Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 18, 2016

A. Sreekantiah
C. FilsFils
S. Previdi
S. Sivabalan
Cisco Systems
P. Mattes
S. Lin
Microsoft

October 16, 2015

Segment Routing Traffic Engineering Policy using BGP
draft-sreekantiah-idr-segment-routing-te-00

Abstract

This document describes mechanisms allowing advertising Segment Routing Traffic Engineering (SRTE) policies using BGP. Through the mechanisms described in this document, a BGP speaker has the ability to trigger, in a remote BGP node, the setup of a SR Encapsulation policy with specific characteristics and an explicit path. Steering mechanisms are also defined to enable the application of the policy on a per BGP route basis.

Table of Contents

1. Introduction

Segment Routing (SR) technology leverages the source routing and tunneling paradigms. [I-D.filsfils-rtgwg-segment-routing] provides an introduction to SR architecture. As defined in the SR architecture draft, "Node-SID" and "Adjacency-SID" denote Node Segment Identifier and Adjacency Segment Identifier respectively.

[I-D.sivabalan-pce-segment-routing] introduced the notion of a segment routed TE path which may may not follow IGP SPT. In this draft, we provide for the definition of SR Encapsulation Policy which can be used to provision such a segment routed TE path using BGP and the corresponding steering mechanism. SR Encapsulation Policy (referred to as SR policy in the rest of the document) is defined to be a weighted-ECMP set of segment lists that are added on the packets steered on the SR Encapsulation policy. Steering onto an SR Encapsulation Policy involves the classification of packets for encapsulation into the specified SR encapsulation policy

Border Gateway protocol (BGP) can also be used in order to propagate SR policy and corresponding steering associated with BGP routes. This document describes extensions to BGP in order to achieve this. This document describes the mechanisms through which a BGP speaker (which can be either a router or a controller) advertises an SR policy in the form of a weighted-ECMP set of segment lists that will result, in the node receiving the SR policy advertisement, in the instantiation of a segment routed TE path. Traffic steering onto the TE path is realized through the use of route coloring (based on BGP extended community Color attribute)

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. Segment Routing Encapsulation SAFI

A new Subsequent Address Family Identifier is defined, 'SR Encapsulation SAFI' (codepoint to be assigned by IANA). The new SAFI will encode SR policy parameters in order to instantiate the SR policy on the receiving BGP speakers. Specifically the NLRI associated with the SAFI will carry the ERO attribute, specifying a weighted set of explicit segment identifier lists (SID lists) representing the explicit SR TE path to the endpoint address encoded with the NLRI.

The NLRI, defined below, is carried in a BGP UPDATE message[RFC4271] using BGP multiprotocol extensions [RFC4760] with an AFI of 1 or 2 (IPv4 or IPv6) [IANA-AF] and a SAFI value to be assigned by the IANA (suggested value 80).

The NLRI is encoded in a format defined in Section 5 of [RFC4760] (a 2-tuple of the form (length, value)). The value field is structured as follows:

```
+-----------------------------------------------+
|           Identifier (Color Value)            |
+-----------------------------------------------+
|          Endpoint Address (Variable)          |
+-----------------------------------------------+
```

- Identifier: A 4-Octet identifier defined by the BGP speaker originating the NLRI, in order to uniquely identify the SR policy in the SR domain. The value of the Identifier field SHOULD be the value used in the color extended community (as defined in [RFC5512]) carried with payload prefixes the use of which is detailed in subsequent sections.

- Endpoint Address: This field identifies the endpoint of the SR TE path being encoded. It is one of the network addresses configured on the endpoint router of the SR TE path. The length of the endpoint address is dependent on the AFI being advertised. If the AFI is set to IPv4 (1), then the endpoint address is a 4-octet IPv4 address, whereas if the AFI is set to IPv6 (2), the endpoint address is a 16-octet IPv6 address.

