC4303, December 2005, <<https://www.rfc-editor.org/info/rfc4303>>.
- [RFC4443] Conta, A., Deering, S., and M. Gupta, Ed., "Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6) Specification", STD 89, RFC 4443, DOI 10.17487/RFC4443, March 2006, <<https://www.rfc-editor.org/info/rfc4443>>.
- [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>>.
- [RFC8200] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", STD 86, RFC 8200, DOI 10.17487/RFC8200, July 2017, <<https://www.rfc-editor.org/info/rfc8200>>.

## 14.2. Informative References

- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, DOI 10.17487/RFC3031, January 2001, <<https://www.rfc-editor.org/info/rfc3031>>.
- [RFC4023] Worster, T., Rekhter, Y., and E. Rosen, Ed., "Encapsulating MPLS in IP or Generic Routing Encapsulation (GRE)", RFC 4023, DOI 10.17487/RFC4023, March 2005, <<https://www.rfc-editor.org/info/rfc4023>>.
- [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>>.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <<https://www.rfc-editor.org/info/rfc4364>>.
- [RFC4659] De Clercq, J., Ooms, D., Carugi, M., and F. Le Faucheur, "BGP-MPLS IP Virtual Private Network (VPN) Extension for IPv6 VPN", RFC 4659, DOI 10.17487/RFC4659, September 2006, <<https://www.rfc-editor.org/info/rfc4659>>.
- [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>>.
- [RFC4761] Kompella, K., Ed. and Y. Rekhter, Ed., "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, DOI 10.17487/RFC4761, January 2007, <<https://www.rfc-editor.org/info/rfc4761>>.
- [RFC4762] Lasserre, M., Ed. and V. Kompella, Ed., "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, DOI 10.17487/RFC4762, January 2007, <<https://www.rfc-editor.org/info/rfc4762>>.
- [RFC4817] Townsley, M., Pignataro, C., Wainner, S., Seely, T., and J. Young, "Encapsulation of MPLS over Layer 2 Tunneling Protocol Version 3", RFC 4817, DOI 10.17487/RFC4817, March 2007, <<https://www.rfc-editor.org/info/rfc4817>>.

- [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>>.
- [RFC6624] Kompella, K., Kothari, B., and R. Cherukuri, "Layer 2 Virtual Private Networks Using BGP for Auto-Discovery and Signaling", RFC 6624, DOI 10.17487/RFC6624, May 2012, <<https://www.rfc-editor.org/info/rfc6624>>.
- [RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <<https://www.rfc-editor.org/info/rfc7432>>.
- [RFC7510] Xu, X., Sheth, N., Yong, L., Callon, R., and D. Black, "Encapsulating MPLS in UDP", RFC 7510, DOI 10.17487/RFC7510, April 2015, <<https://www.rfc-editor.org/info/rfc7510>>.
- [RFC8077] Martini, L., Ed. and G. Heron, Ed., "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", STD 84, RFC 8077, DOI 10.17487/RFC8077, February 2017, <<https://www.rfc-editor.org/info/rfc8077>>.
- [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>>.
- [RFC8317] Sajassi, A., Ed., Salam, S., Drake, J., Uttaro, J., Boutros, S., and J. Rabadan, "Ethernet-Tree (E-Tree) Support in Ethernet VPN (EVPN) and Provider Backbone Bridging EVPN (PBB-EVPN)", RFC 8317, DOI 10.17487/RFC8317, January 2018, <<https://www.rfc-editor.org/info/rfc8317>>.

Authors' Addresses

Srihari Sangli  
Juniper Networks Inc.  
Exora Business Park  
Bangalore, KA 560103  
India  
  
Email: [ssangli@juniper.net](mailto:ssangli@juniper.net)

Ron Bonica  
Juniper Networks Inc.  
2251 Corporate Park Drive  
Herndon, Virginia 20171  
USA

Email: [rbonica@juniper.net](mailto:rbonica@juniper.net)



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

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

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

This Internet-Draft will expire on January 8, 2020.

