IDR Working Group
W. Hao
S. Zhuang
Z. Li
Huawei
Internet Draft
Intended status: Standards Track

Expires: April 2016 October 19, 2015

Dissemination of Flow Specification Rules for NVO3
draft-hao-idr-flowspec-nvo3-02.txt

Abstract

This draft proposes a new subset of component types to support the NVO3 flow-spec application.

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction

BGP Flow-spec is an extension to BGP that allows for the dissemination of traffic flow specification rules. It leverages the BGP Control Plane to simplify the distribution of ACLs, new filter rules can be injected to all BGP peers simultaneously without changing router configuration. The typical application of BGP Flow-spec is to automate the distribution of traffic filter lists to routers for DDOS mitigation.

RFC5575 defines a new BGP Network Layer Reachability Information (NLRI) format used to distribute traffic flow specification rules. NLRI (AFI=1, SAFI=133)is for IPv4 unicast filtering. NLRI (AFI=1, SAFI=134)is for BGP/MPLS VPN filtering. [IPv6-FlowSpec] defines flow-spec extension for IPv6 data packets. [Layer2-FlowSpec] extends the flow-spec rules for layer 2 Ethernet packets.

In cloud computing era, multi-tenancy has become a core requirement for data centers. Since NVO3 can satisfy multi-tenancy key requirements, this technology is being deployed in an increasing number of cloud data center network. NVO3 focuses on the construction of overlay networks that operate over an IP (L3) underlay transport network. It can provide layer 2 bridging and layer 3 IP service for each tenant. VXLAN and NVGRE are two typical NVO3 encapsulations.

[EVPN-Overlays] provides a scalable and efficient multi-tenant solution within the Data Center where VXLAN, NVGRE or MPLS over GRE can be used as possible data plane encapsulation options. It uses EVPN as the control plane. [Inter-Overlays] provides a interconnect solution for EVPN overlay networks.

Both in data center inside or DCI networks, we also have requirements to deploy BGP Flow-spec for DDoS attack traffic mitigation. The Flow specification rules in NVO3 network can be based on inner layer 2 Ethernet header, inner layer 3 IP header, outer layer 2 Ethernet header, outer layer 3 IP header, and/or NVO3 header information. Currently the Flow specification rule [RFC5575] only includes single layer IP information like source/destination prefix, protocol, ports, and etc, the match part lacks layer indicator and NVO3 header information, so it can’t be used for the traffic filtering based on NVO3 header or a specified layer header directly.

This draft proposes a new subset of component types to support the NVO3 flow-spec application.

2. The Flow Specification encoding for NVO3

In default, the current flow-spec rules can only impose on the outer layer header of NVO3 encapsulation data packets. To make traffic filtering based on NVO3 header and inner header of NVO3 packets, a new component type acts as a delimiter is introduced. The delimiter type is used to specify the boundary of the inner or outer layer component types for NVO3 data packets. All the component types defined in [RFC5575],[IPv6-FlowSpec],[Layer2-FlowSpec],and etc can be used between two delimiters.

The NVO3 outer layer address normally belongs to public network, the "Flow Specification" NLRI only for the outer layer header doesn’t need to include Route Distinguisher field (8 bytes).

VNID is the identification for each tenant network, the "Flow Specification" NLRI for NVO3 header part should always include VNID field, Route Distinguisher field doesn’t need to be included.

The inner layer address normally belongs to a VPN, the NLRI format for the inner header should consist of a fixed-length Route Distinguisher field (8 bytes) corresponding to the VPN, the RD is followed by the flow specification for the inner layer. The NLRI length field shall include both the 8 bytes of the Route Distinguisher as well as the subsequent flow specification.

Flow specification rules received via this NLRI apply only to traffic that belongs to the VPN instance(s) in which it is imported.

This document proposes the following extended specifications for NVO3 flow:

Type TBD1 - Delimiter type

Encoding: <type (1 octet), length (1 octet), Value>.

When the delimiter type is present, it indicates the component types for the inner or outer layer of NVO3 packets will be followed immediately. At the same time, it indicates the end of the component types belonging to the former delimiter.

The value field defines encapsulation type and is encoded as:

```
  0   1   2   3   4   5   6   7
+---+---+---+---+---+---+---+---+
|          Encap Type           |
+---+---+---+---+---+---+---+---+
| I | O |        Resv           |
+---+---+---+---+---+---+---+---+
```

This document defines the following Encap types:

- VXLAN: Tunnel Type = 0

- NVGRE: Tunnel Type = 1

I: If I is set to one, it indicates the component types for the inner layer of NVO3 packets will be followed immediately.

O: If O is set to one, it indicates the component types for the outer layer of NVO3 packets will be followed immediately.

For NVO3 header part, the following additional component types are introduced.

Type TBD2 - VNID

Encoding: <type (1 octet), [op, value]+>.

Defines a list of {operation, value} pairs used to match 24-bit VN ID which is used as tenant identification in NVO3 network. For NVGRE encapsulation, the VNID is equivalent to VSID. Values are encoded as 1- to 3-byte quantities.

Type TBD3 - Flow ID

Encoding: <type (1 octet), [op, value]+>

Defines a list of {operation, value} pairs used to match 8-bit Flow id fields which are only useful for NVGRE encapsulation. Values are encoded as 1-byte quantity.

Other types:

The additional types for GENEVE [GENEVE],GUE [GUE] and GPE [GPE] header specific part will be added later.