An update message that carries the MP_REACH_NLRI or MP_UNREACH_NLRI attribute with the Encapsulation SAFI MUST also carry the BGP mandatory attributes: ORIGIN, AS_PATH, and LOCAL_PREF (for IBGP neighbors), as defined in [RFC4271]. In addition, such an update message can also contain any of the BGP optional attributes. The SAFI NLRI also encodes the color value in order to influence an action on the receiving speaker. Specifically the Color extended community SHOULD be propagated with BGP payload prefixes in order to associate with the NLRI of the SAFI prefixes with a given SR policy.

The nexthop address is set based on the AFI in the attribute. For example, if the AFI is set to IPv4 (1), the nexthop is encoded as a 4-byte IPv4 address. If the AFI is set to IPv6 (2), 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

Encapsulation NLRI, will process it only if the NLRI has a best path as per the BGP best path selection algorithm.

3. BGP SR Explicit Route Object (ERO) Attribute

3.1. New Attribute Definition

BGP SR ERO (Explicit Route Object) attribute is a new optional, transitive BGP path attribute. The attribute type code for BGP SR ERO attribute is to be assigned by IANA (suggested value 50).

The value field of the attribute is composed or one or more TLV objects. The following sections describe the SR-ERO, Weight and Binding TLVs.

3.2. ERO TLV Definition

The SR-ERO TLV encodes one segment of the SR policy path (loose or strict) in the form of a SID and its related information. Multiple occurrences of ERO TLV object can be encoded in a single attribute with the objects grouped into multiple sets for load-balancing purposes.