## Copyright Notice

Copyright (c) 2019 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of

publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

|                                                                                                        |    |
|--------------------------------------------------------------------------------------------------------|----|
| 1. Introduction . . . . .                                                                              | 3  |
| 1.1. Requirements Language . . . . .                                                                   | 4  |
| 2. Conventions used in this document . . . . .                                                         | 4  |
| 3. Traffic treatment behavior intent (T-intent) . . . . .                                              | 5  |
| 3.1. COLOR . . . . .                                                                                   | 5  |
| 3.2. COLOR name spaces . . . . .                                                                       | 6  |
| 4. Scaling Consideration . . . . .                                                                     | 6  |
| 5. BGP labeled-colored-unicast NLRI . . . . .                                                          | 6  |
| 5.1. BGP capability negotiation . . . . .                                                              | 6  |
| 5.2. BGP UPDATE message MP_REACH_NLRI . . . . .                                                        | 7  |
| 5.3. BGP explicit WITHDRAWN message . . . . .                                                          | 9  |
| 5.4. Some BGP attribute considerations . . . . .                                                       | 10 |
| 5.4.1. BGP Next-Hop . . . . .                                                                          | 10 |
| 5.4.2. Prefix SID . . . . .                                                                            | 10 |
| 5.4.3. Color Extended Community . . . . .                                                              | 10 |
| 5.4.4. Tunnel encapsulation . . . . .                                                                  | 10 |
| 6. BGP operation . . . . .                                                                             | 11 |
| 6.1. Injection of labeled colored unicast route to BGP . . . . .                                       | 11 |
| 6.1.1. Injecting from colored routes . . . . .                                                         | 11 |
| 6.1.2. Injections from non-colored labeled routes . . . . .                                            | 13 |
| 6.1.3. Injections from non-colored non-labeled routes . . . . .                                        | 13 |
| 6.2. Receiving BGP-LCU from eBGP (single hop) . . . . .                                                | 14 |
| 6.3. Receiving BGP-LCU from iBGP or multihop-eBGP . . . . .                                            | 14 |
| 6.4. Advertising BGP-LCU over eBGP session and iBGP session<br>with BGP NH changed (NH-self) . . . . . | 15 |
| 6.5. Advertising BGP-LCU over iBGP session when BGP NH remain<br>unchanged. . . . .                    | 15 |
| 6.6. Label value assignment procedure . . . . .                                                        | 16 |
| 7. Deployment and Operation Consideration . . . . .                                                    | 16 |
| 7.1. Building label stack . . . . .                                                                    | 16 |
| 7.1.1. Purpose of multiple-label stack . . . . .                                                       | 16 |
| 7.1.2. Ingress recursive resolution . . . . .                                                          | 17 |
| 7.2. Handling BGP-LCU ingress PEs with limited label<br>imposition depth capabilities . . . . .        | 19 |
| 8. Contributors . . . . .                                                                              | 20 |
| 9. IANA Considerations . . . . .                                                                       | 20 |
| 10. Security Considerations . . . . .                                                                  | 20 |
| 11. References . . . . .                                                                               | 21 |
| 11.1. Normative References . . . . .                                                                   | 21 |

|                                        |    |
|----------------------------------------|----|
| 11.2. Informative References . . . . . | 22 |
| Authors' Addresses . . . . .           | 23 |

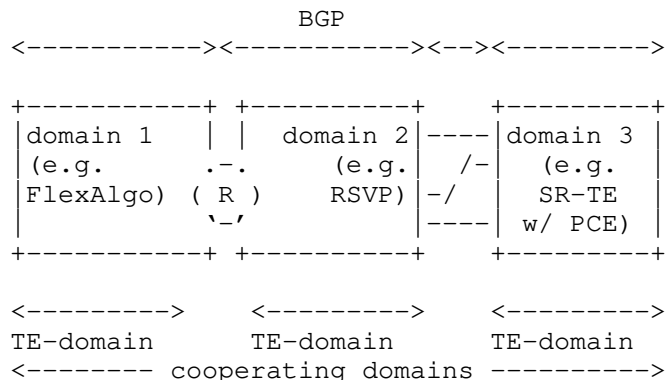
## 1. Introduction

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 IGP (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.



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 disjointed 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.