3. The Flow Specification Traffic Actions for NVO3

The current traffic filtering actions can still be used for NVO3 encapsulation traffic. For Traffic Marking, only the DSCP in outer header can be modified.

4. Security Considerations

No new security issues are introduced to the BGP protocol by this specification.

5. IANA Considerations

IANA is requested to create and maintain a new registry entitled:

"Flow spec NVO3 Component Types":

Type TBD1 - Delimiter type

Type TBD2 - VNID

Type TBD3 - Flow ID

5.1. Normative References

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

[2] [GENEVE] J. Gross, T. Sridhar, etc, " Geneve: Generic Network Virtualization Encapsulation", draft-ietf-nvo3-geneve-00, May 2015.

[3] [GUE] T. Herbert, L. Yong, O. Zia, " Generic UDP Encapsulation", draft-ietf-nvo3-gue-01, Jun 2015.

[4] [GPE] P. Quinn,etc, " Generic Protocol Extension for VXLAN", draft-ietf-nvo3-vxlan-gpe-00, May 2015.

5.2. Informative References

[1] [EVPN-Overlays] A. Sajassi,etc, " A Network Virtualization Overlay Solution using EVPN", draft-ietf-bess-evpn-overlay-01 , work in progress, February, 2014.

[2] [Inter-Overlays] J. Rabadan,etc, " Interconnect Solution for EVPN Overlay networks", draft-ietf-bess-dci-evpn-overlay-01, work in progress, July, 2015.

[3] [RFC7348] M. Mahalingam, etc, "Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", RFC7348, August 2014.

[4] [NVGRE] P. Garg, etc, "NVGRE: Network Virtualization using Generic Routing Encapsulation", draft-sridharan-virtualization-nvgre-08, April 13, 2015.

[5] [IPv6-FlowSpec] R. Raszuk, etc, " Dissemination of Flow Specification Rules for IPv6", draft-ietf-idr-flow-spec-v6-06, November 2014.

[6] [Layer2-FlowSpec] W. Hao, etc, "Dissemination of Flow Specification Rules for L2 VPN", draft-ietf-idr-flowspec-l2vpn-02, August 2015.

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

6. Acknowledgments

The authors wish to acknowledge the important contributions of Susan Hares, Qiandeng Liang, Nan Wu, Yizhou Li, Lucy Yong.

Authors' Addresses

Weiguo Hao
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China
Email: haoweiguo@huawei.com

Shunwan Zhuang
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China
Email: zhuangshunwan@huawei.com

Zhenbin Li
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China
Email: lizhenbin@huawei.com

IDR Working Group W. Hao
S. Zhuang
Z. Li
Internet Draft Huawei
Intended status: Standards Track R.Gu
China Mobile
Expires: June 2016 December 18, 2015

Dissemination of Flow Specification Rules for NVO3
draft-hao-idr-flowspec-nvo3-03.txt

Abstract

This draft proposes a new subset of component types to support the NVO3 flow-spec application.

Table of Contents

1. Introduction

BGP Flow-spec is an extension to BGP that allows for the dissemination of traffic flow specification rules. It leverages the BGP Control Plane to simplify the distribution of ACLs, new filter rules can be injected to all BGP peers simultaneously without changing router configuration. The typical application of BGP Flow-spec is to automate the distribution of traffic filter lists to routers for DDOS mitigation.

RFC5575 defines a new BGP Network Layer Reachability Information (NLRI) format used to distribute traffic flow specification rules. NLRI (AFI=1, SAFI=133)is for IPv4 unicast filtering. NLRI (AFI=1, SAFI=134)is for BGP/MPLS VPN filtering. [IPv6-FlowSpec] and [Layer2-FlowSpec] extend the flow-spec rules for IPv6 and layer 2 Ethernet packets respectively. All these flow specifications match parts only reflect single layer IP/Ethernet information like source/destination MAC, source/destination IP prefix, protocol type, ports, and etc.

In cloud computing era, multi-tenancy has become a core requirement for data centers. Since NVO3 can satisfy multi-tenancy key requirements, this technology is being deployed in an increasing number of cloud data center network. NVO3 is an overlay technology, VXLAN and NVGRE are two typical NVO3 encapsulations. GENEVE [draft-ietf-nvo3-geneve-00],GUE[draft-ietf-nvo3-gue-01] and GPE [draft-ietf-nvo3-vxlan-gpe-00] are three emerging NVO3 encapsulations in progress.

```
                                +--+
                                |CE|
                                +--+
                                  |
                               +----+
                          +----| PE |----+
              +---------+ |    +----+    | +---------+
      +----+  |        +---+            +---+        |  +----+
      |NVE1|--|        |   |            |   |        |--|NVE3|
      +----+  |        |GW1|            |GW3|        |  +----+
              |        +---+            +---+        |
              |  NVO-1   |      MPLS      |   NVO-2  |
              |        +---+            +---+        |
              |        |   |            |   |        |
      +----+  |        |GW2|            |GW4|        |  +----+
      |NVE2|--|        +---+            +---+        |--|NVE4|
      +----+  +---------+ |              | +---------+  +----+
                          +--------------+
            Figure 1 NVO3 data center interconnection
```

The MPLS L2/L3 VPN in the WAN network can be used for NVO3 based data center network interconnection. When the DC and the WAN are operated by the same administrative entity, the Service Provider can decide to integrate the GW and WAN Edge PE functions in the same router for obvious CAPEX and OPEX saving reasons. This is illustrated in Figure 1. There are two interconnection solutions as follows:

1. End to end NVO3 tunnel across different data centers. NVE1 perform NVO3 encapsulation for DCI interconnection with NVE3, the destination VTEP IP is NVE3’s IP. The GW doesn’t perform NVO3 tunnel termination. The DCI WAN is pure underlay network.