Each set defines a distinct SID list to be used in equal-cost or unequal-cost load-balancing. The SR-ERO TLV has the following format (encoding is based on ERO sub-object definition in [I-D.sivabalan-pce-segment-routing]):

```
 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                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   ST Type     |          Flags                 |I|L|N|F|S|C|M|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
//                           SID (32 bits or 128 bits)         //
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
//                         NAI (variable)                      //
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Each SR-ERO TLV object consists of a 64-bit header followed by the SID and the NAI associated with the SID. The SID is a 32-bit value (as specified in [I-D.sivabalan-pce-segment-routing]).The SID can also be a 128 bit value when encoding a IPv6 SID as indicated with the 'I' flag bit set, if this bit is unset, SID defaults to a 32 bit

value. The size of the NAI depends on its respective type, as described in the following sections.

"Type" contains the codepoint of the SR-ERO TLV object. The SR-ERO TLV codepoint is to be assigned by IANA (suggested value 1). In future other types maybe defined to encode characteristics (e.g bandwidth values) relevant to the segment list thus encoded.

"Length" contains the total length of the object in octets, excluding "Type" and "Length" fields. Length MUST be at least 4, and MUST be a multiple of 4. The length value should take into consideration SID or NAI only if they are not null. The flags described below used to indicate whether SID or NAI field is null.

SID Type (ST) indicates the type of the information associated with the SID contained in the object body. The SID-Type values are described later in this document

SID is the Segment Identifier.

NAI contains the Node or Adjacency Identifier associated with the SID. Depending on the value of ST, the NAI can have different formats as described in the following section.

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

* M: 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.

* C: When this bit as well as the M bit 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 its local policy and MPLS forwarding rules.

* S: When this bit is 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.

* F: When this bit is set, the NAI value in the object body is null.

* N: When this bit is set, it specifies the start of a new SID list. Multiple SID lists can be encoded in the ERO attribute signifying a equal-cost or unequal-cost set of multi-path SID lists to be used for load-balancing (leveraging the Weight TLV defined later in this document).

* L:(Loose flag). Indicates whether the encoding represents a loose-hop in the LSP [RFC3209]. If "L" is unset, a BGP speaker MUST NOT overwrite the SID value present in the SR-TLV object. Otherwise, a BGP speaker, based on local policy, MAY expand or replace one or more SID value(s) in the received SR-ERO attribute.

* I:(IPv6 SID flag). Indicates whether the SID encoding represents a 128 bit IPv6 address when set. When unset, the SID defaults to a 32 bit encoding

3.3. NAI Associated with SID

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

3.4. Weight TLV object type.

The Weight TLV specifies the weight associated to the SID list in case of unequal cost multipath.

The Weight 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 = 2                 |      Length                   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                             Weight                            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

"Type" is 2. Length is 4 octets (excluding Type and Length)

When present, the Weight TLV specifies a weight to be associated with the corresponding SID list, for use in unequal-cost multipath. Weights are applied by summing the total value of all of the weights for all SID lists, and then assigning a fraction of the forwarded traffic to each SID list in proportion its weight's fraction of the total."

The Weight TLV, if present, MUST be encoded prior to the encoding of the SR ERO TLV object with the "N" bit set, that is it MUST precede the encoding of a new SID list. The weight, when present, is thus associated with set of SIDs following the weight TLV and there MUST be only one weight TLV encoded for each set of SIDs (SID list) encoded in the attribute. Length is 4 indicating the number of octets used to encode the weight value.

3.5. Binding SID TLV object type.

The Binding TLV allows to request the allocation of a Binding Segment associated to the SID list carried in the SR-ERO TLVs. The Binding 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 = 3                 |      Length                   |
  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
  |                         Value (optional)                      |
  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

"Type" is 3. Value field is optional. Length is 0 or 4 (when the "Value" field is present). When the Length is 0, it implies the Value field is not present.

When present, the TLV instructs the receiver of the message to allocate a Binding SID for the set of SID lists that are encoded. The value field when present, encodes a 32 bit SID value. Also, when the value field is present, the TLV instructs the receiver of the message to use that value for the Binding SID allocated

Further use of the Binding SID is described in a subsequent section. There MUST be only one binding SID TLV encoded in the attribute.

4. Segment Routing Encapsulation SAFI operation

SR Encap SAFI NLRIs are advertised with the ERO attribute. The ERO attribute specifies one (or more for loadbalancing purposes) list of segment identifiers (SIDs), specifying the explicit SR TE path to the endpoint address encoded in the SR Encap SAFI NLRI. Additionally, the SR Encap SAFI NLRI encodes a Color value in order to associate payload prefixes with a SR Policy path definition. The BGP speaker SHOULD then attach a Color Extended Community (as defined in [RFC5512]) to payload prefixes (e.g., IPv4 unicast) in order to select the appropriate SR-TE path created by the SR Encap SAFI

update. Hence, the Color value in the Encap SAFI NLRI allows to implement a classification mechanism.

