

Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: September 17, 2016

J. Dong  
Z. Li  
Huawei Technologies  
J. Tantsura  
Ericsson  
H. Gredler  
Private Contributor  
March 16, 2016

BGP Link-State Extension for Distribution of IP Tunnel Information  
draft-dong-idr-ls-ip-tunnel-00

Abstract

This document specifies extensions to BGP-LS for the collection and distribution of IP tunnel information. Such information can be distributed to external components for service mapping and tunnel selection.

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 <http://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 17, 2016.

Copyright Notice

Copyright (c) 2016 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 . . . . .	2
2. Carrying IP Tunnel Information in BGP-LS . . . . .	3
2.1. IP Tunnel Identifier Information . . . . .	3
2.2. IP Tunnel Parameters TLV . . . . .	6
3. Operational Considerations . . . . .	8
4. IANA Considerations . . . . .	8
4.1. BGP-LS NLRI-Types . . . . .	8
4.2. BGP-LS Protocol-IDs . . . . .	9
4.3. BGP-LS Attribute TLVs . . . . .	9
5. Security Considerations . . . . .	9
6. Acknowledgements . . . . .	10
7. References . . . . .	10
7.1. Normative References . . . . .	10
7.2. Informative References . . . . .	11
Authors' Addresses . . . . .	11

## 1. Introduction

BGP has been extended to distribute the link-state [I-D.ietf-idr-ls-distribution] and TE-LSP information [I-D.ietf-idr-te-lsp-distribution] to external components. When IP tunnel technologies, such as Generic Routing Encapsulation (GRE), Layer Two Tunneling Protocol - Version 3 (L2TPv3), VxLAN, NVGRE, etc., are used in the network, it is necessary to collect the information of IP tunnels in the network and share with the external components. Such information can be distributed to external components for service mapping and tunnel selection. One typical use case of IP tunnel information is described in [I-D.hao-idr-flowspec-redirect-tunnel]. This document specifies extensions to BGP-LS for the collection and distribution of IP tunnel information.

## 2. Carrying IP Tunnel Information in BGP-LS

### 2.1. IP Tunnel Identifier Information

The IP tunnel Identifier information is advertised in BGP UPDATE messages using the MP\_REACH\_NLRI and MP\_UNREACH\_NLRI attributes [RFC4760]. The "Link-State NLRI" defined in [I-D.ietf-idr-ls-distribution] is extended to carry the IP Tunnel information. BGP speakers that wish to exchange IP Tunnel information MUST use the BGP Multiprotocol Extensions Capability Code (1) to advertise the corresponding (AFI, SAFI) pair, as specified in [RFC4760].

The format of "Link-State NLRI" is defined in [I-D.ietf-idr-ls-distribution]. A new "NLRI Type" is defined for IP Tunnel Identifier Information as following:

- o NLRI Type = TBA: IPv4/IPv6 Tunnel NLRI

The IPv4/IPv6 Tunnel NLRI is shown in the following figure:

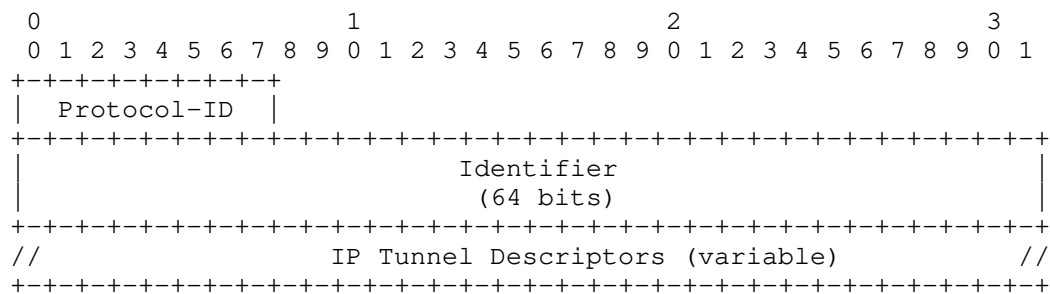


Figure 1. IPv4/IPv6 Tunnel NLRI

Where

The 'Protocol-ID' field is used to identify the source of the advertised NLRI. For IPv4/IPv6 Tunnel NLRI, according to the method of tunnel establishment, the Protocol-ID field can be set to either "Static configuration" or the specific signaling protocol of the IP tunnel. Several new Protocol-IDs are defined as below:

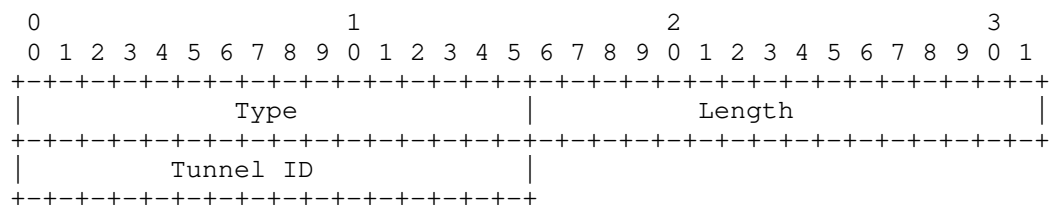
Protocol-ID	NLRI information source protocol
TBD	L2TPv3
TBD	GTPv2-C

As defined in [I-D.ietf-idr-ls-distribution], the 64-Bit 'Identifier' field is used to identify the "routing universe" where the NLRI belongs.

The "IP Tunnel Descriptors" field consists of a set of Descriptor TLVs which together identifies the IP tunnel. The following Descriptor TLVs as defined in [I-D.ietf-idr-te-lsp-distribution] are reused for IPv4/IPv6 Tunnel NLRI:

o Tunnel ID

The Tunnel Identifier TLV contains the Tunnel ID defined in [RFC3209] and has the following format:

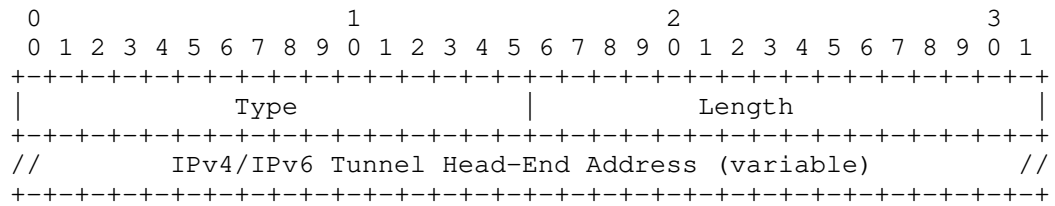


where:

- + Type: To be assigned by IANA (suggested value: 267)
- + Length: 2 octets.
- + Tunnel ID: 2 octets as defined in [RFC3209].

o IPv4/6 Tunnel Head-end address

The IPv4/IPv6 Tunnel Head-End Address TLV contains the Tunnel Head- End Address defined in [RFC3209] and has following format:



where:

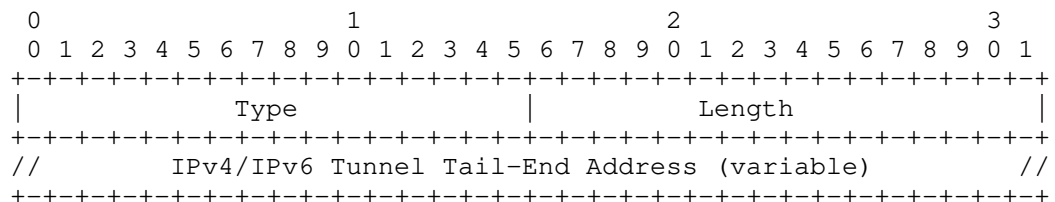
- + Type: To be assigned by IANA (suggested value: 269)
- + Length: 4 or 16 octets.

When the IPv4/IPv6 Tunnel Head-end Address TLV contains an IPv4 address, its length is 4 (octets).

When the IPv4/IPv6 Tunnel Head-end Address TLV contains an IPv6 address, its length is 16 (octets).

#### o IPv4/6 Tunnel Tail-end address

The IPv4/IPv6 Tunnel Tail-End Address TLV contains the Tunnel Tail- End Address defined in [RFC3209] and has following format:



where:

- + Type: To be assigned by IANA (suggested value: 270)
- + Length: 4 or 16 octets.

When the IPv4/IPv6 Tunnel Tail-end Address TLV contains an IPv4 address, its length is 4 (octets).

When the IPv4/IPv6 Tunnel Tail-end Address TLV contains an IPv6 address, its length is 16 (octets).

In addition, a new descriptor TLV called "Tunnel Type TLV" is defined for IP Tunnel as below:

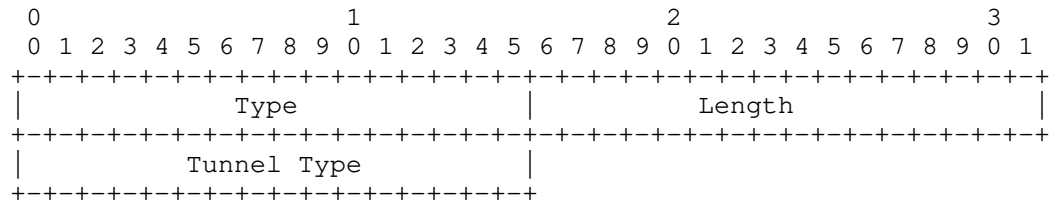


Figure 2. Tunnel Type TLV

- o Type: TBA.
- o Length: 2 octets.
- o Value: The 2-octet Tunnel Type identifies the type of tunneling technology as defined in the "BGP Tunnel Encapsulation Attribute Tunnel Types" registry [RFC5512].

The IPv4/6 Tunnel Head-end address TLV, IPv4/6 Tunnel Tail-end address, Tunnel-type TLV and the Tunnel ID TLV together uniquely identify the IP tunnel.

## 2.2. IP Tunnel Parameters TLV

A new TLV called "IP Tunnel Parameters TLV" is defined to describe the detailed information of the IP tunnels, which is carried in the optional non-transitive BGP Attribute "LINK\_STATE Attribute" defined in [I-D.ietf-idr-ls-distribution]. The IP Tunnel Parameters TLV SHOULD only be used with IPv4/IPv6 Tunnel NLRI.

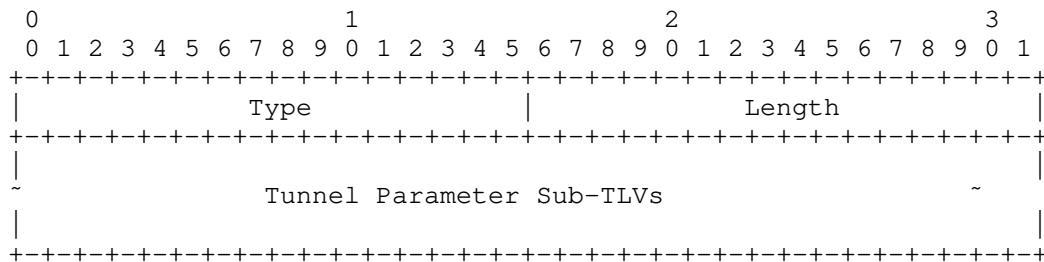
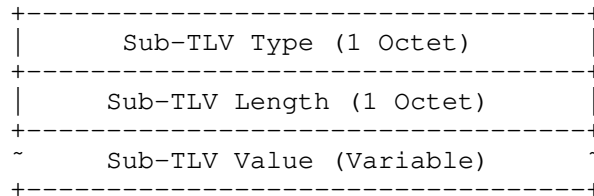


Figure 3. IP Tunnel Parameters TLV

The "Value" field of the IP Tunnel Parameters TLV is composed of a set of sub-TLVs. The sub-TLV is structured as below:



The following sub-TLVs defined in this document can be carried in the Value field of the IP Tunnel Parameters TLV:

- o Tunnel Name sub-TLV:

Type: TBA

Length: Variable

Value: A string identifies the name of the IP tunnel.

- o Description sub-TLV

Type: TBA

Length: Variable

Value: A string which contains the textual description of the IP tunnel.

- o Status sub-TLV:

Type: TBA

Length: 1 octet

Value: 8-bit flags which indicate the status of the IP tunnel. Bit 0 is defined as the Up/Down bit, which SHOULD be set to 1 if there is no available route for the tunnel destination. The other bits are reserved which MUST be set to 0 on transmission and ignored on receipt.

- o Encapsulation sub-TLV:

Type: TBA

Length: Variable

Value: The encapsulation information of the IP tunnel, syntax and semantics of which are determined by the Tunnel Type. The format of Encapsulation sub-TLVs are defined in [RFC5512] and [I-D.ietf-idr-tunnel-encaps].

- o CoS Sub-TLV:

Type: TBA

Length: 1 octet

Value: the class of differentiated services that can be provided by the tunnel. The format is same as the DS field as defined in [RFC2474].

- o MTU sub-TLV:

Type: TBA

Length: 2 octets

Value: the Maximum Transmission Unit (MTU) of the IP tunnel.

### 3. Operational Considerations

The Existing BGP operational procedures apply to this document. No new operation procedures are defined in this document. The operational considerations as specified in [I-D.ietf-idr-ls-distribution] apply to this document.

In general the ingress nodes of the IP Tunnels are responsible for the distribution of the IP tunnel information, while the egress nodes of the IP tunnels MAY report the IP tunnel information if needed.

### 4. IANA Considerations

IANA is requested to administer the assignment of new values defined in this document and summarized in this section.

#### 4.1. BGP-LS NLRI-Types

IANA maintains a registry called "Border Gateway Protocol - Link State (BGP-LS) Parameters" with a sub-registry called "BGP-LS NLRI-Types".

IANA is requested to assign two new NLRI-Types:



Type	NLRI Type	Reference
TBD	IPv4/v6 Tunnel NLRI	this document

#### 4.2. BGP-LS Protocol-IDs

IANA maintains a registry called "Border Gateway Protocol - Link State (BGP-LS) Parameters" with a sub-registry called "BGP-LS Protocol-IDs".

IANA is requested to assign two new Protocol-IDs:

Protocol-ID	NLRI information source protocol	Reference
TBD	L2TPv3	this document
TBD	GTPv2-C	this document

#### 4.3. BGP-LS Attribute 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".

IANA is requested to assign one new TLV code point:

TLV Code Point	Description	IS-IS TLV/ Sub-TLV	Value defined in:
TBD	IP Tunnel Parameters TLV	---	this document

#### 5. 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.

## 6. Acknowledgements

The authors would like to thank Nan Wu, Shunwan Zhuang and Xia Chen for their review and valuable comments.

## 7. References

### 7.1. Normative References

- [I-D.ietf-idr-ls-distribution]  
Gredler, H., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and TE Information using BGP", draft-ietf-idr-ls-distribution-13 (work in progress), October 2015.
- [I-D.ietf-idr-te-lsp-distribution]  
Dong, J., Chen, M., Gredler, H., Previdi, S., and J. Tantsura, "Distribution of MPLS Traffic Engineering (TE) LSP State using BGP", draft-ietf-idr-te-lsp-distribution-04 (work in progress), December 2015.
- [I-D.ietf-idr-tunnel-encaps]  
Rosen, E., Patel, K., and G. Velde, "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-01 (work in progress), December 2015.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<http://www.rfc-editor.org/info/rfc2119>>.
- [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, <<http://www.rfc-editor.org/info/rfc3209>>.
- [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, <<http://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, <<http://www.rfc-editor.org/info/rfc4760>>.

## 7.2. Informative References

- [I-D.hao-idr-flowspec-redirect-tunnel]  
Weiguo, H., Li, Z., and L. Yong, "BGP Flow-Spec Redirect to Tunnel action", draft-hao-idr-flowspec-redirect-tunnel-00 (work in progress), October 2015.
- [RFC2474] Nichols, K., Blake, S., Baker, F., and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474, DOI 10.17487/RFC2474, December 1998, <<http://www.rfc-editor.org/info/rfc2474>>.
- [RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis", RFC 4272, DOI 10.17487/RFC4272, January 2006, <<http://www.rfc-editor.org/info/rfc4272>>.
- [RFC5342] Eastlake 3rd, D., "IANA Considerations and IETF Protocol Usage for IEEE 802 Parameters", RFC 5342, DOI 10.17487/RFC5342, September 2008, <<http://www.rfc-editor.org/info/rfc5342>>.
- [RFC5512] Mohapatra, P. and E. Rosen, "The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", RFC 5512, DOI 10.17487/RFC5512, April 2009, <<http://www.rfc-editor.org/info/rfc5512>>.
- [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, <<http://www.rfc-editor.org/info/rfc6952>>.

## Authors' Addresses

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

Email: [jie.dong@huawei.com](mailto:jie.dong@huawei.com)

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

Email: lizhenbin@huawei.com

Jeff Tantsura  
Ericsson  
300 Holger Way  
San Jose, CA 95134  
US

Email: jeff.tantsura@ericsson.com

Hannes Gredler  
Private Contributor

Email: hannes@gredler.at

Inter-Domain Routing  
Internet-Draft  
Intended status: Standards Track  
Expires: May 3, 2017

S. Previdi, Ed.  
P. Psenak  
C. Filsfils  
Cisco Systems, Inc.  
H. Gredler  
RtBrick Inc.  
M. Chen  
Huawei Technologies  
J. Tantsura  
Individual  
October 30, 2016

BGP Link-State extensions for Segment Routing  
draft-gredler-idr-bgp-ls-segment-routing-ext-04

Abstract

Segment Routing (SR) allows for a flexible definition of end-to-end paths within IGP topologies by encoding paths as sequences of topological sub-paths, called "segments". These segments are advertised by the link-state routing protocols (IS-IS, OSPF and OSPFv3).

This draft defines extensions to the BGP Link-state address-family in order to carry segment information via BGP.

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 <http://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 3, 2017.

#### Copyright Notice

Copyright (c) 2016 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. BGP-LS Extensions for Segment Routing . . . . .	5
2.1. Node Attributes TLVs . . . . .	5
2.1.1. SR-Capabilities TLV . . . . .	5
2.1.2. SR-Algorithm TLV . . . . .	6
2.1.3. SR Local Block TLV . . . . .	7
2.1.4. SRMS Preference TLV . . . . .	7
2.2. Link Attribute TLVs . . . . .	8
2.2.1. Adjacency SID TLV . . . . .	9
2.2.2. LAN Adjacency SID TLV . . . . .	9
2.3. Prefix Attribute TLVs . . . . .	10
2.3.1. Prefix-SID TLV . . . . .	11
2.3.2. IPv6 Prefix-SID TLV . . . . .	12
2.3.3. IGP Prefix Attributes TLV . . . . .	13
2.3.4. Source Router Identifier (Source Router-ID) TLV . . . . .	14
2.3.5. Range TLV . . . . .	14
2.3.6. Binding SID TLV . . . . .	15
2.3.7. Binding SID SubTLVs . . . . .	16
2.4. Equivalent IS-IS Segment Routing TLVs/Sub-TLVs . . . . .	22
2.5. Equivalent OSPF/OSPFv3 Segment Routing TLVs/Sub-TLVs . . . . .	23
3. Procedures . . . . .	25
3.1. Advertisement of a IS-IS Prefix SID TLV . . . . .	25
3.2. Advertisement of a OSPF/OSPFv3 Prefix-SID TLV . . . . .	25
3.3. Advertisement of a range of prefix-to-SID mappings in OSPF . . . . .	26
3.4. Advertisement of a range of IS-IS SR bindings . . . . .	26
3.5. Advertisement of a path and its attributes from IS-IS protocol . . . . .	26
3.6. Advertisement of a path and its attributes from	

OSPFv2/OSPFv3 protocol . . . . .	27
4. IANA Considerations . . . . .	27
4.1. TLV/Sub-TLV Code Points Summary . . . . .	27
5. Manageability Considerations . . . . .	28
5.1. Operational Considerations . . . . .	28
5.1.1. Operations . . . . .	28
6. Security Considerations . . . . .	29
7. Contributors . . . . .	29
8. Acknowledgements . . . . .	29
9. References . . . . .	29
9.1. Normative References . . . . .	29
9.2. Informative References . . . . .	30
9.3. URIs . . . . .	31
Authors' Addresses . . . . .	34

## 1. Introduction

Segment Routing (SR) allows for a flexible definition of end-to-end paths by combining sub-paths called "segments". A segment can represent any instruction, topological or service-based. A segment can have a local semantic to an SR node or global within a domain. Within IGP topologies an SR path is encoded as a sequence of topological sub-paths, called "IGP segments". These segments are advertised by the link-state routing protocols (IS-IS, OSPF and OSPFv3).

Two types of IGP segments are defined, Prefix segments and Adjacency segments. Prefix segments, by default, represent an ECMP-aware shortest-path to a prefix, as per the state of the IGP topology. Adjacency segments represent a hop over a specific adjacency between two nodes in the IGP. A prefix segment is typically a multi-hop path while an adjacency segment, in most of the cases, is a one-hop path. [I-D.ietf-spring-segment-routing].

When Segment Routing is enabled in a IGP domain, segments are advertised in the form of Segment Identifiers (SIDs). The IGP link-state routing protocols have been extended to advertise SIDs and other SR-related information. IGP extensions are described in: IS-IS [I-D.ietf-isis-segment-routing-extensions], OSPFv2 [I-D.ietf-ospf-segment-routing-extensions] and OSPFv3 [I-D.ietf-ospf-ospfv3-segment-routing-extensions]. Using these extensions, Segment Routing can be enabled within an IGP domain.

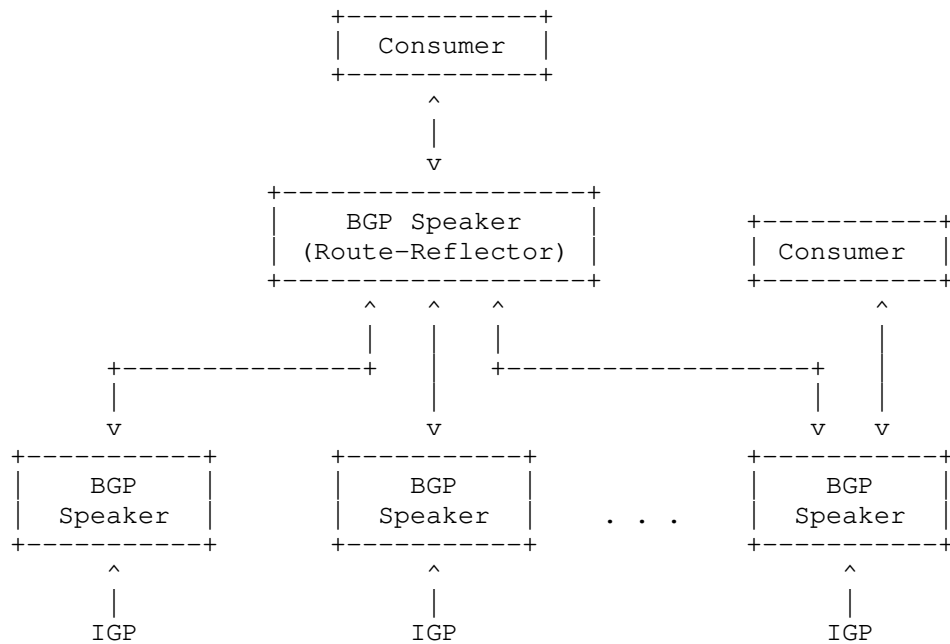


Figure 1: Link State info collection

Segment Routing (SR) allows advertisement of single or multi-hop paths. The flooding scope for the IGP extensions for Segment routing is IGP area-wide. Consequently, the contents of a Link State Database (LSDB) or a Traffic Engineering Database (TED) has the scope of an IGP area and therefore, by using the IGP alone it is not enough to construct segments across multiple IGP Area or AS boundaries.

In order to address the need for applications that require topological visibility across IGP areas, or even across Autonomous Systems (AS), the BGP-LS address-family/sub-address-family have been defined to allow BGP to carry Link-State information. The BGP Network Layer Reachability Information (NLRI) encoding format for BGP-LS and a new BGP Path Attribute called the BGP-LS attribute are defined in [RFC7752]. The identifying key of each Link-State object, namely a node, link, or prefix, is encoded in the NLRI and the properties of the object are encoded in the BGP-LS attribute. Figure Figure 1 describes a typical deployment scenario. In each IGP area, one or more nodes are configured with BGP-LS. These BGP speakers form an IBGP mesh by connecting to one or more route-reflectors. This way, all BGP speakers (specifically the route-reflectors) obtain Link-State information from all IGP areas (and from other ASes from EBGP peers). An external component connects to the route-reflector to obtain this information (perhaps moderated by



a policy regarding what information is or isn't advertised to the external component).

This document describes extensions to BGP-LS to advertise the SR information. An external component (e.g., a controller) then can collect SR information in the "northbound" direction across IGP areas or ASes and construct the end-to-end path (with its associated SIDs) that need to be applied to an incoming packet to achieve the desired end-to-end forwarding.

## 2. BGP-LS Extensions for Segment Routing

This document defines IGP SR extensions BGP-LS TLVs and Sub-TLVs. Section 2.4 and Section 2.5 illustrates the equivalent TLVs and Sub-TLVs in IS-IS, OSPF and OSPFv3 protocols.

BGP-LS [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. BGP-LS [RFC7752] defines the TLVs that map link-state information to BGP-LS NLRI and the BGP-LS attribute. This document adds additional BGP-LS attribute TLVs in order to encode SR information.

### 2.1. Node Attributes TLVs

The following Node Attribute TLVs are defined:

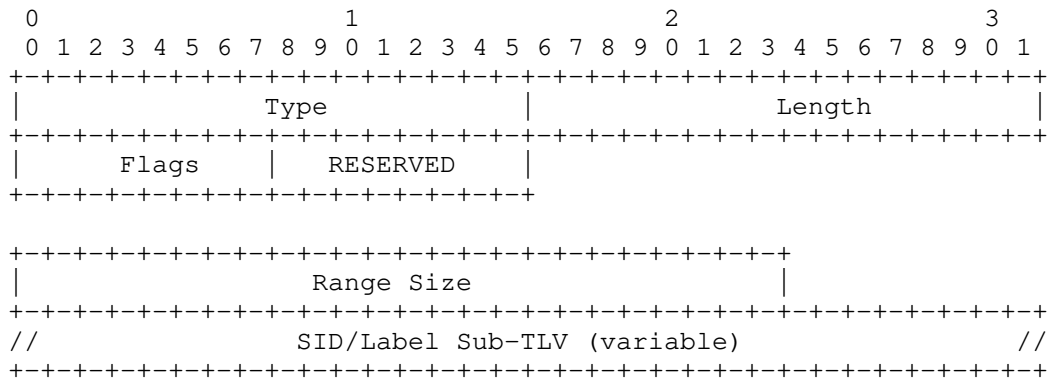
TLV Code Point	Description	Length	Section
1034	SR Capabilities	variable	Section 2.1.1
1035	SR Algorithm	variable	Section 2.1.2
1036	SR Local Block	variable	Section 2.1.3
1037	SRMS Preference	variable	Section 2.1.4

Table 1: Node Attribute TLVs

These TLVs can ONLY be added to the Node Attribute associated with the Node NLRI that originates the corresponding SR TLV.

#### 2.1.1. SR-Capabilities TLV

The SR Capabilities sub-TLV has following format:



Type: TBD, suggested value 1034.

Length: Variable.

Flags: 1 octet of flags as defined in  
 [I-D.ietf-isis-segment-routing-extensions] and  
 [I-D.ietf-ospf-ospfv3-segment-routing-extensions].

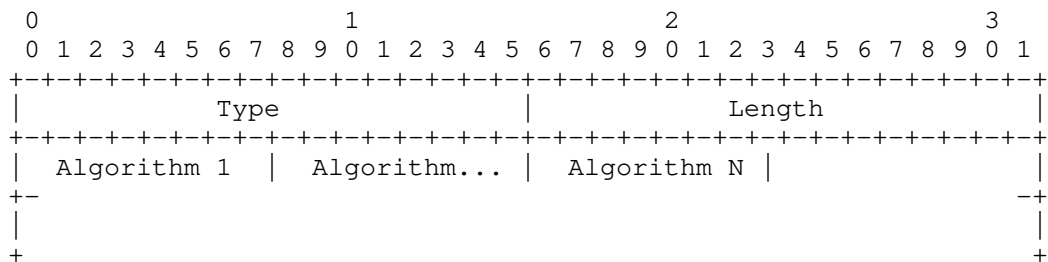
One or more entries, each of which have the following format:

Range Size: 3 octet value indicating the number of labels in  
 the range.

SID/Label sub-TLV (as defined in Section 2.3.7.2).

#### 2.1.2. SR-Algorithm TLV

The SR-Algorithm TLV has the following format:



where:

Type: TBD, suggested value 1035.

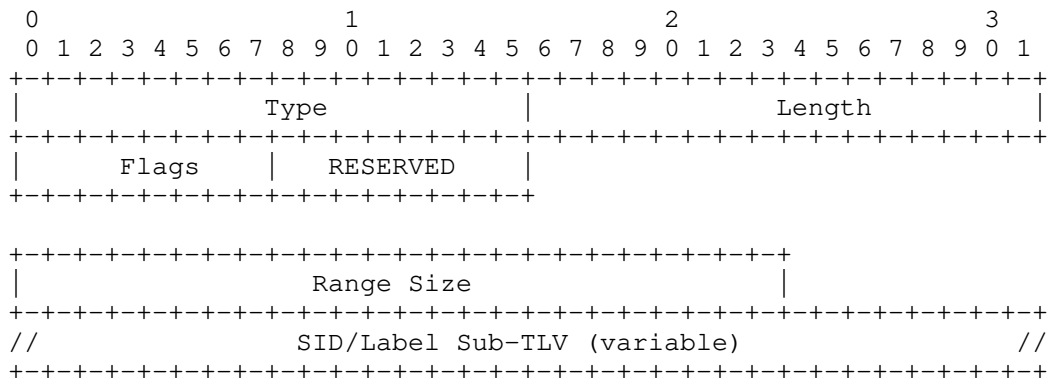
Length: Variable.

Algorithm: 1 octet identifying the algorithm.

### 2.1.3. SR Local Block TLV

The SR Local Block (SRLB) Sub-TLV contains the range of labels the node has reserved for local SIDs. Local SIDs are used, e.g., in IGP (IS-IS, OSPF) for Adjacency-SIDs, and may also be allocated by other components than IGP protocols. As an example, an application or a controller may instruct a node to allocate a specific local SID. Therefore, in order for such applications or controllers to know the range of local SIDs available, it is required that the node advertises its SRLB.

The SRLB TLV has the following format:



Type: TBD, suggested value 1036.

Length: Variable.

Flags: 1 octet of flags. None are defined at this stage.

One or more entries, each of which have the following format:

Range Size: 3 octet value indicating the number of labels in the range.

SID/Label sub-TLV (as defined in Section 2.3.7.2).

### 2.1.4. SRMS Preference TLV

The Segment Routing Mapping Server (SRMS) Preference sub-TLV is used in order to associate a preference with SRMS advertisements from a particular source.

The SRMS Preference sub-TLV has following format:

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|  Type      |      Length      | Preference      |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

Type: TBD, suggested value 1037.

Length: 1.

Preference: 1 octet. Unsigned 8 bit SRMS preference.

The use of the SRMS Preference TLV is defined in [I-D.ietf-isis-segment-routing-extensions].

## 2.2. Link Attribute TLVs

The following Link Attribute TLVs are are defined:

TLV Code Point	Description	Length	Section
1099	Adjacency Segment Identifier (Adj-SID) TLV	variable	Section 2.2.1
1100	LAN Adjacency Segment Identifier (Adj-SID) TLV	variable	Section 2.2.2

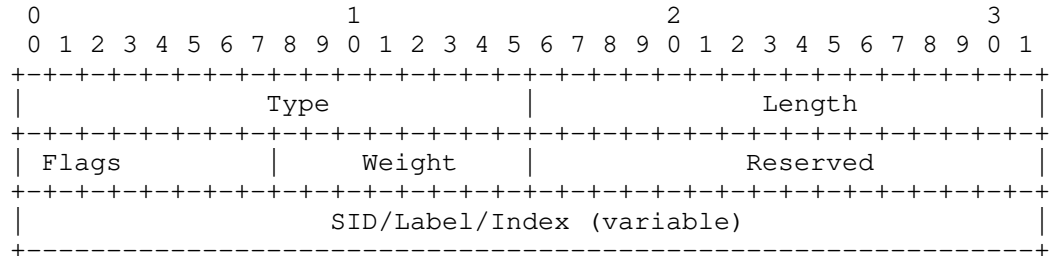
Table 2: Link Attribute TLVs

These TLVs can ONLY be added to the Link Attribute associated with the link whose local node originates the corresponding TLV.

For a LAN, normally a node only announces its adjacency to the IS-IS pseudo-node (or the equivalent OSPF Designated and Backup Designated Routers) [I-D.ietf-isis-segment-routing-extensions]. The LAN Adjacency Segment TLV allows a node to announce adjacencies to all other nodes attached to the LAN in a single instance of the BGP-LS Link NLRI. Without this TLV, the corresponding BGP-LS link NLRI would need to be originated for each additional adjacency in order to advertise the SR TLVs for these neighbor adjacencies.

### 2.2.1. Adjacency SID TLV

The Adjacency SID (Adj-SID) TLV has the following format:



where:

Type: TBD, suggested value 1099.

Length: Variable.

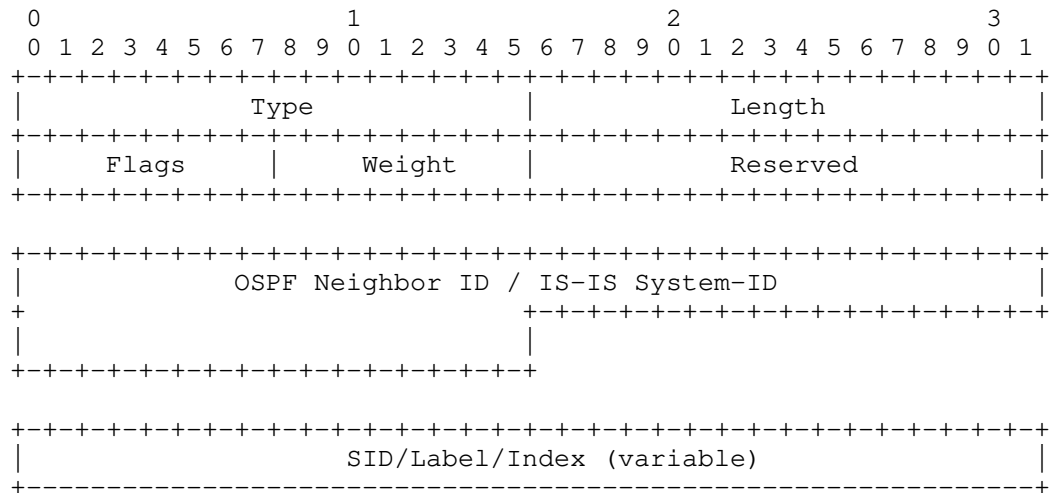
Flags. 1 octet field of following flags as defined in [I-D.ietf-isis-segment-routing-extensions], [I-D.ietf-ospf-segment-routing-extensions] and [I-D.ietf-ospf-ospfv3-segment-routing-extensions].

Weight: Weight used for load-balancing purposes.

SID/Index/Label: Label or index value depending on the flags setting as defined in [I-D.ietf-isis-segment-routing-extensions], [I-D.ietf-ospf-segment-routing-extensions] and [I-D.ietf-ospf-ospfv3-segment-routing-extensions].

### 2.2.2. LAN Adjacency SID TLV

The LAN Adjacency SID (LAN-Adj-SID-SID) has the following format:



where:

Type: TBD, suggested value 1100.

Length: Variable.

Flags. 1 octet field of following flags as defined in [I-D.ietf-isis-segment-routing-extensions], [I-D.ietf-ospf-segment-routing-extensions] and [I-D.ietf-ospf-ospfv3-segment-routing-extensions].

Weight: Weight used for load-balancing purposes.

SID/Index/Label: Label or index value depending on the flags setting as defined in [I-D.ietf-isis-segment-routing-extensions], [I-D.ietf-ospf-segment-routing-extensions] and [I-D.ietf-ospf-ospfv3-segment-routing-extensions].

### 2.3. Prefix Attribute TLVs

The following Prefix Attribute TLVs and Sub-TLVs are defined:

TLV Code Point	Description	Length	Section
1158	Prefix SID	variable	Section 2.3.1
1159	Range	variable	Section 2.3.5
1160	Binding SID	variable	Section 2.3.6
1169	IPv6 Prefix SID	variable	Section 2.3.2
1170	IGP Prefix Attributes	variable	Section 2.3.3
1171	Source Router-ID	variable	Section 2.3.4

Table 3: Prefix Attribute TLVs

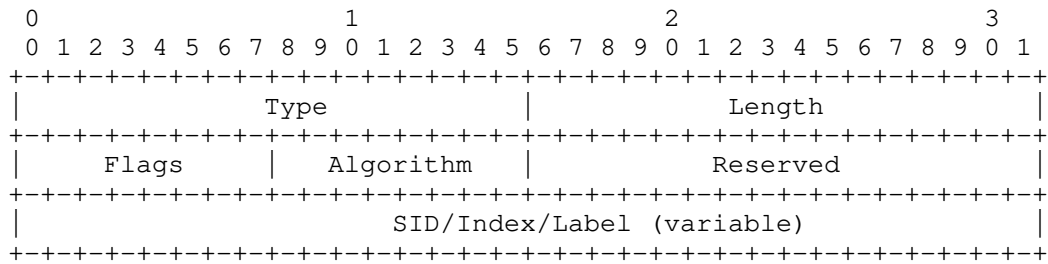
TLV Code Point	Description	Length	Section
1161	SID/Label TLV	variable	Section 2.3.7.2
1162	ERO Metric TLV	4 octets	Section 2.3.7.3
1163	IPv4 ERO TLV	8 octets	Section 2.3.7.4
1164	IPv6 ERO TLV	20 octets	Section 2.3.7.5
1165	Unnumbered Interface ID ERO TLV	12	Section 2.3.7.6
1166	IPv4 Backup ERO TLV	8 octets	Section 2.3.7.7
1167	IPv6 Backup ERO TLV	10 octets	Section 2.3.7.8
1168	Unnumbered Interface ID Backup ERO TLV	12	Section 2.3.7.9

Table 4: Prefix Attribute - Binding SID Sub-TLVs

### 2.3.1. Prefix-SID TLV

The Prefix-SID TLV can ONLY be added to the Prefix Attribute whose local node in the corresponding Prefix NLRI is the node that originates the corresponding SR TLV.