2. Segmented NVO3 tunnels across different data centers. NVE1 doesn’t perform end to end NVO3 encapsulation to NVE3 for DCI interconnection. The GW performs NVO3 tunnel encapsulation termination, and then transmits the inner original traffic through MPLS network to peer data center GW. The peer data center GW

terminates MPLS encapsulation, and then performs NVO3 encapsulation to transmit the traffic to local NVE3.

In the first solution, to differentiate bandwidth and QOS among different tenants or applications, different TE tunnels in the WAN network will be used to carry the end to end NVO3 encapsulation traffic using VN ID, NVO3 outer header DSCP and etc as traffic classification match part. BGP Flow-spec protocol can be used to set the traffic classification on all GWs simultaneously.

In the second solution, a centralized BGP speaker can be deployed for DDOS mitigation in the WAN network. When the analyzer detects abnormal traffic, it will automatically generate Flow-spec rules and distribute it to each GW through BGP Flow-spec protocol, the match part should include inner or outer L2/L3 layer or NVO3 header.

In summary, the Flow specification match part on the GW/PE should include inner layer 2 Ethernet header, inner layer 3 IP header, outer layer 2 Ethernet header, outer layer 3 IP header, and/or NVO3 header information. Because the current match part lacks layer indicator and NVO3 header information, so it can’t be used directly for the traffic filtering based on NVO3 header or a specified layer header directly. This draft will propose a new subset of component types to support the NVO3 flow-spec application.

2. The Flow Specification encoding for NVO3

In default, the current flow-spec rules can only impose on the outer layer header of NVO3 encapsulation data packets. To make traffic filtering based on NVO3 header and inner header of NVO3 packets, a new component type acts as a delimiter is introduced. The delimiter type is used to specify the boundary of the inner or outer layer component types for NVO3 data packets. All the component types defined in [RFC5575],[IPv6-FlowSpec],[Layer2-FlowSpec],and etc can be used between two delimiters.

The NVO3 outer layer address normally belongs to public network, the "Flow Specification" NLRI only for the outer layer header doesn’t need to include Route Distinguisher field (8 bytes). If the outer layer address belongs to a VPN, the NLRI format for the outer header should consist of a fixed-length Route Distinguisher field (8 bytes) corresponding to the VPN, the RD is followed by the detail flow specifications for the outer layer.

VN ID is the identification for each tenant network, the "Flow Specification" NLRI for NVO3 header part should always include VN ID field, Route Distinguisher field doesn’t need to be included.

The inner layer MAC/IP address always associates with a VN ID, the NLRI format for the inner header should consist of a fixed-length VNID field (4 bytes), the VNID is followed by the detail flow specifications for the inner layer. The NLRI length field shall include both the 4 bytes of the VN ID as well as the subsequent flow specification. In NVO3 terminating into VPN scenario, if multiple access VN ID maps to one VPN instance, one share VN ID can be carried in the Flow-Spec rule to enforce the rule to entire VPN instance, the share VN ID and VPN correspondence should be configured on each VPN PE beforehand, the function of the layer3 VN ID is same with Route Distinguisher to act as the identification of VPN instance.

This document proposes the following extended specifications for NVO3 flow:

Type TBD1 - - Delimiter type

Encoding: <type (1 octet), length (1 octet), Value>.

When the delimiter type is present, it indicates the component types for the inner or outer layer of NVO3 packets will be followed immediately. At the same time, it indicates the end of the component types belonging to the former delimiter.

The value field defines encapsulation type and is encoded as:

```
  0   1   2   3   4   5   6   7
+---+---+---+---+---+---+---+---+
|          Encap Type           |
+---+---+---+---+---+---+---+---+
| I | O |        Resv           |
+---+---+---+---+---+---+---+---+
```

This document defines the following Encap types:

- VXLAN: Tunnel Type = 0

- NVGRE: Tunnel Type = 1

I: If I is set to one, it indicates the component types for the inner layer of NVO3 packets will be followed immediately.

O: If O is set to one, it indicates the component types for the outer layer of NVO3 packets will be followed immediately.

For NVO3 header part, the following additional component types are introduced.

Type TBD2 - VNID

Encoding: <type (1 octet), [op, value]+>.

Defines a list of {operation, value} pairs used to match 24-bit VN ID which is used as tenant identification in NVO3 network. For NVGRE encapsulation, the VNID is equivalent to VSID. Values are encoded as 1- to 3-byte quantities.

Type TBD3 - Flow ID

Encoding: <type (1 octet), [op, value]+>

Defines a list of {operation, value} pairs used to match 8-bit Flow id fields which are only useful for NVGRE encapsulation. Values are encoded as 1-byte quantity.

3. The Flow Specification Traffic Actions for NVO3

The current traffic filtering actions can still be used for NVO3 encapsulation traffic. For Traffic Marking, only the DSCP in outer header can be modified.

4. Security Considerations

No new security issues are introduced to the BGP protocol by this specification.