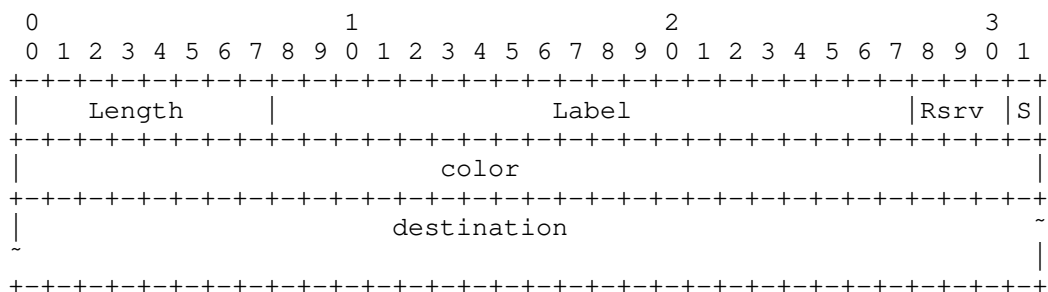
The Multiple Labels Capability is defined in chapter 2.1 of BGP Labeled Unicast [RFC8277]. The BGP speaker supporting BGP-LCU MUST follow procedure defined there.

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

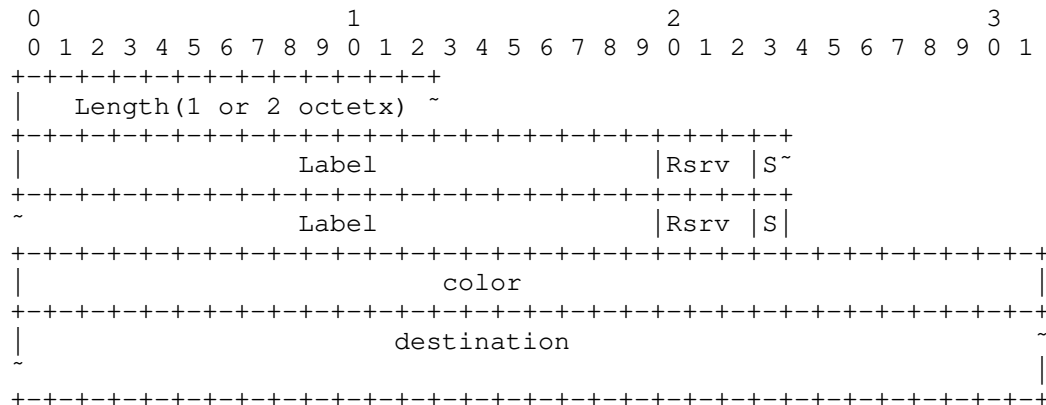
The procedure described in BGP Labeled Unicast [RFC8277] is used to encode the <COLOR,DESTINATION> tuple as prefix into NLRI with exception of NLRI length field. The Length field is encoded in one or two octets, in order to accommodate large sequence of labels. The Length field encoding follows BGP FlowSpec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



NLRI with One Label.

Figure 2



NLRI encoding with more than 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 first 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:

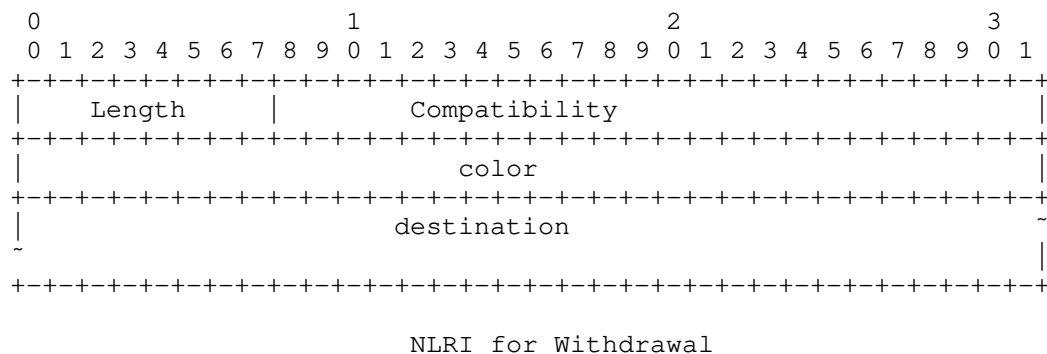


Figure 4

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 specification 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 treat 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.