If a BGP speaker originates an update for prefix P with color C and with itself or a third party as the next hop, then it SHOULD also originate an SR Encap SAFI update containing a NLRI encoded with the prefix P's BGP nexthop (as the Endpoint address) and Color value matching the one of prefix P, and a list of SIDs defining the SR-TE Explicit path in the ERO attribute.

The SR Encap SAFI update in this case MAY also be sent by a controller, in lieu of the originating speaker sending the SAFI update with with the endpoint address set to the originating speaker in the SAFI NLRI. Also the controller can possibly color payload prefixes or originate payload prefixes with a color value in place of the originating BGP speaker.

On reception of a SR Encap SAFI update, a BGP speaker SHOULD initiate the creation of a SR TE explicit path with the Endpoint address in the NLRI as the destination endpoint and the explicit path specified by the lists of SIDs in the ERO TLVs in the attribute.

On the receiving BGP speaker, all payload prefixes that share the same color (as determined by the color extended community on the best path) and the same nexthop are mapped to the same SR-TE explicit path created upon the receipt of the SR Encap SAFI update with the matching color and endpoint addresses in the NLRI.

Similarly, different payload prefixes can be mapped to distinct SR-TE Explicit paths by coloring them differently.

5. Multipath Operation

The ERO attribute in the SR Encap SAFI update MAY contain multiple list of SIDs (instead of a single one) which, in the absence of the Weight TLV, signifies equal cost load-balancing amongst them.

When a weight TLV is encoded for each list, then the weight values SHOULD be used in order to perform a unequal cost load balance amongst the list of SIDs specified. Thus in the general case the ERO attribute in the SR Encap SAFI NLRI can identify a set of SR TE Explicit paths for load balancing operation.

6. Binding SID TLV

When the optional Binding SID TLV is present in the ERO attribute of a SR Encap SAFI update, it indicates an instruction, to the receiving

BGP speaker to allocate a Binding SID for the list of SIDs the Binding TLV is related to.

Any incoming packet with the Binding SID will then be swapped for the list of SIDs specified in the ERO attribute 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 so to, typically, feed a BGP controller with updated topology information.

7. Reception of a BGP ERO Attribute

When a BGP speaker receives a SR Encap SAFI NLRI from a neighbor with an acceptable BGP SR ERO attribute, it SHOULD compute the segment list and equivalent MPLS label stack from the attribute TLVs and program them in the MPLS data plane.

Also, It SHOULD program its MPLS dataplane so that BGP payload prefixes sharing the same Color Extended Community of the SR Encap SAFI NLRI are steered into the SR-Encap policy (defined by the SR Encap SAFI NLRI) instead of using the original BGP payload prefix nexthop label.

In the future, new flags in the ERO attribute TLV MAY be defined in order to support other label operations (such as replacing inner labels associated with the BGP prefix)

When building the MPLS label stack from a ERO attribute, the receiving BGP speaker MUST interpret the list of SR-ERO TLVs as follows.

The first TLV of a SR ERO attribute represents the topmost label. In the receiving BGP speaker, it identifies the first segment the traffic will be directed towards to (along the SR-TE path).

The last TLV of a SR ERO attribute represents the bottommost label.

8. Announcing BGP Segment Routing ERO Attribute

Typically, the value of the SIDs encoded in the ERO attribute TLV is determined by configuration.

A BGP speaker SHOULD follow normal iBGP/eBGP rules to propagate the ERO attribute

Since the BGP SR ERO attribute SID value must be unique within an SR domain, by default an implementation SHOULD NOT advertise the BGP SR ERO attribute outside an SR domain unless it is explicitly configured to do so. To contain distribution of the BGP SR ERO attribute beyond

its intended scope of applicability, attribute filtering MAY be deployed.

9. Flowspec and BGP SR ERO Attribute

The BGP SR ERO attribute can be carried in context of a Flowspec NLRI (RFC 5575). In this case, when the redirect to IP nexthop is specified as in draft-ietf-idr-flowspec-redirect-ip-02, the tunnel to the nexthop is specified by the segment list in the ERO attribute,. The Segment List (i.e.: label stack) is imposed to flows matching the criteria in the Flowspec route in order to steer them towards the nexthop as specified by the ERO attribute.

10. Deployment Considerations

11. Contributors

12. Acknowledgments

13. IANA Considerations

This document defines a new BGP attribute known as BGP SR ERO attribute. This document requests IANA to assign a new attribute code type for BGP SR ERO attribute from the BGP Path Attributes registry.

14. Security Considerations

There are no additional security risks introduced by this design.

15. References

15.1. 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>.

[RFC2858] Bates, T., Rekhter, Y., Chandra, R., and D. Katz, "Multiprotocol Extensions for BGP-4", RFC 2858, DOI 10.17487/RFC2858, June 2000, <http://www.rfc-editor.org/info/rfc2858>.

[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>.

[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>.

15.2. Informational References

[I-D.filsfils-rtgwg-segment-routing]
Filsfils, C., Previdi, S., Bashandy, A., Decraene, B., Litkowski, S., Horneffer, M., Milojevic, I., Shakir, R., Ytti, S., Henderickx, W., Tantsura, J., and E. Crabbe, "Segment Routing Architecture", draft-filsfils-rtgwg-segment-routing-01 (work in progress), October 2013.

[I-D.sivabalan-pce-segment-routing]
Sivabalan, S., Medved, J., Filsfils, C., Crabbe, E., Raszuk, R., Lopez, V., and J. Tantsura, "PCEP Extensions for Segment Routing", draft-sivabalan-pce-segment-routing-03 (work in progress), July 2014.

Authors' Addresses

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

Email: asreekan@cisco.com

Clarence FilsFils
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: cfilsfil@cisco.com

Stefano Previdi
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: sprevidi@cisco.com

Siva Sivabalan
Cisco Systems
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

Steven Lin
Microsoft
One Microsoft Way
Redmond , WA 98052
USA