5. IANA Considerations

IANA is requested to create and maintain a new registry entitled:

"Flow spec NVO3 Component Types":

Type TBD1 - Delimiter type

Type TBD2 - VNID

Type TBD3 - Flow ID

5.1. Normative References

[1] [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate

Requirement Levels", BCP 14, RFC 2119, March 1997.

[2] [GENEVE] J. Gross, T. Sridhar, etc, " Geneve: Generic Network Virtualization Encapsulation", draft-ietf-nvo3-geneve-00, May 2015.

[3] [GUE] T. Herbert, L. Yong, O. Zia, " Generic UDP Encapsulation", draft-ietf-nvo3-gue-01, Jun 2015.

[4] [GPE] P. Quinn,etc, " Generic Protocol Extension for VXLAN", draft-ietf-nvo3-vxlan-gpe-00, May 2015.

5.2. Informative References

[1] [EVPN-Overlays] A. Sajassi,etc, " A Network Virtualization Overlay Solution using EVPN", draft-ietf-bess-evpn-overlay-01 , work in progress, February, 2014.

[2] [Inter-Overlays] J. Rabadan,etc, " Interconnect Solution for EVPN Overlay networks", draft-ietf-bess-dci-evpn-overlay-01, work in progress, July, 2015.

[3] [RFC7348] M. Mahalingam, etc, "Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", RFC7348, August 2014.

[4] [NVGRE] P. Garg, etc, "NVGRE: Network Virtualization using Generic Routing Encapsulation", draft-sridharan-virtualization-nvgre-08, April 13, 2015.

[5] [IPv6-FlowSpec] R. Raszuk, etc, " Dissemination of Flow Specification Rules for IPv6", draft-ietf-idr-flow-spec-v6-06, November 2014.

[6] [Layer2-FlowSpec] W. Hao, etc, "Dissemination of Flow Specification Rules for L2 VPN", draft-ietf-idr-flowspec-l2vpn-02, August 2015.

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

6. Acknowledgments

The authors wish to acknowledge the important contributions of Jeff Haas, Susan Hares, Qiandeng Liang, Nan Wu, Yizhou Li, Lucy Yong.

Authors' Addresses

Weiguo Hao
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China
Email: haoweiguo@huawei.com

Shunwan Zhuang
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China
Email: zhuangshunwan@huawei.com

Zhenbin Li
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China
Email: lizhenbin@huawei.com

Rong Gu
China Mobile
gurong_cmcc@outlook.com

W. Hao
Z. Li
Y. Lucy
Huawei

Internet Draft
Intended status: Standards Track

Expires: March 2016 October 6, 2015

BGP Flow-Spec Redirect to Tunnel action
draft-hao-idr-flowspec-redirect-tunnel-00.txt

Abstract

This draft defines a new flow-spec action, redirect-to-Tunnel, and a new sub-TLV for the redirect extended community to provide redirecting a flow to a tunnel. A BGP UPDATE for a flow-spec NLRI can contain the extended community. When activated, the corresponding flow packets will be encapsulated by a tunnel encapsulation protocol and then be forward to the target IP address. The redirected tunnel information and target IP address are encoded in BGP Path Attribute [TUNNELENCAPS] [MPP] that is carried in the BGP flow-spec UPDATE. The draft expends the tunnel encapsulation attribute to apply to flow-spec SAFI, i.e. 133 and 134.

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction

BGP Flow-spec is an extension to BGP that allows for the dissemination of traffic flow specification rules. It leverages the BGP Control Plane to simplify the distribution of ACLs, new filter rules can be injected to all BGP peers simultaneously without changing router configuration. The typical application of BGP Flow-spec is to automate the distribution of traffic filter lists to routers for DDOS mitigation.

Every flow-spec route consists of a matching part (encoded in the NLRI field) and an action part(encoded in one or more BGP extended communities). The flow-spec standard [RFC 5575] defines widely-used filter actions such as discard and rate limit; it also defines a redirect-to-VRF action for policy-based forwarding. [Redirect to IP] defines a new redirect-to-IP flow-spec action that provides a simpler method of policy-based forwarding. In some cases like service chaining, traffic steering and etc, the traffic needs to be redirected to tunnel directly. Using the redirect-to-VRF action or redirect-to-IP action for this will be complex and cumbersome.

This draft proposes a new redirect-to-tunnel flow-spec action that provides a straightforward solution for policy-based forwarding. The details of the redirected tunnel information are encoded in already existing defined BGP Path Attributes.

2. Redirect to Tunnel Extended Community

To support ''Redirect to Tunnel'', besides the extended communities in below per RFC5575, a new extended community of ''Redirect to Tunnel'' is defined by this draft. This redirect extended community allows the traffic to be redirected to a set of tunnel(s) that are specified by BGP Tunnel Encapsulation Attribute [TUNNELENCAPS] and/or BGP Extended Unicast Tunnel Attribute [MPP].

type	extended community	RFC or Draft
0x8006	traffic-rate	RFC5575
0x8007	traffic-action	RFC5575
0x8008	redirect	RFC5575
0x8009	traffic-marking	RFC5575
TBD	redirect to Tunnel	This draft

The new extended community for ''Redirect to Tunnel'' has a type indicating it is transitive and ''Redirect to Tunnel'' [to be assigned by IANA]. The sub-TLV has following format.

```
40  41  42  43  44  45  46  47
+---+---+---+---+---+---+---+---+
|       reserved            | C |
+---+---+---+---+---+---+---+---+
```

In this value field (6 bytes) 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. All bits other than the 'C' bit MUST be set to 0 by the originating BGP speaker and ignored by the receiving BGP speakers.