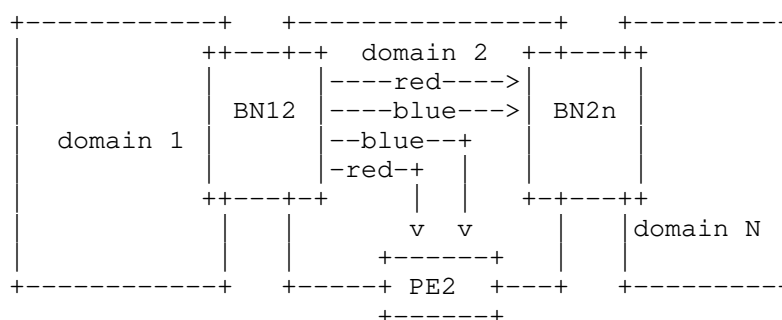
If tunnel encapsulation attribute is attached, it MUST NOT conflict with intent of particular BGP path and its NLRI.

## 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.



## 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 through 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 from 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.

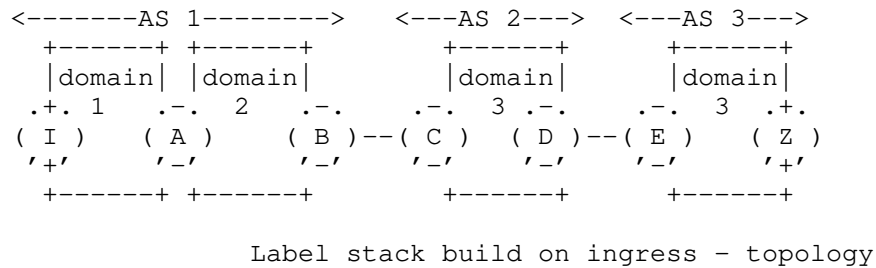


Figure 6

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 from 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: [louis@juniper.net](mailto:louis@juniper.net)

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

Phone: +14089365629  
Email: [rafal@juniper.net](mailto:rafal@juniper.net)

Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: May 20, 2021

Y. Zhu  
China Telecom  
Z. Hu  
S. Peng  
Huawei Technologies  
R. Mwehaire  
MTN Uganda Ltd.  
November 16, 2020

Signaling Maximum Transmission Unit (MTU) using BGP-LS  
draft-zhu-idr-bgp-ls-path-mtu-05

Abstract

BGP Link State (BGP-LS) describes a mechanism by which link-state and TE information can be collected from networks and shared with external components using the BGP routing protocol. The centralized controller (PCE/SDN) completes the service path calculation based on the information transmitted by the BGP-LS and delivers the result to the Path Computation Client (PCC) through the PCEP or BGP protocol.

Segment Routing (SR) leverages the source routing paradigm, which can be directly applied to the MPLS architecture with no change on the forwarding plane and applied to the IPv6 architecture, with a new type of routing header, called SRH. The SR uses the IGP protocol as the control protocol. Compared to the MPLS tunneling technology, the SR does not require additional signaling. Therefore, the SR does not support the negotiation of the Path MTU. Since multiple labels or SRv6 SIDs are pushed in the packets, it is more likely that the packet size exceeds the path mtu of SR tunnel.

This document specifies the extensions to BGP Link State (BGP-LS) to carry maximum transmission unit (MTU) messages of link. The PCE/SDN calculates the Path MTU while completing the service path calculation based on the information transmitted by the BGP-LS.

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].

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.



Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

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

This Internet-Draft will expire on May 20, 2021.

#### Copyright Notice

Copyright (c) 2020 IETF Trust and the persons identified as the document authors. All rights reserved.

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

#### Table of Contents

|                                             |   |
|---------------------------------------------|---|
| 1. Introduction . . . . .                   | 2 |
| 2. Terminology . . . . .                    | 4 |
| 3. Deploying scenarios . . . . .            | 5 |
| 4. BGP_LS Extensions for Link MTU . . . . . | 6 |
| 5. IANA Considerations . . . . .            | 6 |
| 6. Security Considerations . . . . .        | 6 |
| 7. Acknowledgements . . . . .               | 7 |
| 8. Contributors . . . . .                   | 7 |
| 9. References . . . . .                     | 7 |
| 9.1. Normative References . . . . .         | 7 |
| 9.2. Informative References . . . . .       | 7 |
| Authors' Addresses . . . . .                | 8 |

#### 1. Introduction

[RFC7752] describes the implementation mechanism of BGP-LS by which link-state and TE information can be collected from networks and shared with external components using the BGP routing protocol [RFC4271]. BGP-LS allows the necessary Link-State Database (LSDB)

and Traffic Engineering Database (TEDB) information to be collected from the IGP within the network, filtered according to configurable policy, and distributed to the PCE as necessary.