The Prefix-SID has the following format:



where:

Type: TBD, suggested value 1158.

Length: Variable

Algorithm: 1 octet value identify the algorithm.

SID/Index/Label: Label or index value depending on the flags setting as defined in [I-D.ietf-isis-segment-routing-extensions], [I-D.ietf-ospf-segment-routing-extensions] and [I-D.ietf-ospf-ospfv3-segment-routing-extensions].

The Prefix-SID TLV includes a Flags field. In the context of BGP-LS, the Flags field format and the semantic of each individual flag MUST be taken from the corresponding source protocol (i.e.: the protocol of origin of the Prefix-SID being advertised in BGP-LS).

IS-IS Prefix-SID flags are defined in [I-D.ietf-isis-segment-routing-extensions] section 2.1.

OSPF Prefix-SID flags are defined in [I-D.ietf-ospf-segment-routing-extensions] section 5.

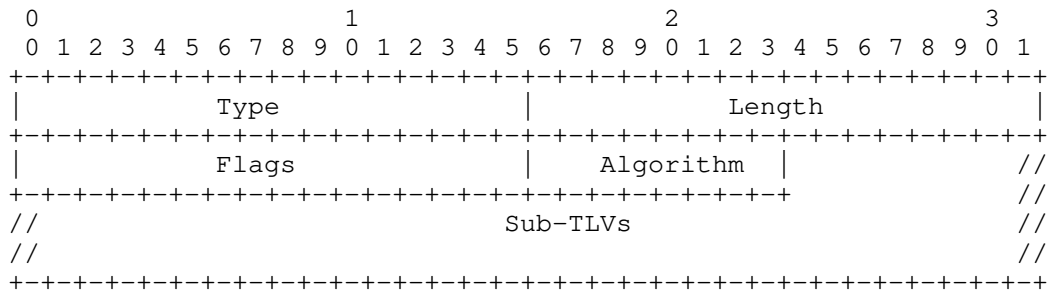
OSPFv3 Prefix-SID flags are defined in [I-D.ietf-ospf-segment-routing-extensions] section 5.

### 2.3.2. IPv6 Prefix-SID TLV

The IPv6 Prefix-SID TLV can ONLY be added to the Prefix Attribute whose local node in the corresponding Prefix NLRI is the node that originates the corresponding SR TLV.

The IPv6 Prefix-SID has the following format:





where:

Type: TBD, suggested value 1169.

Length: 3 + length of Sub-TLVs.

Flags: 2 octet field of flags. None of them is defined at this stage.

Algorithm: 1 octet value identify the algorithm as defined in [I-D.previdi-isis-ipv6-prefix-sid].

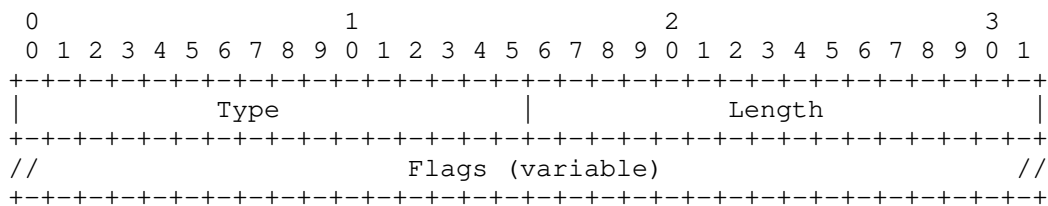
Sub-TLVs: additional information encoded into the IPv6 Prefix-SID Sub-TLV as defined in [I-D.previdi-isis-ipv6-prefix-sid].

The IPv6 Prefix-SID TLV is defined in [I-D.previdi-isis-ipv6-prefix-sid].

### 2.3.3. IGP Prefix Attributes TLV

The IGP Prefix Attribute TLV carries IPv4/IPv6 prefix attribute flags as defined in [RFC7684] and [RFC7794].

The IGP Prefix Attribute TLV has the following format:



where:

Type: TBD, suggested value 1170.

Length: variable.

Flags: a variable length flag field (according to the length field). Flags are routing protocol specific (OSPF and IS-IS). OSPF flags are defined in [RFC7684] and IS-IS flags are defined in [RFC7794]. The receiver of the BGP-LS update, when inspecting the IGP Prefix Attribute TLV, MUST check the Protocol-ID of the NLRI and refer to the protocol specification in order to parse the flags.

#### 2.3.4. Source Router Identifier (Source Router-ID) TLV

The Source Router-ID TLV contains the IPv4 or IPv6 Router-ID of the originator as defined in [RFC7794]. While defined in the IS-IS protocol, the Source Router-ID TLV may be used to carry the OSPF Router-ID of the prefix originator.

The Source Router-ID TLV has the following format:

```

0                               1                               2                               3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                                     |                                     |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
//                               IPv4/IPv6 Address (Router-ID)                               //
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

where:

Type: TBD, suggested value 1171.

Length: 4 or 16.

IPv4/IPv6 Address: 4 octet IPv4 address or 16 octet IPv6 address.

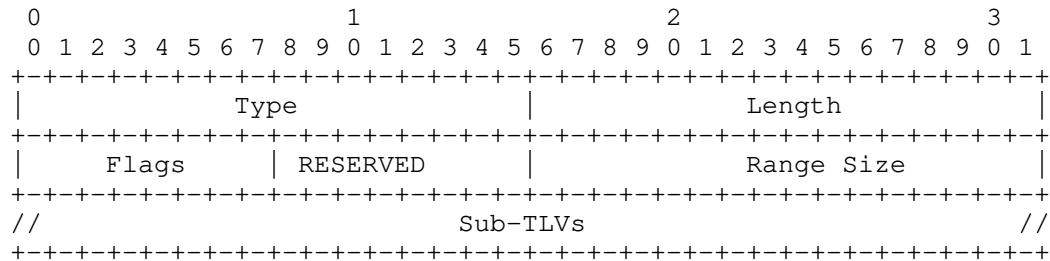
The semantic of the Source Router-ID TLV is defined in [RFC7794].

#### 2.3.5. Range TLV

The Range TLV can ONLY be added to the Prefix Attribute whose local node in the corresponding Prefix NLRI is the node that originates the corresponding SR TLV.

When the range TLV is used in order to advertise a path to a prefix or a range of prefix-to-SID mappings, the Prefix-NLRI the Range TLV is attached to MUST be advertised as a non-routing prefix where no IGP metric TLV (TLV 1095) is attached.

The format of the Range TLV is as follows:



where:

Figure 2: Range TLV format

Type: 1159

Length is 4.

Flags: Only used when the source protocol is OSPF and defined in [I-D.ietf-ospf-segment-routing-extensions] section 4 and [I-D.ietf-ospf-ospfv3-segment-routing-extensions] section 4.

Range Size: 2 octets as defined in [I-D.ietf-ospf-segment-routing-extensions] section 4.

Within the Range TLV, the following SubTLVs are may be present:

Binding SID TLV, defined in Section 2.3.6

Prefix-SID TLV, defined in Section 2.3.1

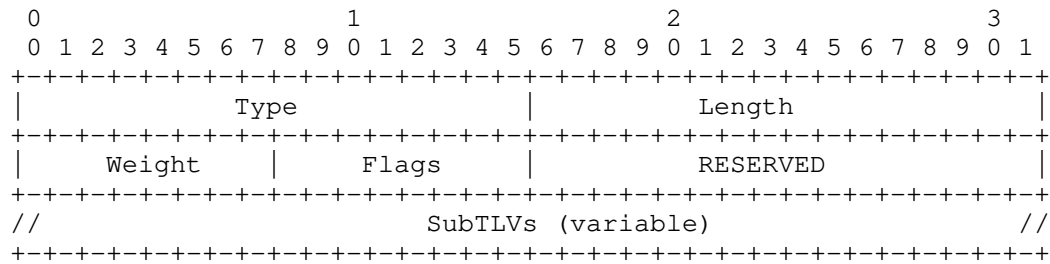
SID/Label TLV, defined in Section 2.3.7.2

### 2.3.6. Binding SID TLV

The Binding SID TLV can be used in two ways:

- o as a sub-TLV of the Range TLV
- o as a Prefix Attribute TLV

The format of the Binding SID TLV is as follows:



where:

Figure 3: Binding SID Sub-TLV format

Type is 1160

Length is variable

Weight and Flags are mapped to Weight and Flags defined in [I-D.ietf-isis-segment-routing-extensions] section 2.4, [I-D.ietf-ospf-segment-routing-extensions] section 4 and [I-D.ietf-ospf-ospfv3-segment-routing-extensions] section 4.

Sub-TLVs are defined in the following sections.

#### 2.3.7. Binding SID SubTLVs

This section defines the Binding SID Sub-TLVs in BGP-LS to encode the equivalent Sub-TLVs defined in [I-D.ietf-isis-segment-routing-extensions], [I-D.ietf-ospf-segment-routing-extensions] and [I-D.ietf-ospf-ospfv3-segment-routing-extensions].

All ERO (Explicit Route Object) Sub-TLVs must immediately follow the (SID)/Label Sub-TLV.

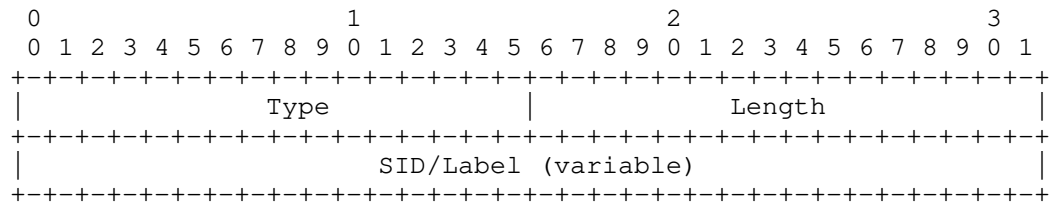
All Backup ERO Sub-TLVs must immediately follow the last ERO Sub-TLV.

##### 2.3.7.1. Binding SID Prefix-SID Sub-TLV

When encoding IS-IS Mapping Server entries as defined in [I-D.ietf-isis-segment-routing-extensions] the Prefix-SID TLV defined in Section 2.3.1 is used as Sub-TLV in the Binding TLV.

### 2.3.7.2. SID/Label Sub-TLV

The SID/Label TLV has following format:



where:

Type: TBD, suggested value 1161.

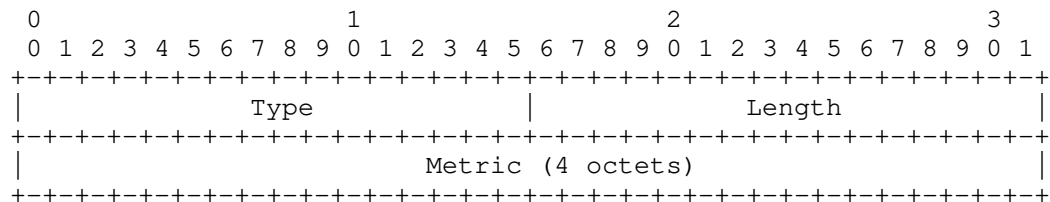
Length: Variable, 3 or 4 bytes

SID/Label: If length is set to 3, then the 20 rightmost bits represent a label. If length is set to 4, then the value represents a 32 bit SID.

The receiving router MUST ignore the SID/Label Sub-TLV if the length is other then 3 or 4.

### 2.3.7.3. ERO Metric Sub-TLV

The ERO Metric Sub-TLV has following format:



ERO Metric Sub-TLV format

where:

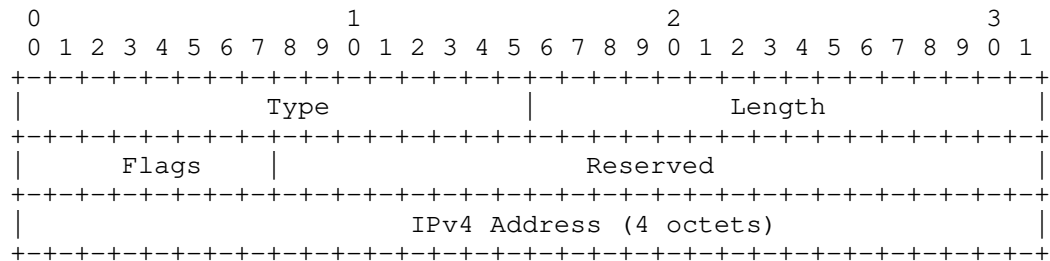
Type: TBD, suggested value 1162.

Length: Always 4

Metric: A 4 octet metric representing the aggregate IGP or TE path cost.

#### 2.3.7.4. IPv4 ERO Sub-TLV

The ERO Sub-TLV has following format:



IPv4 ERO Sub-TLV format

where:

Type: TBD, suggested value 1163

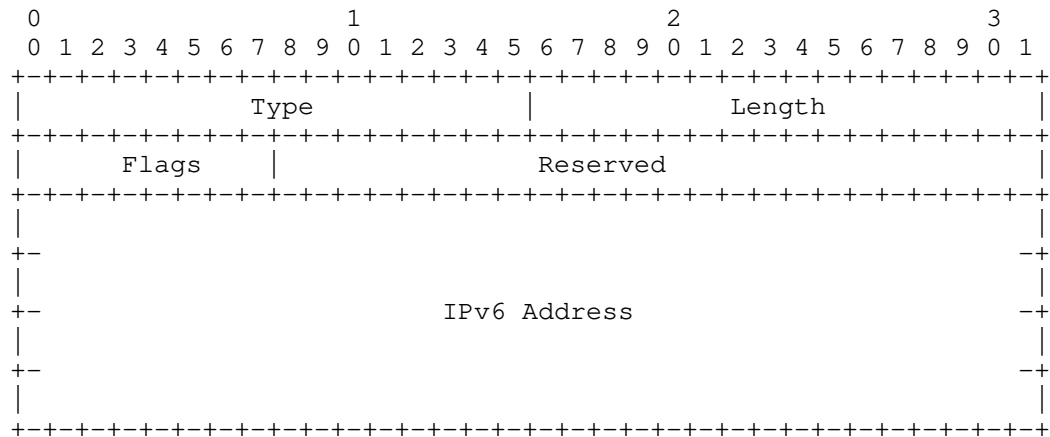
Length: 8 octets

Flags: 1 octet of flags as defined in:  
 [I-D.ietf-isis-segment-routing-extensions],  
 [I-D.ietf-ospf-segment-routing-extensions] and  
 [I-D.ietf-ospf-ospfv3-segment-routing-extensions].

IPv4 Address - the address of the explicit route hop.

#### 2.3.7.5. IPv6 ERO Sub-TLV

The IPv6 ERO Sub-TLV has following format:



IPv6 ERO Sub-TLV format

where:

Type: TBD, suggested value 1164

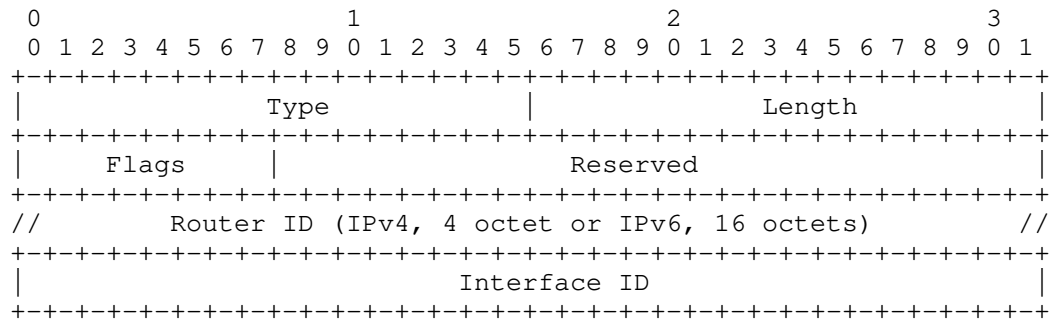
Length: 20 octets

Flags: 1 octet of flags as defined in:  
 [I-D.ietf-isis-segment-routing-extensions],  
 [I-D.ietf-ospf-segment-routing-extensions] and  
 [I-D.ietf-ospf-ospfv3-segment-routing-extensions].

IPv6 Address - the address of the explicit route hop.

#### 2.3.7.6. Unnumbered Interface ID ERO Sub-TLV

The Unnumbered Interface-ID ERO Sub-TLV has following format:



where:

Unnumbered Interface ID ERO Sub-TLV format

Type: TBD, suggested value 1165.

Length: Variable (12 for IPv4 Router-ID or 24 for IPv6 Router-ID).

Flags: 1 octet of flags as defined in:

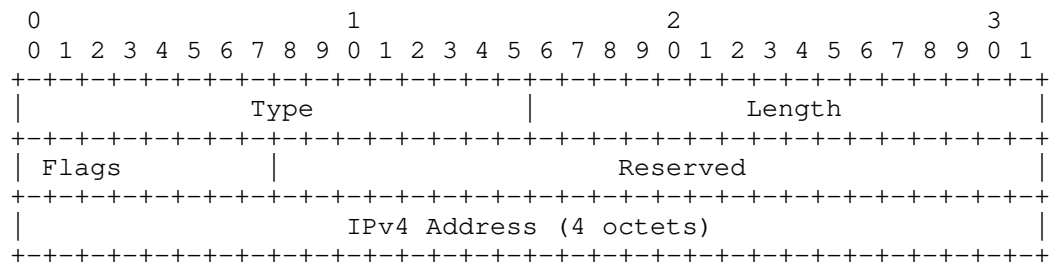
[I-D.ietf-isis-segment-routing-extensions],  
[I-D.ietf-ospf-segment-routing-extensions] and  
[I-D.ietf-ospf-ospfv3-segment-routing-extensions].

Router-ID: Router-ID of the next-hop.

Interface ID: is the identifier assigned to the link by the router specified by the Router-ID.

#### 2.3.7.7. IPv4 Backup ERO Sub-TLV

The IPv4 Backup ERO Sub-TLV has following format:



IPv4 Backup ERO Sub-TLV format

where:



Type: TBD, suggested value 1166.

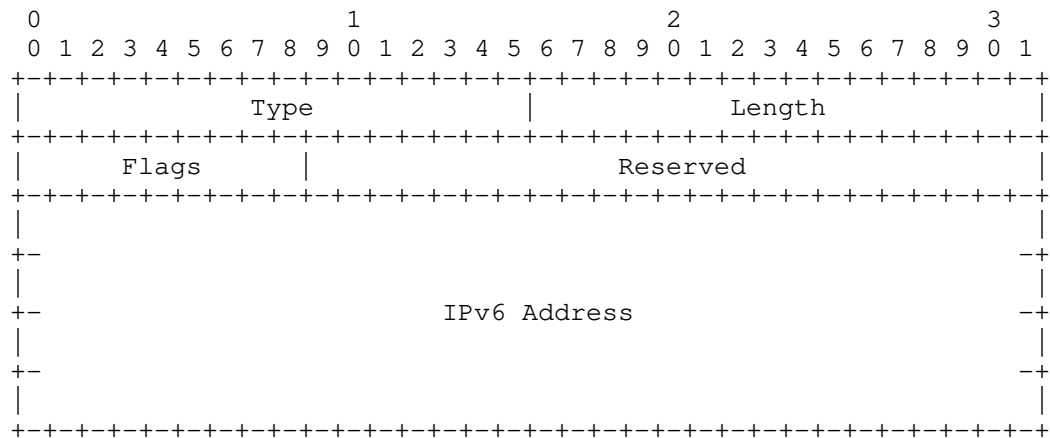
Length: 8 octets

Flags: 1 octet of flags as defined in:  
 [I-D.ietf-isis-segment-routing-extensions],  
 [I-D.ietf-ospf-segment-routing-extensions] and  
 [I-D.ietf-ospf-ospfv3-segment-routing-extensions].

IPv4 Address: Address of the explicit route hop.

#### 2.3.7.8. IPv6 Backup ERO Sub-TLV

The IPv6 Backup ERO Sub-TLV has following format:



IPv6 Backup ERO Sub-TLV format

where:

Type: TBD, suggested value 1167.

Length: 8 octets

Flags: 1 octet of flags as defined in:  
 [I-D.ietf-isis-segment-routing-extensions],  
 [I-D.ietf-ospf-segment-routing-extensions] and  
 [I-D.ietf-ospf-ospfv3-segment-routing-extensions].

IPv6 Address: Address of the explicit route hop.



TLV Code Point	Description	Length	IS-IS TLV /Sub-TLV
1034	SR Capabilities	variable	2 [1]
1035	SR Algorithm	variable	19 [2]
1099	Adjacency Segment Identifier (Adj-SID) TLV	variable	31 [3]
1100	LAN Adjacency Segment Identifier (LAN-Adj-SID) TLV	variable	32 [4]
1158	Prefix SID	variable	3 [5]
1160	Binding SID	variable	149 [6]
1161	SID/Label TLV	variable	1 [7]
1162	ERO Metric TLV	4 octets	10 [8]
1163	IPv4 ERO TLV	5 octets	11 [9]
1164	IPv6 ERO TLV	17 octets	12 [10]
1165	Unnumbered Interface ID ERO TLV	variable	13 [11]
1166	IPv4 Backup ERO TLV	5 octets	14 [12]
1167	IPv6 Backup ERO TLV	17 octets	15 [13]
1168	Unnumbered Interface ID Backup ERO TLV	variable	16 [14]
1169	IPv6 Prefix SID	variable	5 [15]
1170	IGP Prefix Attributes	variable	4 [16]
1171	Source Router ID	variable	11/12 [17]

Table 5: IS-IS Segment Routing Extensions TLVs/Sub-TLVs

## 2.5. Equivalent OSPF/OSPFv3 Segment Routing TLVs/Sub-TLVs

This section illustrate the OSPF and OSPFv3 Segment Routing Extensions TLVs and Sub-TLVs mapped to the ones defined in this document.

The following table, illustrates for each BGP-LS TLV, its equivalence in OSPF and OSPFv3.

TLV Code Point	Description	Length	OSPF TLV /Sub-TLV
1034	SR Capabilities	variable	9 [18]
1035	SR Algorithm	variable	8 [19]
1099	Adjacency Segment Identifier (Adj-SID) TLV	variable	2 [20]
1100	LAN Adjacency Segment Identifier (Adj-SID) TLV	variable	3 [21]
1158	Prefix SID	variable	2 [22]
1161	SID/Label TLV	variable	1 [23]
1162	ERO Metric TLV	4 octets	8 [24]
1163	IPv4 ERO TLV	8 octets	4 [25]
1165	Unnumbered Interface ID ERO TLV	12 octets	5 [26]
1166	IPv4 Backup ERO TLV	8 octets	6 [27]
1167	Unnumbered Interface ID Backup ERO TLV	12 octets	7 [28]
1167	Unnumbered Interface ID Backup ERO TLV	12 octets	7 [29]

Table 6: OSPF Segment Routing Extensions TLVs/Sub-TLVs

TLV Code Point	Description	Length	OSPFv3 TLV /Sub-TLV
1034	SR Capabilities	variable	9 [30]
1035	SR Algorithm	variable	8 [31]
1099	Adjacency Segment Identifier (Adj-SID) TLV	variable	5 [32]
1100	LAN Adjacency Segment Identifier (Adj-SID) TLV	variable	6 [33]
1158	Prefix SID	variable	4 [34]
1161	SID/Label TLV	variable	3 [35]
1162	ERO Metric TLV	4 octets	8 [36]
1163	IPv4 ERO TLV	8 octets	9 [37]
1164	IPv6 ERO TLV	20 octets	8 [38]
1165	Unnumbered Interface ID ERO TLV	12 octets	11 [39]
1166	IPv4 Backup ERO TLV	8 octets	12 [40]
1167	IPv6 Backup ERO TLV	20 octets	13 [41]
1167	Unnumbered Interface ID Backup ERO TLV	12 octets	14 [42]

Table 7: OSPFv3 Segment Routing Extensions TLVs/Sub-TLVs

### 3. Procedures

The following sections describe the different operations for the propagation of SR TLVs into BGP-LS.

#### 3.1. Advertisement of a IS-IS Prefix SID TLV

The advertisement of a IS-IS Prefix SID TLV has following rules:

The IS-IS Prefix-SID is encoded in the BGP-LS Prefix Attribute Prefix-SID as defined in Section 2.3.1. The flags in the Prefix-SID TLV have the semantic defined in [I-D.ietf-isis-segment-routing-extensions] section 2.1.

#### 3.2. Advertisement of a OSPF/OSPFv3 Prefix-SID TLV

The advertisement of a OSPF/OSPFv3 Prefix-SID TLV has following rules:

The OSPF (or OSPFv3) Prefix-SID is encoded in the BGP-LS Prefix Attribute Prefix-SID as defined in Section 2.3.1. The flags in

the Prefix-SID TLV have the semantic defined in [I-D.ietf-ospf-segment-routing-extensions] section 5 or [I-D.ietf-ospf-ospfv3-segment-routing-extensions] section 5.

### 3.3. Advertisement of a range of prefix-to-SID mappings in OSPF

The advertisement of a range of prefix-to-SID mappings in OSPF has following rules:

The OSPF/OSPFv3 Extended Prefix Range TLV is encoded in the BGP-LS Prefix Attribute Range TLV as defined in Section 2.3.5. The flags of the Range TLV have the semantic mapped to the definition in [I-D.ietf-ospf-segment-routing-extensions] section 4 or [I-D.ietf-ospf-ospfv3-segment-routing-extensions] section 4. The Prefix-SID from the original OSPF Prefix SID Sub-TLV is encoded using the BGP-LS Prefix Attribute Prefix-SID as defined in Section 2.3.1 with the flags set according to the definition in [I-D.ietf-ospf-segment-routing-extensions] section 5 or [I-D.ietf-ospf-ospfv3-segment-routing-extensions] section 5.

### 3.4. Advertisement of a range of IS-IS SR bindings

The advertisement of a range of IS-IS SR bindings has following rules:

In IS-IS the Mapping Server binding ranges are advertised using the Binding TLV. The IS-IS Binding TLV is encoded in the BGP-LS Prefix Attribute Range TLV as defined in Section 2.3.5 using the Binding Sub-TLV as defined in Section 2.3.6. The flags in the Range TLV are all set to zero on transmit and ignored on reception. The range value from the original IS-IS Binding TLV is encoded in the Range TLV "Range" field.

### 3.5. Advertisement of a path and its attributes from IS-IS protocol

The advertisement of a Path and its attributes is described in [I-D.ietf-isis-segment-routing-extensions] section 2.4 and has following rules:

The original Binding SID TLV (from IS-IS) is encoded into the BGP-LS Range TLV defined in Section 2.3.5 using the Binding Sub-TLV as defined in Section 2.3.6. The set of Sub-TLVs from the original IS-IS Binding TLV are encoded as Sub-TLVs of the BGP-LS Binding TLV as defined in Section 2.3.6. This includes the SID/Label TLV defined in Section 2.3.

### 3.6. Advertisement of a path and its attributes from OSPFv2/OSPFv3 protocol

The advertisement of a Path and its attributes is described in [I-D.ietf-ospf-segment-routing-extensions] section 6 and [I-D.ietf-ospf-ospfv3-segment-routing-extensions] section 6 and has following rules:

Advertisement of a path for a single prefix: the original Binding SID TLV (from OSPFv2/OSPFv3) is encoded into the BGP-LS Prefix Attribute Binding TLV as defined in Section 2.3.6. The set of Sub-TLVs from the original OSPFv2/OSPFv3 Binding TLV are encoded as Sub-TLVs of the BGP-LS Binding TLV as defined in Section 2.3.6. This includes the SID/Label TLV defined in Section 2.3.

Advertisement of an SR path for range of prefixes: the OSPF/OSPFv3 Extended Prefix Range TLV is encoded in the BGP-LS Prefix Attribute Range TLV as defined in Section 2.3.5. The original OSPFv2/OSPFv3 Binding SID TLV is encoded into the BGP-LS Binding Sub-TLV as defined in Section 2.3.6. The set of Sub-TLVs from the original OSPFv2/OSPFv3 Binding TLV are encoded as Sub-TLVs of the BGP-LS Binding TLV as defined in Section 2.3.6. This includes the SID/Label TLV defined in Section 2.3.

## 4. IANA Considerations

This document requests assigning code-points from the registry for BGP-LS attribute TLVs based on table Table 8.

### 4.1. 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	Length	Section
1034	SR Capabilities	variable	Section 2.1.1
1035	SR Algorithm	variable	Section 2.1.2
1036	SR Local Block	variable	Section 2.1.3
1037	SRMS Preference	variable	Section 2.1.4
1099	Adjacency Segment Identifier (Adj-SID) TLV	variable	Section 2.2.1
1100	LAN Adjacency Segment Identifier (Adj-SID) TLV	variable	Section 2.2.2
1158	Prefix SID	variable	Section 2.3.1
1159	Range	variable	Section 2.3.5
1160	Binding SID	variable	Section 2.3.6
1161	SID/Label TLV	variable	Section 2.3.7.2
1162	ERO Metric TLV	4 octets	1 [43]
1163	IPv4 ERO TLV	8 octets	1 [44]
1164	IPv6 ERO TLV	20 octets	1 [45]
1165	Unnumbered Interface ID ERO TLV	12 octets	1 [46]
1166	IPv4 Backup ERO TLV	8 octets	1 [47]
1167	IPv6 Backup ERO TLV	20 octets	1 [48]
1168	Unnumbered Interface ID Backup ERO TLV	12 octets	1 [49]
1169	IPv6 Prefix SID	variable	Section 2.3.2
1170	IGP Prefix Attributes	variable	Section 2.3.3
1171	Source Router-ID	variable	Section 2.3.4

Table 8: Summary Table of TLV/Sub-TLV Codepoints

## 5. Manageability Considerations

This section is structured as recommended in [RFC5706].

### 5.1. Operational Considerations

#### 5.1.1. Operations

Existing BGP and BGP-LS operational procedures apply. No additional operation procedures are defined in this document.



## 6. 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.

## 7. Contributors

The following people have substantially contributed to the editing of this document:

Acee Lindem  
Cisco Systems  
Email: [acee@cisco.com](mailto:acee@cisco.com)

Saikat Ray  
Individual  
Email: [raysaikat@gmail.com](mailto:raysaikat@gmail.com)

## 8. Acknowledgements

The authors would like to thank Les Ginsberg for the review of this document.

## 9. References

### 9.1. Normative References

- [I-D.ietf-isis-segment-routing-extensions]  
Previdi, S., Filsfils, C., Bashandy, A., Gredler, H., Litkowski, S., Decraene, B., and j. jeffrant@gmail.com, "IS-IS Extensions for Segment Routing", draft-ietf-isis-segment-routing-extensions-08 (work in progress), October 2016.
- [I-D.ietf-ospf-ospfv3-segment-routing-extensions]  
Psenak, P., Previdi, S., Filsfils, C., Gredler, H., Shakir, R., Henderickx, W., and J. Tantsura, "OSPFv3 Extensions for Segment Routing", draft-ietf-ospf-ospfv3-segment-routing-extensions-07 (work in progress), October 2016.
- [I-D.ietf-ospf-segment-routing-extensions]  
Psenak, P., Previdi, S., Filsfils, C., Gredler, H., Shakir, R., Henderickx, W., and J. Tantsura, "OSPF Extensions for Segment Routing", draft-ietf-ospf-segment-routing-extensions-10 (work in progress), October 2016.

- [I-D.previdi-isis-ipv6-prefix-sid]  
Previdi, S., Ginsberg, L., and C. Filsfils, "Segment Routing IPv6 Prefix-SID", draft-previdi-isis-ipv6-prefix-sid-02 (work in progress), May 2016.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<http://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, <<http://www.rfc-editor.org/info/rfc4271>>.
- [RFC7684] Psenak, P., Gredler, H., Shakir, R., Henderickx, W., Tantsura, J., and A. Lindem, "OSPFv2 Prefix/Link Attribute Advertisement", RFC 7684, DOI 10.17487/RFC7684, November 2015, <<http://www.rfc-editor.org/info/rfc7684>>.
- [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, <<http://www.rfc-editor.org/info/rfc7752>>.
- [RFC7794] Ginsberg, L., Ed., Decraene, B., Previdi, S., Xu, X., and U. Chunduri, "IS-IS Prefix Attributes for Extended IPv4 and IPv6 Reachability", RFC 7794, DOI 10.17487/RFC7794, March 2016, <<http://www.rfc-editor.org/info/rfc7794>>.

## 9.2. Informative References

- [I-D.ietf-spring-segment-routing]  
Filsfils, C., Previdi, S., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", draft-ietf-spring-segment-routing-09 (work in progress), July 2016.
- [RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis", RFC 4272, DOI 10.17487/RFC4272, January 2006, <<http://www.rfc-editor.org/info/rfc4272>>.
- [RFC5706] Harrington, D., "Guidelines for Considering Operations and Management of New Protocols and Protocol Extensions", RFC 5706, DOI 10.17487/RFC5706, November 2009, <<http://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, <<http://www.rfc-editor.org/info/rfc6952>>.

### 9.3. URIs

- [1] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-3.1>
- [2] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-3.2>
- [3] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-2.2.1>
- [4] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-2.2.2>
- [5] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-2.1>
- [6] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-2.4>
- [7] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-2.3>
- [8] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-2.4.7>
- [9] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-2.4.8>
- [10] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-2.4.9>
- [11] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-2.4.10>
- [12] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-2.4.11>
- [13] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-2.4.12>

- [14] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-2.4.13>
- [15] <http://tools.ietf.org/html/draft-previdi-isis-ipv6-prefix-sid-01>
- [16] <http://tools.ietf.org/html/RFC7794>
- [17] <http://tools.ietf.org/html/RFC7794>
- [18] <http://tools.ietf.org/html/draft-ietf-ospf-segment-routing-extensions-05#section-3.2>
- [19] <http://tools.ietf.org/html/draft-ietf-ospf-segment-routing-extensions-05#section-3.1>
- [20] <http://tools.ietf.org/html/draft-ietf-ospf-segment-routing-extensions-05#section-7.1>
- [21] <http://tools.ietf.org/html/draft-ietf-ospf-segment-routing-extensions-05#section-7.2>
- [22] <http://tools.ietf.org/html/draft-ietf-ospf-segment-routing-extensions-05#section-5>
- [23] <http://tools.ietf.org/html/draft-ietf-ospf-segment-routing-extensions-05#section-2.1>
- [24] <http://tools.ietf.org/html/draft-ietf-ospf-segment-routing-extensions-05#section-6.1>
- [25] <http://tools.ietf.org/html/draft-ietf-ospf-segment-routing-extensions-05#section-6.2.1>
- [26] <http://tools.ietf.org/html/draft-ietf-ospf-segment-routing-extensions-05#section-6.2.2>
- [27] <http://tools.ietf.org/html/draft-ietf-ospf-segment-routing-extensions-05#section-6.2.3>
- [28] <http://tools.ietf.org/html/draft-ietf-ospf-segment-routing-extensions-05#section-6.2.4>
- [29] <http://tools.ietf.org/html/draft-ietf-ospf-segment-routing-extensions-05#section-6.2.4>
- [30] <http://tools.ietf.org/html/draft-ietf-ospf-ospfv3-segment-routing-extensions-05#section-3.2>

- [31] <http://tools.ietf.org/html/draft-ietf-ospf-ospfv3-segment-routing-extensions-05#section-3.1>
- [32] <http://tools.ietf.org/html/draft-ietf-ospf-ospfv3-segment-routing-extensions-05#section-7.1>
- [33] <http://tools.ietf.org/html/draft-ietf-ospf-ospfv3-segment-routing-extensions-05#section-7.2>
- [34] <http://tools.ietf.org/html/draft-ietf-ospf-ospfv3-segment-routing-extensions-05#section-5>
- [35] <http://tools.ietf.org/html/draft-ietf-ospf-ospfv3-segment-routing-extensions-05#section-2.1>
- [36] <http://tools.ietf.org/html/draft-ietf-ospf-ospfv3-segment-routing-extensions-05#section-6.1>
- [37] <http://tools.ietf.org/html/draft-ietf-ospf-ospfv3-segment-routing-extensions-05#section-6.2.1>
- [38] <http://tools.ietf.org/html/draft-ietf-ospf-ospfv3-segment-routing-extensions-05#section-6.2.2>
- [39] <http://tools.ietf.org/html/draft-ietf-ospf-ospfv3-segment-routing-extensions-05#section-6.2.3>
- [40] <http://tools.ietf.org/html/draft-ietf-ospf-ospfv3-segment-routing-extensions-05#section-6.2.4>
- [41] <http://tools.ietf.org/html/draft-ietf-ospf-ospfv3-segment-routing-extensions-05#section-6.2.5>
- [42] <http://tools.ietf.org/html/draft-ietf-ospf-ospfv3-segment-routing-extensions-05#section-6.2.6>
- [43] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-2.4.7>
- [44] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-2.4.8>
- [45] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-2.4.9>
- [46] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-2.4.10>

- [47] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-2.4.11>
- [48] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-2.4.12>
- [49] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-2.4.13>

#### Authors' Addresses

Stefano Previdi (editor)  
Cisco Systems, Inc.  
Via Del Serafico, 200  
Rome 00142  
Italy

Email: [sprevidi@cisco.com](mailto:sprevidi@cisco.com)

Peter Psenak  
Cisco Systems, Inc.  
Apollo Business Center  
Mlynske nivy 43  
Bratislava 821 09  
Slovakia

Email: [ppsenak@cisco.com](mailto:ppsenak@cisco.com)

Clarence Filsfils  
Cisco Systems, Inc.  
Brussels  
Belgium

Email: [cfilsfil@cisco.com](mailto:cfilsfil@cisco.com)

Hannes Gredler  
RtBrick Inc.

Email: [hannes@rtbrick.com](mailto:hannes@rtbrick.com)

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

Email: mach.chen@huawei.com

Jeff Tantsura  
Individual

Email: jefftant@gmail.com

Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: June 23, 2016

S. Previdi, Ed.  
C. Filsfils  
Cisco Systems, Inc.  
S. Ray  
Individual Contributor  
K. Patel  
Cisco Systems, Inc.  
J. Dong  
M. Chen  
Huawei Technologies  
December 21, 2015

Segment Routing Egress Peer Engineering BGP-LS Extensions  
draft-ietf-idr-bgpls-segment-routing-epe-02

Abstract

Segment Routing (SR) leverages source routing. A node steers a packet through a controlled set of instructions, called segments, by prepending the packet with an SR header. A segment can represent any instruction, topological or service-based. SR allows to enforce a flow through any topological path and service chain while maintaining per-flow state only at the ingress node of the SR domain.

The Segment Routing architecture can be directly applied to the MPLS dataplane with no change on the forwarding plane. It requires minor extension to the existing link-state routing protocols.

This document outline a BGP-LS extension for exporting BGP egress point topology information (including its peers, interfaces and peering ASs) in a way that is exploitable in order to compute efficient Egress Point Engineering policies and strategies.

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 <http://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 June 23, 2016.

#### Copyright Notice

Copyright (c) 2015 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. Segment Routing Documents . . . . .	3
3. BGP Peering Segments . . . . .	4
4. Link NLRI for EPE Connectivity Description . . . . .	5
4.1. BGP Router ID and Member ASN . . . . .	5
4.2. EPE Node Descriptors . . . . .	6
4.3. Link Attributes . . . . .	7
5. Peer Node and Peer Adjacency Segments . . . . .	9
5.1. Peer Node Segment (Peer-Node-SID) . . . . .	9
5.2. Peer Adjacency Segment (Peer-Adj-SID) . . . . .	10
5.3. Peer Set Segment . . . . .	11
6. Illustration . . . . .	12
6.1. Reference Diagram . . . . .	12
6.2. Peer Node Segment for Node D . . . . .	14
6.3. Peer Node Segment for Node H . . . . .	14
6.4. Peer Node Segment for Node E . . . . .	14
6.5. Peer Adjacency Segment for Node E, Link 1 . . . . .	15
6.6. Peer Adjacency Segment for Node E, Link 2 . . . . .	15
7. IANA Considerations . . . . .	16
8. Manageability Considerations . . . . .	16
9. Security Considerations . . . . .	16

10. Contributors . . . . .	17
11. Acknowledgements . . . . .	17
12. References . . . . .	17
12.1. Normative References . . . . .	17
12.2. Informative References . . . . .	17
Authors' Addresses . . . . .	18

## 1. Introduction

Segment Routing (SR) leverages source routing. A node steers a packet through a controlled set of instructions, called segments, by prepending the packet with an SR header. A segment can represent any instruction, topological or service-based. SR allows to enforce a flow through any topological path and service chain while maintaining per-flow state only at the ingress node of the SR domain.

The Segment Routing architecture can be directly applied to the MPLS dataplane with no change on the forwarding plane. It requires minor extension to the existing link-state routing protocols.

This document outline a BGP-LS extension for exporting BGP egress point topology information (including its peers, interfaces and peering ASs) in a way that is exploitable in order to compute efficient Egress Point Engineering policies and strategies.

This document defines new types of segments: a Peer Node segment describing the BGP session between two nodes; a Peer Adjacency Segment describing the link (one or more) that is used by the BGP session; the Peer Set Segment describing an arbitrary set of sessions or links between the local BGP node and its peers.

While an egress point topology usually refers to eBGP sessions between external peers, there's nothing in the extensions defined in this document that would prevent the use of these extensions in the context of iBGP sessions.

## 2. Segment Routing Documents

The main reference for this document is the SR architecture defined in [I-D.ietf-spring-segment-routing].

The Segment Routing Egress Peer Engineering architecture is described in [I-D.ietf-spring-segment-routing-central-epe].

### 3. BGP Peering Segments

As defined in [I-D.ietf-spring-segment-routing-central-epe], an EPE enabled Egress PE node MAY advertise segments corresponding to its attached peers. These segments are called BGP peering segments or BGP Peering SIDs. They enable the expression of source-routed inter-domain paths.

An ingress border router of an AS may compose a list of segments to steer a flow along a selected path within the AS, towards a selected egress border router C of the AS and through a specific peer. At minimum, a BGP Peering Engineering policy applied at an ingress PE involves two segments: the Node SID of the chosen egress PE and then the BGP Peering Segment for the chosen egress PE peer or peering interface.

This document defines the BGP EPE Peering Segments:

- o Peer Node Segment (Peer-Node-SID)
- o Peer Adjacency Segment (Peer-Adj-SID)
- o Peer Set Segment (Peer-Set-SID)

Each BGP session MUST be described by a Peer Node Segment. The description of the BGP session MAY be augmented by additional Adjacency Segments. Finally, each Peer Node Segment and Peer Adjacency Segment MAY be part of the same group/set so to be able to group EPE resources under a common Peer-Set Segment Identifier (SID).

Therefore, when the extensions defined in this document are applied to the use case defined in [I-D.ietf-spring-segment-routing-central-epe]:

- o One Peer Node Segment MUST be present.
- o One or more Peer Adjacency Segments MAY be present.
- o Each of the Peer Node and Peer Adjacency Segment MAY use the same Peer-Set.

While an egress point topology usually refers to eBGP sessions between external peers, there's nothing in the extensions defined in this document that would prevent the use of these extensions in the context of iBGP sessions.

#### 4. Link NLRI for EPE Connectivity Description

This section describes the NLRI used for describing the connectivity of the BGP Egress router. The connectivity is based on links and remote peers/ASs and therefore the existing Link-Type NLRI (defined in [I-D.ietf-idr-ls-distribution]) is used. A new Protocol ID is used (codepoint to be assigned by IANA, suggested value 7).

The use of a new Protocol-ID allows separation and differentiation between the NLRIs carrying BGP-EPE descriptors from the NLRIs carrying IGP link-state information as defined in [I-D.ietf-idr-ls-distribution]. The Link NLRI Type uses descriptors and attributes already defined in [I-D.ietf-idr-ls-distribution] in addition to new TLVs defined in the following sections of this document.

The extensions defined in this document apply to both internal and external BGP-LS EPE advertisements.

[I-D.ietf-idr-ls-distribution] defines Link NLRI Type is as follows:

```

      0                               1                               2                               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+
| Protocol-ID |
+-----+-----+-----+-----+
|                               Identifier                               |
|                               (64 bits)                               |
+-----+-----+-----+-----+
//      Local Node Descriptors      //
+-----+-----+-----+-----+
//      Remote Node Descriptors      //
+-----+-----+-----+-----+
//      Link Descriptors              //
+-----+-----+-----+-----+

```

Node Descriptors and Link Descriptors are defined in [I-D.ietf-idr-ls-distribution].

##### 4.1. BGP Router ID and Member ASN

Two new Node Descriptors Sub-TLVs are defined in this document:

- o BGP Router Identifier (BGP Router-ID):

Type: TBA (suggested value 516).

Length: 4 octets

Value: 4 octet unsigned integer representing the BGP Identifier as defined in [RFC4271] and [RFC6286].

- o Confederation Member ASN (Member-ASN)

Type: TBA (suggested value 517).

Length: 4 octets

Value: 4 octet unsigned integer representing the Member ASN inside the Confederation.[RFC5065].

#### 4.2. EPE Node Descriptors

The following Node Descriptors Sub-TLVs MUST appear in the Link NLRI as Local Node Descriptors:

- o BGP Router ID, which contains the BGP Identifier of the local BGP EPE node.
- o Autonomous System Number, which contains the local ASN or local confederation identifier (ASN) if confederations are used.
- o BGP-LS Identifier.

It has to be noted that [RFC6286] (section 2.1) requires the BGP identifier (router-id) to be unique within an Autonomous System. Therefore, the <ASN, BGP identifier> tuple is globally unique.

The following Node Descriptors Sub-TLVs MAY appear in the Link NLRI as Local Node Descriptors:

- o Member-ASN, which contains the ASN of the confederation member (when BGP confederations are used).
- o Node Descriptors as defined in [I-D.ietf-idr-ls-distribution].

The following Node Descriptors Sub-TLVs MUST appear in the Link NLRI as Remote Node Descriptors:

- o BGP Router ID, which contains the BGP Identifier of the peer node.
- o Autonomous System Number, which contains the peer ASN or the peer confederation identifier (ASN), if confederations are used.

The following Node Descriptors Sub-TLVs MAY appear in the Link NLRI as Remote Node Descriptors:

- o Member-ASN, which contains the ASN of the confederation member (when BGP confederations are used).
- o Node Descriptors as defined in defined in [I-D.ietf-idr-ls-distribution].

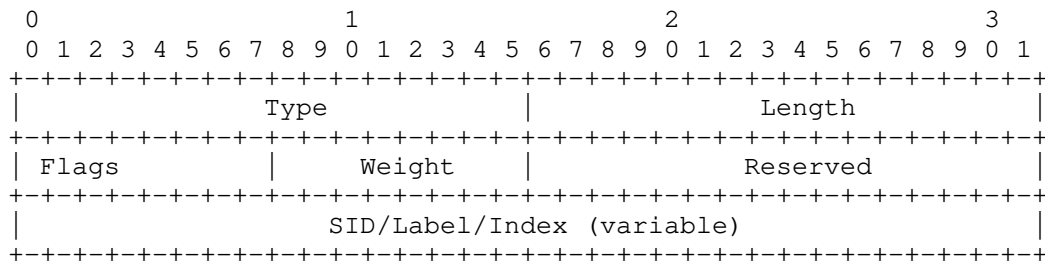
#### 4.3. Link Attributes

The following BGP-LS Link attributes TLVs are used with the Link NLRI:

TLV Code Point	Description	Length
1101	Peer Node Segment Identifier (Peer-Node-SID)	variable
1102	Peer Adjacency Segment Identifier (Peer-Adj-SID)	variable
1103	Peer Set Segment Identifier (Peer-Set-SID)	variable

Figure 1: TLV code points for BGP-LS EPE

Peer-Node-SID, Peer-Adj-SID and Peer-Set-SID have all the same format defined here below:



where:

Figure 2

- o Type: To be assigned by IANA. The suggested values are defined in Figure 1.
- o Length: variable.
- o Flags: following flags have been defined:

```

  0 1 2 3 4 5 6 7
+---+---+---+---+
|V|L|           |
+---+---+---+---+

```

where:

- \* V-Flag: Value flag. If set, then the Adj-SID carries a value. By default the flag is SET.
- \* L-Flag: Local Flag. If set, then the value/index carried by the Adj-SID has local significance. By default the flag is SET.
- \* Other bits: MUST be zero when originated and ignored when received.
- o Weight: 1 octet. The value represents the weight of the SID for the purpose of load balancing. An example use of the weight is described in [I-D.ietf-spring-segment-routing].
- o SID/Index/Label. According to the TLV length and to the V and L flags settings, it contains either:
  - \* A 3 octet local label where the 20 rightmost bits are used for encoding the label value. In this case the V and L flags MUST be set.
  - \* A 4 octet index defining the offset in the SID/Label space advertised by this router using the encodings defined in Section 3.1. In this case V and L flags MUST be unset.
  - \* A 16 octet IPv6 address. In this case the V flag MUST be set. The L flag MUST be unset if the IPv6 address is globally unique.

The values of the Peer-Node-SID, Peer-Adj-SID and Peer-Set-SID Sub-TLVs SHOULD be persistent across router restart.

The Peer-Node-SID MUST be present when BGP-LS is used for the use case described in [I-D.ietf-spring-segment-routing-central-epe] and MAY be omitted for other use cases.

The Peer-Adj-SID and Peer-Set-SID SubTLVs MAY be present when BGP-LS is used for the use case described in [I-D.ietf-spring-segment-routing-central-epe] and MAY be omitted for other use cases.

In addition, BGP-LS Nodes and Link Attributes, as defined in [I-D.ietf-idr-ls-distribution] MAY be inserted in order to advertise the characteristics of the link.

## 5. Peer Node and Peer Adjacency Segments

In this section the following Peer Segments are defined:

Peer Node Segment (Peer-Node-SID)

Peer Adjacency Segment (Peer-Adj-SID)

Peer Set Segment (Peer-Set-SID)

The Peer Node, Peer Adjacency and Peer Set segments can be either a local or a global segment (depending on the setting of the V and L flags defined in Figure 2. For example, when EPE is used in the context of a SR network over the IPv6 dataplane, it is likely the case that the IPv6 addresses used as SIDs will be global.

### 5.1. Peer Node Segment (Peer-Node-SID)

The Peer Node Segment describes the BGP session peer (neighbor). It MUST be present when describing an EPE topology as defined in [I-D.ietf-spring-segment-routing-central-epe]. The Peer Node Segment is encoded within the BGP-LS Link NLRI specified in Section 4.

The Peer Node Segment, at the BGP node advertising it, has the following semantic:

- o SR header operation: NEXT (as defined in [I-D.ietf-spring-segment-routing]).
- o Next-Hop: the connected peering node to which the segment is related.

The Peer Node Segment is advertised with a Link NLRI, where:

- o Local Node Descriptors contains

Local BGP Router ID of the EPE enabled egress PE.  
Local ASN.  
BGP-LS Identifier.

- o Remote Node Descriptors contains

Peer BGP Router ID (i.e.: the peer BGP ID used in the BGP session).  
Peer ASN.



- o Link Descriptors Sub-TLVs, as defined in [I-D.ietf-idr-ls-distribution], contain the addresses used by the BGP session:
  - \* IPv4 Interface Address (Sub-TLV 259) contains the BGP session IPv4 local address.
  - \* IPv4 Neighbor Address (Sub-TLV 260) contains the BGP session IPv4 peer address.
  - \* IPv6 Interface Address (Sub-TLV 261) contains the BGP session IPv6 local address.
  - \* IPv6 Neighbor Address (Sub-TLV 262) contains the BGP session IPv6 peer address.
- o Link Attribute contains the Peer-Node-SID TLV as defined in Section 4.3.
- o In addition, BGP-LS Link Attributes, as defined in [I-D.ietf-idr-ls-distribution], MAY be inserted in order to advertise the characteristics of the link.

#### 5.2. Peer Adjacency Segment (Peer-Adj-SID)

The Peer Adjacency Segment, at the BGP node advertising it, has the following semantic:

- o SR header operation: NEXT (as defined in [I-D.ietf-spring-segment-routing]).
- o Next-Hop: the interface peer address.

The Peer Adjacency Segment is advertised with a Link NLRI, where:

- o Local Node Descriptors contains
  - Local BGP Router ID of the EPE enabled egress PE.
  - Local ASN.
  - BGP-LS Identifier.
- o Remote Node Descriptors contains
  - Peer BGP Router ID (i.e.: the peer BGP ID used in the BGP session).
  - Peer ASN.
- o Link Descriptors Sub-TLVs, as defined in [I-D.ietf-idr-ls-distribution], MUST contain the following TLVs:

- \* Link Local/Remote Identifiers (Sub-TLV 258) contains the 4-octet Link Local Identifier followed by the 4-octet value 0 indicating the Link Remote Identifier in unknown [RFC5307].
- o In addition, Link Descriptors Sub-TLVs, as defined in [I-D.ietf-idr-ls-distribution], MAY contain the following TLVs:
  - \* IPv4 Interface Address (Sub-TLV 259) contains the address of the local interface through which the BGP session is established.
  - \* IPv6 Interface Address (Sub-TLV 261) contains the address of the local interface through which the BGP session is established.
  - \* IPv4 Neighbor Address (Sub-TLV 260) contains the IPv4 address of the peer interface used by the BGP session.
  - \* IPv6 Neighbor Address (Sub-TLV 262) contains the IPv6 address of the peer interface used by the BGP session.
- o Link attribute used with the Peer-Adj-SID contains the TLV as defined in Section 4.3.

In addition, BGP-LS Link Attributes, as defined in [I-D.ietf-idr-ls-distribution], MAY be inserted in order to advertise the characteristics of the link.

### 5.3. Peer Set Segment

The Peer Adjacency Segment, at the BGP node advertising it, has the following semantic:

- o SR header operation: NEXT (as defined in [I-D.ietf-spring-segment-routing]).
- o Next-Hop: load balance across any connected interface to any peer in the related set.

The Peer Set Segment is advertised within a Link NLRI (describing a Peer Node Segment or a Peer Adjacency segment) as a BGP-LS attribute.

The Peer Set Attribute contains the Peer-Set-SID TLV, defined in Section 4.3 identifying the set of which the Peer Node Segment or Peer Adjacency Segment is a member.

## 6. Illustration

### 6.1. Reference Diagram

The following reference diagram is used throughout this document. The solution is illustrated for IPv4 with MPLS-based segments and the EPE topology is based on eBGP sessions between external peers.

As stated in Section 3, the solution illustrated hereafter is equally applicable to an iBGP session topology. In other words, the solution also applies to the case where C, D, H, and E are in the same AS and run iBGP sessions between each other.

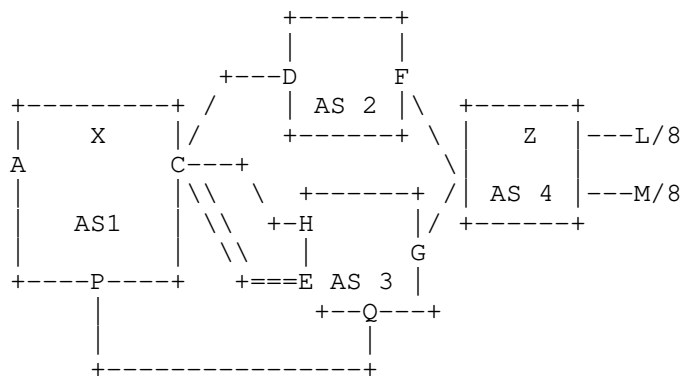


Figure 3: Reference Diagram

IPv4 addressing:

- o C's IPv4 address of interface to D: 1.0.1.1/24, D's interface: 1.0.1.2/24
- o C's IPv4 address of interface to H: 1.0.2.1/24, H's interface: 1.0.2.2/24
- o C's IPv4 address of upper interface to E: 1.0.3.1, E's interface: 1.0.3.2/24
- o C's local identifier of upper interface to E: 0.0.0.1.0.0.0.0
- o C's IPv4 address of lower interface to E: 1.0.4.1/24, E's interface: 1.0.4.2/24
- o C's local identifier of lower interface to E: 0.0.0.2.0.0.0.0
- o Loopback of E used for eBGP multi-hop peering to C: 1.0.5.2/32

- o C's loopback is 3.3.3.3/32 with SID 64

BGP Router-IDs are C, D, H and E.

- o C's BGP Router-ID: 3.3.3.3
- o D's BGP Router-ID: 4.4.4.4
- o E's BGP Router-ID: 5.5.5.5
- o H's BGP Router-ID: 6.6.6.6

C's BGP peering:

- o Single-hop eBGP peering with neighbor 1.0.1.2 (D)
- o Single-hop eBGP peering with neighbor 1.0.2.2 (H)
- o Multi-hop eBGP peering with E on ip address 1.0.5.2 (E)

C's resolution of the multi-hop eBGP session to E:

- o Static route 1.0.5.2/32 via 1.0.3.2
- o Static route 1.0.5.2/32 via 1.0.4.2

Node C configuration is such that:

- o A Peer Node segment (Peer-Node-SID) is allocated to each peer (D, H and E).
- o An Peer Adjacency segment (Peer-Adj-SID) is defined for each recursing interface to a multi-hop peer (CE upper and lower interfaces).
- o A Peer Set segment (Peer-Set-SID) is defined to include all peers in AS3 (peers H and E).

Local BGP-LS Identifier in router C is set to 10000.

The Link NLRI Type is used in order to encode C's connectivity. The Link NLRI uses the new Protocol-ID value (to be assigned by IANA)

Once the BGP-LS update is originated by C, it may be advertised to internal (iBGP) as well as external (eBGP) neighbors supporting the BGP-LS EPE extensions defined in this document.

## 6.2. Peer Node Segment for Node D

### Descriptors:

- o Local Node Descriptors (BGP Router-ID, local ASN, BGP-LS Identifier): 3.3.3.3 , AS1, 10000
- o Remote Node Descriptors (BGP Router-ID, peer ASN): 4.4.4.4, AS2
- o Link Descriptors (BGP session IPv4 local address, BGP session IPv4 neighbor address): 1.0.1.1, 1.0.1.2

### Attributes:

- o Peer-Node-SID: 1012
- o Link Attributes: see section 3.3.2 of [I-D.ietf-idr-ls-distribution]

## 6.3. Peer Node Segment for Node H

### Descriptors:

- o Local Node Descriptors (BGP Router-ID, ASN, BGPL Identifier): 3.3.3.3 , AS1, 10000
- o Remote Node Descriptors (BGP Router-ID ASN): 6.6.6.6, AS3
- o Link Descriptors (BGP session IPv4 local address, BGP session IPv4 peer address): 1.0.2.1, 1.0.2.2

### Attributes:

- o Peer-Node-SID: 1022
- o Peer-Set-SID: 1060
- o Link Attributes: see section 3.3.2 of [I-D.ietf-idr-ls-distribution]

## 6.4. Peer Node Segment for Node E

### Descriptors:

- o Local Node Descriptors (BGP Router-ID, ASN, BGP-LS Identifier): 3.3.3.3 , AS1, 10000
- o Remote Node Descriptors (BGP Router-ID, ASN): 5.5.5.5, AS3

- o Link Descriptors (BGP session IPv4 local address, BGP session IPv4 peer address): 3.3.3.3, 1.0.5.2

Attributes:

- o Peer-Node-SID: 1052
- o Peer-Set-SID: 1060

#### 6.5. Peer Adjacency Segment for Node E, Link 1

Descriptors:

- o Local Node Descriptors (BGP Router-ID, ASN, BGP-LS Identifier): 3.3.3.3 , AS1, 10000
- o Remote Node Descriptors (BGP Router-ID, ASN): 5.5.5.5, AS3
- o Link Descriptors (local interface identifier, IPv4 peer interface address): 0.0.0.1.0.0.0.0 , 1.0.3.2

Attributes:

- o Peer-Adj-SID: 1032
- o LinkAttributes: see section 3.3.2 of [I-D.ietf-idr-ls-distribution]

#### 6.6. Peer Adjacency Segment for Node E, Link 2

Descriptors:

- o Local Node Descriptors (BGP Router-ID, ASN, BGP-LS Identifier): 3.3.3.3 , AS1, 10000
- o Remote Node Descriptors (BGP Router-ID, ASN): 5.5.5.5, AS3
- o Link Descriptors (local interface identifier, IPv4 peer interface address): 0.0.0.2.0.0.0.0 , 1.0.4.2

Attributes:

- o Peer-Adj-SID: 1042
- o LinkAttributes: see section 3.3.2 of [I-D.ietf-idr-ls-distribution]

## 7. IANA Considerations

This document defines:

Two new Node Descriptors Sub-TLVs: BGP-Router-ID and BGP Confederation Member.

A new Protocol-ID for EPE: BGP-EPE.

Three new BGP-LS Attribute Sub-TLVs: Peer-Node-SID, Peer-Adj-SID and Peer-Set-SID.

The codepoints are to be assigned by IANA. The following are the suggested values:

Suggested Codepoint	Description	Defined in:
7	Protocol-ID	Section 4
516	BGP Router ID	Section 4.1
517	BGP Confederation Member	Section 4.1
1101	Peer-Node-SID	Section 4.3
1102	Peer-Adj-SID	Section 4.3
1103	Peer-Set-SID	Section 4.3

Table 1: Summary Table of BGP-LS EPE Codepoints

## 8. Manageability Considerations

TBD

## 9. Security Considerations

[I-D.ietf-idr-ls-distribution] defines BGP-LS NLRIs to which the extensions defined in this document apply.

The Security Section of [I-D.ietf-idr-ls-distribution] also applies to:

- o New Node Descriptors Sub-TLVs: BGP-Router-ID and BGP-Confederation-Member;
- o New BGP-LS Attributes TLVs: Peer-Node-SID, Peer-Adj-SID and Peer-Set-SID.

## 10. Contributors

Acee Lindem gave a substantial contribution to this document.

## 11. Acknowledgements

The authors would like to thank Jakob Heitz, Howard Yang, Hannes Gredler, Peter Psenak, Ketan Jivan Talaulikar, and Arjun Sreekantiah for their feedback and comments.

## 12. References

### 12.1. 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, <<http://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, <<http://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, <<http://www.rfc-editor.org/info/rfc5065>>.
- [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, <<http://www.rfc-editor.org/info/rfc5307>>.
- [RFC6286] Chen, E. and J. Yuan, "Autonomous-System-Wide Unique BGP Identifier for BGP-4", RFC 6286, DOI 10.17487/RFC6286, June 2011, <<http://www.rfc-editor.org/info/rfc6286>>.

### 12.2. Informative References

- [I-D.ietf-idr-ls-distribution] Gredler, H., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and TE Information using BGP", draft-ietf-idr-ls-distribution-13 (work in progress), October 2015.



[I-D.ietf-spring-segment-routing]  
Filsfils, C., Previdi, S., Decraene, B., Litkowski, S.,  
and r. rjs@rob.sh, "Segment Routing Architecture", draft-  
ietf-spring-segment-routing-07 (work in progress),  
December 2015.

[I-D.ietf-spring-segment-routing-central-epe]  
Filsfils, C., Previdi, S., Ginsburg, D., and D. Afanasiev,  
"Segment Routing Centralized Egress Peer Engineering",  
draft-ietf-spring-segment-routing-central-epe-00 (work in  
progress), October 2015.

#### Authors' Addresses

Stefano Previdi (editor)  
Cisco Systems, Inc.  
Via Del Serafico, 200  
Rome 00142  
Italy

Email: sprevidi@cisco.com

Clarence Filsfils  
Cisco Systems, Inc.  
Brussels  
BE

Email: cfilsfil@cisco.com

Saikat Ray  
Individual Contributor

Email: raysaikat@gmail.com

Keyur Patel  
Cisco Systems, Inc.  
170, West Tasman Drive  
San Jose, CA 95134  
US

Email: keyupate@cisco.com

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

Email: jie.dong@huawei.com

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

Email: mach.chen@huawei.com

IDR and SIDR  
Internet-Draft  
Intended status: Standards Track  
Expires: September 15, 2016

K. Sriram  
D. Montgomery  
US NIST  
B. Dickson

K. Patel  
Cisco  
A. Robachevsky  
Internet Society  
March 14, 2016

Methods for Detection and Mitigation of BGP Route Leaks  
draft-ietf-idr-route-leak-detection-mitigation-02

Abstract

In [I-D.ietf-grow-route-leak-problem-definition], the authors have provided a definition of the route leak problem, and also enumerated several types of route leaks. In this document, we first examine which of those route-leak types are detected and mitigated by the existing origin validation (OV) [RFC 6811]. It is recognized that OV offers a limited detection and mitigation capability against route leaks. This document proposes an enhancement that significantly extends the route-leak detection and mitigation capabilities of BGP. The solution involves carrying a per-hop route-leak protection (RLP) field in BGP updates. The RLP field is proposed be carried in an optional transitive path attribute. The solution is meant to be initially implemented as an enhancement of BGP without requiring BGPsec [I-D.ietf-sidr-bgpsec-protocol]. However, when BGPsec is deployed in the future, the solution can be incorporated in BGPsec, enabling cryptographic protection for the RLP field. That would be one way of implementing the proposed solution in a secure way. It is not claimed that the solution detects all possible types of route leaks but it detects several types, especially considering some significant route-leak occurrences that have been observed in recent years. The document also includes a stopgap method for detection and mitigation of route leaks for an intermediate phase when OV is deployed but BGP protocol on the wire is unchanged.

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 <http://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 15, 2016.

#### Copyright Notice

Copyright (c) 2016 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. Related Prior Work . . . . .	3
3. Mechanisms for Detection and Mitigation of Route Leaks . . .	4
3.1. Route-Leak Protection (RLP) Field Encoding by Sending Router . . . . .	6
3.2. Recommended Actions at a Receiving Router for Detection of Route Leaks . . . . .	8
3.3. Possible Actions at a Receiving Router for Mitigation . .	9
4. Stopgap Solution when Only Origin Validation is Deployed . .	9
5. Design Rationale and Discussion . . . . .	10
5.1. Is route-leak solution without cryptographic protection a serious attack vector? . . . . .	10
5.2. Combining results of route-leak detection, OV and BGPsec validation for path selection decision . . . . .	12
5.3. Are there cases when valley-free violations can be considered legitimate? . . . . .	12
5.4. Comparison with other methods, routing security BCP . . .	13
6. Summary . . . . .	13
7. Security Considerations . . . . .	14
8. IANA Considerations . . . . .	14
9. Acknowledgements . . . . .	14

10. References	14
10.1. Normative References	14
10.2. Informative References	14
Authors' Addresses	18

## 1. Introduction

In [I-D.ietf-grow-route-leak-problem-definition], the authors have provided a definition of the route leak problem, and also enumerated several types of route leaks. In this document, we first examine which of those route-leak types are detected and mitigated by the existing Origin Validation (OV) [RFC6811] method. OV and BGPsec path validation [I-D.ietf-sidr-bgpsec-protocol] together offer mechanisms to protect against re-originations and hijacks of IP prefixes as well as man-in-the-middle (MITM) AS path modifications. Route leaks (see [I-D.ietf-grow-route-leak-problem-definition] and references cited at the back) are another type of vulnerability in the global BGP routing system against which OV offers only partial protection. BGPsec (i.e. path validation) provides cryptographic protection for some aspects of BGP update messages, but in its current form BGPsec doesn't offer any protection against route leaks.

For the types of route leaks enumerated in [I-D.ietf-grow-route-leak-problem-definition], where the current OV method doesn't offer a solution, this document proposes an enhancement that would significantly extend the detection and mitigation capabilities of BGP. The solution involves carrying a per-hop route-leak protection (RLP) field in BGP updates. The RLP field is proposed be carried in an optional transitive path attribute. The solution is meant to be initially implemented as an enhancement of BGP without requiring BGPsec. However, when BGPsec is deployed in the future, the solution can be incorporated in BGPsec, enabling cryptographic protection for the RLP field. That would be one way of implementing the proposed solution in a secure way. It is not claimed that the solution detects all possible types of route leaks but it detects several types, especially considering some significant route-leak occurrences that have been observed in recent years. The document also includes a stopgap method (in Section 4) for detection and mitigation of route leaks for an intermediate phase when OV is deployed but BGP protocol on the wire is unchanged.

## 2. Related Prior Work

The basic idea and mechanism embodied in the proposed solution is based on setting an attribute in BGP route announcement to manage the transmission/receipt of the announcement based on the type of neighbor (e.g. customer, transit provider, etc.). Documented prior work related to said basic idea and mechanism dates back to at least

the 1980's. Some examples of prior work are: (1) Information flow rules described in [proceedings-sixth-ietf] (see pp. 195-196); (2) Link Type described in [RFC1105-obsolete] (see pp. 4-5); (3) Hierarchical Recording described in [draft-kunzinger-idrp-ISO10747-01] (see Section 6.3.1.12). The problem of route leaks and possible solution mechanisms based on encoding peering-link type information, e.g. P2C (i.e. Transit-Provider to Customer), C2P (i.e. Customer to Transit-Provider), p2p (i.e. peer to peer) etc., in BGPsec updates and protecting the same under BGPsec path signatures have been discussed in IETF SIDR WG at least since 2011. Dickson developed the initial Internet draft of these mechanisms in a BGPsec context; see [draft-dickson-sidr-route-leak-solns]. The draft expired in 2012. [draft-dickson-sidr-route-leak-solns] defined neighbor relationships on a per link basis, but in the current draft the relationship is encoded per prefix, as routes for prefixes with different business models are often sent over the same link. Also [draft-dickson-sidr-route-leak-solns] proposed a second signature block for the link type encoding, separate from the path signature block in BGPsec. By contrast, in the current draft when BGPsec-based solution is considered, cryptographic protection is provided for Route-Leak Protection (RLP) encoding using the same signature block as that for path signatures (see Section 3.1).

### 3. Mechanisms for Detection and Mitigation of Route Leaks

Referring to the enumeration of route leaks discussed in [I-D.ietf-grow-route-leak-problem-definition], Table 1 summarizes the route-leak detection capability offered by OV and BGPsec for different types of route leaks. (Note: Prefix filtering is not considered here in this table. Please see Section 4.)

A detailed explanation of the contents of Table 1 is as follows. It is readily observed that route leaks of Types 1, 2, 3, and 4 are not detected by OV or BGPsec in its current form. Clearly, Type 5 route leak involves re-origination or hijacking, and hence can be detected by OV. In the case of Type 5 route leak, there would be no existing ROAs to validate a re-originated prefix or more specific, but instead a covering ROA would normally exist with the legitimate AS, and hence the update will be considered Invalid by OV.

Type of Route Leak	Current State of Detection Coverage
Type 1: Hairpin Turn with Full Prefix	Neither OV nor BGPsec (in its current form) detects Type 1.
Type 2: Lateral ISP-ISP-ISP Leak	Neither OV nor BGPsec (in its current form) detects Type 2.
Type 3: Leak of Transit-Provider Prefixes to Peer	Neither OV nor BGPsec (in its current form) detects Type 3.
Type 4: Leak of Peer Prefixes to Transit Provider	Neither OV nor BGPsec (in its current form) detects Type 4.
Type 5: Prefix Re-Origination with Data Path to Legitimate Origin	OV detects Type 5.
Type 6: Accidental Leak of Internal Prefixes and More Specifics	For internal prefixes never meant to be routed on the Internet, OV helps detect their leak; they might either have no covering ROA or have an AS0-ROA to always filter them. In the case of accidental leak of more specifics, OV may offer some detection due to ROA maxLength.

Table 1: Examination of Route-Leak Detection Capability of Origin Validation and Current BGPsec Path Validation

In the case of Type 6 leaks involving internal prefixes that are not meant to be routed in the Internet, they are likely to be detected by OV. That is because such prefixes might either have no covering ROA or have an AS0-ROA to always filter them. In the case of Type 6 leaks that are due to accidental leak of more specifics, they may be detected due to violation of ROA maxLength. BGPsec (i.e. path validation) in its current form does not detect Type 6. However, route leaks of Type 6 are least problematic due to the following reasons. In the case of leak of more specifics, the offending AS is itself the legitimate destination of the leaked more-specific prefixes. Hence, in most cases of this type, the data traffic is neither misrouted nor denied service. Also, leaked announcements of Type 6 are short-lived and typically withdrawn quickly following the announcements. Further, the MaxPrefix limit may kick-in in some

receiving routers and that helps limit the propagation of sometimes large number of leaked routes of Type 6.

Realistically, BGPsec may take a much longer time being deployed than OV. Hence solution proposals for route leaks should consider both scenarios: (A) OV only (without BGPsec) and (B) OV plus BGPsec. Assuming an initial scenario A, and based on the above discussion and Table 1, it is evident that in our proposed solution method, we need to focus primarily on route leaks of Types 1, 2, 3, and 4. In Section 3.1 and Section 3.2, we describe a simple addition to BGP that facilitates detection of route leaks of Types 1, 2, 3, and 4. The simple addition involves a Route-Leak Protection (RLP) field, which is carried in an optional transitive path attribute in BGP. When BGPsec is deployed, the RLP field will be accommodated in the existing Flags field (see [I-D.ietf-sidr-bgpsec-protocol]) which is cryptographically protected under path signatures.

### 3.1. Route-Leak Protection (RLP) Field Encoding by Sending Router

The key principle is that, in the event of a route leak, a receiving router in a transit-provider AS (e.g. referring to Figure 1, ISP2 (AS2) router) should be able to detect from the update message that its customer AS (e.g. AS3 in Figure 1) SHOULD NOT have forwarded the update (towards the transit-provider AS). This means that at least one of the ASes in the AS path of the update has indicated that it sent the update to its customer or lateral (i.e. non-transit) peer, but forbade any subsequent 'Up' forwarding (i.e. from a customer AS to its transit-provider AS). For this purpose, a Route-Leak Protection (RLP) field to be set by a sending router is proposed to be used for each AS hop.



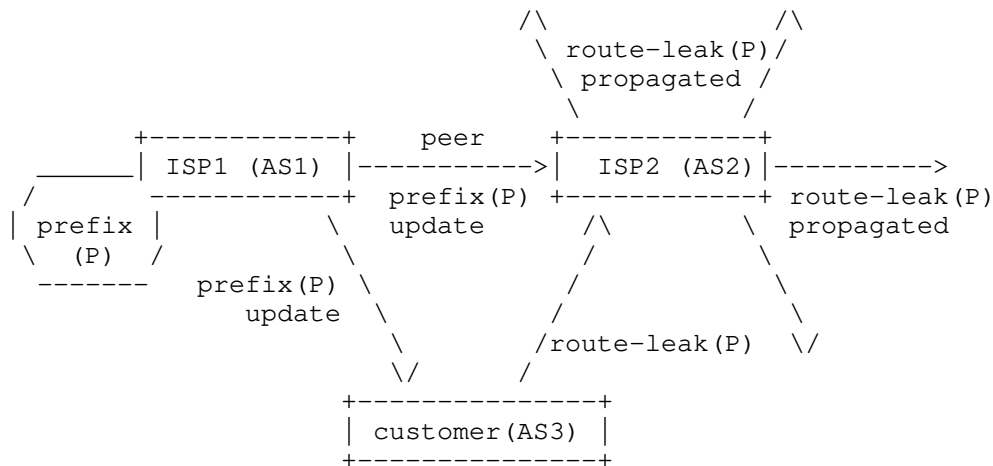


Figure 1: Illustration of the basic notion of a route leak.

For the purpose of route-leak detection and mitigation proposed in this document, the RLP field value SHOULD be set to one of two values as follows:

- o 00: This is the default value (i.e. "nothing specified"),
- o 01: This is the 'Do not Propagate Up or Lateral' indication; sender indicating that the route SHOULD NOT be forwarded 'Up' towards a transit-provider AS or to a lateral (i.e. non-transit) peer AS.

The RLP indications SHOULD be set on a per prefix and per neighbor AS basis. This is because updates for prefixes with different business models are often sent over the same link between ASes.

There are two different scenarios when a sending AS SHOULD set the '01' indication in an update: (1) when sending the update to a customer AS, and (2) when sending the update to a lateral peer (i.e. non-transit) AS. In essence, in both scenarios, the intent of '01' indication is that the neighbor AS and any receiving AS along the subsequent AS path SHOULD NOT forward the update 'Up' towards its (receiving AS's) transit-provider AS or laterally towards its peer (i.e. non-transit) AS. When sending an update 'Up' to a transit-provider AS, the RLP encoding should be set to the default value of '00'. When a sending AS sets the RLP encoding to '00', it is indicating to the receiving AS that the update can be propagated in any direction (i.e. towards transit-provider, customer, or lateral peer). This two-state specification in the RLP field can be shown to

work for detection and mitigation of route leaks of Types 1, 2, 3, and 4 which are the focus here (see Section 3.2 and Section 3.3). The '10' and '11' values in the RLP field (assuming that two bits are used to encode it) are currently unassigned, and reserved for possible future use.

The proposed RLP encoding SHOULD be carried in BGP-4 [RFC4271] updates in an optional transitive path attribute. In BGPsec enabled routers, the RLP encoding SHOULD be accommodated in the existing Flags field in BGPsec updates. The Flags field is part of the Secure\_Path Segment in BGPsec updates [I-D.ietf-sidr-bgpsec-protocol]. It is one octet long, and one Flags field is available for each AS hop, and currently only the first bit is used in BGPsec. So there are 7 bits that are currently unused in the Flags field. Two (or more if needed) of these bits can be designated for the RLP field. Since the BGPsec protocol specification requires a sending AS to include the Flags field in the data that are signed over, the RLP field for each hop (assuming it would be part of the Flags field) will be protected under the sending AS's signature.

### 3.2. Recommended Actions at a Receiving Router for Detection of Route Leaks

We provide here an example set of receiver actions that work to detect and mitigate route leaks of Types 1, 2, 3, and 4. This example algorithm serves as a proof of concept. However, other receiver algorithms or procedures can be designed (based on the same sender specification as in Section 3.1) and may perform with greater efficacy, and are by no means excluded.

A recommended receiver algorithm for detecting a route leak is as follows:

A receiving router SHOULD mark an update as a 'Route Leak' if ALL of the following conditions hold true:

1. The update is received from a customer or lateral peer AS.
2. The update has the RLP field set to '01' (i.e. 'Do not Propagate Up or Lateral') indication for one or more hops (excluding the most recent) in the AS path.

The reason for stating "excluding the most recent" in the above algorithm is as follows. An ISP should look at RLP values set by ASes preceding the immediate sending AS in order to ascertain a leak. The receiving router already knows that the most recent hop in the update is from its customer or lateral-peer AS to itself, and it does

not need to rely on the RLP field value set by said AS for detection of route leaks.

If the RLP encoding is secured by BGPsec (see Section 3.1) and hence protected against tampering by intermediate ASes, then there would be added certainty in the route-leak detection algorithm described above (see discussions in Section 5.1 and Section 5.2).

A receiving router expects the RLP field value for any hop in the AS path to be either 00 or 01. However, if a different value (say, 10 or 11) is found in the RLP field, then an error condition will get flagged, and any further action is TBD.

### 3.3. Possible Actions at a Receiving Router for Mitigation

After applying the above detection algorithm, a receiving router may use any policy-based algorithm of its own choosing to mitigate any detected route leaks. An example receiver algorithm for mitigating a route leak is as follows:

- o If an update from a customer or lateral peer AS is marked as a 'Route Leak', then the receiving router SHOULD prefer an alternate unmarked route if available.
- o If no alternate unmarked route is available, then the route marked as a 'Route Leak' MAY be accepted.

A basic principle here is that the presence of '01' value in the RLP field corresponding to one or more AS hops in the AS path of an update coming from a customer AS informs a receiving router in a transit-provider AS that a route leak is likely occurring. The transit-provider AS then overrides the "prefer customer route" policy, and instead prefers an alternate 'clean' route learned from another customer, a lateral peer, or a transit provider over the 'marked' route from the customer in question.

### 4. Stopgap Solution when Only Origin Validation is Deployed

Here we describe a stopgap method for detection and mitigation of route leaks for the intermediate phase when OV is deployed but BGP protocol on the wire is unchanged. The stopgap solution can be in the form of construction of a prefix filter list from ROAs. A suggested procedure for constructing such a list comprises of the following steps:

- o ISP makes a list of all the ASes (Cust\_AS\_List) that are in its customer cone (ISP's own AS is also included in the list). (Some

of the ASes in Cust\_AS\_List may be multi-homed to another ISP and that is OK.)

- o ISP downloads from the RPKI repositories a complete list (Cust\_ROA\_List) of valid ROAs that contain any of the ASes in Cust\_AS\_List.
- o ISP creates a list of all the prefixes (Cust\_Prfx\_List) that are contained in any of the ROAs in Cust\_ROA\_List.
- o Cust\_Prfx\_List is the allowed list of prefixes that is permitted by the ISP's AS, and will be forwarded by the ISP to upstream ISPs, customers, and peers.
- o A route for a prefix that is not in Cust\_Prfx\_List but announced by one of ISP's customers is 'marked' as a potential route leak. Further, the ISP's router SHOULD prefer an alternate route that is Valid (i.e. valid according to origin validation) and 'clean' (i.e. not marked) over the 'marked' route. The alternate route may be from a peer, transit provider, or different customer.

Special considerations with regard to the above procedure may be needed for DDoS mitigation service providers. They typically originate or announce a DDoS victim's prefix to their own ISP on a short notice during a DDoS emergency. Some provisions would need to be made for such cases, and they can be determined with the help of inputs from DDoS mitigation service providers.

For developing a list of all the ASes (Cust\_AS\_List) that are in the customer cone of an ISP, the AS path based Outbound Route Filter (ORF) technique [draft-ietf-idr-aspath-orf] can be helpful (see discussion in Section 5.4).

## 5. Design Rationale and Discussion

In this section, we will try to provide design justifications for the methodology specified in Section 3, and also answer some questions that are anticipated or have been raised in IETF IDR/SIDR meetings.

### 5.1. Is route-leak solution without cryptographic protection a serious attack vector?

It has been asked if a route-leak solution without BGPsec, i.e. when RLP bits are not protected, can turn into a serious new attack vector. The answer seems to be: not really! Even the NLRI and AS\_PATH in BGP updates are attack vectors, and RPKI/OV/BGPsec seek to fix that. Consider the following. Say, if 99% of route leaks are accidental and 1% are malicious, and if route-leak solution without

BGPsec eliminates the 99%, then perhaps it is worth it (step in the right direction). When BGPsec comes into deployment, the route-leak protection (RLP) bits can be mapped into BGPsec (using the Flags field) and then necessary security will be in place as well (within each BGPsec island as and when they emerge).

Further, let us consider the worst-case damage that can be caused by maliciously manipulating the RLP bits in an implementation without cryptographic protection (i.e. sans BGPsec). Manipulation of the RLP bits can result in one of two types of attacks: (a) Upgrade attack and (b) Downgrade attack. Descriptions and discussions about these attacks follow. In what follows, P2C stands for transit provider to customer (Down); C2P stands for customer to transit provider (Up), and p2p stands for peer to peer (lateral or non-transit relationship).

(a) Upgrade attack: An AS that wants to intentionally leak a route would alter the RLP encodings for the preceding hops from '01' (i.e. 'Do not Propagate Up or Lateral') to '00' (default) wherever applicable. This poses no problem for a route that keeps propagating in the 'Down' (P2C) direction. However, for a route that propagates 'Up' (C2P) or 'Lateral' (p2p), the worst that can happen is that a route leak goes undetected. That is, a receiving router would not be able to detect the leak for the route in question by the RLP mechanism described here. However, the receiving router may still detect and mitigate it in some cases by applying other means such as prefix filters [RFC7454]. If some malicious leaks go undetected (when RLP is deployed without BGPsec) that is possibly a small price to pay for the ability to detect the bulk of route leaks that are accidental.

(b) Downgrade attack: RLP encoding is set to '01' (i.e. 'Do not Propagate Up or Lateral') when it should be set to '00' (default). This would result in a route being mis-detected and marked as a route leak. By default RLP encoding is set to '00', and that helps reduce errors of this kind (i.e. accidental downgrade incidents). Every AS or ISP wants reachability for prefixes it originates and for its customer prefixes. So an AS or ISP is not likely to change an RLP value '00' to '01' intentionally. If a route leak is detected (due to intentional or accidental downgrade) by a receiving router, it would prefer an alternate 'clean' route from a transit provider or peer over a 'marked' route from a customer. It may end up with a suboptimal path. In order to have reachability, the receiving router would accept a 'marked' route if there is no alternative that is 'clean'. So RLP downgrade attacks (intentional or accidental) would be quite rare, and the consequences do not appear to be grave.

## 5.2. Combining results of route-leak detection, OV and BGPsec validation for path selection decision

Combining the results of route-leak detection, OV, and BGPsec validation for path selection decision is up to local policy in a receiving router. As an example, a router may always give precedence to outcomes of OV and BGPsec validation over that of route-leak detection. That is, if an update fails OV or BGPsec validation, then the update is not considered a candidate for path selection. Instead, an alternate update is chosen that passed OV and BGPsec validation and additionally was not marked as route leak.

If only OV is deployed (and not BGPsec), then there are six possible combinations between OV and route-leak detection outcomes. Because there are three possible outcomes for OV (NotFound, Valid, and Invalid) and two possible outcomes for route-leak detection (marked as leak and not marked). If OV and BGPsec are both deployed, then there are twelve possible combinations between OV, BGPsec validation, and route-leak detection outcomes. As stated earlier, since BGPsec protects the RLP encoding, there would be added certainty in route-leak detection outcome if an update is BGPsec valid (see Section 5.1).

## 5.3. Are there cases when valley-free violations can be considered legitimate?

There are studies in the literature [Anwar] [Giotsas] [Wijchers] observing and analyzing the behavior of routes announced in BGP updates using data gathered from the Internet. In particular, the studies have focused on how often there appear to be valley-free (e.g. Gao-Rexford [Gao] model) violations, and if they can be explained [Anwar]. One important consideration for explanation of violations is per-prefix routing policies, i.e. routes for prefixes with different business models are often sent over the same link. One encouraging result reported in [Anwar] is that when per-prefix routing policies are taken into consideration in the data analysis, more than 80% of the observed routing decisions fit the valley-free model (see Section 4.3 and SPA-1 data in Figure 2). The authors in [Anwar] also observe, "it is well known that this model [the basic Gao-Rexford model and some variations of it] fails to capture many aspects of the interdomain routing system. These aspects include AS relationships that vary based on the geographic region or destination prefix, and traffic engineering via hot-potato routing or load balancing." So there may be potential for explaining the remaining (20% or less) violations of valley-free as well.

One major design factor in the methodology described in this document is that the Route-Leak Protection (RLP) encoding is per prefix. So

the proposed solution is consistent with ISPs' per-prefix routing policies. Large global and other major ISPs will be the likely early adopters, and they are expected to have expertise in configuring policies (including per prefix policies, if applicable), and make proper use of the RLP indications on a per prefix basis. When said large ISPs participate in this solution deployment, it is envisioned that they would form a ring of protection against route leaks, and co-operatively avoid many of the common types of route leaks that are observed. Route leaks may still happen occasionally within the customer cones (if some customer ASes are not participating or not diligently implementing RLP), but said leaks would be much less likely to propagate from one large participating ISP to another.

#### 5.4. Comparison with other methods, routing security BCP

It is reasonable to ask if techniques considered in BCPs such as [RFC7454] (BGP Operations and Security) and [NIST-800-54] may be adequate to address route leaks. The prefix filtering recommendations in the BCPs may be complementary but not adequate. The difficulty is in ISPs' ability to construct prefix filters that represent their customer cones (CC) accurately, especially when there are many levels in the hierarchy within the CC. In the RLP-encoding based solution described here, AS operators signal for each route propagated, if it SHOULD NOT be subsequently propagated to a transit provider or peer.

AS path based Outbound Route Filter (ORF) described in [draft-ietf-idr-aspath-orf] is also an interesting complementary technique. It can be used as an automated collaborative messaging system (implemented in BGP) for ISPs to try to develop a complete view of the ASes and AS paths in their CCs. Once an ISP has that view, then AS path filters can be possibly used to detect route leaks. One limitation of this technique is that it cannot duly take into account the fact that routes for prefixes with different business models are often sent over the same link between ASes. Also, the success of AS path based ORF depends on whether ASes at all levels of the hierarchy in a CC participate and provide accurate information (in the ORF messages) about the AS paths they expect to have in their BGP updates.

#### 6. Summary

It should be emphasized once again that the proposed route-leak detection method using the RLP encoding is not intended to cover all forms of route leaks. However, we feel that the solution covers several important types of route leaks, especially considering some significant route-leak attacks or occurrences that have been frequently observed in recent years. The solution can be implemented

in BGP without necessarily tying it to BGPsec. The proposed solution without BGPsec can detect and mitigate accidental route leaks, and the same with BGPsec can detect and mitigate both accidental and malicious route leaks. Carrying the proposed RLP encoding in an optional transitive path attribute in BGP is proposed, but in order to prevent abuse, the RLP encoding would require cryptographic protection. Incorporating the RLP encoding in the BGPsec Flags field is one way of implementing it securely using an already existing protection mechanism provided in BGPsec path signatures.

## 7. Security Considerations

The proposed Route-Leak Protection (RLP) field requires cryptographic protection in order to prevent malicious route leaks. Since it is proposed that the RLP field be included in the Flags field in the Secure\_Path Segment in BGPsec updates, the cryptographic security mechanisms in BGPsec are expected to also apply to the RLP field. The reader is therefore directed to the security considerations provided in [I-D.ietf-sidr-bgpsec-protocol].

## 8. IANA Considerations

No updates to the registries are suggested by this document.

## 9. Acknowledgements

The authors wish to thank Danny McPherson and Eric Osterweil for discussions related to this work. Also, thanks are due to Jared Mauch, Jeff Haas, Warren Kumari, Amogh Dhamdhere, Jakob Heitz, Geoff Huston, Randy Bush, Ruediger Volk, Sue Hares, Wes George, Chris Morrow, and Sandy Murphy for comments, suggestions, and critique. The authors are also thankful to Padma Krishnaswamy, Oliver Borchert, and Okhee Kim for their comments and review.

## 10. References

### 10.1. Normative References

- [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, <<http://www.rfc-editor.org/info/rfc4271>>.

### 10.2. Informative References



- [Anwar] Anwar, R., Niaz, H., Choffnes, D., Cunha, I., Gill, P., and N. Katz-Bassett, "Investigating Interdomain Routing Policies in the Wild", ACM Internet Measurement Conference (IMC), October 2015, <<http://www.cs.usc.edu/assets/007/94928.pdf>>.
- [Cowie2010] Cowie, J., "China's 18 Minute Mystery", Dyn Research/Renesys Blog, November 2010, <<http://research.dyn.com/2010/11/chinas-18-minute-mystery/>>.
- [Cowie2013] Cowie, J., "The New Threat: Targeted Internet Traffic Misdirection", Dyn Research/Renesys Blog, November 2013, <<http://research.dyn.com/2013/11/mitm-internet-hijacking/>>.
- [draft-dickson-sidr-route-leak-solns] Dickson, B., "Route Leaks -- Proposed Solutions", IETF Internet Draft (expired), March 2012, <<https://tools.ietf.org/html/draft-dickson-sidr-route-leak-solns-01>>.
- [draft-ietf-idr-aspath-orf] Patel, K. and S. Hares, "AS path Based Outbound Route Filter for BGP-4", IETF Internet Draft (expired), August 2007, <<https://tools.ietf.org/html/draft-ietf-idr-aspath-orf-09>>.
- [draft-kunzinger-idrp-ISO10747-01] Kunzinger, C., "Inter-Domain Routing Protocol (IDRP)", IETF Internet Draft (expired), November 1994, <<https://tools.ietf.org/pdf/draft-kunzinger-idrp-ISO10747-01.pdf>>.
- [Gao] Gao, L. and J. Rexford, "Stable Internet routing without global coordination", IEEE/ACM Transactions on Networking, December 2001, <<http://www.cs.princeton.edu/~jrex/papers/sigmetrics00.long.pdf>>.
- [Gill] Gill, P., Schapira, M., and S. Goldberg, "A Survey of Interdomain Routing Policies", ACM SIGCOMM Computer Communication Review, January 2014, <<https://www.cs.bu.edu/~goldbe/papers/survey.pdf>>.

- [Giotsas] Giotsas, V. and S. Zhou, "Valley-free violation in Internet routing - Analysis based on BGP Community data", IEEE ICC 2012, June 2012.
- [Hiran] Hiran, R., Carlsson, N., and P. Gill, "Characterizing Large-scale Routing Anomalies: A Case Study of the China Telecom Incident", PAM 2013, March 2013, <<http://www3.cs.stonybrook.edu/~phillipa/papers/CTelecom.html>>.
- [Huston2012] Huston, G., "Leaking Routes", March 2012, <<http://labs.apnic.net/blabs/?p=139/>>.
- [Huston2014] Huston, G., "What's so special about 512?", September 2014, <<http://labs.apnic.net/blabs/?p=520/>>.
- [I-D.ietf-grow-route-leak-problem-definition] Sriram, K., Montgomery, D., McPherson, D., Osterweil, E., and B. Dickson, "Problem Definition and Classification of BGP Route Leaks", draft-ietf-grow-route-leak-problem-definition-04 (work in progress), February 2016.
- [I-D.ietf-sidr-bgpsec-protocol] Lepinski, M., "BGPsec Protocol Specification", draft-ietf-sidr-bgpsec-protocol-14 (work in progress), December 2015.
- [Kapela-Pilosov] Pilosov, A. and T. Kapela, "Stealing the Internet: An Internet-Scale Man in the Middle Attack", DEFCON-16 Las Vegas, NV, USA, August 2008, <<https://www.defcon.org/images/defcon-16/dc16-presentations/defcon-16-pilosov-kapela.pdf>>.
- [Kephart] Kephart, N., "Route Leak Causes Amazon and AWS Outage", ThousandEyes Blog, June 2015, <<https://blog.thousandeyes.com/route-leak-causes-amazon-and-aws-outage>>.
- [Khare] Khare, V., Ju, Q., and B. Zhang, "Concurrent Prefix Hijacks: Occurrence and Impacts", IMC 2012, Boston, MA, November 2012, <<http://www.cs.arizona.edu/~bzhang/paper/12-imc-hijack.pdf>>.

- [Labovitz] Labovitz, C., "Additional Discussion of the April China BGP Hijack Incident", Arbor Networks IT Security Blog, November 2010,  
<<http://www.arbornetworks.com/asert/2010/11/additional-discussion-of-the-april-china-bgp-hijack-incident/>>.
- [LRL] Khare, V., Ju, Q., and B. Zhang, "Large Route Leaks", Project web page, 2012,  
<<http://nrl.cs.arizona.edu/projects/lrsl-events-from-2003-to-2009/>>.
- [Luckie] Luckie, M., Huffaker, B., Dhamdhere, A., Giotsas, V., and kc. claffy, "AS Relationships, Customer Cones, and Validation", IMC 2013, October 2013,  
<<http://www.caida.org/~amogh/papers/asrank-IMC13.pdf>>.
- [Madory] Madory, D., "Why Far-Flung Parts of the Internet Broke Today", Dyn Research/Renesys Blog, September 2014,  
<<http://research.dyn.com/2014/09/why-the-internet-broke-today/>>.
- [Mauch] Mauch, J., "BGP Routing Leak Detection System", Project web page, 2014,  
<<http://puck.nether.net/bgp/leakinfo.cgi/>>.
- [Mauch-nanog] Mauch, J., "Detecting Routing Leaks by Counting", NANOG-41 Albuquerque, NM, USA, October 2007,  
<<https://www.nanog.org/meetings/nanog41/presentations/mauch-lightning.pdf>>.
- [NIST-800-54] Kuhn, D., Sriram, K., and D. Montgomery, "Border Gateway Protocol Security", NIST Special Publication 800-54, July 2007, <<http://csrc.nist.gov/publications/nistpubs/800-54/SP800-54.pdf>>.
- [Paseka] Paseka, T., "Why Google Went Offline Today and a Bit about How the Internet Works", CloudFare Blog, November 2012,  
<<http://blog.cloudflare.com/why-google-went-offline-today-and-a-bit-about/>>.
- [proceedings-sixth-ietf] Gross, P., "Proceedings of the April 22-24, 1987 Internet Engineering Task Force", April 1987,  
<<https://www.ietf.org/proceedings/06.pdf>>.

- [RFC1105-obsolete]  
Lougheed, K. and Y. Rekhter, "A Border Gateway Protocol (BGP)", IETF RFC (obsolete), June 1989, <<https://tools.ietf.org/html/rfc1105>>.
- [RFC6811] Mohapatra, P., Scudder, J., Ward, D., Bush, R., and R. Austein, "BGP Prefix Origin Validation", RFC 6811, DOI 10.17487/RFC6811, January 2013, <<http://www.rfc-editor.org/info/rfc6811>>.
- [RFC7454] Durand, J., Pepelnjak, I., and G. Doering, "BGP Operations and Security", BCP 194, RFC 7454, DOI 10.17487/RFC7454, February 2015, <<http://www.rfc-editor.org/info/rfc7454>>.
- [Toonk] Toonk, A., "What Caused Today's Internet Hiccup", August 2014, <<http://www.bgpmon.net/what-caused-todays-internet-hiccup/>>.
- [Toonk2015-A] Toonk, A., "What caused the Google service interruption", March 2015, <<http://www.bgpmon.net/what-caused-the-google-service-interruption/>>.
- [Toonk2015-B] Toonk, A., "Massive route leak causes Internet slowdown", June 2015, <<http://www.bgpmon.net/massive-route-leak-cause-internet-slowdown/>>.
- [Wijchers] Wijchers, B. and B. Overeinder, "Quantitative Analysis of BGP Route Leaks", RIPE-69, November 2014, <<https://ripe69.ripe.net/presentations/157-RIPE-69-Routing-WG.pdf>>.
- [Zmijewski] Zmijewski, E., "Indonesia Hijacks the World", Dyn Research/Renesys Blog, April 2014, <<http://research.dyn.com/2014/04/indonesia-hijacks-world/>>.

## Authors' Addresses

Kotikalapudi Sriram  
US NIST

Email: [ksriram@nist.gov](mailto:ksriram@nist.gov)

Doug Montgomery  
US NIST

Email: dougm@nist.gov

Brian Dickson

Email: brian.peter.dickson@gmail.com

Keyur Patel  
Cisco

Email: keyupate@cisco.com

Andrei Robachevsky  
Internet Society

Email: robachevsky@isoc.org

Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: June 5, 2016

J. Dong  
M. Chen  
Huawei Technologies  
H. Gredler  
Individual Contributor  
S. Previdi  
Cisco Systems, Inc.  
J. Tantsura  
Ericsson  
December 3, 2015

Distribution of MPLS Traffic Engineering (TE) LSP State using BGP  
draft-ietf-idr-te-lsp-distribution-04

## Abstract

This document describes a mechanism to collect the Traffic Engineering (TE) LSP information using BGP. Such information can be used by external components for path reoptimization, service placement, and network visualization.

## 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 <http://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 June 5, 2016.

## Copyright Notice

Copyright (c) 2015 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 . . . . .	2
2. Carrying LSP State Information in BGP . . . . .	4
2.1. MPLS TE LSP Information . . . . .	4
2.2. IPv4/IPv6 MPLS TE LSP NLRI . . . . .	5
2.2.1. MPLS TE LSP Descriptors . . . . .	6
2.3. LSP State Information . . . . .	8
2.3.1. RSVP Objects . . . . .	10
2.3.2. PCE Objects . . . . .	11
2.3.3. SR Encap TLVs . . . . .	11
3. Operational Considerations . . . . .	12
4. IANA Considerations . . . . .	12
4.1. BGP-LS NLRI-Types . . . . .	12
4.2. BGP-LS Protocol-IDs . . . . .	12
4.3. BGP-LS Descriptors TLVs . . . . .	13
4.4. BGP-LS LSP-State TLV Protocol Origin . . . . .	13
5. Security Considerations . . . . .	14
6. Acknowledgements . . . . .	14
7. References . . . . .	14
7.1. Normative References . . . . .	14
7.2. Informative References . . . . .	15
Authors' Addresses . . . . .	16

## 1. Introduction

In some network environments, the state of established Multi-Protocol Label Switching (MPLS) Traffic Engineering (TE) Label Switched Paths (LSPs) and Tunnels in the network are required by components external to the network domain. Usually this information is directly maintained by the ingress Label Edge Routers (LERs) of the MPLS TE LSPs.

One example of using the LSP information is stateful Path Computation Element (PCE) [I-D.ietf-pce-stateful-pce], which could provide benefits in path reoptimization. While some extensions are proposed in Path Computation Element Communication Protocol (PCEP) for the Path Computation Clients (PCCs) to report the LSP states to the PCE, this mechanism may not be applicable in a management-based PCE architecture as specified in section 5.5 of [RFC4655]. As illustrated in the figure below, the PCC is not an LSR in the routing domain, thus the head-end nodes of the TE-LSPs may not implement the PCEP protocol. In this case a general mechanism to collect the TE-LSP states from the ingress LERs is needed. This document proposes an LSP state collection mechanism complementary to the mechanism defined in [I-D.ietf-pce-stateful-pce].

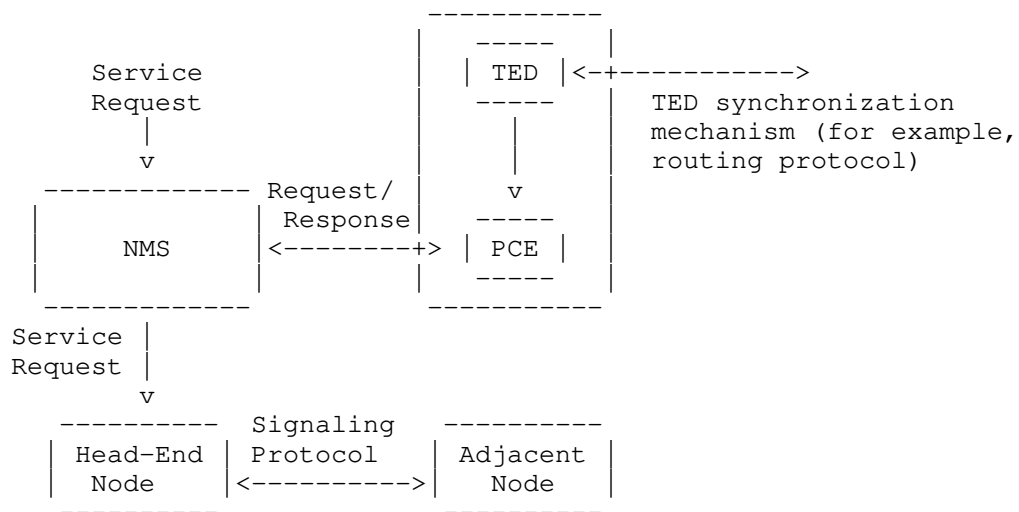


Figure 1. Management-Based PCE Usage

In networks with composite PCE nodes as specified in section 5.1 of [RFC4655], PCE is implemented on several routers in the network, and the PCCs in the network can use the mechanism described in [I-D.ietf-pce-stateful-pce] to report the LSP information to the PCE nodes. An external component may also need to collect the LSP information from all the PCEs in the network to obtain a global view of the LSP state in the network.

In multi-area or multi-AS scenarios, each area or AS can have a child PCE to collect the LSP state in its own domain, in addition, a parent PCE needs to collect LSP information from multiple child PCEs to obtain a global view of LSPs inside and across the domains involved.



In another network scenario, a centralized controller is used for service placement. Obtaining the TE LSP state information is quite important for making appropriate service placement decisions with the purpose to both meet the application's requirements and utilize network resources efficiently.

The Network Management System (NMS) may need to provide global visibility of the TE LSPs in the network as part of the network visualization function.

BGP has been extended to distribute link-state and traffic engineering information to external components [I-D.ietf-idr-ls-distribution]. Using the same protocol to collect TE LSP information is desirable for these external components since this avoids introducing multiple protocols for network information collection. This document describes a mechanism to distribute TE LSP information to external components using BGP.

## 2. Carrying LSP State Information in BGP

## 2.1. MPLS TE LSP Information

The MPLS TE LSP information is advertised in BGP UPDATE messages using the MP\_REACH\_NLRI and MP\_UNREACH\_NLRI attributes [RFC4760]. The "Link-State NLRI" defined in [I-D.ietf-idr-ls-distribution] is extended to carry the MPLS TE LSP information. BGP speakers that wish to exchange MPLS TE LSP information MUST use the BGP Multiprotocol Extensions Capability Code (1) to advertise the corresponding (AFI, SAFI) pair, as specified in [RFC4760].

The format of "Link-State NLRI" is defined in [I-D.ietf-idr-ls-distribution]. A new "NLRI Type" is defined for MPLS TE LSP Information as following:

- o NLRI Type: IPv4/IPv6 MPLS TE LSP NLRI (suggested codepoint value 5, to be assigned by IANA).

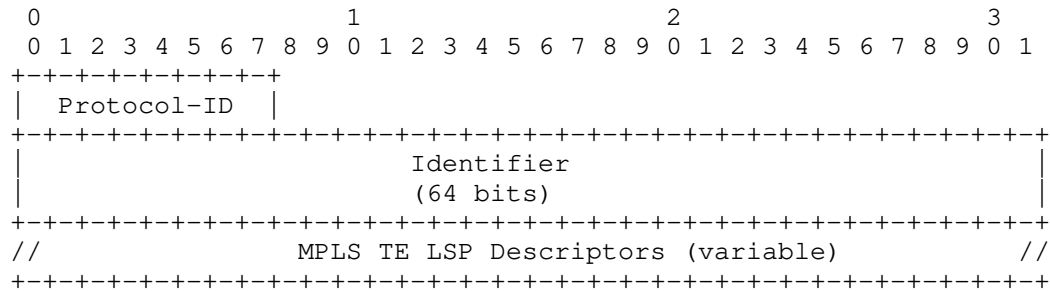
[I-D.ietf-idr-ls-distribution] defines the BGP-LS NLRI as follows:

[illegible]

This document defines a new NLRI-Type and its format: the IPv4/IPv6 MPLS TE LSP NLRI defined in the following section.

## 2.2. IPv4/IPv6 MPLS TE LSP NLRI

The IPv4/IPv6 MPLS TE LSP NLRI (NLRI Type 5. Suggested value, to be assigned by IANA) is shown in the following figure:



where:

- o Protocol-ID field specifies the type of signaling of the MPLS TE LSP. The following Protocol-IDs are defined (suggested values, to be assigned by IANA) and apply to the IPv4/IPv6 MPLS TE LSP NLRI:

Protocol-ID	NLRI information source protocol
7	RSVP-TE
8	Segment Routing

- o "Identifier" is an 8 octet value as defined in [I-D.ietf-idr-ls-distribution].
- o Following MPLS TE LSP Descriptors are defined:

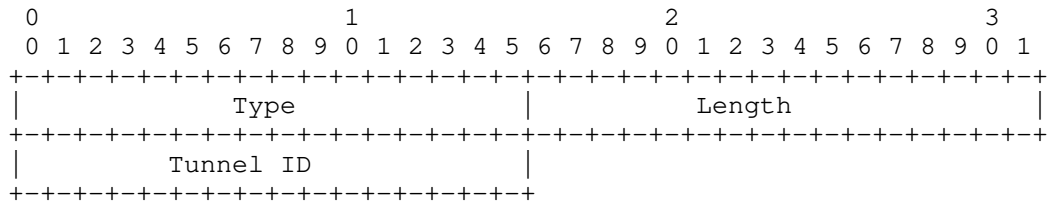
Codepoint	Descriptor TLV
267	Tunnel ID
268	LSP ID
269	IPv4/6 Tunnel Head-end address
270	IPv4/6 Tunnel Tail-end address
271	SR-ENCAP Identifier

### 2.2.1. MPLS TE LSP Descriptors

This sections defines the MPLS TE Descriptors TLVs.

#### 2.2.1.1. Tunnel Identifier (Tunnel ID)

The Tunnel Identifier TLV contains the Tunnel ID defined in [RFC3209] and has the following format:

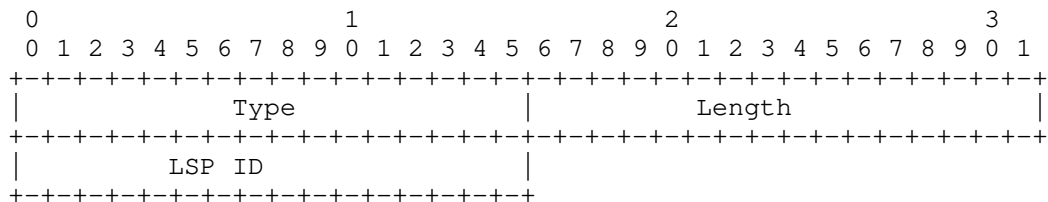


where:

- o Type: To be assigned by IANA (suggested value: 267)
- o Length: 2 octets.
- o Tunnel ID: 2 octets as defined in [RFC3209].

#### 2.2.1.2. LSP Identifier (LSP ID)

The LSP Identifier TLV contains the LSP ID defined in [RFC3209] and has the following format:



where:

- o Type: To be assigned by IANA (suggested value: 268)
- o Length: 2 octets.
- o LSP ID: 2 octets as defined in [RFC3209].

## 2.2.1.3. IPv4/IPv6 Tunnel Head-End Address

The IPv4/IPv6 Tunnel Head-End Address TLV contains the Tunnel Head-End Address defined in [RFC3209] and has following format:

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                                     |                                     |
|                                     Type                                Length
|                                     |                                     |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
//                                IPv4/IPv6 Tunnel Head-End Address (variable)                                //
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

where:

- o Type: To be assigned by IANA (suggested value: 269)
- o Length: 4 or 16 octets.

When the IPv4/IPv6 Tunnel Head-end Address TLV contains an IPv4 address, its length is 4 (octets).

When the IPv4/IPv6 Tunnel Head-end Address TLV contains an IPv6 address, its length is 16 (octets).

## 2.2.1.4. IPv4/IPv6 Tunnel Tail-End Address

The IPv4/IPv6 Tunnel Tail-End Address TLV contains the Tunnel Tail-End Address defined in [RFC3209] and has following format:

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                                     |                                     |
|                                     Type                                Length
|                                     |                                     |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
//                                IPv4/IPv6 Tunnel Tail-End Address (variable)                                //
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

where:

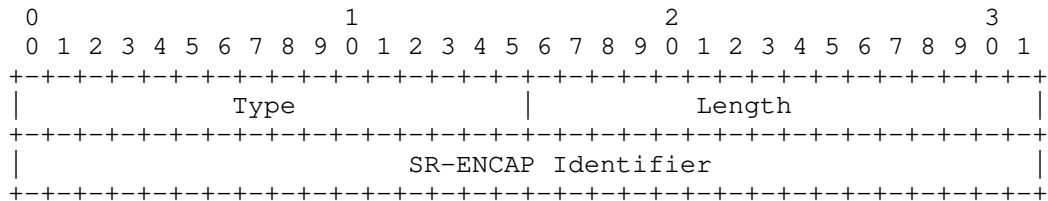
- o Type: To be assigned by IANA (suggested value: 270)
- o Length: 4 or 16 octets.

When the IPv4/IPv6 Tunnel Tail-end Address TLV contains an IPv4 address, its length is 4 (octets).

When the IPv4/IPv6 Tunnel Tail-end Address TLV contains an IPv6 address, its length is 16 (octets).

#### 2.2.1.5. SR-Encap TLV

The SR-ENCAP TLV contains the Identifier defined in [I-D.sreekantiah-idr-segment-routing-te] and has the following format:

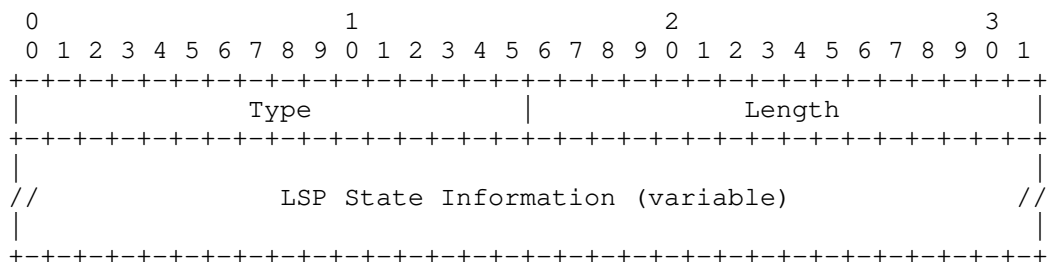


where:

- o Type: To be assigned by IANA (suggested value: 271)
- o Length: 4 octets.
- o SR-ENCAP Identifier: 4 octets as defined in [I-D.sreekantiah-idr-segment-routing-te].

#### 2.3. LSP State Information

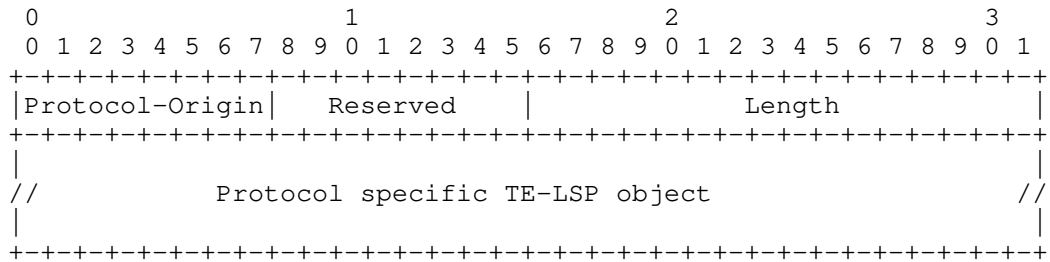
A new TLV called "LSP State TLV" (codepoint to be assigned by IANA), is used to describe the characteristics of the MPLS TE LSPs, which is carried in the optional non-transitive BGP Attribute "LINK\_STATE Attribute" defined in [I-D.ietf-idr-ls-distribution]. These MPLS TE LSP characteristics include the switching technology of the LSP, Quality of Service (QoS) parameters, route information, the protection mechanisms, etc.



LSP State TLV

Type: Suggested value 1158 (to be assigned by IANA)

LSP State Information: Consists of a set of TE-LSP objects as defined in [RFC3209],[RFC3473] and [RFC5440]. Rather than replicating all MPLS TE LSP related objects in this document, the semantics and encodings of the MPLS TE LSP objects are reused. These MPLS TE LSP objects are carried in the "LSP State Information" with the following format.



#### LSP State Information

The Protocol-Origin field identifies the protocol from which the contained MPLS TE LSP object originated. This allows for MPLS TE LSP objects defined in different protocols to be collected while avoiding the possible code collisions among these protocols. Three Protocol-Origins are defined in this document (suggested values, to be assigned by IANA)

Protocol Origin	LSP Object Origin
1	RSVP-TE
2	PCE
3	SR ENCAP

The 8-bit Reserved field SHOULD be set to 0 on transmission and ignored on receipt.

The Length field is set to the Length of the value field, which is the total length of the contained MPLS TE LSP object.

The Valued field is a MPLS-TE LSP object which is defined in the protocol identified by the Protocol-Origin field.

### 2.3.1. RSVP Objects

RSVP-TE objects are encoded in the "Value" field of the LSP State TLV and consists of MPLS TE LSP objects defined in RSVP-TE [RFC3209] [RFC3473]. Rather than replicating all MPLS TE LSP related objects in this document, the semantics and encodings of the MPLS TE LSP objects are re-used. These MPLS TE LSP objects are carried in the LSP State TLV.

When carrying RSVP-TE objects, the "Protocol-Origin" field is set to "RSVP-TE" (suggested value 1, to be assigned by IANA).

The following RSVP-TE Objects are defined:

- o SENDER\_TSPEC and FLOW\_SPEC [RFC2205]
- o SESSION\_ATTRIBUTE [RFC3209]
- o EXPLICIT\_ROUTE Object (ERO) [RFC3209]
- o ROUTE\_RECORD Object (RRO) [RFC3209]
- o FAST\_REROUTE Object [RFC4090]
- o DETOUR Object [RFC4090]
- o EXCLUDE\_ROUTE Object (XRO) [RFC4874]
- o SECONDARY\_EXPLICIT\_ROUTE Object (SERO) [RFC4873]
- o SECONDARY\_RECORD\_ROUTE (SRRO) [RFC4873]
- o LSP\_ATTRIBUTES Object [RFC5420]
- o LSP\_REQUIRED\_ATTRIBUTES Object [RFC5420]
- o PROTECTION Object [RFC3473] [RFC4872] [RFC4873]
- o ASSOCIATION Object [RFC4872]
- o PRIMARY\_PATH\_ROUTE Object [RFC4872]
- o ADMIN\_STATUS Object [RFC3473]
- o LABEL\_REQUEST Object [RFC3209] [RFC3473]

For the MPLS TE LSP Objects listed above, the corresponding sub-objects are also applicable to this mechanism. Note that this list

is not exhaustive, other MPLS TE LSP objects which reflect specific characteristics of the MPLS TE LSP can also be carried in the LSP state TLV.

#### 2.3.2. PCE Objects

PCE objects are encoded in the "Value" field of the MPLS TE LSP State TLV and consists of PCE objects defined in [RFC5440]. Rather than replicating all MPLS TE LSP related objects in this document, the semantics and encodings of the MPLS TE LSP objects are re-used. These MPLS TE LSP objects are carried in the LSP State TLV.

When carrying PCE objects, the "Protocol-Origin" field is set to "PCE" (suggested value 2, to be assigned by IANA).

The following PCE Objects are defined:

- o METRIC Object [RFC5440]
- o BANDWIDTH Object [RFC5440]

For the MPLS TE LSP Objects listed above, the corresponding sub-objects are also applicable to this mechanism. Note that this list is not exhaustive, other MPLS TE LSP objects which reflect specific characteristics of the MPLS TE LSP can also be carried in the LSP state TLV.

#### 2.3.3. SR Encap TLVs

SR-ENCAP objects are encoded in the "Value" field of the LSP State TLV and consists of SR-ENCAP objects defined in [I-D.sreekantiah-idr-segment-routing-te]. Rather than replicating all MPLS TE LSP related objects in this document, the semantics and encodings of the MPLS TE LSP objects are re-used. These MPLS TE LSP objects are carried in the LSP State TLV.

When carrying SR-ENCAP objects, the "Protocol-Origin" field is set to "SR-ENCAP" (suggested value 3, to be assigned by IANA).

The following SR-ENCAP Objects are defined:

- o ERO TLV [I-D.sreekantiah-idr-segment-routing-te]
- o Weight TLV [I-D.sreekantiah-idr-segment-routing-te]
- o Binding SID TLV [I-D.sreekantiah-idr-segment-routing-te]



For the MPLS TE LSP Objects listed above, the corresponding sub-objects are also applicable to this mechanism. Note that this list is not exhaustive, other MPLS TE LSP objects which reflect specific characteristics of the MPLS TE LSP can also be carried in the LSP state TLV.

### 3. Operational Considerations

The Existing BGP operational procedures apply to this document. No new operation procedures are defined in this document. The operational considerations as specified in [I-D.ietf-idr-ls-distribution] apply to this document.

In general the ingress nodes of the MPLS TE LSPs are responsible for the distribution of LSP state information, while other nodes on the LSP path MAY report the LSP information when needed. For example, the border routers in the inter-domain case will also distribute LSP state information since the ingress node may not have the complete information for the end-to-end path.

### 4. IANA Considerations

This document requires new IANA assigned codepoints.

#### 4.1. BGP-LS NLRI-Types

IANA maintains a registry called "Border Gateway Protocol - Link State (BGP-LS) Parameters" with a sub-registry called "BGP-LS NLRI-Types".

The following codepoints is suggested (to be assigned by IANA):

Type	NLRI Type	Reference
5	IPv4/IPv6 MPLS TE LSP NLRI	this document

#### 4.2. BGP-LS Protocol-IDs

IANA maintains a registry called "Border Gateway Protocol - Link State (BGP-LS) Parameters" with a sub-registry called "BGP-LS Protocol-IDs".

The following Protocol-ID codepoints are suggested (to be assigned by IANA):

Protocol-ID	NLRI information source protocol	Reference
7	RSVP-TE	this document
8	Segment Routing	this document

#### 4.3. BGP-LS Descriptors 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 (to be assigned by IANA):

TLV Code Point	Description	Value defined in
1158	LSP State TLV	this document
267	Tunnel ID TLV	this document
268	LSP ID TLV	this document
269	IPv4/6 Tunnel Head-end address TLV	this document
270	IPv4/6 Tunnel Tail-end address TLV	this document
271	SR-ENCAP Identifier TLV	this document

#### 4.4. BGP-LS LSP-State TLV Protocol Origin

This document requests IANA to maintain a new sub-registry under "Border Gateway Protocol - Link State (BGP-LS) Parameters". The new registry is called "Protocol Origin" and contains the codepoints allocated to the "Protocol Origin" field defined in Section 2.3. The registry contains the following codepoints (suggested values, to be assigned by IANA):

Protocol Origin	Description
1	RSVP-TE
2	PCE
3	SR-ENCAP

## 5. Security Considerations

Procedures and protocol extensions defined in this document do not affect the BGP security model. See [RFC6952] for details.

## 6. Acknowledgements

The authors would like to thank Dhruv Dhody, Mohammed Abdul Aziz Khalid, Lou Berger, Acee Lindem, Siva Sivabalan and Arjun Sreekantiah for their review and valuable comments.

## 7. References

### 7.1. Normative References

- [I-D.ietf-idr-ls-distribution]  
Gredler, H., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and TE Information using BGP", draft-ietf-idr-ls-distribution-13 (work in progress), October 2015.
- [I-D.sreekantiah-idr-segment-routing-te]  
Sreekantiah, A., Filsfils, C., Previdi, S., Sivabalan, S., Mattes, P., and J. Marcon, "Segment Routing Traffic Engineering Policy using BGP", draft-sreekantiah-idr-segment-routing-te-00 (work in progress), October 2015.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<http://www.rfc-editor.org/info/rfc2119>>.
- [RFC2205] Braden, R., Ed., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, DOI 10.17487/RFC2205, September 1997, <<http://www.rfc-editor.org/info/rfc2205>>.
- [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, <<http://www.rfc-editor.org/info/rfc3209>>.
- [RFC3473] Berger, L., Ed., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, DOI 10.17487/RFC3473, January 2003, <<http://www.rfc-editor.org/info/rfc3473>>.

- [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, <<http://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, <<http://www.rfc-editor.org/info/rfc4760>>.
- [RFC4872]   Lang, J., Ed., Rekhter, Y., Ed., and D. Papadimitriou, Ed., "RSVP-TE Extensions in Support of End-to-End Generalized Multi-Protocol Label Switching (GMPLS) Recovery", RFC 4872, DOI 10.17487/RFC4872, May 2007, <<http://www.rfc-editor.org/info/rfc4872>>.
- [RFC4873]   Berger, L., Bryskin, I., Papadimitriou, D., and A. Farrel, "GMPLS Segment Recovery", RFC 4873, DOI 10.17487/RFC4873, May 2007, <<http://www.rfc-editor.org/info/rfc4873>>.
- [RFC4874]   Lee, CY., Farrel, A., and S. De Cnodder, "Exclude Routes - Extension to Resource ReserVation Protocol-Traffic Engineering (RSVP-TE)", RFC 4874, DOI 10.17487/RFC4874, April 2007, <<http://www.rfc-editor.org/info/rfc4874>>.
- [RFC5420]   Farrel, A., Ed., Papadimitriou, D., Vasseur, JP., and A. Ayyangarps, "Encoding of Attributes for MPLS LSP Establishment Using Resource Reservation Protocol Traffic Engineering (RSVP-TE)", RFC 5420, DOI 10.17487/RFC5420, February 2009, <<http://www.rfc-editor.org/info/rfc5420>>.
- [RFC5440]   Vasseur, JP., Ed. and JL. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, DOI 10.17487/RFC5440, March 2009, <<http://www.rfc-editor.org/info/rfc5440>>.

## 7.2. Informative References

- [I-D.ietf-pce-stateful-pce]   Crabbe, E., Minei, I., Medved, J., and R. Varga, "PCEP Extensions for Stateful PCE", draft-ietf-pce-stateful-pce-13 (work in progress), December 2015.
- [RFC4655]   Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, DOI 10.17487/RFC4655, August 2006, <<http://www.rfc-editor.org/info/rfc4655>>.

[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, <<http://www.rfc-editor.org/info/rfc6952>>.

Authors' Addresses

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

Email: [jie.dong@huawei.com](mailto:jie.dong@huawei.com)

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

Email: [mach.chen@huawei.com](mailto:mach.chen@huawei.com)

Hannes Gredler  
Individual Contributor  
Austria

Email: [hannes@gredler.at](mailto:hannes@gredler.at)

Stefano Previdi  
Cisco Systems, Inc.  
Via Del Serafico, 200  
Rome 00142  
Italy

Email: [sprevidi@cisco.com](mailto:sprevidi@cisco.com)

Jeff Tantsura  
Ericsson  
300 Holger Way  
San Jose, CA 95134  
US

Email: [jeff.tantsura@ericsson.com](mailto:jeff.tantsura@ericsson.com)

Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: April 20, 2016

K. Patel  
Cisco Systems  
J. Uttaro  
ATT  
B. Decraene  
Orange  
W. Henderickx  
Alcatel Lucent  
October 18, 2015

Constrain Attribute announcement within BGP  
draft-keyupate-idr-bgp-attribute-announcement-00.txt

Abstract

[RFC4271] defines four different categories of BGP Path attributes. The different Path attribute categories can be identified by the attribute flag values. These flags help identify if an attribute is optional or well-known, Transitive or non-Transitive, Partial, or of an Extended length type. BGP attribute announcement depends on whether an attribute is a well-known or optional, and whether an attribute is a transitive or non-transitive. BGP implementations MUST recognize all well-known attributes. The well-known attributes are always Transitive. It is not required for BGP implementations to recognise all the Optional attributes. The Optional attributes could be Transitive or Non-Transitive. BGP implementations MUST store and forward any Unknown Optional Transitive attributes and ignore and drop any Unknown Optional Non-Transitive attributes.

Currently, there is no way to confine the scope of Path attributes within a given Autonomous System (AS) or a given BGP member-AS in Confederation. This draft defines two new attribute categories that help confine the scope of Optional attributes within a given AS or a given BGP member-AS in Confederation

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 <http://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 20, 2016.

#### Copyright Notice

Copyright (c) 2015 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.

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

#### Table of Contents

1. Introduction . . . . .	3
1.1. Requirements Language . . . . .	3
2. Path Attribute Flags . . . . .	4
3. Operation . . . . .	5
4. IANA Considerations . . . . .	6
5. Security Considerations . . . . .	6
5.1. Acknowledgements . . . . .	6
6. References . . . . .	6
6.1. Normative References . . . . .	7
6.2. Information References . . . . .	7
Authors' Addresses . . . . .	7



## 1. Introduction

[RFC4271] defines four different categories of BGP Path attributes. The different Path attribute categories can be identified by the attribute flag values. These flags help identify if an attribute is optional or well-known, Transitive or non-Transitive, Partial, or of an Extended length type. BGP attribute announcement depends on whether an attribute is a well-known or optional, and whether an attribute is a transitive or non-transitive. BGP implementations MUST recognize all well-known attributes. The well-known attributes are always Transitive. It is not required for BGP implementations to recognise all the Optional attributes. The Optional attributes could be Transitive or Non-Transitive. BGP implementations MUST store and forward any Unknown Optional Transitive attributes and ignore and drop any Unknown Optional Non-Transitive attributes.

Optional Transitive attributes help foster partial deployments of newer BGP features. Alternatively, Optional Non-Transitive attributes are drop by BGP speakers that do not recognise the attribute. The optional attributes in their current definition do not provide any automated attribute level filtering to control the scope of announcements within a given AS or a BGP member-AS in Confederation. Scoped announcements of attributes may be needed in certain scenarios. Announcing attributes beyond their intended scope MAY result in breakage of functionalities or leaking of any undesired information.

This draft defines new attribute categories that help confine the scope of Path attributes; in particular Optional attributes within a given Autonomous System or a given BGP member-AS in confederation or a given Administrative domain. Note that "BGP Member-AS in Confederation" and "Member-AS" are used entirely interchangeably throughout this document. The newly defined attribute scoping is specifically for newer attributes that explicitly state their use of such scoping bits. These newly defined attributes would be either an Optional transitive attributes (recognized and unrecognized) or any recognized optional non-transitive attributes. For any well-known attributes or unrecognized optional non-transitive attributes, the standard rules mentioned in [RFC4271] applies.

### 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. Path Attribute Flags

[RFC4271] defines four type of BGP Path attributes using the attribute Flags field. This draft introduces three more flags fields as follows:

Path Attribute flags:

```

    0 1 2 3 4 5 6 7
+---+---+---+---+---+---+
|O|T|P|E|A|C| R |   (R = MUST Be Zero)
+---+---+---+---+---+---+

```

- O    Optional or a Well-known as defined in [RFC4271]
- T    Transitive or Non-Transitive as defined in [RFC4271]
- P    Partial as defined in [RFC4271]
- E    Extended Length type as defined in [RFC4271]
- A    AS Wide Scope
- C    Member-AS in Confederation Scope
- M    Multi-AS Scope

The fifth most significant bit ("A") is defined as the AS Wide Scope bit, which is used to indicate that an optional attribute cannot be announced outside a given AS boundary. When set, the given optional attribute MUST be filtered by the sending BGP speaker at an AS boundary. If the "A" bit is set then the "O" bit MUST be set. Otherwise a BGP speaker MUST consider an attribute as an error and malformed.

The Sixth most significant bit ("C") is defined as the Member-AS Scope bit, which is used to indicate that an optional attribute cannot be announced outside a given Member-AS boundary. When set, the given optional attribute MUST be filtered by the sending BGP speaker at a Member-AS boundary. If the "C" bit is set then the "O" bit MUST be set. Otherwise a BGP speaker MUST consider an attribute as an error and malformed. "C" bit SHOULD only be set when an Autonomous System is configured as a BGP Confederation. A BGP speaker MUST not transmit an attribute with "C" bit set to peers that are not members of the local confederation. Otherwise a BGP speaker MUST consider an attribute as an error and malformed.

Both the fifth and the sixth most significant bit together is defined as the Multiple AS Scope within a Single Administration. When both the fifth and the sixth bits are set, optional attribute can be traversed across multiple AS and filtered by the sending BGP speaker at the Administration boundary.

The handling of malformed attributes SHOULD follow the procedures mentioned in [RFC7606]. For any malformed attribute that is handled by the "attribute discard" instead of the "treat-as-withdraw" approach, it is critical to consider the potential impact. In particular, if the attribute has an impact on the route selection or installation process, then the presumption is that "attribute discard" is unsafe and "treat-as-withdraw" procedure SHOULD be considered. Otherwise, "attribute discard" procedure SHOULD be used.

### 3. Operation

When originating an optional Path attribute, a BGP speaker SHOULD use and set AS Wide Scope bit if it wants to restrict the announcement within a AS. Similarly, when originating an optional Path attribute, a BGP speaker SHOULD use and set Member-AS Scope bit if it wants to restrict the announcement with a Member-AS. When originating an optional Path attribute, a BGP speaker SHOULD use and set both Member-AS Scope bit and AS Wide Scope bit if it wants to restrict the announcement within a single administration composed of multiple ASes.

When a BGP speaker receives or originates a route that includes an optional Path attribute with a AS Wide Scope bit set and a Member-AS Scope bit cleared, it MUST remove that Path attribute when announcing the route to any of its EBGp speakers. To deal with partial deployments it is suggested that a BGP speaker SHOULD quietly ignore and not pass along to other BGP peers any Path attribute received from its EBGp peers with a AS Wide Scope bit set and a Member-AS Scope bit cleared unless configured explicitly using a policy.

When a BGP speaker receives or originates a route that includes an optional Path attribute with a Member-AS Scope bit set and a AS Wide Scope bit cleared, it MUST remove that Path attribute when announcing the route to any of its BGP speakers outside its Member-AS. To deal with partial deployments it is suggested that a BGP speaker SHOULD quietly ignore and not pass along to other BGP peers any Path attribute received from its BGP peers with a Member-AS Scope bit set and a AS Wide Scope bit cleared unless configured explicitly as a policy.

When a BGP speaker receives or originates a route with an optional path attribute that has both, the AS Wide Scope bit set and the

Member-AS Scope bit set, it MUST announce it to all its EBGP peers within its administrative domain. Such an attribute MUST be filtered when the attribute is announced outside its administrative domain. The BGP peering boundaries for an administrative domain is a matter of a policy and is set by the operators.

Any implementation that supports the extensions defined in this draft MUST support the Enhanced Error handling defined in [RFC7606]. Enhanced Error handling allows any error condition that MAY occur during the parsing and processing of new attribute flags to be treated according to the procedures of [RFC7606]. Furthermore, it is assumed that the BGP network is enabled with Enhanced Error Handling feature. This allows BGP speakers not implementing the draft extensions to apply the procedures of [RFC7606].

#### 4. IANA Considerations

This draft defines two new Path attribute flags. We request IANA to create a new registry for BGP Path Attribute Flags under BGP Path attributes as follows:

Under "Border Gateway Protocol (BGP) Parameters" registry, "BGP Path Attributes Flags" Reference: draft-keyupate-idr-bgp-attribute-flags-00 Registration Procedures as follows:

Bit Value (MSB)	Type	Reference
1	Optional/Mandatory	RFC4271
2	Transitive/NonTransitive	RFC4271
3	Partial	RFC4271
4	Extended Length Type	RFC4271
5	AS Wide Scope	Current Draft
6	Member-AS in Confederation	Current Draft

#### 5. Security Considerations

This extension to BGP does not change the underlying security issues inherent in the existing [RFC4724] and [RFC4271].

##### 5.1. Acknowledgements

The authors would like to thank John Scudder, Jakob Heitz, Shyam Seturam, Juan Alcaide and Acee Lindem for the review and comments.

#### 6. References

## 6.1. 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, <<http://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, <<http://www.rfc-editor.org/info/rfc4271>>.
- [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, <<http://www.rfc-editor.org/info/rfc7606>>.

## 6.2. Information References

- [RFC3392] Chandra, R. and J. Scudder, "Capabilities Advertisement with BGP-4", RFC 3392, DOI 10.17487/RFC3392, November 2002, <<http://www.rfc-editor.org/info/rfc3392>>.
- [RFC4486] Chen, E. and V. Gillet, "Subcodes for BGP Cease Notification Message", RFC 4486, DOI 10.17487/RFC4486, April 2006, <<http://www.rfc-editor.org/info/rfc4486>>.
- [RFC4724] Sangli, S., Chen, E., Fernando, R., Scudder, J., and Y. Rekhter, "Graceful Restart Mechanism for BGP", RFC 4724, DOI 10.17487/RFC4724, January 2007, <<http://www.rfc-editor.org/info/rfc4724>>.

## Authors' Addresses

Keyur Patel  
Cisco Systems  
170 W. Tasman Drive  
San Jose, CA 95134  
USA

Email: [keyupate@cisco.com](mailto:keyupate@cisco.com)

James Uttaro  
ATT  
200 S. Laurel Ave  
Middletown, NJ 07748  
USA

Email: [uttaro@att.com](mailto:uttaro@att.com)

Bruno Decraene  
Orange

Email: [bruno.decraene@orange.com](mailto:bruno.decraene@orange.com)

Wim Henderickx  
Alcatel Lucent  
Copernicuslaan 50  
Antwerp 2018  
Belgium

Email: [wim.henderickx@alcatel-lucent.com](mailto:wim.henderickx@alcatel-lucent.com)

Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: January 9, 2017

K. Patel  
A. Lindem  
Cisco Systems  
L. Jalil  
Verizon  
July 8, 2016

Selective Advertisement of Multiple Paths within BGP  
draft-keyupate-idr-bgp-selective-add-paths-01.txt

Abstract

[draft-ietf-idr-add-paths] defines a BGP extension that allows the advertisement of multiple paths for the same address prefix without the new paths implicitly replacing any previous ones. The essence of the extension is that each path is identified by a path identifier in addition to the address prefix. This draft augments functionality defined in [draft-ietf-idr-add-paths] to facilitate advertisement of multiple paths for a subset of prefixes in a given address family. Prefixes are selected through specification of a well-known BGP extended community.

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 <http://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, 2017.

Copyright Notice

Copyright (c) 2016 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.

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

#### Table of Contents

1. Introduction . . . . .	2
1.1. Requirements Language . . . . .	3
2. Selective Add-Path Capability . . . . .	3
3. Selective Add-Path Community . . . . .	4
4. Selective Add-Path Use Case . . . . .	5
5. IANA Considerations . . . . .	5
6. Security Considerations . . . . .	5
6.1. Acknowledgements . . . . .	5
7. References . . . . .	5
7.1. Normative References . . . . .	6
7.2. Information References . . . . .	6
Authors' Addresses . . . . .	6

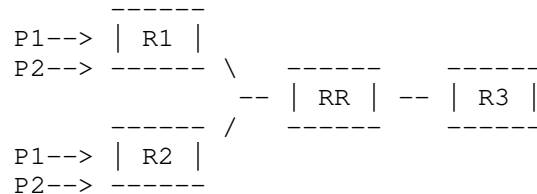
#### 1. Introduction

[I-D.ietf-idr-add-paths] defines a BGP extension that allows the advertisement of multiple paths for the same address prefix without the new paths implicitly replacing any previous ones. The essence of the extension is that each path is identified by a path identifier in addition to the address prefix. This document augments functionality defined in defined in [I-D.ietf-idr-add-paths] to facilitate advertisement of multiple paths for a subset of prefixes in a given address family. Prefixes are selected through specification of a reserved BGP extended community.

This draft defines a capability to limit the scope of BGP multiple path advertisement to a subset prefixes in a given address family.



Prefixes are selected through specification of a reserved BGP extended community [RFC4360].



As an example, suppose that RR is a route reflector that doesn't change nexthops of the prefixes it reflects, with clients R1, R2 and R3. Suppose R1 sends RR an UPDATE: <NLRI=P1, NH=R1> and <NLRI=P2, NH=R1>. Suppose R2 sends RR an UPDATE: <NLRI=P1, NH=R2> and <NLRI=P2, NH=R2>. R1, R2, and R3 would like selective ADDPATHs for Prefix P1 and not for Prefix P2. R1, R2, and R3 exchange selective the ADDPATH capability with RR. R1, R2, R3 are configured with the reserved selective ADDPATHs community that they attach to prefixes that need selective ADDPATHs. RR now has two paths to P1 and P2. RR announces P2 with bestpath to all its clients while RR announces P1 with additional paths. The number of additional paths with its best path and its additional paths is a matter of local policy configured on RR.

### 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. Selective Add-Path Capability

The ADD-PATH Capability is a new BGP capability [RFC5492]. The Capability Code for this capability is allocated by IANA as specified in the Section 5. The Capability Length field of this capability is variable. The Capability Value field consists of one or more of the following tuples:

```
+-----+
| Address Family Identifier (2 octets) |
+-----+
| Subsequent Address Family Identifier (1 octet) |
+-----+
```

The meaning and use of the fields are as follows:

Address Family Identifier (AFI):

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

Subsequent Address Family Identifier (SAFI):

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

A BGP Speaker that wishes to announce or receive multiple paths **MUST** exchange the add-path capability defined in [I-D.ietf-idr-add-paths]. A BGP Speaker that wishes to announce or receive multiple paths for selected prefixes **MUST** exchange the selective add-path capability defined in this draft. A BGP speaker wanting to advertise selective add-path capability **MUST** also advertise the add-path capability defined in [I-D.ietf-idr-add-paths].

In processing a received selective add-path capability from a peer, a BGP speaker **MUST** ensure that it also received the add-path capability defined in [I-D.ietf-idr-add-paths]. Otherwise, the BGP speaker should ignore the received selective add-path capability and follow the error handling rules for unsupported add-path capabilities in [RFC5492].

### 3. Selective Add-Path Community

Upon successful Selective Add-Path capability negotiation, a BGP speaker **MUST NOT** announce multiple paths for any AFI/SAFI prefix unless it has received at least one UPDATE for that prefix that includes the Selective Add-Path well-known community in its attributes. The community is a Transitive Opaque Extended Community with the sub-type value IANA-TBD.

If Selective Add-Path capability negotiation for a given AFI/SAFI has not taken place and the Selective Add-Path Community is included with a prefix advertised for the same AFI/SAFI, the Selective Add-Path Community will be ignored. However, the occurrence of the unexpected community **SHOULD** be logged.

#### 4. Selective Add-Path Use Case

A use case is a BGP deployment where underlay and overlay routes are associated with the same AFI/SAFI and, due to scaling, only multiple paths are only advertised and installed for underlay routes. For direct BGP sessions, the ingress routers would only advertise multiple paths for the underlay routes. However, if the topology includes BGP Router Reflectors [RFC4456], it is likely that multiple ingress routers will advertise the same overlay routes. In this case, the mechanism describe herein would be useful in limiting multi-path best-path computation and advertisement to the underlay routes.

As a second usecase, many times a service provider will carry both customer traffic and internal services (e.g., VOIP) on the same backbone network using routes in the same BGP address families. In this situation, the number of customer routes and paths greatly exceed the number of routes and paths for internal services. However, the service provider desires the faster failover and convergence provided by BGP Add-Paths [I-D.ietf-idr-add-paths]. In this scenario, the Selective Add-Path functionality described herein can be leveraged for routes corresponding to internal services without the overhead incurred if multiple paths were advertised for the customer routes.

#### 5. IANA Considerations

This document defines a new capability for BGP. We request IANA to assign BGP capability number from BGP Capabilities Registry.

This document also defines a new extended community for BGP. We request IANA to assign a BGP well-known extended community from the Transitive Opaque Extended Community Sub-Types Registry.

#### 6. Security Considerations

This extension to BGP does not change the underlying security issues inherent in the existing [RFC4724] and [RFC4271].

##### 6.1. Acknowledgements

The authors would like to thank .... for the review and comments.

#### 7. References

## 7.1. Normative References

- [I-D.ietf-idr-add-paths]  
Walton, D., Retana, A., Chen, E., and J. Scudder,  
"Advertisement of Multiple Paths in BGP", draft-ietf-idr-  
add-paths-13 (work in progress), December 2015.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate  
Requirement Levels", BCP 14, RFC 2119,  
DOI 10.17487/RFC2119, March 1997,  
<<http://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,  
<<http://www.rfc-editor.org/info/rfc4271>>.
- [RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended  
Communities Attribute", RFC 4360, DOI 10.17487/RFC4360,  
February 2006, <<http://www.rfc-editor.org/info/rfc4360>>.
- [RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement  
with BGP-4", RFC 5492, DOI 10.17487/RFC5492, February  
2009, <<http://www.rfc-editor.org/info/rfc5492>>.

## 7.2. Information References

- [RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route  
Reflection: An Alternative to Full Mesh Internal BGP  
(IBGP)", RFC 4456, DOI 10.17487/RFC4456, April 2006,  
<<http://www.rfc-editor.org/info/rfc4456>>.
- [RFC4724] Sangli, S., Chen, E., Fernando, R., Scudder, J., and Y.  
Rekhter, "Graceful Restart Mechanism for BGP", RFC 4724,  
DOI 10.17487/RFC4724, January 2007,  
<<http://www.rfc-editor.org/info/rfc4724>>.

## Authors' Addresses

Keyur Patel  
Cisco Systems  
170 W. Tasman Drive  
San Jose, CA 95134  
USA

Email: [keyupate@cisco.com](mailto:keyupate@cisco.com)

Acee Lindem  
Cisco Systems  
170 W. Tasman Drive  
San Jose, CA 95134  
USA

Email: [acee@cisco.com](mailto:acee@cisco.com)

Luay Jalil  
Verizon  
400 International Parkway  
Richardson, Tx 75081  
USA

Email: [luay.jalil@verizon.com](mailto:luay.jalil@verizon.com)

Inter-Domain Routing  
Internet-Draft  
Intended status: Standards Track  
Expires: September 13, 2016

P. Lapukhov  
Facebook  
March 12, 2016

Use of BGP for dissemination of ILA mapping information  
draft-lapukhov-bgp-ila-afi-00

Abstract

Identifier-Locator Addressing [I-D.herbert-nvo3-ila] relies on splitting the 128-bit IPv6 address into identifier and locator parts to implement identifier mobility, and network virtualization. This document proposes a method for distributing the identifier to locator mapping information using Multiprotocol Extensions for BGP-4 [RFC4760].

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 <http://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 13, 2016.

Copyright Notice

Copyright (c) 2016 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 . . . . .	2
2. BGP ILA AFI . . . . .	2
3. Capability Advertisement . . . . .	3
4. Disseminating Identifier-Locator mapping information . . . . .	3
4.1. Advertising ILA mapping information . . . . .	3
4.2. Withdrawing ILA mapping information . . . . .	3
5. Interpreting the mapping information . . . . .	4
5.1. Unicast SAFI . . . . .	4
5.2. Multicast SAFI . . . . .	4
6. IANA Considerations . . . . .	5
7. Manageability Considerations . . . . .	5
8. Security Considerations . . . . .	5
9. Acknowledgements . . . . .	5
10. References . . . . .	5
10.1. Normative References . . . . .	5
10.2. Informative References . . . . .	5
Author's Address . . . . .	6

## 1. Introduction

Under the ILA proposal, the IPv6 address is split in 64-bit identifier (lower address bits) and locator (higher address bits) portions. The locator part is determined from a mapping table that maintains associations between the location-independent identifiers and topologically significant locators. The hosts that collectively implement and maintain such mappings are referred to as "ILA domain" in this document. This document proposes a new address family identifier (AFI) for the purpose of disseminating the locator-identifier mappings among the nodes participating in the ILA domain.

## 2. BGP ILA AFI

This document introduces a new AFI known as a "Identifier-Locator AFI" with the actual value to be assigned by IANA. The purpose of this AFI is disseminating the mapping information between identifiers and locators in ILA domain. This document defines the use of SAFI values of "1" (unicast) and "2" (multicast) only.

### 3. Capability Advertisement

A BGP speaker that wishes to exchange ILA mapping information MUST use the Multiprotocol Extensions Capability Code, as defined in [RFC4760], to advertise the corresponding AFI/SAFI pair.

### 4. Disseminating Identifier-Locator mapping information

#### 4.1. Advertising ILA mapping information

For the purpose of ILA mapping encoding, the 8-octet locator field SHALL be encoded in the "next-hop address" field. The "length of the next-hop address" MUST be set to "8" (64-bit). The identifiers bound to the locator SHALL be encoded within the NLRI portion of MP\_REACH\_NLRI attribute. The NLRI portion of MP\_REACH\_NLRI starts with the two-octet "Length of identifiers" field. The rest of the NLRI is a collection of 8-octet (64-bit) identifiers that are bound to the locator specified in the "next-hop address" field.

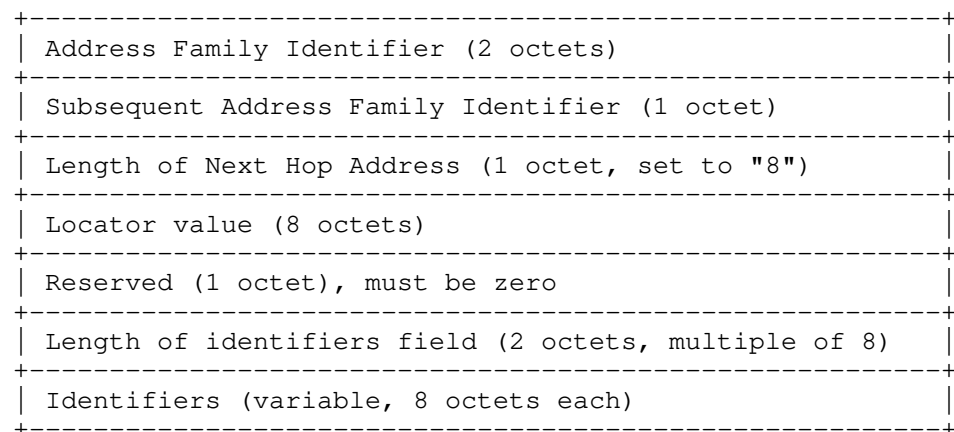


Figure 1: MP\_REACH\_NLRI Layout

#### 4.2. Withdrawing ILA mapping information

Withdrawal of ILA mapping information is performed via an MP\_UNREACH\_NLRI attribute advertisement organized as following:



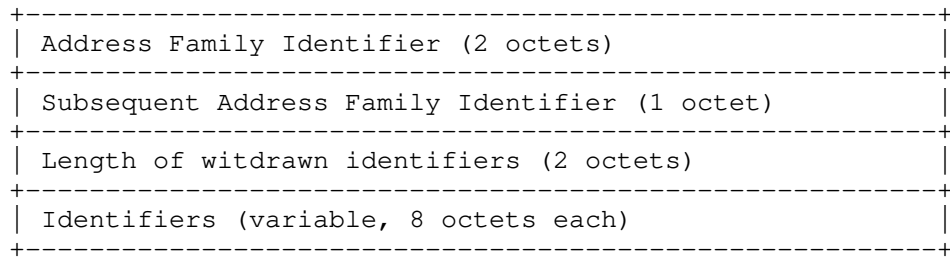


Figure 2: MP\_UNREACH\_NLRI Layout

## 5. Interpreting the mapping information

### 5.1. Unicast SAFI

Only the locator part of ILA address is used for packet routing, and every node that hosts an identifier MUST have a unique routable /64 prefix within the scope of the ILA domain. The identifiers advertised under the ILA AFI are expected to be used by the data-plane implementation to perform match on a full IPv6 address and decide whether the locator portion of the address needs a re-write. It is up to the implementation to decide which full 128-bit IPv6 addresses need a rewrite, e.g. by matching on a Standard Identifier Representation (SIR) prefix as defined in [I-D.herbert-nvo3-ila].

The locator rewrite information comes from the next-hop "address" associated with the identifier. The next-hop field of MP\_REACH\_NLRI attribute MUST NOT be used for any routing resolutions/lookups by the BGP process itself. It should be used purely to create a rewrite rule in the data-plane forwarding table. The actual forwarding decision is then based on subsequent lookup in the forwarding table to find the next hop to send the packet to.

### 5.2. Multicast SAFI

For multicast packets, the RPF check process SHALL be modified for use with ILA source addresses. Specifically, source ILA IPv6 addresses with the identifier portion matching the mapping table SHALL be mapped to proper locator, prior to performing the RPF check. The ILA source addresses need to be identified by some means specific to ILA implementation, e.g. by matching on configured SIR prefixes. The ILA addresses that do not match any mapping entry SHALL be considered as failing the RPF check.

## 6. IANA Considerations

For the purpose of this work, IANA would be asked to allocate values for the new AFI.

## 7. Manageability Considerations

TBD

## 8. Security Considerations

This document does not introduce any changes in terms of BGP security. Defining ILA security model is outside of scope of this document.

## 9. Acknowledgements

The author would like to thank Doug Porter for the initial idea suggestion and discussion of this proposal.

## 10. References

### 10.1. 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, <<http://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, <<http://www.rfc-editor.org/info/rfc4271>>.
- [I-D.herbert-nvo3-ila] Herbert, T., "Identifier-locator addressing for network virtualization", draft-herbert-nvo3-ila-01 (work in progress), October 2015.

### 10.2. Informative References

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

Author's Address

Petr Lapukhov  
Facebook  
1 Hacker Way  
Menlo Park, CA 94025  
US

Email: petr@fb.com

Inter-Domain Routing  
Internet-Draft  
Intended status: Standards Track  
Expires: August 6, 2016

P. Lapukhov  
Facebook  
E. Aries, Ed.  
P. Marques  
Juniper Networks  
E. Nkposong  
Salesforce.com Inc  
February 3, 2016

Use of BGP for Opaque Signaling  
draft-lapukhov-bgp-opaque-signaling-01

Abstract

Border Gateway Protocol with multi-protocol extensions (MP-BGP) enables the use of the protocol for dissemination of virtually any information. This document proposes a new Address Family/Subsequent Address Family along with new optional transitive attribute to be used for distribution of opaque data. This functionality is intended to be used by applications other than BGP for exchange of their own data on top of BGP mesh. The structure of such data MAY to be interpreted by the regular BGP speakers, rather the goal is to use BGP purely as a convenient and scalable communication system.

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 <http://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 August 6, 2016.

## Copyright Notice

Copyright (c) 2016 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 . . . . .	2
2. BGP Opaque Data AFI . . . . .	3
3. BGP Key-Value SAFI . . . . .	3
4. Capability Advertisement . . . . .	3
5. Disseminating Key-Value bindings . . . . .	3
5.1. Publishing a Key-Value binding . . . . .	4
5.2. Removing a Key-Value binding . . . . .	5
5.3. Propagating multiple values for the same key . . . . .	6
6. Message filtering . . . . .	6
6.1. Automated filtering . . . . .	6
6.2. Filtering via policy . . . . .	6
7. IANA Considerations . . . . .	7
8. Manageability Considerations . . . . .	7
9. Security Considerations . . . . .	7
10. Acknowledgements . . . . .	7
11. References . . . . .	7
11.1. Normative References . . . . .	7
11.2. Informative References . . . . .	7
Authors' Addresses . . . . .	8

## 1. Introduction

Implementation of Multiprotocol Extensions for BGP-4 [RFC4760] gives the ability to pass arbitrary data in BGP protocol messages. This capability has been leveraged by many for dissemination of non-routing related information over BGP (e.g. "Dissemination of Flow Specification Rules" [RFC5575] as well as "North-Bound Distribution of Link-State and TE Information using BGP" [I-D.ietf-idr-ls-distribution]). However, there has been no channel defined explicitly to disseminate data with arbitrary payload. The intended use case is for applications other than BGP to leverage the

protocol machinery for distribution (broadcasting) of their own state in the network domain. Publishers and consumers will use BGP UPDATE messages to submit and receive opaque data. It is up to the BGP implementation to provide a custom API for message producers or consumers if needed.

One application of this extension could be auto-discovery of various services in the data-center network that uses BGP as the routing protocol of choice ([I-D.ietf-rtgwg-bgp-routing-large-dc]).

Another application is building and testing new routing protocols or BGP extensions within existing BGP implementation. The new protocol/extension may influence routing either by directly communicating to the RIB/FIB of the router it runs on, or by overriding BGP paths via BGP route injection. An example of such BGP extension could be [WISER]

## 2. BGP Opaque Data AFI

This document introduces a new AFI known as a "BGP Opaque Data AFI" with the actual value to be assigned by IANA. The purpose of this AFI is to exchange opaque information within a BGP network.

## 3. BGP Key-Value SAFI

This document introduces a new SAFI known as "BGP Key-Value SAFI" with the actual value to be assigned by IANA. The purpose of this SAFI is exchange of opaque information structured as a Key-Value binding.

## 4. Capability Advertisement

A BGP speaker that wishes to exchange Opaque Data MUST use the Multiprotocol Extensions Capability Code, as defined in [RFC4760], to advertise the corresponding AFI/SAFI pair.

## 5. Disseminating Key-Value bindings

This document proposes to implement a distributed, eventually consistent Key-Value store on top of existing BGP protocol mechanics. The "Key" portion is to be encoded as the NLRI part of MP\_REACH\_NLRI attribute and "Value" encoded using a new optional transitive attribute.

- o Publishers, acting as BGP speakers, advertise keys along with associated values into the routing domain. The BGP network synchronizes that state by propagating the encoded data following regular BGP protocol operations.

- o Consumers, acting as BGP speakers, receive the information via BGP protocol UPDATE messages. Only publishers and consumers of the opaque data are supposed to interpret its contents - the rest of the BGP network acts merely as a dissemination system.

Multiple publishers can advertise the same key (NLRI) bound to different values. It is also possible for the advertised binding to have the same Key-Value pairs but differ in some other BGP attributes. In that case, BGP would follow the best-path selection logic to prevent duplicate information in the network. A consumer will receive the value created by the publisher "closest" in terms of BGP best-path selection logic, based on the policies that exist in the routing domain. This document does not propose any method of achieving global consensus for all published values for a given key.

#### 5.1. Publishing a Key-Value binding

The encoding scheme proposed below follows the semantics of a Key-Value bindings. The "Key" is stored in the NLRI section of the MP\_REACH\_NLRI attribute, as shown on Figure 1.

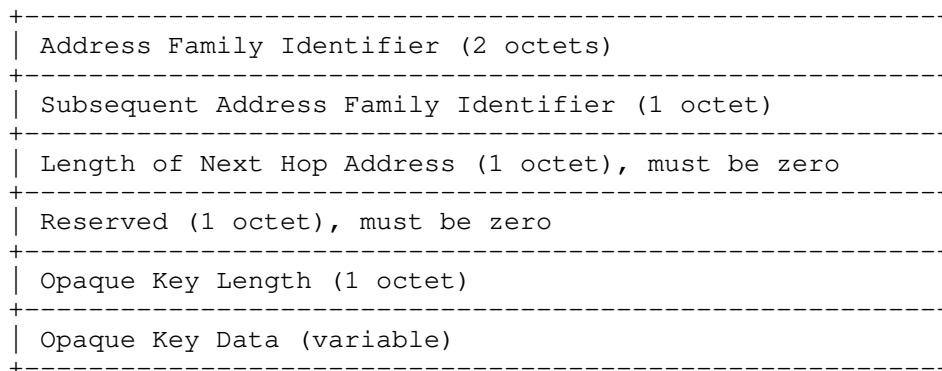


Figure 1: MP\_REACH\_NLRI Layout

- o The AFI/SAFI values are to be allocated by IANA.
- o Length of Next Hop Address: must be zero, since no information is encoded in the next-hop address field.
- o Opaque Key Length: identifies the size of the Key field. If field is set to zero, the implementation MUST ignore the advertisement.
- o Opaque Key Data: the byte string representing the opaque key contents. This portion SHOULD NOT be interpreted by BGP implementation.

The "Value" portion of a published binding is to be encoded in a new optional transitive attribute as shown on Figure 2:

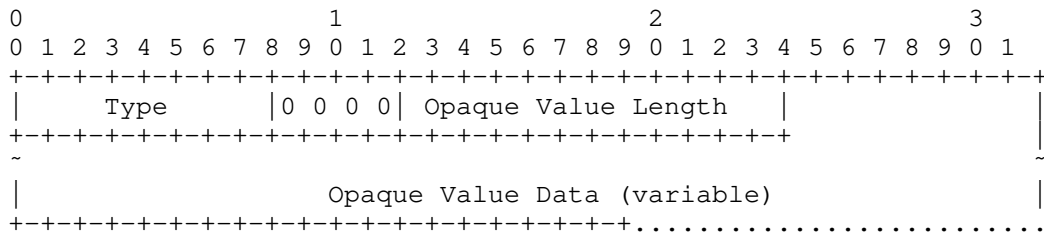


Figure 2: OPAQUE\_VALUE attribute layout

- o Type: Identifies the new OPAQUE\_VALUE attribute, with the value to be allocated by IANA.
- o Opaque Value Length: Two octets encoding the total length of the attribute in octets, including the Type and Length fields. The length is encoded as an unsigned binary integer. The four most significant bits of this field MUST be set to zero, due to the limit imposed by maximum BGP message size. Note that the minimum length is 3, indicating that no Opaque Value Data field is present. Such binding, in presence of non-zero length key is still valid, as it informs the consumers that the key "exists".
- o Opaque Value Data: A field containing zero or more octets. This portion SHOULD NOT be interpreted by BGP implementations.

Even when the OPAQUE\_VALUE optional transitive attribute is not present in BGP advertisement, the BGP implementation MUST still retain Opaque Key (NLRI) in its LocRIB and propagate it further as usual. This case is to be interpreted as an announcement of the key existence.

## 5.2. Removing a Key-Value binding

The removal procedure follows the regular MP-BGP route withdrawal, using the MP\_UNREACH\_NLRI attribute. This section defines the attribute structure for the new AFI/SAFI.

The message shown on Figure 3 instructs the receiving BGP speaker to delete the N bindings corresponding to Key 1, Key 2 ... Key N if the keys have been previously learned from the withdrawing speaker. If any of the Keys is not found in the LocRIB or has not been previously received from the withdrawing BGP peer, such key removal request MUST be ignored.



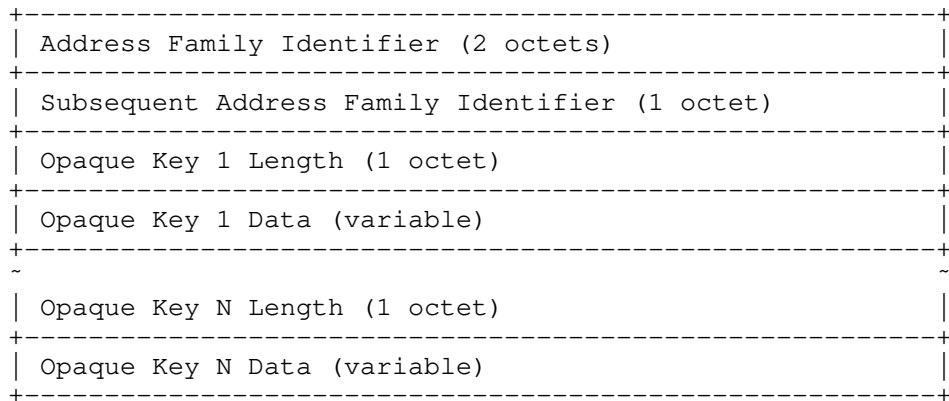


Figure 3: MP\_UNREACH\_NLRI attribute layout

### 5.3. Propagating multiple values for the same key

It is possible to propagate multiple values associated with the same key using the Add-Path extension defined in [I-D.ietf-idr-add-paths]. However, this document recommends that instead unique key values be used for this purpose. It is up to the consumers and publishers of the opaque data to settle on single unique value using some kind of consensus protocol.

## 6. Message filtering

Limiting the scope of opaque information flooding is an important operational concern. BGP already has the mechanisms needed to control this process, and these mechanisms are briefly reviewed below.

### 6.1. Automated filtering

One can leverage mechanics presented in [RFC4684] and use the router-target extended community attribute to identify "channels" where key-value bindings are published. The consumers would signal their interest in particular "channel" by advertising the corresponding router-target membership. The publications then need to contain the router-target extended community attribute to constrain information propagation.

### 6.2. Filtering via policy

Ad-doc message filtering could be implemented using BGP standard (see [RFC4271]) or extended community attributes (see [RFC4360]). The semantic of these attributes is to determined by the policy and

publishers/consumers. Filtering could be done locally on receiving speaker, or on remote speaker, by using outbound route filtering feature defined in [RFC5291].

## 7. IANA Considerations

For the purpose of this work, IANA would be asked to allocate values for the new AFI and SAFI, as well as a value for the new optional transitive attribute.

## 8. Manageability Considerations

TBD

## 9. Security Considerations

This document does not introduce any changes in terms of BGP security.

## 10. Acknowledgements

TBD

## 11. References

### 11.1. 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, <<http://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, <<http://www.rfc-editor.org/info/rfc4271>>.

### 11.2. Informative References

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

- [RFC4684] Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk, R., Patel, K., and J. Guichard, "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, DOI 10.17487/RFC4684, November 2006, <<http://www.rfc-editor.org/info/rfc4684>>.
- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <<http://www.rfc-editor.org/info/rfc4760>>.
- [RFC5291] Chen, E. and Y. Rekhter, "Outbound Route Filtering Capability for BGP-4", RFC 5291, DOI 10.17487/RFC5291, August 2008, <<http://www.rfc-editor.org/info/rfc5291>>.
- [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, <<http://www.rfc-editor.org/info/rfc5575>>.
- [I-D.ietf-idr-add-paths]  
Walton, D., Retana, A., Chen, E., and J. Scudder,  
"Advertisement of Multiple Paths in BGP", draft-ietf-idr-add-paths-13 (work in progress), December 2015.
- [I-D.ietf-idr-ls-distribution]  
Gredler, H., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and TE Information using BGP", draft-ietf-idr-ls-distribution-13 (work in progress), October 2015.
- [I-D.ietf-rtgwg-bgp-routing-large-dc]  
Lapukhov, P., Premji, A., and J. Mitchell, "Use of BGP for routing in large-scale data centers", draft-ietf-rtgwg-bgp-routing-large-dc-07 (work in progress), August 2015.
- [WISER] Mahajan, R., Wetherall, D., and T. Anderson, "Mutually Controlled Routing with Independent ISPs", 2007, <<http://research.microsoft.com/en-us/um/people/ratul/papers/nsdi2007-wiser.pdf>>.

Authors' Addresses

Petr Lapukhov  
Facebook  
1 Hacker Way  
Menlo Park, CA 94025  
US

Email: petr@fb.com

Ebben Aries (editor)  
Juniper Networks  
1133 Innovation Way  
Sunnyvale, CA 94089  
US

Email: exa@juniper.net

Pedro Marques  
Juniper Networks  
1194 N. Mathilda Ave  
Sunnyvale, CA 94089  
US

Email: roque@juniper.net

Edet Nkposong  
Salesforce.com Inc  
The Landmark @ One Market, ST 300  
San Francisco, CA 94105  
US

Email: enkposong@salesforce.com

IDR  
Internet-Draft  
Intended status: Standards Track  
Expires: September 16, 2016

Z. Li  
China Mobile  
J. Dong  
Huawei Technologies  
March 15, 2016

Carry congestion status in BGP extended community  
draft-li-idr-congestion-status-extended-community-00

Abstract

A new extended community is introduced in this document to carry the link congestion status, especially for the exit link of one AS. We call this extended community congestion status community, which can be used by the BGP routers to steer the Internet-access traffic among the exit links by deploying policy routing.

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 [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 <http://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 16, 2016.

Copyright Notice

Copyright (c) 2016 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 . . . . .	2
2. Congestion Status Extended Community . . . . .	4
2.1. Congestion Status Extended Community for Two-Octet AS . .	4
2.2. Congestion Status Extended Community for Four-Octet AS .	5
3. Security Considerations . . . . .	6
4. IANA Considerations . . . . .	6
5. Normative References . . . . .	6
Authors' Addresses . . . . .	6

## 1. Introduction

typically the architecture of a large scale ISP's network is multi-layered, as illustrated in Figure 1. The national backbone network has its own AS, and each of the province or state network has a specific AS. Backbone network connects all the province or state networks together and has several exit links to access the Internet. In some circumstances, the province or state network may have direct exit links to the Internet. The total bandwidth of the backbone exit links is usually much bigger than that of the direct exit links in the province or state networks. Thus, the Internet-access traffic is mainly transported through the backbone exit links by deploying route policies on the ASBR routers in the province or state networks. The ASBR routers in the province or state networks, for example, prefer the routes learned from the backbone by setting higher local preference for those routes. However, when the backbone exit links are congested due to traffic increasing or delay of the capacity expansion, the ASBR routers in the province or state networks do not know this, and still delivery Internet-access traffic to the backbone. The customer experience deteriorates, the operator, in turn, will receive more and more complaints for its bad network performance. Then, the operator has to steer some Internet-access traffic to the direct exit links in the province or state networks by deploying route policy on the ASBR routers. This kind of policy should be removed when the capacity expansion of the backbone exit links is done. The ASBR routers do not know this again.

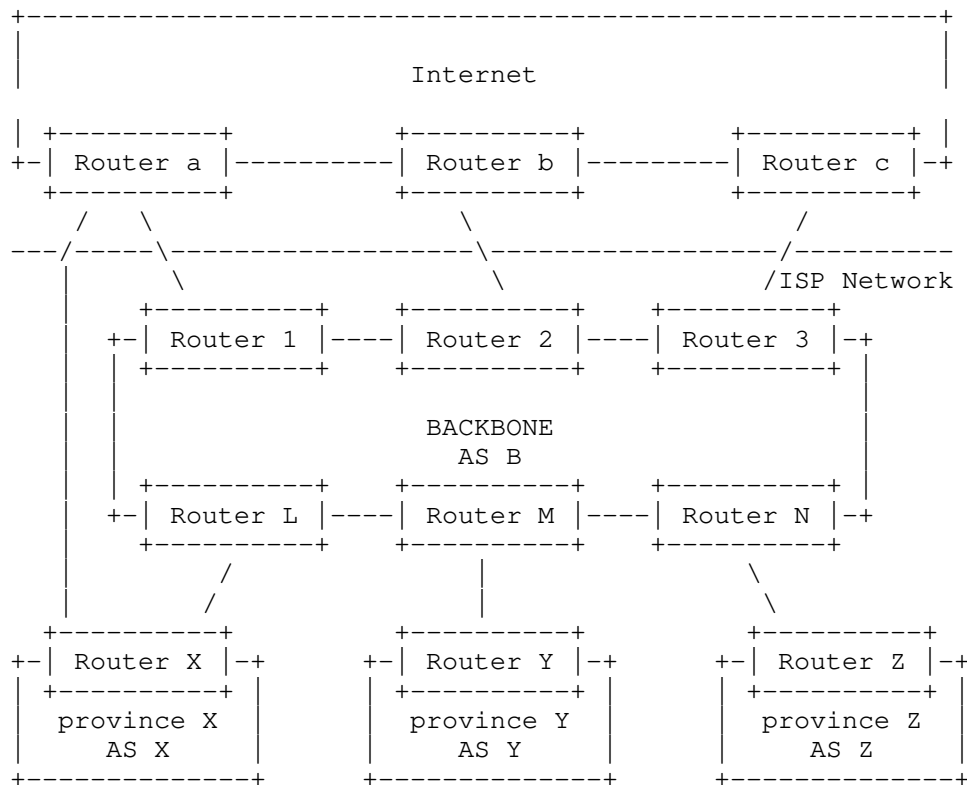


Figure 1

This document introduces a new extended community [RFC4360] to delivery the congestion status of the exit link to other BGP peers. The BGP receiver can then use this community to deploy route policy, thus steer Internet-access traffic according to the congestion status of the exit link. Router X in the above figure, for example, can steer some Internet-access traffic to the direct exit link when it knows the backbone exit link is congested. The introduced community is called congestion status extended community.

Congestion status extended community is good not only to the ASBRs in other AS, but also to the BGP peers within one AS. For instance, Router M in backbone AS chooses Router 2 to transport the Internet-access traffic by default. When Router M receives congestion status extended communities from Router 1,2,3, which indicate the utilization of the exit link of Router 1,2,3 is 90%, 70%, and 50% respectively, it can choose Router 3 to transport some Internet-access traffic using route policy.

## 2. Congestion Status Extended Community

As described in [RFC4360], the extended community attribute is an 8-octet value with the first one or two octets to indicate the type of this attribute. Since congestion status extended community needs to be delivered from on AS to other ASes, and used by the BGP speakers both in other ASes and within the same AS as the sender, it MUST be a transitive extended community, i.e. the T bit in the first octet MUST be zero.

Congestion status extended community has two encoding formats, one is for two-octet AS, the other is for four-octet AS.

### 2.1. Congestion Status Extended Community for Two-Octet AS

Congestion status extended community for two-octet AS is a sub-type allocated from Transitive Two-Octet AS-Specific Extended Community Sub-Types defined in section 5.2.2 of [RFC7153]. Its format is as Figure 2.

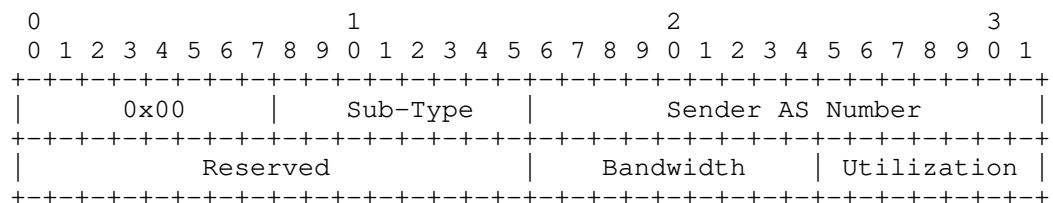


Figure 2

The "Type" field MUST be 0x00, which indicate this is a Transitive Two-Octet AS-Specific Extended Community.

The "Sub-Type" field is used to indicate this is a Congestion Status Extended Community. Its value is to be assigned by IANA. 0x06 is suggested.

The "Sender AS Number" field is 2 octets. Its value is the AS number of the BGP speaker who generates this congestion status extended community. The generator MUST have 2-octet AS number.

The "Reserved" field is 2 octets. This field is used to align with the Congestion Status Extended Community for Four-Octet AS defined in the next section of this document. Its value SHOULD be zero. The BGP peers who receive this community MUST ignore this field.



The "Bandwidth" field is 1 octet. Its value is the bandwidth of the exit link in unit of gbps (gigabits per second).

The "Utilization" field is 1 octet. Its value is the utilization of the exit link in unit of percent. We can use the "Utilization" field together with the "Bandwidth" field to calculate the traffic load that we can further steer to this exit link.

## 2.2. Congestion Status Extended Community for Four-Octet AS

Congestion status extended community for four-octet AS is a sub-type allocated from Transitive Four-Octet AS-Specific Extended Community Sub-Types defined in section 5.2.4 of [RFC7153]. Its format is as Figure 3.

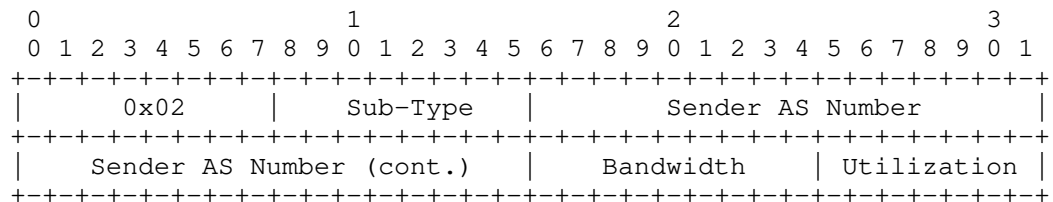


Figure 3

The "Type" field MUST be 0x02, which indicate this is a Transitive Four-Octet AS-Specific Extended Community.

The "Sub-Type" field is used to indicate this is a Congestion Status Extended Community. Its value is to be assigned by IANA. 0x06 is suggested.

The "Sender AS Number" field is 4 octets. Its value is the AS number of the BGP speaker who generates this congestion status extended community. The generator MUST have 4-octet AS number.

The "Bandwidth" field is 1 octet. Its value is the bandwidth of the exit link in unit of gbps (gigabits per second).

The "Utilization" field is 1 octet. Its value is the utilization of the exit link in unit of percent. We can use the "Utilization" field together with the "Bandwidth" field to calculate the traffic load that we can further steer to this exit link.

### 3. Security Considerations

Malicious router may use the congestion status extended community to interfere the traffic steering decision of the BGP receiver. BGP peers SHOULD use MD5 for authentication [RFC4360]. BGP receiver SHOULD only accept the congestion status community or extended community delivered from BGP peers with MD5 authentication.

### 4. IANA Considerations

One sub-type is solicited to be assigned from Transitive Two-Octet AS-Specific Extended Community Sub-Types registry to indicate the extended community with Type 0x00 is a Congestion Status Extended Community for Two-Octet AS. 0x06 is suggested.

One sub-type is solicited to be assigned from Transitive Four-Octet AS-Specific Extended Community Sub-Types registry to indicate the extended community with Type 0x02 is a Congestion Status Extended Community for Four-Octet AS. 0x06 is suggested.

### 5. 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, <<http://www.rfc-editor.org/info/rfc2119>>.
- [RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, DOI 10.17487/RFC4360, February 2006, <<http://www.rfc-editor.org/info/rfc4360>>.
- [RFC7153] Rosen, E. and Y. Rekhter, "IANA Registries for BGP Extended Communities", RFC 7153, DOI 10.17487/RFC7153, March 2014, <<http://www.rfc-editor.org/info/rfc7153>>.

### Authors' Addresses

Zhenqiang Li  
China Mobile  
No.32 Xuanwumenxi Ave., Xicheng District  
Beijing 100032  
P.R. China

Email: [li\\_zhenqiang@hotmail.com](mailto:li_zhenqiang@hotmail.com)

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

Email: jie.dong@huawei.com

IDR Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: September 22, 2016

Q. Liang  
S. Hares  
J. You  
Huawei  
R. Raszuk  
Bloomberg LP  
D. Ma  
Cisco  
March 21, 2016

BGP Flow Specification MPLS action  
draft-liang-idr-flowspec-mpls-action-00

Abstract

This document specifies a BGP Flow specification policy action to push/pop/swap MPLS labels.

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 <http://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 22, 2016.

Copyright Notice

Copyright (c) 2016 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 . . . . .	2
1.1. Background . . . . .	2
1.2. MPLS Flow Specification Deployment . . . . .	3
2. Terminology . . . . .	3
2.1. Requirements Language . . . . .	3
3. Overview of Proposal . . . . .	3
4. Protocol Extensions . . . . .	4
5. Deployment Examples . . . . .	6
5.1. Example 1 - MPLS Filter + MPLS Action . . . . .	6
5.2. Example 2 - IP filter + MPLS action . . . . .	7
6. Security Considerations . . . . .	9
7. IANA Considerations . . . . .	9
8. Acknowledgement . . . . .	10
9. References . . . . .	10
9.1. Normative References . . . . .	10
9.2. Informative References . . . . .	11
Authors' Addresses . . . . .	11

## 1. Introduction

This section provides the background for proposing a new action for BGP Flow specification [RFC5575] that push/pops MPLS labels or swaps MPLS labels. For those familiar with BGP Flow specification ([RFC5575], [RFC7674] [I-D.ietf-idr-flow-spec-v6], [I-D.ietf-idr-flowspec-l2vpn], and MPLS ([RFC3107]) can skip this background section.

### 1.1. Background

[RFC5575] defines the flow specification (FlowSpec) that is an n-tuple consisting of several matching criteria that can be applied to IP traffic. The matching criteria can include elements such as source and destination address prefixes, IP protocol, and transport protocol port numbers. A given IP packet is said to match the defined flow if it matches all the specified criteria. [RFC5575] also defines a set of filtering actions, such as rate limit, redirect, marking, associated with each flow specification. A new Border Gateway Protocol ([RFC4271]) Network Layer Reachability Information (BGP NLRI) (AFI/SAFI: 1/133 for IPv4, AFI/SAFI: 1/134 for VPNv4) encoding format is used to distribute traffic flow specifications.

[RFC3107] specifies the way in which the label mapping information for a particular route is piggybacked in the same Border Gateway Protocol Update message that is used to distribute the route itself. Label mapping information is carried as part of the Network Layer Reachability Information (NLRI) in the Multiprotocol Extensions attributes. The Network Layer Reachability Information is encoded as one or more triples of the form <length, label, prefix>. The NLRI contains a label is indicated by using Subsequent Address Family Identifier (SAFI) value 4.

[RFC4364] describes a method in which each route within a Virtual Private Network (VPN) is assigned a Multiprotocol Label Switching (MPLS) label. If the Address Family Identifier (AFI) field is set to 1, and the SAFI field is set to 128, the NLRI is an MPLS-labeled VPN-IPv4 address.

## 1.2. MPLS Flow Specification Deployment

In BGP VPN/MPLS networks when flow specification policy rules exist on multiple forwarding devices in the network bound with labels from one or more LSPs, only the ingress LSR (Label Switching Router) needs to identify a particular traffic flow based on the matching criteria for flow. Once the flow is match by the ingress LSR, the ingress LSR steers the packet to a corresponding LSP (Label Switched Path). Other LSRs of the LSP just need to forward the packet according to the label carried in it.

## 2. Terminology

Flow Specification (FlowSpec): A flow specification is an n-tuple consisting of several matching criteria that can be applied to IP traffic, including filters and actions. Each FlowSpec consists of a set of filters and a set of actions.

### 2.1. 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 [RFC2119].

## 3. Overview of Proposal

This document proposes adding a BGP-FS action in an extended community alters the label switch path associated with a matched flow. If the match does not have a label switch path, this action is skipped.

The BGP flow specification (BGP-FS) policy rule could match on the destination prefix and then utilize a BGP-FS action to adjust the label path associated with it (push/pop/swap tags.) Or a BGP-FS policy rule could match on any set of BGP-FS match conditions associated with a BGP-FS action that adjust the label switch path (push/pop/swap).

draft-ietf-yong-flowspec-mpls-match provides a match BGP-FS that may be used with this action to match and direct MPLS packets.

#### 4. Protocol Extensions

A new label-action is defined as BGP extended community value based on Section 7 of [RFC5575].

type	extended community	encoding
TBD1	label-action	MPLS tag

Figure 1

Label-action is described below:

0										1										2										3									
0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1								
Type (TBD1)										OpCode										Reserve					order														
Label										Exp										S					TTL					Label Stack Entry									

The use and the meaning of these fields are as follows:

Type: the same as defined in [RFC4360]

Figure 2

OpCode: Operation code

OpCode	Function
0	Push the MPLS tag
1	Pop the outermost MPLS tag in the packet
2	Swap the MPLS tag with the outermost MPLS tag in the packet
3~15	Reserved

\* where:

- \* When the Opcode field is set to 0, the label stack entry Should be pushed on the MPLS label stack.
- \* When the Opcode field is set to 1, the label stack entry is invalid, and the router SHOULD pop the existing outermost MPLS tag in the packet.
- \* When the Opcode field is set to 2, the router SHOULD swap the label stack entry with the existing outermost MPLS tag in the packet. If the packet has no MPLS tag, it just pushes the label stack entry.
- \* Note-1: The Opcode 0 or 1 may be used in some SDN networks, such as the scenario described in [I-D.filsfils-spring-segment-routing-central-epe].
- \* The Opcode 2 can be used in traditional BGP MPLS/VPN networks.

Reserved: all zeros

Order: within multiple label actions A FlowSpec rule MAY be associated with one or more ordering label-action each in an extended community. If multiple label-actions occur, this field gives the order of this action within that group. If two MPLs actions arrive with the same order the last mpls action received for an order will be used.

Label: the same as defined in [RFC3032]

Bottom of Stack (S): the same as defined in [RFC3032]. It SHOULD be invalid, and set to zero by default. It MAY be modified by the forwarding router locally.



Time to Live (TTL): the same as defined in [RFC3032]. It MAY be modified by the forwarding router locally.

Experimental Use (Exp): the same as defined in [RFC3032]. It MAY be modified by the forwarding router according to the local routing policy.

## 5. Deployment Examples

### 5.1. Exampel 1 - MPLS Filter + MPLS Action

Forwarding information for the traffic  
for source: IP2, Destination: IP1

Purpose of BGP-FS filters: send DDoS traffic to IDS/IPS server

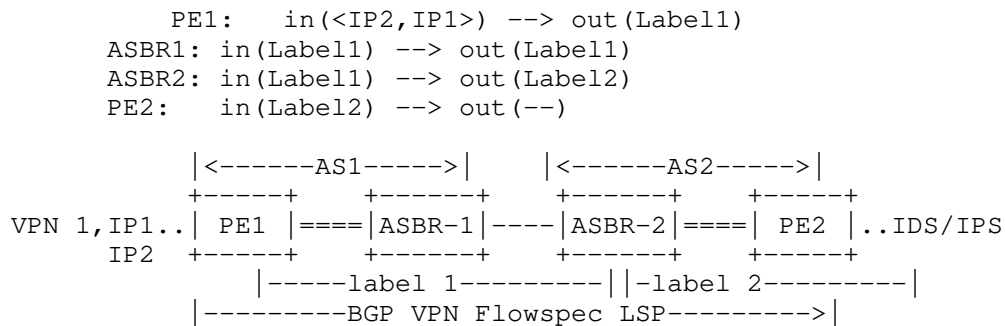


Figure 1 - Forwarding Diagram

locally configured filters

Filters:

destination ip prefix:IP2/32

source ip prefix:IP1/32

Action:

put on LSP with Label 1

PE-2 Installs:

BGP-FS Filter:

MPLS filter for Label 1 and label 2

BGP-FS Actions:

Traffic-Rate limit

MPLS POP

PE-2 Sends to ASBR-2

BGP-FS Filter

MPLS filter for label 1 and Label 2

BGP-FS Actions:

Traffic-Rate limit

Label SWAP 1 to 2

PE-1 Sends to ASBR 1

BGP-FS filter

MPLS filter for label 1

BGP-FS Actions

Traffic-Rate limit

## 5.2. Example 2 - IP filter + MPLS action

Forwarding information for the traffic from IP1 to IP2 in the Routers:

```
PE1:   in(<IP2,IP1>) --> out(Label2)
ASBR1: in(Label2) --> out(Label3)
ASBR2: in(Label3) --> out(Label4)
PE2:   in(Label4) --> out(--)
```

Labels allocated by Flow policy process

```
Label4 allocated by PE2
Label3 allocated by ASBR2
Label2 allocated by ASBR1
```

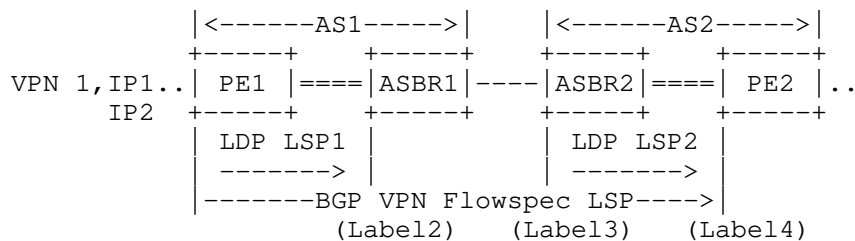


Figure 1 - Forwarding Diagram

BGP-FS rule1 (locally configured)

Filters:

```
destination ip prefix:IP2/32
source      ip prefix:IP1/32
```

Actions:

```
traffic-marking: 1
MPLS POP
```

Note:

```
The following Extended Communities are added/deleted
[rule-1a] BGP-FS action MPLS POP [used on PE2]
[rule-1b] BGP-FS action SWAP 4   [used on ASBR-2]
[rule-1c] BGP-FS action SWAP 3   [used on ASBR-1]
[rule-1d] BGP-FS action push 2   [used on PE1]
```

BGP Filter rules

PE-2 Changes BGP-FS rule-1a to rule-1b prior to sending  
 Clears Extended Community: BGP-FS action MPLS POP  
 Adds Extended Community: BGP-FS action MPLS SWAP 4

ASBR-2 receives BGP-FS rule-1b (NLRI + 2 Extended Community)  
 Installs the BGP-FS rule-1b (MPLS SWAP 4, traffic-marking)  
 Changes BGP-FS rule-1b to rule-1c prior to sending to ASBR1  
 Clear Extended Community: BGP-FS action MPLS SWAP 4  
 Adds Extended Community: BGP-FS action MPLS SWAP 3

ASBR-1 Receives BGP-FS rule-1c (NLRI + 2 Extended Community)  
 Installs the BGP-FS rule-1c (MPLS SWAP 3, traffic-marking)  
 Changes BGP-FS rule-1c to rule-1d prior to sending to PE-2  
 Clear Extended Community: BGP-FS action MPLS SWAP 3  
 Adds Extended Community: BGP-FS action MPLS SWAP 2

PE-1 Receives BGP-FS rule-1d (NLRI + 2 Extended Communities)  
 Installs BGP-FS rule-1d action [MPLS SWAP 2, traffic-marking]

## 6. Security Considerations

The validation of BGP Flow Specification policy in NLRI is considered in [I-D.hares-idr-flowspec-combo] for option 1, and for option 2. Additional security has been proposed in [I-D.ietf-idr-bgp-flowspec-oid]. A BGP5575bis document will consider the revised security.

For Option 1, the MPLS Match can be one of the match filters, and the final match is an "AND" of all the filters. Match filters are tested in the order specified in [I-D.hares-idr-flowspec-combo] and/or an RFC5575bis document.

[I-D.hares-idr-flowspec-combo] suggests a default order for filters and for the BGP-FS action proposed after [RFC5575], and this document discusses how conflicts between action are handled.

## 7. IANA Considerations

This section complies with [RFC7153]

IANA is requested to a new entry in "Flow Spec action types registry" with the following values:

Value Name:	Value	Reference
=====	=====	=====
Lable Action	TBD	[this document]

## 8. Acknowledgement

The authors would like to thank Shunwan Zhuang, Zhenbin Li, Peng Zhou and Jeff Haas for their comments.

## 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, <<http://www.rfc-editor.org/info/rfc2119>>.
- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, DOI 10.17487/RFC3031, January 2001, <<http://www.rfc-editor.org/info/rfc3031>>.
- [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, <<http://www.rfc-editor.org/info/rfc3032>>.
- [RFC3107] Rekhter, Y. and E. Rosen, "Carrying Label Information in BGP-4", RFC 3107, DOI 10.17487/RFC3107, May 2001, <<http://www.rfc-editor.org/info/rfc3107>>.
- [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, <<http://www.rfc-editor.org/info/rfc4271>>.
- [RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, DOI 10.17487/RFC4360, February 2006, <<http://www.rfc-editor.org/info/rfc4360>>.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <<http://www.rfc-editor.org/info/rfc4364>>.
- [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, <<http://www.rfc-editor.org/info/rfc5575>>.

- [RFC7153] Rosen, E. and Y. Rekhter, "IANA Registries for BGP Extended Communities", RFC 7153, DOI 10.17487/RFC7153, March 2014, <<http://www.rfc-editor.org/info/rfc7153>>.
- [RFC7674] Haas, J., Ed., "Clarification of the Flowspec Redirect Extended Community", RFC 7674, DOI 10.17487/RFC7674, October 2015, <<http://www.rfc-editor.org/info/rfc7674>>.

## 9.2. Informative References

- [I-D.filsfils-spring-segment-routing-central-epe]  
Filsfils, C., Previdi, S., Patel, K., Shaw, S., Ginsburg, D., and D. Afanasiev, "Segment Routing Centralized Egress Peer Engineering", draft-filsfils-spring-segment-routing-central-epe-05 (work in progress), August 2015.
- [I-D.hares-idr-flowspec-combo]  
Hares, S., "An Information Model for Basic Network Policy and Filter Rules", draft-hares-idr-flowspec-combo-01 (work in progress), March 2016.
- [I-D.ietf-idr-bgp-flowspec-oid]  
Uttaro, J., Filsfils, C., Smith, D., Alcaide, J., and P. Mohapatra, "Revised Validation Procedure for BGP Flow Specifications", draft-ietf-idr-bgp-flowspec-oid-02 (work in progress), January 2014.
- [I-D.ietf-idr-flow-spec-v6]  
McPherson, D., Raszuk, R., Pithawala, B., Andy, A., and S. Hares, "Dissemination of Flow Specification Rules for IPv6", draft-ietf-idr-flow-spec-v6-07 (work in progress), March 2016.
- [I-D.ietf-idr-flowspec-l2vpn]  
Weiguo, H., Litkowski, S., and S. Zhuang, "Dissemination of Flow Specification Rules for L2 VPN", draft-ietf-idr-flowspec-l2vpn-03 (work in progress), November 2015.
- [I-D.yong-idr-flowspec-mpls-match]  
Yong, L., Hares, S., Liang, Q., and J. You, "BGP Flow Specification Filter for MPLS Label", March 2016.

Authors' Addresses

Qiandeng Liang  
Huawei  
101 Software Avenue, Yuhuatai District  
Nanjing 210012  
China

Email: liangqiandeng@huawei.com

Susan Hares  
Huawei  
7453 Hickory Hill  
Saline, MI 48176  
USA

Email: shares@ndzh.com

Jianjie You  
Huawei  
101 Software Avenue, Yuhuatai District  
Nanjing 210012  
China

Email: youjianjie@huawei.com

Robert Raszuk  
Bloomberg LP  
731 Lexington Ave  
New York City, NY 10022  
USA

Email: robert@raszuk.net

Dan Ma  
Cisco

Email: danma@cisco.com

Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: September 19, 2016

S. Previdi, Ed.  
C. Filsfils  
A. Sreekantiah  
S. Sivabalan  
Cisco Systems, Inc.  
P. Mattes  
Microsoft  
March 18, 2016

Advertising Segment Routing Traffic Engineering Policies in BGP  
draft-previdi-idr-segment-routing-te-policy-00

Abstract

This document defines a new BGP SAFI with a new NLRI in order to advertise a Segment Routing Traffic Engineering Policy (SR TE Policy). The SR TE Policy is advertised along with the Tunnel Encapsulation Attribute for which this document also defines new sub-TLVs. An SR TE policy is advertised with the information that will be used by the node receiving the advertisement in order to instantiate the policy in its forwarding table and to steer traffic according to the policy.

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 <http://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 19, 2016.

Copyright Notice

Copyright (c) 2016 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 . . . . .	2
1.1. Requirements Language . . . . .	4
2. SR TE Policy Encoding . . . . .	4
2.1. SR TE Policy SAFI and NLRI . . . . .	4
2.1.1. SR TE Policies and Add-Paths . . . . .	5
2.2. SR TE Policy and Tunnel Encapsulation Attribute . . . . .	5
2.3. Remote Endpoint and Color . . . . .	6
2.4. SR TE Policy Sub-TLVs . . . . .	7
2.4.1. SR TE Binding SID Sub-TLV . . . . .	7
2.4.2. Weight Sub-TLV . . . . .	8
2.4.3. Segment List Sub-TLV . . . . .	9
2.4.4. Segment Sub-TLV . . . . .	9
3. SR TE Policy Operations . . . . .	11
3.1. Multipath Operation . . . . .	12
3.2. Binding SID TLV . . . . .	12
3.3. Reception of an SR TE Policy . . . . .	13
3.4. Announcing BGP SR TE Policies . . . . .	14
3.5. Flowspec and SR TE Policies . . . . .	14
4. Acknowledgments . . . . .	15
5. IANA Considerations . . . . .	15
6. Security Considerations . . . . .	15
7. References . . . . .	15
7.1. Normative References . . . . .	15
7.2. Informational References . . . . .	17
Authors' Addresses . . . . .	17

## 1. Introduction

Segment Routing (SR) technology leverages the source routing and tunneling paradigms. [I-D.ietf-spring-segment-routing] describes the SR architecture. [I-D.ietf-spring-segment-routing-mpls] describes its instantiation on the MPLS data plane and [I-D.ietf-6man-segment-routing-header] describes the Segment Routing instantiation over the IPv6 data plane.

This document defines the Segment Routing Traffic Engineering Policy (SR TE Policy) as a set of weighted equal cost multi path (WECMP)

segment lists (representing explicit paths) as well as the mechanism allowing a router to steer traffic into an SR TE Policy.

The SR TE Policy is advertised in the Border Gateway Protocol (BGP) by the BGP speaker being a router or a controller and using extensions defined in this document. Among the information encoded in the BGP message and representing the SR TE Policy, the steering mechanism makes also use of the Extended Color Community currently defined in [I-D.ietf-idr-tunnel-encaps]

Typically, a controller defines the set of policies and advertise them to BGP routers (typically ingress routers). The policy advertisement uses BGP extensions defined in this document. The policy advertisement is, in most but not all of the cases, tailored for the receiver. In other words, a policy advertised to a given BGP speaker has significance only for that particular router and is not intended to be propagated anywhere else. Then, the receiver of the policy instantiate the policy in its routing and forwarding tables and steer traffic into it based on both the policy and destination prefix color and next-hop.

Alternatively, a router (i.e.: an BGP egress router) advertises SR TE Policies representing paths to itself. These advertisements are sent to BGP ingress nodes who instantiate these policies and steer traffic into them according to the color and endpoint/BGP next-hop of both the policy and the destination prefix.

An SR TE Policy being intended only for the receiver of the advertisement, the SR TE Policies are sent directly to each receiver and, in most of the cases will not traverse any Route Reflector (RR, [RFC4456]).

However, in the case where the same SR TE Policy is intended for a group of nodes, nothing prevents the originator to rely on one or more RRs in order to distribute the SR TE Policy to multiple receivers. The encoding of the SR TE Policy defined in this document supports both propagation schemes: direct BGP session and Route Reflectors.

The BGP extensions for the advertisement of SR TE Policies include following components:

- o A new Subsequent Address Family Identifier (SAFI) identifying the content of the BGP message (i.e.: the SR TE Policy).
- o A new NLRI identifying the SR TE Policy.

- o A set of new TLVs to be inserted into the Tunnel Encapsulation Attribute (as defined in [I-D.ietf-idr-tunnel-encaps]) and describing the SR TE Policy.
- o The Extended Color Community (as defined in [I-D.ietf-idr-tunnel-encaps]) and used in order to steer traffic into an SR TE Policy.

### 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. SR TE Policy Encoding

### 2.1. SR TE Policy SAFI and NLRI

A new SAFI is defined: the SR TE Policy SAFI (codepoint suggested value 73, to be assigned by IANA).

The SR TE Policy SAFI uses a new NLRI defined as follows:

```

+-----+
|           Policy Color (4 octets)           |
+-----+
|           Endpoint (4 or 16 octets)         |
+-----+
```

where:

- o Policy Color: 4-octet value identifying (with the endpoint) the policy. The color is used to match the color of the destination prefixes in order to steer traffic into the SR TE 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 or a summary address). The Endpoint may be an IPv4 (4-octet) address or an IPv6 (16-octet) address according to the AFI of the NLRI.

The NLRI containing the SR TE Policy is carried in a BGP UPDATE message [RFC4271] using BGP multiprotocol extensions [RFC4760] with an AFI of 1 or 2 (IPv4 or IPv6) and with a SAFI of 73 (suggested value, to be assigned by IANA).

An update message that carries the MP\_REACH\_NLRI or MP\_UNREACH\_NLRI attribute with the SR TE Policy SAFI MUST also carry the BGP

mandatory attributes: NEXT\_HOP, ORIGIN, AS\_PATH, and LOCAL\_PREF (for IBGP neighbors), as defined in [RFC4271]. In addition, the BGP update message MAY also contain any of the BGP optional attributes.

The NEXT\_HOP attribute of the SR TE Policy SAFI NLRI is set based on the AFI. For example, if the AFI is set to IPv4 (1), then the nexthop is encoded as a 4-byte IPv4 address. If the AFI is set to IPv6 (2), then the nexthop is encoded as a 16-byte IPv6 address of the router. It is important to note that any BGP speaker receiving a BGP message with an SR TE Policy NLRI, will process it only if the NLRI is a best path as per the BGP best path selection algorithm.

The NEXT\_HOP attribute of the SR TE Policy SAFI NLRI MUST be set as one of the local addresses of the BGP speaker originating and advertising the SR TE Policy (either the controller or the BGP egress node).

#### 2.1.1. SR TE Policies and Add-Paths

The SR TE Policy SAFI NLRI MAY use the Add Paths extension ([I-D.ietf-idr-add-paths]) when the same policy (identified by the same Color and Endpoint) is to be advertised by multiple originators (e.g.: multiple controllers) and all advertisements need to be advertised to a group of receivers (hence these advertisements need to be preserved from a RR selection process).

In such case, each controller will use a different path identifier in the advertisement of the SR TE Policy.

When Add-Paths extensions is to be used, it MUST be signaled in the BGP capability according to ([I-D.ietf-idr-add-paths]).

#### 2.2. SR TE Policy and Tunnel Encapsulation Attribute

The content of the SR TE Policy is encoded in the Tunnel Encapsulation Attribute originally defined in [I-D.ietf-idr-tunnel-encaps] using a new Tunnel-Type TLV (suggested codepoint is 14, to be assigned by IANA).

The SR TE Policy Encoding structure is as follows:

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

Attributes:

  Tunnel Encaps Attribute (23)

    Tunnel Type: SR TE Policy

      Binding SID

      Segment List

        Weight

        Segment (sid/nai/flags)

        Segment (sid/nai/flags)

        ...

    ...

where:

- o SR TE Policy SAFI NLRI is defined in Section 2.1.
- o Tunnel Encapsulation Attribute is defined in [I-D.ietf-idr-tunnel-encaps].
- o Tunnel-Type is set to a suggested value of 14 (to be assigned by IANA).
- o Binding SID, Weight, Segment and Segment-List are new sub-TLVs defined in this document.
- o Additional sub-TLVs may be defined in the future.

A single occurrence of "Tunnel Type: SR TE Policy" MUST be encoded within the same Tunnel Encapsulation Attribute.

Multiple occurrences of "Segment List" MAY be encoded within the same SR TE Policy.

Multiple occurrences of "Segment" MAY be encoded within the same Segment List.

### 2.3. Remote Endpoint and Color

The Remote Endpoint and Color sub-TLVs, as defined in [I-D.ietf-idr-tunnel-encaps], MAY also be present in the SR TE Policy encodings.

If present, the Remote Endpoint sub-TLV MUST match the Endpoint of the SR TE Policy SAFI NLRI. If they don't match, the SR TE Policy advertisement MUST be considered as invalid.

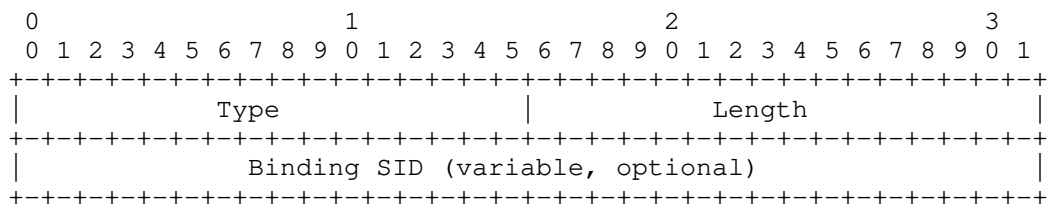
If present, the Color sub-TLV MUST match the Policy Color of the SR TE Policy SAFI NLRI. If they don't match, the SR TE Policy advertisement MUST be considered as invalid.

## 2.4. SR TE Policy Sub-TLVs

This section defines the SR TE Policy sub-TLVs.

### 2.4.1. SR TE Binding SID Sub-TLV

The Binding SID sub-TLV requests the allocation of a Binding Segment identifier associated with the SR TE Policy. The Binding SID sub-TLV has the following format:



where:

- o Type: to be assigned by IANA (suggested value is 6).
- o Length: specifies the length of the value field not including Type and Length fields. Can be 0 or 4 or 16.
- o Binding SID: if length is 0, then no field is present. If length is 4 then the Binding SID contains a 4-octet SID. If length is 16 then the Binding SID contains a 16-octet IPv6 SID.

The Binding SID sub-TLV is used to instruct the receiver of the BGP message to allocate a Binding SID to the SR TE Policy. The allocation of the Binding SID in the receiver is done according to following rules:

- o If length is 0 (no value field is present), then the receiver MUST allocate a local Binding SID whose value is chosen by the receiver.
- o If length is 4, then the value field contains the 4-octet Binding SID value the receiver SHOULD allocate.
- o If length is 16, then the value field contains the 16-octet Binding SID value the receiver SHOULD allocate.

The Binding SID sub-TLV is mandatory and MUST NOT appear more than once on an SR TE Policy Advertisement.

When a controller is used in order to define and advertise SR TE Policies and when the Binding SID is allocated by the receiver, such Binding SID SHOULD be reported to the controller. The mechanisms and/or APIs used for the reporting of the Binding SID are outside the scope of this document.

Further use of the Binding SID is described in a subsequent section.

#### 2.4.2. Weight Sub-TLV

The Weight sub-TLV specifies the weight associated to a given path (i.e.: a given segment list). The weight is used in order to apply weighted-ECMP mechanism when steering traffic into a policy that includes multiple paths (i.e.: multiple segment lists).

The Weight sub-TLV has the following format:

0								1								2								3							
0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1
Type																Length															
Weight																															

where:

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

Length: 4.

The Weight sub-TLV is optional and MAY appear only once in the Segment List sub-TLV.

When present, the Weight sub-TLV specifies a weight to be associated with the corresponding Segment List, for use in unequal-cost multi path. Weights are applied by summing the total value of all of the weights for all Segment Lists, and then assigning a fraction of the forwarded traffic to each Segment List in proportion its weight's fraction of the total.

### 2.4.3. Segment List Sub-TLV

The Segment List sub-TLV is used in order to encode a single explicit path towards the endpoint. The Segment List sub-TLV includes the elements of the paths (i.e.: segments) as well as an optional Weight TLV.

The Segment List sub-TLV has the following format:

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|          Type          |          Length          |
+-----+-----+-----+-----+-----+-----+-----+-----+
//                          sub-TLVs                          //
+-----+-----+-----+-----+-----+-----+-----+-----+

```

where:

- o Type: to be assigned by IANA (suggested value is 8).
- o Length: the total length (not including the Type and Length fields) of the sub-TLVs encoded within the Segment List sub-TLV.
- o sub-TLVs:
  - \* An optional single Weight sub-TLV.
  - \* One or more Segment sub-TLVs.

The Segment List sub-TLV is mandatory.

Multiple occurrences of the Segment List sub-TLV MAY appear in the SR TE Policy.

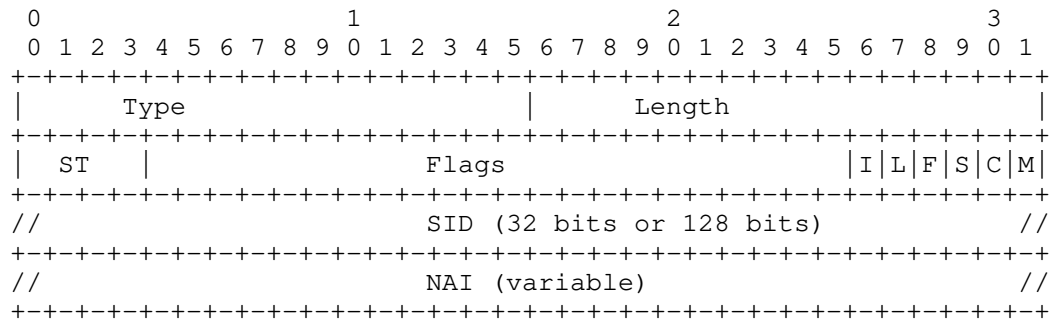
When multiple occurrences of the Segment List sub-TLV appear in the SR TE Policy, the traffic is load-balanced across them either through an ECMP scheme (if no Weight sub-TLV is present) or through a W-ECMP scheme according to Section 2.4.2.

### 2.4.4. Segment Sub-TLV

The Segment sub-TLV describes a single segment in a segment list (i.e.: a single element of the explicit path). Multiple Segment sub-TLVs constitute an explicit path of the SR TE Policy.

The encoding format of the Segment sub-TLV is based on the ERO sub-object definition described in [I-D.ietf-pce-segment-routing]):





where:

- o Type: to be assigned by IANA (suggested value is 9).
- o Length: the length of the Segment sub-TLV not including the Type and Length fields.

SID Type (ST) indicates the type of the information associated with the SID and NAI contained in the sub-TLV. ST is defined in [I-D.ietf-pce-segment-routing].

SID is the Segment Identifier as defined in [I-D.ietf-pce-segment-routing].

NAI (Node and Adjacency Identifier) contains the NAI associated with the SID. Depending on the value of ST, the NAI can have different formats as described in [I-D.ietf-pce-segment-routing].

Flags carry any additional information related to the SID. Currently, the following flags are defined:

I-Flag: IPv6 SID flag. When set, it indicates that the SID is encoded as a 16-octet IPv6 SID (IPv6 SIDs are defined in [I-D.ietf-6man-segment-routing-header]). When clear, the SID is encoded as a 4-octet SID.

L-Flag: Loose flag. Indicates whether the encoding represents a loose-hop in the LSP ([RFC3209]). If L-Flag is clear, a BGP speaker MUST NOT overwrite the SID value present in the Segment sub-TLV. Otherwise, a BGP speaker, based on local policy, MAY expand or replace the SID value in the received Segment sub-TLV.

F-flag: when set, the NAI value in the object body is null.

S-Flag: when set, the SID value in the object body is null. In this case, the receiving BGP speaker is responsible for choosing

the SID value, e.g., by looking up its Tunnel DB using the NAI which, in this case, MUST be present in the object.

C-Flag: when this flag as well as the M-flag are set, then the SID value represents an MPLS label stack entry as specified in [RFC5462], where all the entry's fields (Label, TC, S, and TTL) are specified by the sending BGP speaker. However, a receiving BGP speaker MAY choose to override TC, S, and TTL values according to its local policy and MPLS forwarding rules.

M-Flag: when this bit is set, the SID value represents an MPLS label stack entry as specified in [RFC5462] where only the label value is specified by the BGP speaker. Other label fields (i.e: TC, S, and TTL) fields MUST be ignored, and receiving BGP speaker MUST set these fields according to its local policy and MPLS forwarding rules.

Other flags may be defined in the future.

The NAI encoding is as per corresponding sub-TLV definition in [I-D.ietf-pce-segment-routing]

### 3. SR TE Policy Operations

SR TE Policies are advertised in the Tunnel Encapsulation Attribute defined in [I-D.ietf-idr-tunnel-encaps]. The SR TE Policy TLVs specify one (or more for load balancing purposes) list of segment identifiers (SIDs), that define the set of explicit SR TE paths towards the endpoint address encoded in the NLRI.

The Color field of the NLRI allows association of destination prefixes with a given SR TE Policy. The BGP speaker SHOULD then attach a Color Extended Community (as defined in [RFC5512]) to destination prefixes (e.g.: IPv4/IPv6 unicast prefixes) in order to allow the receiver of the SR TE Policy and of the destination prefix to steer traffic into the SR TE Policy if the destination prefix:

- o Has a BGP next-hop attribute matching the SR TE Policy SAFI NLRI Endpoint and
- o Has an attached Extended Color Community with the same value as the color of the SR TE Policy NLRI Color.

A SR TE Policy MAY also be sent by a controller, in lieu of the originating speaker. The controller sends the SR TE Policy SAFI NLRI with a Policy Color and an Endpoint identifying the Policy, where:

The Policy Color is to be used in order to steer traffic into the policy in the node receiving the SR TE Policy.

The Endpoint (with the Color) identifies the policy. Endpoint is used to match the BGP next-hop attribute of the destination prefix when steering traffic in the node receiving the SR TE Policy.

On reception of an SR TE Policy, a BGP speaker SHOULD instantiate the SR TE Policy in its routing and forwarding table with the set of segment lists (i.e.: explicit paths) included in the policy and taking into account the Binding SID and Weight sub-TLVs.

On the receiving BGP speaker, all destination prefixes that share the same Extended Color Community value and the same BGP next-hop attribute are steered to the corresponding SR TE Policy that has been instantiated and which matches the Color and Endpoint NLRI values.

Similarly, different destination prefixes can be steered into distinct SR TE Policies by coloring them differently.

### 3.1. Multipath Operation

The SR TE Policy MAY contain multiple Segment Lists which, in the absence of the Weight TLV, signifies equal cost load balancing amongst them.

When a weight sub-TLV is encoded in each Segment List TLV, then the weight value SHOULD be used in order to perform an unequal cost load balance amongst the Segment Lists as specified in Section 2.4.2.

### 3.2. Binding SID TLV

When the optional Binding SID sub-TLV is present, it indicates an instruction, to the receiving BGP speaker to allocate a Binding SID for the list of SIDs the Binding sub-TLV is related to.

Any incoming packet with the Binding SID as active segment (according to the terminology described in [I-D.ietf-spring-segment-routing]) will then have the Binding SID swapped with the list of SIDs specified in the Segment List sub-TLVs on the allocating BGP speaker. The allocated Binding SID MAY be then advertised by the BGP speaker that created it, through, e.g., BGP-LS in order to, typically, feed a controller with the updated topology and SR TE Policy information.

### 3.3. Reception of an SR TE Policy

When a BGP speaker receives an SR TE Policy from a neighbor it has to determine if the SR TE Policy advertisement is acceptable. The following applies:

- o The SR TE Policy NLRI MUST have a color value and MAY have an Endpoint value.
- o The Tunnel Encapsulation Attribute MUST be attached to the BGP Update and MUST have the Tunnel Type set to SR TE Policy (value to be assigned by IANA).
- o Within the SR TE Policy, at least one Segment List sub-TLV MUST be present.
- o Within the Segment List sub-TLV at least one Segment sub-TLV MUST be present.
- o Within Segment sub-TLV it is not required that both SID and NAI are encoded however, at least one of the two MUST be present.

Any segment (in the segment list sub-TLV) being advertised with an NAI MUST be validated by the receiver. The validation consists of resolving the SID using the NAI information, i.e., the receiver does a lookup in its local table and finds the SID value corresponding to the NAI information. The type of information carried in the NAI is related to the settings of the ST bits in the segment sub-TLV and described in [I-D.ietf-pce-segment-routing].

When a BGP speaker receives an SR TE Policy from a neighbor and according to [I-D.ietf-pce-segment-routing], the receiver MUST check the validity of the first SID of each Segment List sub-TLV of the SR TE Policy. The first SID MUST be known in the receiver local table either as a label (in the case the SID encodes a label value) or as an IPv6 address.

When a BGP speaker receives an SR TE Policy from a neighbor with an acceptable SR TE Policy SAFI NLRI and with the I-flag clear, it SHOULD compute the segment list and equivalent MPLS label from the Segment List sub-TLVs and program them in the MPLS data plane.

When a BGP speaker receives an SR TE Policy from a neighbor with an acceptable SR TE Policy SAFI NLRI and with the I-flag set, it SHOULD compute the segment list and equivalent IPv6 segment list from the Segment List sub-TLVs and program them in the IPv6 data plane according to [I-D.ietf-6man-segment-routing-header].

Also, the receiver SHOULD program its MPLS or IPv6 data planes so that BGP destination prefixes matching their Extended Color Community and BGP next-hop with the SR TE Policy SAFI NLRI Color and Endpoint are steered into the SR TE Policy and forwarded accordingly.

When building the MPLS label stack or the IPv6 Segment list from the Segment List sub-TLV, the receiving BGP speaker MUST interpret the set of Segment sub-TLVs as follows:

- o The first Segment sub-TLV represents the topmost label or the first IPv6 segment. In the receiving BGP speaker, it identifies the first segment the traffic will be directed towards to (along the SR TE explicit path).
- o The last Segment sub-TLV represents the bottommost label or the last IPv6 segment.

### 3.4. Announcing BGP SR TE Policies

Typically, the value of the SIDs encoded in the Segment sub-TLVs is determined by configuration/provisioning either in the controller or in the node originating the SR TE Policy.

A BGP speaker SHOULD follow normal iBGP/eBGP rules to propagate the SR TE Policy. The Add-Paths capability in the SR TE Policy SAFI NLRI allows the propagation of each individual policy through one or more Route Reflectors (RR) without incurring the case where one or more policies are dropped due to RR selection process.

Since the SR TE Policies are unique within an SR domain and intended only for the receiver of the SR TE Policy advertisement, a BGP speaker receiving an SR TE Policy, by default, MUST NOT propagate such policy unless explicitly configured to do so.

In order to prevent propagation of SR TE Policy advertisement, BGP filters MAY be deployed in addition to the use of the NO\_ADVERTISE community ([RFC1997]) that MAY be attached to the advertisement.

### 3.5. Flowspec and SR TE Policies

The SR TE Policy can be carried in context of a Flowspec NLRI ([RFC5575]). In this case, when the redirect to IP nexthop is specified as in [I-D.ietf-idr-flowspec-redirect-ip], the tunnel to the nexthop is specified by the segment list in the Segment List sub-TLVs. The Segment List (e.g.: label stack or IPv6 segment list) is imposed to flows matching the criteria in the Flowspec route in order to steer them towards the nexthop as specified in the SR TE Policy SAFI NLRI.

#### 4. Acknowledgments

The authors of this document would like to thank Eric Rosen for his review of this document.

#### 5. IANA Considerations

This document defines:

- o a new SAFI in the registry "Subsequent Address Family Identifiers (SAFI) Parameters":

Suggested Value	Description	Reference
73	SR TE Policy SAFI	This document

- o a new Tunnel-Type in the registry "BGP Tunnel Encapsulation Attribute Tunnel Types":

Suggested Value	Description	Reference
14	SR TE Policy Type	This document

- o new sub-TLVs in the registry "BGP Tunnel Encapsulation Attribute sub-TLVs":

Suggested Value	Description	Reference
6	Binding SID sub-TLV	This document
7	Weight sub-TLV	This document
8	Segment List sub-TLV	This document
9	Segment sub-TLV	This document

#### 6. Security Considerations

TBD.

#### 7. References

##### 7.1. Normative References

[I-D.ietf-idr-tunnel-encaps]  
 Rosen, E., Patel, K., and G. Velde, "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-01 (work in progress), December 2015.

- [I-D.ietf-pce-segment-routing]  
Sivabalan, S., Medved, J., Filsfils, C., Crabbe, E.,  
Lopez, V., Tantsura, J., Henderickx, W., and J. Hardwick,  
"PCEP Extensions for Segment Routing", draft-ietf-pce-  
segment-routing-06 (work in progress), August 2015.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate  
Requirement Levels", BCP 14, RFC 2119,  
DOI 10.17487/RFC2119, March 1997,  
<<http://www.rfc-editor.org/info/rfc2119>>.
- [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,  
<<http://www.rfc-editor.org/info/rfc3209>>.
- [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,  
<<http://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, <<http://www.rfc-editor.org/info/rfc4364>>.
- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter,  
"Multiprotocol Extensions for BGP-4", RFC 4760,  
DOI 10.17487/RFC4760, January 2007,  
<<http://www.rfc-editor.org/info/rfc4760>>.
- [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, <<http://www.rfc-editor.org/info/rfc5462>>.
- [RFC5512] Mohapatra, P. and E. Rosen, "The BGP Encapsulation  
Subsequent Address Family Identifier (SAFI) and the BGP  
Tunnel Encapsulation Attribute", RFC 5512,  
DOI 10.17487/RFC5512, April 2009,  
<<http://www.rfc-editor.org/info/rfc5512>>.
- [RFC5575] Marques, P., Sheth, N., Raszuk, R., Greene, B., Mauch, J.,  
and D. McPherson, "Dissemination of Flow Specification  
Rules", RFC 5575, DOI 10.17487/RFC5575, August 2009,  
<<http://www.rfc-editor.org/info/rfc5575>>.

## 7.2. Informational References

- [I-D.ietf-6man-segment-routing-header]  
Previdi, S., Filsfils, C., Field, B., Leung, I., Linkova, J., Kosugi, T., Vyncke, E., and D. Lebrun, "IPv6 Segment Routing Header (SRH)", draft-ietf-6man-segment-routing-header-00 (work in progress), December 2015.
- [I-D.ietf-idr-add-paths]  
Walton, D., Retana, A., Chen, E., and J. Scudder, "Advertisement of Multiple Paths in BGP", draft-ietf-idr-add-paths-13 (work in progress), December 2015.
- [I-D.ietf-idr-flowspec-redirect-ip]  
Uttaro, J., Haas, J., Texier, M., Andy, A., Ray, S., Simpson, A., and W. Henderickx, "BGP Flow-Spec Redirect to IP Action", draft-ietf-idr-flowspec-redirect-ip-02 (work in progress), February 2015.
- [I-D.ietf-spring-segment-routing]  
Filsfils, C., Previdi, S., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", draft-ietf-spring-segment-routing-07 (work in progress), December 2015.
- [I-D.ietf-spring-segment-routing-mpls]  
Filsfils, C., Previdi, S., Bashandy, A., Decraene, B., Litkowski, S., Horneffer, M., Shakir, R., Tantsura, J., and E. Crabbe, "Segment Routing with MPLS data plane", draft-ietf-spring-segment-routing-mpls-03 (work in progress), February 2016.
- [RFC1997] Chandra, R., Traina, P., and T. Li, "BGP Communities Attribute", RFC 1997, DOI 10.17487/RFC1997, August 1996, <<http://www.rfc-editor.org/info/rfc1997>>.
- [RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, DOI 10.17487/RFC4456, April 2006, <<http://www.rfc-editor.org/info/rfc4456>>.

Authors' Addresses



Stefano Previdi (editor)  
Cisco Systems, Inc.  
Via Del Serafico, 200  
Rome 00142  
Italy

Email: sprevidi@cisco.com

Clarence Filsfils  
Cisco Systems, Inc.  
Brussels  
BE

Email: cfilsfil@cisco.com

Arjun Sreekantiah  
Cisco Systems, Inc.  
170 W. Tasman Drive  
San Jose, CA 95134  
USA

Email: asreekan@cisco.com

Siva Sivabalan  
Cisco Systems, Inc.  
170 W. Tasman Drive  
San Jose, CA 95134  
USA

Email: msiva@cisco.com

Paul Mattes  
Microsoft  
One Microsoft Way  
Redmond, WA 98052  
USA

Email: pamattes@microsoft.com

IDR Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: September 18, 2016

G. Van de Velde, Ed.  
W. Henderickx  
Alcatel-Lucent  
K. Patel  
A. Sreekantiah  
Cisco Systems  
March 17, 2016

Flowspec Indirection-id Redirect  
draft-vandevelde-idr-flowspec-path-redirect-02

Abstract

Flow-spec is an extension to BGP that allows for the dissemination of traffic flow specification rules. This has many possible applications but the primary one for many network operators is the distribution of traffic filtering actions for DDoS mitigation. The flow-spec standard RFC5575 [2] defines a redirect-to-VRF action for policy-based forwarding but this mechanism is not always sufficient, particular if the redirected traffic needs to be steered into an engineered path or into a service plane.

This document defines a new redirect-to-INDIRECTION\_ID (32-bit or 128-bit) flow-spec action to provide advanced redirection capabilities. When activated, the flowspec Indirection-id is used to identify the next-hop redirect information within a router localized Indirection-id table. This allows a flowspec controller to signal redirection towards a next-hop IP address, a shortest path tunnel, a traffic engineered tunnel or a next-next-hop engineered tunnel interface.

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 [1].

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 <http://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 18, 2016.

#### Copyright Notice

Copyright (c) 2016 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 . . . . .	2
2. INDIRECTION_ID and INDIRECTION_ID Table . . . . .	3
3. Use Case Scenarios . . . . .	4
3.1. Redirection shortest Path tunnel . . . . .	4
3.2. Redirection to path-engineered tunnels . . . . .	4
3.3. Redirection to Next-next-hop tunnels . . . . .	5
4. Redirect to INDIRECTION-ID Communities . . . . .	6
5. Redirect using localized INDIRECTION_ID Router Mapping . . . . .	7
6. Validation Procedures . . . . .	8
7. Security Considerations . . . . .	8
8. Acknowledgements . . . . .	8
9. IANA Considerations . . . . .	8
10. References . . . . .	8
10.1. Normative References . . . . .	9
10.2. Informative References . . . . .	9
Authors' Addresses . . . . .	9

#### 1. Introduction

Flow-spec RFC5575 [2] is an extension to BGP that allows for the dissemination of traffic flow specification rules. This has many possible applications but the primary one for many network operators is the distribution of traffic filtering actions for DDoS mitigation.

Every flow-spec route is effectively a rule, consisting of a matching part (encoded in the NLRI field) and an action part (encoded in one or more BGP extended community). The flow-spec standard RFC5575 [2] defines widely-used filter actions such as discard and rate limit; it also defines a redirect-to-VRF action for policy-based forwarding. Using the redirect-to-VRF action for redirecting traffic towards an alternate destination is useful for DDoS mitigation but using this technology can be cumbersome when there is need to redirect the traffic onto an engineered traffic path.

This draft proposes a new redirect-to-Indirection-id flow-spec action facilitating an anchor point for policy-based forwarding onto an engineered path or into a service plane. The router consuming and utilizing the flowspec rule makes a local mapping between the flowspec signalled redirect Indirection-id and locally available redirection information referenced by the Indirection-id. This locally available redirection information is derived from out-of-band programming or signalling.