This draft extends BGP Tunnel Encapsulation Attribute to apply to BGP flow-spec SAFI, i.e., SAFI=133,134. When a tunnel is specified by BGP Tunnel Encapsulation Attribute, the tunnel type and encapsulation information such as VXLAN, NVGRE, VXLAN-GPE are encoded in the Tunnel Encapsulation Attribute Sub-TLVs. When applying it to flow-spec safi, the target IP address, IPv4 or IPv6 MUST be encoded in the Remote Endpiont Sub-TLV with the corresponding AFI. The AS number in the sub-TLV MUST be the number of the AS to which the target IP address in the sub-TLV belongs. If the redirect to tunnel end point is the BGP next hop, the AFI in the sub-TLV should be filled with zero, and the address in the sub-TLV should be omitted, and AS field should be filled with zero.

When a tunnel is specified by BGP Extended Unicast Tunnel Attribute [MPP], the tunnel type and encapsulation information such as RSVP-TE, LDP, Segment Routing Path are encoded in BGP Extended Unicast Tunnel Attributes ([MPP]).

The flow-spec UPDATE carries the ’’Redirect to Tunnel’’ extended community MUST have at least one BGP Path Attribute that specifies a set of tunnel(s) that the flow packets can be redirected to.

The following of this Section specifies a flow-spec to be redirect to the tunnel that is specified by BGP tunnel encapsulation attribute [TUNNELENCAPS]. A flow-spec to be redirected to a tunnel that is specified by the BGP extended unicast tunnel attribute will be addressed in future version.

When a BGP speaker receives a flow-spec route with a ’redirect to Tunnel’ 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 target IPv4 or IPv6 address encoded in Remote Endpoint sub-TLV of Tunnel Encapsulation Attribute. The BGP speaker is expected to do a longest-prefix-match lookup of the ’target address’ in its forwarding information base (FIB) and forward the tunneled redirected/copied packets based on the resulting route (the ’target route’). If the ’target address’ is invalid or unreachable then the extended community SHOULD be ignored.

If a BGP speaker receives a flow-spec route with one ’Redirect to Tunnel’ extended community and one BGP Tunnel Encapsulation Attribute that represents a set of tunnels to the same target address, and all of them are considered best and usable paths according to the BGP speaker’s multipath configuration, the BGP speaker SHOULD load-share the redirected packets across all the tunnels. If the BGP speaker is not capable of redirecting and copying the same packet it SHOULD ignore the extended communities with C=0. If the BGP speaker is not capable of redirecting/copying a packet towards multiple tunnels it SHOULD deterministically select one tunnel to the ’target address’ and ignore the others.

If a BGP speaker receives multiple flow-spec routes for the same flow-spec NLRI and all of them are considered best and usable paths according to the BGP speaker’s multipath configuration and each one carries one ’Redirect to Tunnel’ extended community and one Tunnel Encapsulation Attribute, the BGP speaker SHOULD load-share the tunneled redirected/copied packets across all the tunnels, with the same fallback rules as discussed in the previous paragraph. Note that this situation does not require the BGP speaker to have multiple peers - i.e. Add-Paths could be used for the flow-spec address family.

If a BGP speaker receives a flow-spec route with one ’Redirect to Tunnel’’ and one or more ’redirect to IP’ extended communities; local policy determines which ’redirect’ should be used.

If a BGP speaker receives a flow-spec route with one ’Redirect to Tunnel’’ and one or more ’redirect to VRF’ extended communities, and this route represents the one and only best path, the ’Redirect to Tunnel’ actions described above should be applied in the context of the ’target VRF’ matching the ’redirect to VRF’ extended community - i.e. the ’target addresses’ should be looked up in the FIB of the ’target VRF’. If there are multiple ’redirect to VRF’ extended communities in the route the ’target VRF’ SHOULD be the one that matches the ’redirect to VRF’ extended community with the highest numerical value. If the BGP speaker is not capable of ’redirect to VRF’ followed by ’Redirect to Tunnel’ then it SHOULD give preference to performing the ’redirect to VRF’ action and doing only longest-prefix-match forwarding in the ’target VRF’.

If a BGP speaker receives multiple flow-spec routes for the same flow-spec NLRI and all of them are considered best and usable paths according to the BGP speaker’s multipath configuration and they carry a combination of ’Redirect to Tunnel’ and ’redirect to VRF’ extended communities, the BGP speaker SHOULD apply the ’Redirect to Tunnel’ actions in the context of the ’target VRF’ as described above. Note that this situation does not require the BGP speaker to have multiple peers - i.e. Add-Paths could be used for the flow-spec address family.

The redirected/copied flow packets will be encapsulated first. The outer src address on the encapsulated packets MUST be filled with the IP address of the forwarding router; the outer dst address on the packets MUST be filled with the target IP address. If the flow has multiple tunnels that have the ’target address’ as remote tunnel endpoint, the redirected/copied packets MAY be encapsulated according to tunnel type and be load-shared across these tunnels according to the router’s ECMP configuration.

If the ’target route’ has one or more tunnel next-hops then, in turn, the tunneled redirect/copy packets SHOULD be encapsulated appropriately again.