The appropriate MTU size guarantees efficient data transmission. If the MTU size is too small and the packet size is large, fragmentation may occur too much and packets are discarded by the QoS queue. If the MTU configuration is too large, packet transmission may be slow. Path MTU is the maximum length of a packet that can pass through a path without fragmentation. [RFC1191] describes a technique for dynamically discovering the maximum transmission unit (MTU) of an arbitrary internet path.

The traditional MPLS tunneling technology has signaling for establishing a path. [RFC3988] defines the mechanism for automatically discovering the Path MTU of LSPs. For a certain FEC, the LSR compares the MTU advertised by all downstream devices with the MTU of the FEC output interface in the local device, and calculates the minimum value for the upstream device.

[RFC3209] specify the mechanism of MTU signaling in RSVP-TE. The ingress node of the RSVP-TE tunnel sends a Path message to the downstream device. The Adspec object in the Path message carries the MTU. Each node along the tunnel receives a Path message, compares the MTU value in the Adspec object with the interface MTU value and MPLS MTU configured on the physical output interface of the local tunnel, obtains the minimum MTU value, and puts it into the newly constructed Path message and continues to send it to the downstream equipment. Thus, the MTU carried in the Path message received by the Egress node is the minimum value of the path MTU. The Egress node brings the negotiated Path MTU back to the Ingress node through the Resv message.

Segment Routing (SR) described in [RFC8402] leverages the source routing paradigm. Segment Routing can be directly applied to the MPLS architecture with no change on the forwarding plane [RFC8660] and applied to the IPv6 architecture with a new type of routing header called the SR header (SRH) [RFC8754].

[I-D.ietf-idr-bgp-ls-segment-routing-ext] defines SR extensions to BGP-LS and specifies the TLVs and sub-TLVs for advertising SR information. Based on the SR information reported by the BGP-LS, the SDN can calculate the end-to-end explicit SR-TE paths or SR Policies.

Nevertheless, Segment Routing is a tunneling technology based on the IGP protocol as the control protocol, and there is no additional signaling for establishing the path. so the Segment Routing tunnel cannot currently support the negotiation mechanism of the MTU. Multiple labels or SRv6 SIDs are pushed in the packets. This causes

the length of the packets encapsulated in the Segment Routing tunnel to increase during packet forwarding. This is more likely to cause packet size exceed the traditional MPLS packet size.

This document specify the extension to BGP Link State (BGP-LS) to carry link maximum transmission unit (MTU) messages.

## 2. Terminology

This draft refers to the terms defined in [RFC8201], [RFC4821] and [RFC3988].

**MTU:** Maximum Transmission Unit, the size in bytes of the largest IP packet, including the IP header and payload, that can be transmitted on a link or path. Note that this could more properly be called the IP MTU, to be consistent with how other standards organizations use the acronym MTU.

**Link MTU:** The Maximum Transmission Unit, i.e., maximum IP packet size in bytes, that can be conveyed in one piece over a link. Be aware that this definition is different from the definition used by other standards organizations.

For IETF documents, link MTU is uniformly defined as the IP MTU over the link. This includes the IP header, but excludes link layer headers and other framing that is not part of IP or the IP payload.

Be aware that other standards organizations generally define link MTU to include the link layer headers.

For the MPLS data plane, this size includes the IP header and data (or other payload) and the label stack but does not include any lower-layer headers. A link may be an interface (such as Ethernet or Packet-over-SONET), a tunnel (such as GRE or IPsec), or an LSP.