The redirect-to-Indirection-id is encoded in a newly defined BGP extended Indirection\_ID community.

The construction of the Indirection-id redirect table and the technology used to create an engineered path are out-of-scope of this document.

## 2. INDIRECTION\_ID and INDIRECTION\_ID Table

An INDIRECTION\_ID is an abstract number (32-bit or 128-bit value) used as identifier for a localised redirection decision. e.g. When a BGP flowspec controller intends to redirect a flow using the redirect-to-INDIRECTION\_ID action then it has the ability to redirect the flow to a destination abstracted as the INDIRECTION\_ID. The device receiving the BGP flowspec rule will use the INDIRECTION\_ID to identify the next-hop and the relevant tunnel encapsulations that need to be pushed by a localised recursive lookup using information located within the INDIRECTION\_ID table.

The INDIRECTION\_ID Table is a router localised table. The table content is constructed out of INDIRECTION\_IDs and corresponding redirect information which may be of recursive or non-recursive nature. When the redirect information is non-recursive, then the represented information MUST be sufficient to identify the local egress interface and the corresponding required encapsulations. However, if the information is recursive, then the represented information MUST be sufficient to identify the local egress interface and corresponding encapsulations using additional recursions.

### 3. Use Case Scenarios

This section describes use-case scenarios when deploying redirect-to-INDIRECTION\_ID.

#### 3.1. Redirection shortest Path tunnel

A first use-case is allowing a BGP Flowspec controller send a single flowspec policy message (redirect-to-INDIRECTION\_ID#1) to many BGP flowspec consuming routers. This message is instructing the Flowspec recipient routers to redirect traffic onto a tunnel to a single IP destination address.

For this use-case scenario, each flowspec recipient router has a tunnel configured following the shortest path towards a tunnel IP destination address. Each tunnel can have its own unique encapsulation associated. Each tunnel is associated with an INDIRECTION\_ID, and for this example it is on all recipient routers INDIRECTION\_ID#1. Both manual and orchestrated tunnel provisioning is supported, however for large scale deployment automation is advisable.

When using this setup, a BGP flowspec controller can send a single BGP Flowspec NLRI with redirect-to-INDIRECTION\_ID#1. This BGP Flowspec NLRI is received by all recipient routers. Each of the recipient routers performs a localised recursive lookup for INDIRECTION\_ID#1 in the INDIRECTION\_ID Table and identifies the corresponding localised tunnel redirect information. This localised tunnel information is now used to redirect traffic matching the Flowspec policy towards a tunnel, each potentially using its own unique tunnel encapsulation.

#### 3.2. Redirection to path-engineered tunnels

A second use-case is allowing a BGP Flowspec controller send a single flowspec policy message (redirect-to-INDIRECTION\_ID#2) to many BGP flowspec consuming routers. This message is instructing the Flowspec recipient routers to redirect traffic onto a path engineered tunnel.

For this use-case scenario, each flowspec recipient router has a path engineered tunnel configured. Each tunnel can have its own unique encapsulation and engineered path associated. Each tunnel is associated with an INDIRECTION\_ID, and for this example it is on all recipient routers INDIRECTION\_ID#2. Both manual and orchestrated tunnel provisioning is supported, however for large scale deployment automation is advisable.

A first example using this setup, is when a BGP flowspec controller sends a single BGP Flowspec NLRI with redirect-to-INDIRECTION\_ID#2. This BGP Flowspec NLRI is received by all recipient routers. Each of the recipient routers performs a localised recursive lookup for INDIRECTION\_ID#2 in the INDIRECTION\_ID Table and identifies the corresponding localised tunnel redirect information. This localised tunnel information is now used to redirect traffic matching the Flowspec policy towards a path engineered tunnel.

A second example can be found in segment routed networks where path engineered tunnels can be setup by means of a controller signaling explicit paths to peering routers. In such a case, the INDIRECTION\_ID references to a Segment Routing Binding SID. The Binding SID is a segment identifier value (as per segment routing definitions in [I-D.draft-ietf-spring-segment-routing] [6]) used to associate with a explicit path and can be setup by a controller using BGP as specified in [I-D.sreekantiah-idr-segment-routing-te] [5] or using PCE as detailed in draft-ietf-pce-segment-routing [7]. When a BGP speaker receives a flow-spec route with a 'redirect to Binding SID' extended community, it installs a traffic filtering rule that matches the packets described by the NLRI field and redirects them to the explicit path associated with the Binding SID. The explicit path is specified as a set/stack of segment identifiers as detailed in the previous documents. The stack of segment identifiers is now imposed on packets matching the flow-spec rule to perform redirection as per the explicit path setup prior. The encoding of the Binding SID value is specified in section 4, with the indirection field now encoding the associated value for the binding SID.

### 3.3. Redirection to Next-next-hop tunnels

A Third use-case is allowing a BGP Flowspec controller send a single flowspec policy message (redirect-to-INDIRECTION\_ID#3) to many BGP flowspec consuming routers. This message is instructing the Flowspec recipient routers to redirect traffic onto a shortest or engineered path to a tunnel end-point and onwards to the next-hop-interface on this end-point. This type of tunnel is used for example for Segment Routing Central Egress Path Engineering [4].

For this use-case scenario, each flowspec recipient router constructs redirect information using two tunnel components. The first component is a shortest or engineered path towards a network egress router. The second component is the interface used on this network egress router to which the redirected traffic needs to be steered upon. The combination of these two components allows steering towards the next-hop of the egress router, allowing for example the Central Egress Path Engineering using BGP Flowspec [4].

The redirection towards a next-next-hop tunnel can be done by using either a single INDIRECTION\_ID representing the combined path to the egress router and steering the traffic to the egress interface, or by using individual INDIRECTION\_IDs each representing a tunnel component (One INDIRECTION\_ID value to identify the path towards the egress router and another INDIRECTION\_ID value to identify the egress interface on this egress router towards the next-next-hop). When using individual INDIRECTION\_IDs it is required to use INDIRECTION\_ID community Tunnel IDs (TID) each identifying a component of the complete redirect path attached to the NLRI.

i.e. when using next-next-hop tunnels, a BGP flowspec controller can send a single BGP Flowspec NLRI with redirect-to-INDIRECTION\_ID#3. This BGP Flowspec NLRI is received by all recipient routers. Each of the recipient routers performs a localised recursive lookup for INDIRECTION\_ID#3 in the INDIRECTION\_ID Table and identifies the corresponding localised tunnel redirect information (=path to the egress router and the next-hop egress interface on this router). Traffic matching the flowspec policy is redirected towards the recursively found redirection information.

#### 4. Redirect to INDIRECTION-ID Communities

This document defines a new BGP extended community. The extended communities have a type indicating they are transitive and IPv4-address-specific or IPv6-address-specific, depending on whether the INDIRECTION\_ID is 32-bit or 128-bit. The sub-type value [to be assigned by IANA] indicates that the global administrator and local administrator fields encode a flow-spec 'redirect to INDIRECTION\_ID' action. In the new extended community the 4-byte or 16-byte global administrator field encodes the 32-bit or 128-bit INDIRECTION\_ID's providing the INDIRECTION\_ID to allow a local to the router mapping reference to an engineered Path. The 2-byte local administrator field is formatted as shown in Figure 1.

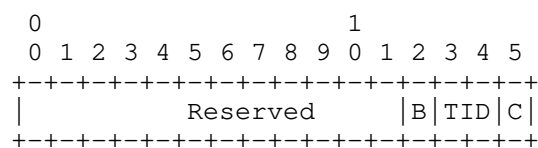


Figure 1

In the local administrator field the least-significant bit is defined as the 'C' (or copy) bit. When the 'C' bit is set the redirection

applies to copies of the matching packets and not to the original traffic stream.

The 'TID' field identifies a 2 bit Table-id field. This field is used to provide the router consuming the flowspec rule an indication how and where to use the INDIRECTION\_ID when redirecting traffic.

Bit 2 is defined to be the 'B' (Binding) bit. When the 'B' bit is set, the value encoded in the global administrator field is a Binding Segment Identifier value the use of which is detailed in section 3.2.

All bits other than the 'C', 'TID' and 'B' bits in the local administrator field MUST be set to 0 by the originating BGP speaker and ignored by receiving BGP speakers.

#### 5. Redirect using localized INDIRECTION\_ID Router Mapping

When a BGP speaker receives a flow-spec route with a 'redirect to INDIRECTION\_ID' extended community and this route represents the one and only best path, it installs a traffic filtering rule that matches the packets described by the NLRI field and redirects them (C=0) or copies them (C=1) towards the INDIRECTION\_ID local recursed Path. The BGP speaker is expected to do a local INDIRECTION\_ID Table lookup to identify the redirection information.

The router local INDIRECTION\_ID table contains a list of INDIRECTION\_ID's each mapped to redirect information. The redirect information can be non-recursive (i.e. there is only one option available in the INDIRECTION\_ID Table) or recursive (i.e. L3 VPN, L2 VPN, a pre-programmed routing topology, etc... ).

- o When the redirect information is non-recursive then the packet is redirected based upon the information found in the Table.
- o In case of a next-hop tunnel, the traffic matching the flowspec rule is redirected to the next-hop tunnel. This tunnel could be instantiated through various means (i.e. manual configuration, PCEP, RSVP-TE, WAN Controller, Segment Routing, etc...).
- o In case of redirection to a local next-hop interface, the traffic matching the flowspec rule is redirected to the local next-hop interface.
- o In case the INDIRECTION\_ID Table lookup results in redirect information identifying an additional layer of recursion, then this will trigger the flow to be redirected based upon an additional routing lookup within the realm of the additional layer of recursion.



## 6. Validation Procedures

The validation check described in RFC5575 [2] and revised in [3] SHOULD be applied by default to received flow-spec routes with a 'redirect to INDIRECTION\_ID' extended community. This means that a flow-spec route with a destination prefix subcomponent SHOULD NOT be accepted from an EBGp peer unless that peer also advertised the best path for the matching unicast route.

It is possible from a semantics perspective to have multiple redirect actions defined within a single flowspec rule. When a BGP flowspec NLRI has a 'redirect to INDIRECTION\_ID' extended community attached resulting in valid redirection then it MUST take priority above all other redirect actions imposed. However, if the 'redirect to INDIRECTION\_ID' does not result in a valid redirection, then the flowspec rule must be processed as if the 'redirect to INDIRECTION\_ID' community was not attached to the flowspec route and MUST provide an indication within the BGP routing table that the 'redirect to INDIRECTION\_ID' resulted in an invalid redirection action.

## 7. Security Considerations

A system using 'redirect-to-INDIRECTION\_ID' extended community can cause during the redirect mitigation of a DDoS attack result in an overflow of traffic being received by the mitigation infrastructure.

## 8. Acknowledgements

This document received valuable comments and input from IDR working group including Adam Simpson, Mustapha Aissaoui, Jan Mertens, Robert Raszuk, Jeff Haas, Susan Hares and Lucy Yong

## 9. IANA Considerations

This document requests a new sub-type from the "Transitive IPv4-Address-Specific" extended community registry. The sub-type name shall be 'Flow-spec Redirect to 32-bit Path-id'.

This document requests a new sub-type from the "Transitive IPv6-Address-Specific" extended community registry. The sub-type name shall be 'Flow-spec Redirect to 128-bit Path-id'.

## 10. References

## 10.1. Normative References

- [1] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997, <<http://xml.resource.org/public/rfc/html/rfc2119.html>>.
- [2] 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, <<http://www.rfc-editor.org/info/rfc5575>>.

## 10.2. Informative References

- [3] Uttaro, J., Filsfils, C., Alcaide, J., and P. Mohapatra, "Revised Validation Procedure for BGP Flow Specifications", January 2014.
- [4] Filsfils, C., Previdi, S., Aries, E., Ginsburg, D., and D. Afanasiev, "Segment Routing Centralized Egress Peer Engineering", October 2015.
- [5] Sreekantiah, A., Filsfils, C., Previdi, S., Sivabalan, S., Mattes, P., and S. Lin, "Segment Routing Traffic Engineering Policy using BGP", October 2015.
- [6] Filsfils, C., Previdi, S., Decraene, B., Litkowski, S., Shakir, R., Bashandy, A., Horneffer, M., Henderickx, W., Tantsura, J., Crabbe, E., Milojevic, I., and S. Ytti, "Segment Routing Architecture", December 2015.
- [7] Sivabalan, S., Medved, M., Filsfils, C., Litkowski, S., Raszuk, R., Bashandy, A., Lopez, V., Tantsura, J., Henderickx, W., Hardwick, J., Milojevic, I., and S. Ytti, "PCEP Extensions for Segment Routing", December 2015.

## Authors' Addresses

Gunter Van de Velde (editor)  
Alcatel-Lucent  
Antwerp  
BE

Email: [gunter.van\\_de\\_velde@alcatel-lucent.com](mailto:gunter.van_de_velde@alcatel-lucent.com)

Wim Henderickx  
Alcatel-Lucent  
Antwerp  
BE

Email: [wim.henderickx@alcatel-lucent.com](mailto:wim.henderickx@alcatel-lucent.com)

Keyur Patel  
Cisco Systems  
170 W. Tasman Drive  
San Jose, CA 95134  
USA

Email: [keyupate@cisco.com](mailto:keyupate@cisco.com)

Arjun Sreekantiah  
Cisco Systems  
170 W. Tasman Drive  
San Jose, CA 95134  
USA

Email: [asreekan@cisco.com](mailto:asreekan@cisco.com)

Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: September 22, 2016

A. Azimov  
E. Bogomazov  
Qrator Labs  
R. Bush  
Internet Initiative Japan  
March 21, 2016

Route Leak Detection and Filtering using Roles in Update and Open  
messages  
draft-ymbk-idr-bgp-open-policy-00

Abstract

Route Leaks are propagation of BGP prefixes which violate assumptions of BGP topology relationships; e.g. passing a route learned from one peer to another peer or to a transit provider, passing a route learned from one transit provider to another transit provider or to a peer. Today, approaches to leak prevention rely on marking routes according to some configuration options without any check of the configuration corresponds to that of the BGP neighbor, or enforcement that the two BGP speakers agree on the relationship. This document enhances BGP Open to establish agreement of the (peer, customer, provider, internal) relationship of two BGP neighboring speakers to enforce appropriate configuration on both sides. Propagated routes are then marked with a flag according to agreed relationship allowing detection and mitigation of route leaks.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" are to be interpreted as described in RFC 2119 [RFC2119] only when they appear in all upper case. They may also appear in lower or mixed case as English words, without normative meaning.

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 <http://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 22, 2016.

#### Copyright Notice

Copyright (c) 2016 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 . . . . .	2
2. BGP Role . . . . .	3
3. Role capability . . . . .	4
4. Role correctness . . . . .	4
4.1. Strict mode . . . . .	5
5. BGP Only To Customer attribute . . . . .	5
5.1. Attribute usage . . . . .	5
6. Additional Considerations . . . . .	6
7. IANA Considerations . . . . .	6
8. Security Considerations . . . . .	7
9. References . . . . .	7
9.1. Normative References . . . . .	7
9.2. Informative References . . . . .	8
Authors' Addresses . . . . .	8

#### 1. Introduction

For the purposes of this document BGP route leaks are when a BGP route was learned from transit provider or peer and is announced to another provider or peer. See [I-D.ietf-grow-route-leak-problem-definition]. These are usually the result of misconfigured or absent BGP route filtering or lack of coordination between two BGP speakers.

[I-D.ietf-idr-route-leak-detection-mitigation] describes a method of marking and detecting leaks which relies on operator maintained

markings. Unfortunately, in most cases, a leaking router will likely also be misconfigured to mark incorrectly. The proposed mechanism provides an opportunity to detect route leaks made by third parties but provides no support to avoid route leak creation. The leak avoidance still relies on communities which are optional and often missed due to mistakes or misunderstanding of the BGP configuration process.

It has been suggested to use white list filtering, relying on knowing the prefixes in the customer cone as import filtering, in order to detect route leaks. Unfortunately, a large number of incidents is created by leaking routes defined in customer cone; a lot of medium size transit operators use a single prefix list as only the ACL for export filtering, without community tagging and paying attention to the source of a learned route. So, if they learn a customer's route from their provider or peer - they will announce it in all directions, including other providers or peers. This misconfiguration affects a limited number of prefixes; but such route leaks will obviously bypass customer cone import filtering made by upper level upstream providers.

Also, route tagging which relies on operator maintained policy configuration is too easily and too often misconfigured.

This document specifies a new BGP Capability Code, [RFC5492] Sec 4, which two BGP speakers MAY use to ensure that they MUST agree on their relationship; i.e. customer and provider or peers. Either or both may optionally be configured to require that this option be exchanged for the BGP Open to succeed.

Also this document specifies a way to mark routes according to BGP Roles and a way to create double-boundary filters for detection and preventing propagation of route leaks via a new BGP Path Attribute.

## 2. BGP Role

BGP Role is new mandatory configuration option. It reflects the real-world agreement between two BGP speakers about their business relationship.

Allowed Role values are:

- o Provider - sender is a transit provider to neighbor;
- o Customer - sender is customer of neighbor;
- o Peer - sender and neighbor are peers;

- o Internal - sender is part of an internal AS of an organization which has multiple ASs, is a confederation, ...

Since BGP Role reflects the relationship between two BGP speakers, it could also be used for more than route leak mitigation.

### 3. Role capability

The TLV (type, length, value) of the BGP Role capability are:

- o Type - <TBD1>;
- o Length - 1 (octet);
- o Value - integer corresponding to speaker' BGP Role.

Value	Role name
0	Undefined
1	Sender is Peer
2	Sender is Provider
3	Sender is Customer
4	Sender is Internal

Table 1: Predefined BGP Role Values

### 4. Role correctness

Section 2 described how BGP Role is a reflection of the relationship between two BGP speakers. But the mere presence of BGP Role doesn't automatically guarantee role agreement between two BGP peers.

To enforce correctness, use the BGP Role check with a set of constrains on how speakers' BGP Roles MUST corresponded. Of course, each speaker MUST announce and accept the BGP Role capability in the BGP OPEN message exchange.

If a speaker receives a BGP Role capability, it SHOULD check value of the received capability with its own BGP Role. The allowed pairings are (first a sender's Role, second the receiver's Role):

Sender Role	Receiver Role
Peer	Peer
Provider	Customer
Customer	Provider
Internal	Internal

Table 2: Allowed Role Capabilities

In all other cases speaker MUST send a Role Mismatch Notification (code 2, sub-code <TBD2>).

#### 4.1. Strict mode

A new BGP configuration option "strict mode" is defined with values of true or false. If set to true, then the speaker MUST refuse to establish a BGP session with peers which do not announce BGP Role capability in their OPEN message. If a speaker rejects a connection, it MUST send a Connection Rejected Notification [RFC4486] (Notification with error code 6, subcode 5). By default strict mode SHOULD be set to false for backward compatibility with BGP speakers, that do not yet support this mechanism.

#### 5. BGP Only To Customer attribute

The Only To Customer (OTC) attribute is a new optional, transitive BGP Path attribute with the Type Code <TBD3>. This attribute has zero length as it used only as a flag.

##### 5.1. Attribute usage

There are four rules for setting the OTC attribute:

1. The OTC attribute SHOULD be added to all incoming routes if the receiver's Role is Customer or Peer;
2. Routes with the OTC attribute set MUST NOT be announced to a neighbor which is a Provider or a Peer;
3. The OTC attribute SHOULD be added to all outgoing routes if the neighbor has not sent the Role capability in the OPEN message and sender's Role is Provider or Peer;
4. If the receiver's Role is Provider or Peer, incoming routes with the OTC attribute set SHOULD be given a lower local preference, or they MAY be dropped.



The first two rules avoid creation of route leaks by an AS. The last two rules provide the data to allow detection of route leaks made by some other AS in the AS Path.

The OTC attribute will be added automatically if at least one of the speakers correctly sets its role. Additionally, the OTC attribute will be checked if at least one of the speakers set its role correctly. In other words, this double-checks at borders to prevent route leaks.

## 6. Additional Considerations

As BGP Role reflects the relationship between neighbors, it can also have other uses. As an example, BGP Role might affect route priority, or be used to distinguish borders of a network if a network consists of multiple AS.

Though such uses may be worthwhile, they are not the goal of this document. Note that such uses would require local policy control.

## 7. IANA Considerations

This document defines a new Capability Codes option [to be removed upon publication: <http://www.iana.org/assignments/capability-codes/capability-codes.xhtml>] [RFC5492], named "BGP Role", assigned value <TBD1> . The length of this capability is 1.

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". Assignments consist of Value and corresponding Role name. Initially this registry is to be populated with the data in Table 1. Future assignments may be made by a standard action procedure [RFC5226].

This document defines new subcode, "Role Mismatch", assigned value <TBD2> in the OPEN Message Error subcodes registry [to be removed upon publication: <http://www.iana.org/assignments/bgp-parameters/bgp-parameters.xhtml#bgp-parameters-6>] [RFC4271].

This document defines a new optional, transitive BGP Path Attributes option, named "Only To Customer", assigned value <TBD3> [To be removed upon publication: <http://www.iana.org/assignments/bgp-parameters/bgp-parameters.xhtml#bgp-parameters-2>] [RFC4271]. The length of this attribute is 0.

## 8. Security Considerations

This document proposes a mechanism for avoiding route leaks, that are the result of BGP policy misconfiguration. That includes preventing route leaks created inside an AS (company), and route leak detection, if a route was leaked by third party.

Deliberate sending of a known conflicting BGP Role could be used to sabotage a BGP connection. This is easily detectable.

Deliberate mis-marking of the OTC flag could be used to sabotage a route's propagation.

BGP Role is disclosed only to an immediate BGP speaker, so it will not itself reveal any sensitive information to third parties.

On the other hand, OTC is a transitive BGP AS\_PATH attribute which reveals a bit about a BGP speaker's business relationship. It will give a strong hint that some link isn't customer to provider, but will not help to distinguish if it is provider to customer or peer to peer. If OTC is BGPsec signed, it can not be removed for business confidentiality.

## 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, <<http://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, <<http://www.rfc-editor.org/info/rfc4271>>.
- [RFC4486] Chen, E. and V. Gillet, "Subcodes for BGP Cease Notification Message", RFC 4486, DOI 10.17487/RFC4486, April 2006, <<http://www.rfc-editor.org/info/rfc4486>>.
- [RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, DOI 10.17487/RFC5492, February 2009, <<http://www.rfc-editor.org/info/rfc5492>>.

## 9.2. Informative References

- [I-D.ietf-grow-route-leak-problem-definition]  
Sriram, K., Montgomery, D., McPherson, D., Osterweil, E.,  
and B. Dickson, "Problem Definition and Classification of  
BGP Route Leaks", draft-ietf-grow-route-leak-problem-  
definition-03 (work in progress), October 2015.
- [I-D.ietf-idr-route-leak-detection-mitigation]  
Sriram, K., Montgomery, D., Dickson, B., Patel, K., and A.  
Robachevsky, "Methods for Detection and Mitigation of BGP  
Route Leaks", draft-ietf-idr-route-leak-detection-  
mitigation-01 (work in progress), October 2015.
- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an  
IANA Considerations Section in RFCs", BCP 26, RFC 5226,  
DOI 10.17487/RFC5226, May 2008,  
<<http://www.rfc-editor.org/info/rfc5226>>.

## Authors' Addresses

Alexander Azimov  
Qrator Labs

Email: [aa@qrator.net](mailto:aa@qrator.net)

Eugene Bogomazov  
Qrator Labs

Email: [eb@qrator.net](mailto:eb@qrator.net)

Randy Bush  
Internet Initiative Japan

Email: [randy@psg.com](mailto:randy@psg.com)

Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: September 18, 2016

Y. Lucy  
S. Zhuang  
W. Hao  
Huawei Technologies  
March 17, 2016

BGP Flowspec Redirect to VPN RD Extended Community  
draft-yong-idr-flowspec-redirect-vpn-rd-00

Abstract

This document defines a new type of the redirect extended community, called as Redirect to VPN RD Extended Community. When activated, the Redirect to VPN RD Extended Community is used to identify the unique VPN instance within a router.

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 <http://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 18, 2016.

Copyright Notice

Copyright (c) 2016 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 . . . . .	2
2. Operation Concerns in Redirect VRF Action . . . . .	3
3. Redirect to VPN RD Extended Community Format . . . . .	5
4. Using Redirect VPN RD Extended Community . . . . .	6
5. IANA Considerations . . . . .	7
6. Security Considerations . . . . .	7
7. References . . . . .	7
7.1. Normative References . . . . .	7
7.2. Informative References . . . . .	8
Authors' Addresses . . . . .	8

## 1. Introduction

"Dissemination of Flow Specification Rules" [RFC5575], commonly known as BGP Flowspec, provided for a BGP Extended Community [RFC4360][RFC4360] that served to redirect traffic that matched the flow specification's Network Layer Reachability Information (NLRI) to a Virtual Routing and Forwarding (VRF) instance that lists the specified route-target in its import policy. In that RFC, the Redirect Extended Community was documented as follows:

type	extended community	encoding
0x8008	redirect	6-byte Route Target

[...]

Redirect: The redirect extended community allows the traffic to be redirected to a VRF routing instance that lists the specified route-target in its import policy. If several local instances match this criteria, the choice between them is a local matter (for example, the instance with the lowest Route Distinguisher value can be elected). This extended community uses the same encoding as the Route Target extended community [RFC4360].

[...]

11. IANA Considerations

[...]

The following traffic filtering flow specification rules have been allocated by IANA from the "BGP Extended Communities Type - Experimental Use" registry as follows:

[...]

0x8008 - Flow spec redirect

[RFC7674] updates RFC 5575 ("Dissemination of Flow Specification Rules") to clarify the formatting of the BGP Flowspec Redirect Extended Community. This document defines the following redirect extended communities:

type	extended community	encoding
0x8008	redirect AS-2byte	2-octet AS, 4-octet Value
0x8108	redirect IPv4	4-octet IPv4 Address, 2-octet Value
0x8208	redirect AS-4byte	4-octet AS, 2-octet Value

## 2. Operation Concerns in Redirect VRF Action

Following example is a case used in a backbone network.

Traffic Analyzer is installed at the edge of the backbone to detect the attack.

Scrubbing Center is installed at the edge of the backbone tackle the attack.

VRF scrubbing-vpn is configured on R1 and R2. A default route in R1's scrubbing-vpn VRF is configured to reach the Scrubbing Center, and MP-BGP is configured to advertise the default route from VRF scrubbing-vpn to the remote router R2.

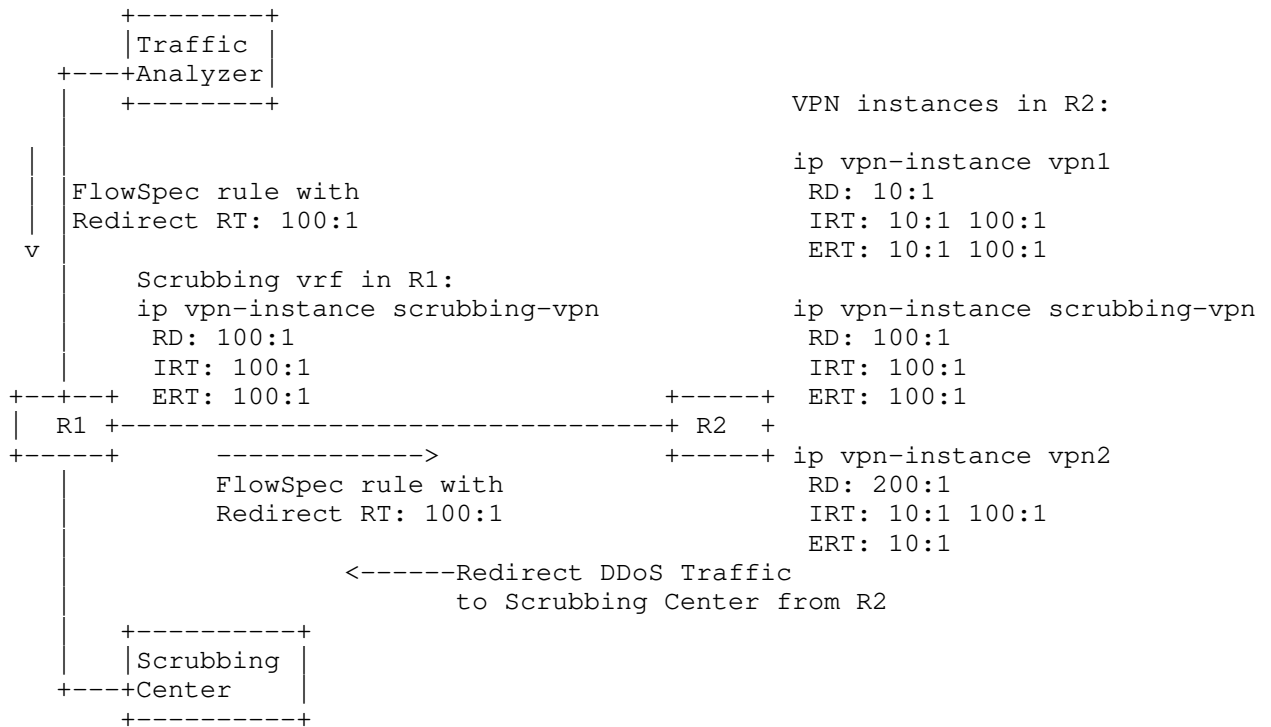


Figure 1 Redirect DDoS Traffic to Scrubbing Center Using Redirect VPN RT

Upon detecting the attack target to the user of the backbone network, Traffic Analyzer will push a Flowspec rule to R1 with Redirect RT: 100:1.

R1 will advertise the receiving Flowspec rule to R2.

If the VRF scrubbing-vpn on R2 is the only VRF routing instance, then the receiving Flowspec rule from R1 can be imported by the VRF routing instance scrubbing-vpn. The attack traffic that matches the Flowspec rule on R2 will be redirected to the VRF scrubbing-vpn and sent to the Scrubbing Center.

However in this case, there are several local instances on R2 can match the Redirect RT: 100:1(as shown in following table). To make it work, according to RFC 5575, an operator has to configure R2 so that 'Redirect to VPN' will point to the scrubbing-vpn, which introduces operation complex and/or prone to an error. To avoid this configuration, a unique RT value for BGP FS 'Redirect to VPN' action has to be selected, which can be an operation complex in a large network.

VRF	IRT	RD
vpn1	10:1 100:1	10:1
scrubbing-vpn	100:1	100:1
vpn2	10:1 100:1	200:1

The reason for the above issue is that the IRT isn't unique on one router, for example, IRT 100:1 can be assigned to multiple VRF instances: vpn1, scrubbing-vpn and vpn2.

The Route Distinguisher is unique on one router, In order to address this operational concern, this document introduces a new type of the redirect extended community, called as Redirect to VPN RD Extended Community, When activated, the Redirect to VPN RD Extended Community is used to identify the unique VPN instance within a router.

### 3. Redirect to VPN RD Extended Community Format

This document defines a new type of the redirect extended community, called as Redirect to VPN RD Extended Community. This extended community is a new transitive extended community with the Sub-Type field is TBD. The IANA registry of BGP Extended Communities clearly identifies communities of specific formats: "Two-octet AS Specific Extended Community" [RFC4360], "Four-octet AS Specific Extended Community" [RFC5668], and "IPv4 Address Specific Extended Community" [RFC4360]. Route Targets [RFC4360] identify this format in the high-order (Type) octet of the Extended Community, Redirect to VPN RD Extended Community uses the same mechanism

This document defines the following VPN RD Extended Communities:



Type	Sub-Type	Extended Community	Encoding
0x80	TBD	AS-2byte RD	2-octet AS, 4-octet Value
0x81	TBD	IPv4 RD	4-octet IPv4 Address, 2-octet Value
0x82	TBD	AS-4byte RD	4-octet AS, 2-octet Value

Figure 2: VPN RD Extended Communities

It should be noted that the low-order nibble of the Redirect's Type field corresponds to the Route Target Extended Community format field (Type). (See Sections 3.1, 3.2, and 4 of [RFC4360] plus Section 2 of [RFC5668].) The low-order octet (Sub-Type) of the Redirect to VPN RD Extended Community is TBD, in contrast to 0x02 for Route Targets and 0x08 for Redirect to VPN RT Extended Community.

#### 4. Using Redirect VPN RD Extended Community

Upon detecting the attack target to the user of the backbone network, Traffic Analyzer will push a Flowspec rule to R1 with Redirect VPN RD: 100:1.

R1 will advertise the receiving Flowspec rule to R2.

In R2, the receiving Flowspec rule from R1 can be imported by the VRF routing instance scrubbing-vpn. The attack traffic that matches the Flowspec rule on R2 will be correctly redirected to the VRF scrubbing-vpn and sent to the Scrubbing Center.

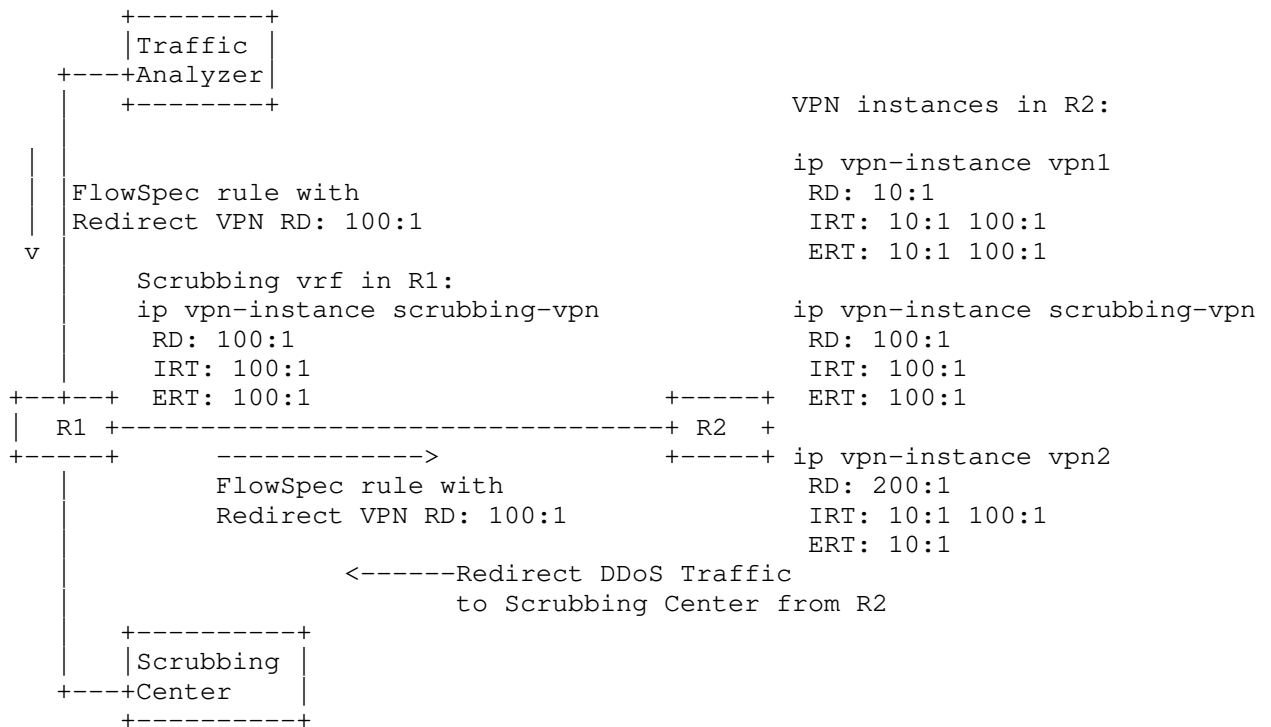


Figure 3: Redirect DDoS Traffic to Scrubbing Center Using Redirect VPN RD

The above procedures assume that all PEs are upgraded to support the Redirect to VPN RD Extended Community.

## 5. IANA Considerations

TBD.

## 6. Security Considerations

TBD.

## 7. References

### 7.1. 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, <<http://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, <<http://www.rfc-editor.org/info/rfc4271>>.
- [RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, DOI 10.17487/RFC4360, February 2006, <<http://www.rfc-editor.org/info/rfc4360>>.
- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <<http://www.rfc-editor.org/info/rfc4760>>.
- [RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, DOI 10.17487/RFC5492, February 2009, <<http://www.rfc-editor.org/info/rfc5492>>.
- [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, <<http://www.rfc-editor.org/info/rfc5575>>.
- [RFC5668] Rekhter, Y., Sangli, S., and D. Tappan, "4-Octet AS Specific BGP Extended Community", RFC 5668, DOI 10.17487/RFC5668, October 2009, <<http://www.rfc-editor.org/info/rfc5668>>.

## 7.2. Informative References

- [RFC7674] Haas, J., Ed., "Clarification of the Flowspec Redirect Extended Community", RFC 7674, DOI 10.17487/RFC7674, October 2015, <<http://www.rfc-editor.org/info/rfc7674>>.

## Authors' Addresses

lucy.yong  
Huawei Technologies  
  
Email: [lucy.yong@huawei.com](mailto:lucy.yong@huawei.com)

Shunwan Zhuang  
Huawei Technologies  
Huawei Bld., No.156 Beiqing Rd.  
Beijing 100095  
China

Email: zhuangshunwan@huawei.com

Weiguo Hao  
Huawei Technologies  
101 Software Avenue,  
Nanjing 210012  
China

Email: haoweiguo@huawei.com