2.1. Validation Procedures

The validation check described in [RFC 5575] and revised in [VALIDATE] SHOULD be applied by default to received flow-spec routes with a ’redirect to tunnel’ extended community, as it is to all types of flow-spec routes and the validation check described in [TUNNELENCAPS] SHOULD be applied to the tunnel encapsulation attribute. 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.

BGP speakers that support the extended communities defined in this draft MUST also, by default, enforce the following check when receiving a flow-spec route from an EBGP peer: if the received flow-spec route has a ’redirect to tunnel’ extended community with a ’target address’ X (in the remote endpoint sub-TLV) and the best matching route to X is not a BGP route with origin AS matching the peer AS then the extended community should be discarded and not propagated along with the flow-spec route to other peers. It MUST be possible to disable this additional validation check on a per-EBGP session basis.

3. Security Considerations

A system that originates a flow-spec route with a ’redirect to tunnel’ extended community can cause many receivers of the flow-spec route to send traffic to a single next-hop, overwhelming that next-hop and resulting in inadvertent or deliberate denial-of-service. This is particularly a concern when the ’redirect to tunnel’ extended community is allowed to cross AS boundaries. The validation check described in section 2.1 significantly reduces this risk.

4. IANA Considerations

IANA is requested to update the reference for the following assignment in the "BGP Extended Communities Type/sub-Type for ’Redirect to Tunnel’ that is specified in this draft.

4.1. Normative References

[1] [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate

Requirement Levels", BCP 14, RFC 2119, March 1997.

4.2. Informative References

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

[2] [Redirect to IP] J.Uttaro, etc, " BGP Flow-Spec Redirect to IP Action ", draft-ietf-idr-flowspec-redirect-ip-02, February 2015.

[3] [TUNNELENCAPS] E. Rosen, etc, " Using the BGP Tunnel Encapsulation Attribute without the BGP Encapsulation SAFI ", draft-rosen-idr-tunnel-encaps-00, June 2015.

[4] [MPP] Z. Li, etc, " BGP Extensions for Service-Oriented MPLS Path Programming (MPP) ", draft-li-idr-mpls-path-programming-01, March 2015.

5. Acknowledgments

The authors wish to acknowledge the important contributions of Shunwan Zhuang, Qiandeng Liang.

Authors' Addresses

Weiguo Hao
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China
Email: haoweiguo@huawei.com

Zhenbin Li
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China
Email: lizhenbin@huawei.com

Lucy Yong
Huawei Technologies
Phone: +1-918-808-1918
Email: lucy.yong@huawei.com

IDR Working Group W. Hao
Z. Li
L. Yong
Internet Draft Huawei
Intended status: Standards Track

Expires: September 2016 March 18, 2016

BGP Flow-Spec Redirect to Tunnel Action
draft-hao-idr-flowspec-redirect-tunnel-01.txt

Abstract

This draft defines a new flow-spec action, Redirect-to-Tunnel, and a new sub-TLV for Redirect-to-Tunnel extended community. A BGP UPDATE for a flow-spec NLRI can contain the extended community. When activated, the corresponding flow packets will be encapsulated and carried via a tunnel. The redirect tunnel information is encoded in BGP Path Attribute or extended community [TUNNELENCAPS][MPP] that is carried in the BGP flow-spec UPDATE. The draft expends the tunnel encapsulation attribute [TUNNELENCAPS] to apply to flow-spec SAFI, i.e., 133 and 134.

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction

BGP Flow-spec is an extension to BGP that allows for the dissemination of traffic flow specification rules. It leverages the BGP Control Plane to simplify the distribution of ACLs, new filter rules can be injected to all BGP peers simultaneously without changing router configuration. The typical application of BGP Flow-spec is to automate the distribution of traffic filter lists to routers for DDOS mitigation.

Every flow-spec route consists of a matching part (encoded in the NLRI field) and an action part(encoded in one or more BGP extended communities). The flow-spec standard [RFC 5575] defines widely-used filter actions such as discard and rate limit; it also defines a redirect-to-VRF action for policy-based forwarding. [Redirect to IP] defines a new redirect-to-IP flow-spec action that provides a simpler method of policy-based forwarding. In some cases like service chaining, traffic steering and etc, the traffic needs to be redirected to a tunnel directly. Redirect-to-VRF action or redirect-to-IP action can’t service this purpose. .

This draft proposes a new redirect-to-tunnel flow-spec action that provides a straightforward policy-based forwarding. The details of the redirect tunnel information are encoded in BGP Path Attributes or extended communities.

2. Redirect-to-Tunnel Extended Community

To support Redirect-to-Tunnel action, besides the extended communities in below per RFC5575, a Redirect-to-Tunnel extended community is defined by this draft. This extended community conveys redirecting tunnel action; the tunnel information is specified in BGP Tunnel Encapsulation Attribute [TUNNELENCAPS] and/or BGP Extended Unicast Tunnel Attribute [MPP].

type	extended community	RFC or Draft
0x8006	traffic-rate	RFC5575
0x8007	traffic-action	RFC5575
0x8008	redirect	RFC5575
0x8009	traffic-marking	RFC5575
TBD	redirect-to-tunnel	This draft

The Redirect-to-Tunnel extended community has a type indicating it is transitive and Redirect-to-Tunnel [to be assigned by IANA]. The sub-TLV has following format.

```
40  41  42  43  44  45  46  47
+---+---+---+---+---+---+---+---+
|        reserved          | C |
+---+---+---+---+---+---+---+---+
```

In this value field (6 bytes) 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. All bits other than the ’C’ bit MUST be set to 0 by the originating BGP speaker and ignored by the receiving BGP speakers.