**Path:** The set of links traversed by a packet between a source node and a destination node.

**Path MTU, or PMTU:** The minimum link MTU of all the links in a path between a source node and a destination node.

### 3. Deploying scenarios

This document suggests a solution to extension to BGP Link State (BGP-LS) to carry maximum transmission unit (MTU) messages. The MTU information of the link is acquired through the process of collecting link state and TE information by BGP-LS. Concretely, a router maintains one or more databases for storing link-state information about nodes and links in any given area. The router's BGP process can retrieve topology from these IGP, BGP and other sources, and distribute it to a consumer, either directly or via a peer BGP speaker (typically a dedicated Route Reflector). [RFC7176] specifies a possible way of using the ISIS mechanism and extensions for link MTU Sub-TLV. In the case of inter-AS scenario (e.g., BGP EPE), the link MTU of the inter-AS link can be collected via BGP-LS directly.

As per [RFC7752], the collection of link-state and TE information and its distribution to consumers is shown in the following figure.

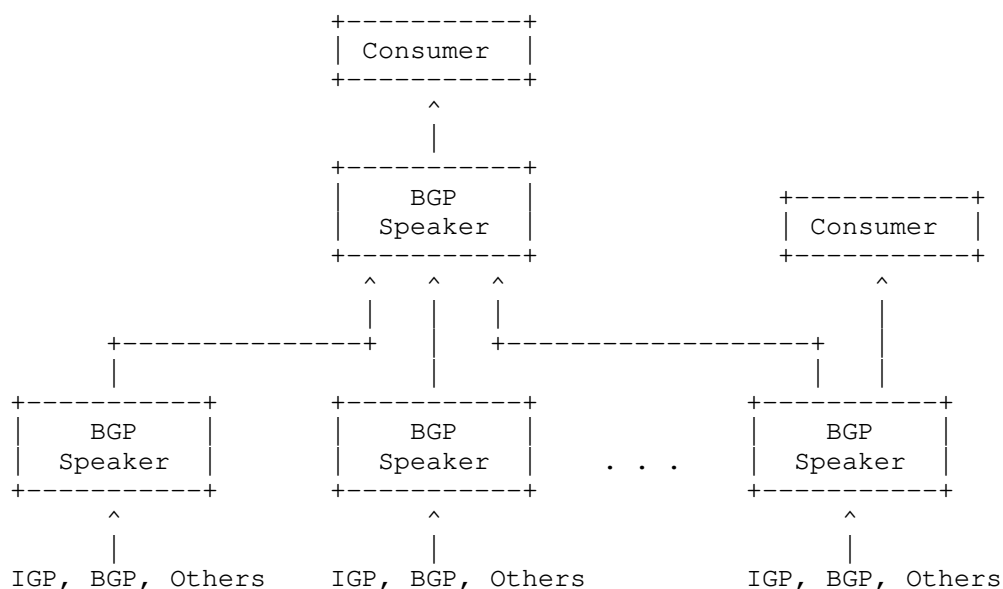


Figure 1: Collection of Link-State and TE Information

Please note that this signaled MTU may be different from the actual MTU, which is usually from configuration mismatches in a control plane and a data plane component.

#### 4. BGP\_LS Extensions for Link MTU

[RFC7752] defines the BGP-LS NLRI that can be a Node NLRI, a Link NLRI or a Prefix NLRI. The corresponding BGP-LS attribute is a Node Attribute, a Link Attribute or a Prefix Attribute. [RFC7752] defines the TLVs that map link-state information to BGP-LS NLRI and the BGP-LS attribute. Therefore, according to this document, a new sub-TLV is added to the Link Attribute TLV. It is an independent attribute TLV that can be used for the link NLRI advertised with all the Protocol IDs.

The format of the sub-TLV is as shown below.

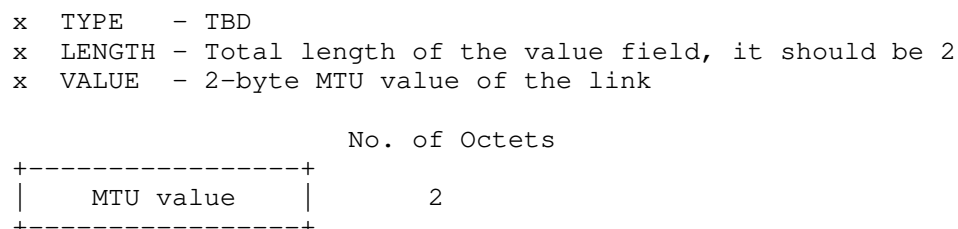


Figure 2. Sub-TLV Format for Link MTU

Whenever there is a change in MTU value represented by Link Attribute TLV, BGP-LS should re-originate the respective TLV with the new MTU value.

#### 5. IANA Considerations

This document requests assigning a new code-point from the BGP-LS Link Descriptor and Attribute TLVs registry as specified in section 4.

| Value | Description | Reference     |
|-------|-------------|---------------|
| TBD   | Link MTU    | This document |

#### 6. Security Considerations

This document does not introduce security issues beyond those discussed in RFC7752.

## 7. Acknowledgements

## 8. Contributors

Gang Yan  
Huawei  
China

Email:yangang@huawei.com

Junda Yao  
Huawei  
China

Email:yaojunda@huawei.com

## 9. References

### 9.1. Normative References

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

### 9.2. Informative References

- [I-D.ietf-idr-bgp-ls-segment-routing-ext] Previdi, S., Talaulikar, K., Filsfils, C., Gredler, H., and M. Chen, "BGP Link-State extensions for Segment Routing", draft-ietf-idr-bgp-ls-segment-routing-ext-16 (work in progress), June 2019.
- [RFC1191] Mogul, J. and S. Deering, "Path MTU discovery", RFC 1191, DOI 10.17487/RFC1191, November 1990, <<https://www.rfc-editor.org/info/rfc1191>>.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, DOI 10.17487/RFC3209, December 2001, <<https://www.rfc-editor.org/info/rfc3209>>.
- [RFC3988] Black, B. and K. Kompella, "Maximum Transmission Unit Signalling Extensions for the Label Distribution Protocol", RFC 3988, DOI 10.17487/RFC3988, January 2005, <<https://www.rfc-editor.org/info/rfc3988>>.

- [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>>.
- [RFC4821] Mathis, M. and J. Heffner, "Packetization Layer Path MTU Discovery", RFC 4821, DOI 10.17487/RFC4821, March 2007, <<https://www.rfc-editor.org/info/rfc4821>>.
- [RFC7176] Eastlake 3rd, D., Senevirathne, T., Ghanwani, A., Dutt, D., and A. Banerjee, "Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS", RFC 7176, DOI 10.17487/RFC7176, May 2014, <<https://www.rfc-editor.org/info/rfc7176>>.
- [RFC7752] Gredler, H., Ed., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and Traffic Engineering (TE) Information Using BGP", RFC 7752, DOI 10.17487/RFC7752, March 2016, <<https://www.rfc-editor.org/info/rfc7752>>.
- [RFC8201] McCann, J., Deering, S., Mogul, J., and R. Hinden, Ed., "Path MTU Discovery for IP version 6", STD 87, RFC 8201, DOI 10.17487/RFC8201, July 2017, <<https://www.rfc-editor.org/info/rfc8201>>.
- [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>>.
- [RFC8660] Bashandy, A., Ed., Filsfils, C., Ed., Previdi, S., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing with the MPLS Data Plane", RFC 8660, DOI 10.17487/RFC8660, December 2019, <<https://www.rfc-editor.org/info/rfc8660>>.
- [RFC8754] Filsfils, C., Ed., Dukes, D., Ed., Previdi, S., Leddy, J., Matsushima, S., and D. Voyer, "IPv6 Segment Routing Header (SRH)", RFC 8754, DOI 10.17487/RFC8754, March 2020, <<https://www.rfc-editor.org/info/rfc8754>>.

Authors' Addresses

Yongqing Zhu  
China Telecom  
109, West Zhongshan Road, Tianhe District.  
Guangzhou 510000  
China

Email: zhuyq8@chinatelecom.cn

Zhibo Hu  
Huawei Technologies  
Huawei Bld., No.156 Beiqing Rd.  
Beijing 100095  
China

Email: huzhibo@huawei.com

Shuping Peng  
Huawei Technologies  
Huawei Bld., No.156 Beiqing Rd.  
Beijing 100095  
China

Email: pengshuping@huawei.com

Robbins Mwehaire  
MTN Uganda Ltd.  
Uganda

Email: Robbins.Mwehair@mtn.com