This draft extends BGP Tunnel Encapsulation Attribute to apply to BGP flow-spec SAFI, i.e., SAFI=133,134. When a tunnel is specified by BGP Tunnel Encapsulation Attribute [TUNNELENCAPs], the tunnel type and encapsulation information such as VXLAN, NVGRE, VXLAN-GPE are encoded in the Tunnel Encapsulation Attribute Sub-TLVs. When applying it to flow-spec safi, the target IP address, IPv4 or IPv6 MUST be encoded in the Remote Endpiont Sub-TLV with the corresponding AFI. The AS number in the sub-TLV MUST be the number of the AS to which the target IP address in the sub-TLV belongs. If the redirect to tunnel end point is the BGP next hop, the AFI in the sub-TLV should be filled with zero, and the address in the sub-TLV should be omitted, and AS field should be filled with zero.

When a tunnel is specified by BGP Extended Unicast Tunnel Attribute [MPP], the tunnel type such as RSVP-TE, LDP, Segment Routing Path and encapsulation information are encoded in BGP Extended Unicast Tunnel Attributes (See section 5.1 of [MPP]). Note that BGP Extended Unicast Tunnel Attribute is used in Centralized Controller Environment [MPP].

The flow-spec UPDATE carries the Redirect-to-Tunnel extended community MUST have at least one BGP Path Attribute that specifies a set of tunnel(s) that the flow packets can be redirected to.

When a BGP speaker receives a flow-spec route with a Redirect-to-Tunnel extended community and BGP tunnel encapsulation attribute [TUNNELENCSPS], if this route represents the one and only best path, it installs a traffic filtering rule that matches the packets described by the NLRI field; the packets matching the rules will be redirected (C=0) or copied (C=1) via the IP tunnel with remote endpoint address encoded in Remote Endpoint sub-TLV of Tunnel Encapsulation Attribute. If the ’target address’ is invalid or unreachable then the extended community and the tunnel attribute SHOULD be ignored.

When a BGP speaker receives a flow-spec route with a Redirect-to-Tunnel extended community and extended unicast tunnel attribute, it installs traffic filtering rules that matches the packets described by the NLRI field and the tunnel info. If BGP speaker can’t resolve the tunnel locally according to the unicast tunnel attribute, then the extended community and the tunnel attribute SHOULD be ignored.

If a BGP speaker receives a flow-spec route with one Redirect-to-Tunnel extended community and one BGP Tunnel Encapsulation Attribute that represents a set of tunnels to the same target address, and all of them are considered best and usable paths according to the BGP speaker’s multipath configuration, the BGP speaker SHOULD load-share the redirected packets across all the tunnels. If the BGP speaker is not capable of redirecting and copying the same packet it SHOULD ignore the extended community with C=0. If the BGP speaker is not capable of redirecting/copying a packet towards multiple tunnels it SHOULD deterministically select one tunnel to the 'target address' and ignore the others.

If a BGP speaker receives multiple flow-spec routes for the same flow-spec NLRI and all of them are considered best and usable paths according to the BGP speaker's multipath configuration and each one carries one Redirect-to-Tunnel extended community and one Tunnel Encapsulation Attribute, the BGP speaker SHOULD load-share the tunneled redirected/copied packets across all the tunnels, with the same fallback rules as discussed in the previous paragraph. Note that this situation does not require the BGP speaker to have multiple peers - i.e. Add-Paths could be used for the flow-spec address family.

If a BGP speaker receives a flow-spec route with one Redirect-to-Tunnel and one 'redirect to VRF' extended community, and this route represents the one and only best path, the Redirect-to-Tunnel actions described above should be applied in the context of the 'target VRF' matching the 'redirect to VRF' extended community, i.e. the 'target addresses' should be looked up in the FIB of the 'target VRF'. If the BGP speaker is not capable of 'redirect to VRF' followed by Redirect-to-Tunnel then it SHOULD give preference to performing the 'redirect to VRF' action and doing only longest-prefix-match forwarding in the 'target VRF'.

If a BGP speaker receives multiple flow-spec routes for the same flow-spec NLRI and all of them are considered best and usable paths according to the BGP speaker's multipath configuration and they carry a combination of Redirect-to-Tunnel and 'redirect to VRF' extended communities, the BGP speaker SHOULD apply the Redirect-to-Tunnel actions in the context of the 'target VRF' as described above. Note that this situation does not require the BGP speaker to have multiple peers - i.e. Add-Paths could be used for the flow-spec address family.

The redirected/copied flow packets will be encapsulated first. The outer src address on the encapsulated packets MUST be filled with the IP address of the forwarding router; the outer dst address on the packets MUST be filled with the target IP address. If the flow has multiple tunnels that have the 'target address' as remote tunnel endpoint, the redirected/copied packets MAY be encapsulated according to tunnel type and be load-shared across these tunnels according to the router's ECMP configuration.

If the 'target route' has one or more tunnel next-hops then, in turn, the tunneled redirect/copy packets SHOULD be encapsulated appropriately again.

3. Usage Rules for Redirect-to-Tunnel Action

3.1. Matching Filters for Redirect Tunnel Action

Redirect-to-Tunnel action can apply to different types of flow spec rules described in the NLRI field. Here are the types of flow spec rules that can have the Redirect-to-Tunnel action. Applicability for Other types of flow spec rules are for further study.

o IPv4 or IPv6

o L3VPN

o L2VPN

o NVO3

3.2. Other Actions Considerations

Flow spec rules in a NLRI can associate with one or more actions that are specified in extended communities, which means that flow spec packets will subject to a sequence of actions.[COMBO] specified default ordering precedence of actions. However some actions do not make a sense to be used with Redirect-to-Tunnel action, i.e. they have to be used in mutually exclusive.

In general Redirect-to-Tunnel action can work with traffic rate in Bite, traffic rate in packet, traffic action, redirect to VRF, interface set, time actions. The use cases for Redirect-to-Tunnel action to work with other actions are for further study. Note that: the two actions that can be used with Redirect-to-Tunnel action may be in mutually exclusive usage.

Memo: need a standard way to document these rules for a flow spec action.

3.3. Validation Procedures

The validation check described in [RFC 5575] and revised in [VALIDATE] SHOULD be applied by default to received flow-spec routes with a Redirect-to-Tunnel extended community, as it is to all types of flow-spec routes and the validation check described in

[TUNNELENCAPS] SHOULD be applied to the tunnel encapsulation attribute. 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.

BGP speakers that support the extended community defined in this draft MUST also, by default, enforce the following check when receiving a flow-spec route from an EBGP peer: if the received flow-spec route has a Redirect-to-Tunnel extended community with a 'target address' X (in the remote endpoint sub-TLV) and the best matching route to X is not a BGP route with origin AS matching the peer AS then the extended community should be discarded and not propagated along with the flow-spec route to other peers. It MUST be possible to disable this additional validation check on a per-EBGP session basis.

4. Security Considerations

A system that originates a flow-spec route with a 'redirect to tunnel' extended community can cause many receivers of the flow-spec route to send traffic to a single next-hop, overwhelming that next-hop and resulting in inadvertent or deliberate denial-of-service. This is particularly a concern when the 'redirect to tunnel' extended community is allowed to cross AS boundaries. The validation check described in section 2.1 significantly reduces this risk.

5. IANA Considerations

IANA is requested to update the reference for the following assignment in the "BGP Extended Communities Type/sub-Type for Redirect-to-Tunnel that is specified in this draft.

5.1. Normative References

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

[TUNNELENCAPS] E. Rosen, et al, "Using the BGP Tunnel Encapsulation Attribute without the BGP Encapsulation SAFI", draft-rosen-idr-tunnel-encaps-00, June 2015.

[MPP] Z. Li, et al, "BGP Extensions for Service-Oriented MPLS Path Programming (MPP) ", draft-li-idr-mpls-path-programming-01, March 2015.

[COMBO] S. Hares, ’’An Information Model for Basic Network Policy and Filter Rules’’, draft-hares-ide-flowspec-combo-01, March 2016.

5.2. Informative References

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

[Redirect to IP] J.Uttaro, et al, "BGP Flow-Spec Redirect to IP Action ", draft-ietf-idr-flowspec-redirect-ip-02, February 2015.

6. Acknowledgments

The authors wish to acknowledge the important contributions of Shunwan Zhuang, Qiandeng Liang.

Authors' Addresses

Weiguo Hao
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China
Email: haoweiguo@huawei.com

Zhenbin Li
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China
Email: lizhenbin@huawei.com

Lucy Yong
Huawei Technologies
Phone: +1-918-808-1918
Email: lucy.yong@huawei.com

Network Working Group K. Patel
Internet-Draft Cisco Systems
Intended status: Standards Track J. Uttaro
Expires: September 16, 2016 ATT
B. Decraene
Orange
W. Henderickx
Alcatel Lucent
J. Haas
Juniper Networks
March 15, 2016

Constrain Attribute announcement within BGP
draft-keyupate-idr-bgp-attribute-announcement-01.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 attribute extensions that help confine the scope of Optional attributes within a given AS or a given BGP member-AS in Confederation

Table of Contents

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 extensions 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 thoughout this document.

As part of new attribute extensions, this draft defines a new attribute format to incorporate the scoping information. The new attribute format applies to all the new attribute types that will be defined moving forward. 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-transtitive attributes. For any well-known attributes or unrecognized optional non-transtive 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 Format

[RFC4271] defines path attribute format as a triple [attribute type, attribute length, attribute value] of a variable length. The attribute value field is of a variable length. This draft augments the path attribute value field and reserves first four bytes of path attribute value field as path attribute extended flags field. All the path attributes carrying extended flags field will have a minumum attribute length of 4 bytes. The augmented path attribute format applies to all the current undefined attributes types (30-39, 41-127, 129-254). Any attribute specific data follows the path attribute extended flags field.

3. Extended Path Attribute Flags

[RFC4271] defines four type of BGP Path attributes using the attribute Flags field as follows:

Path Attribute flags:

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

O   Optional or a Well-known as defined in [RFC4271]

T   Transitive or Non-Transtive as defined in [RFC4271]

P   Partial as defined in [RFC4271]

E   Extended Length type as defined in [RFC4271]

This draft introduces new Flags field known as Extended Path Attribute Flags. The Extended Path Attribute Flags field is defined as first 4 bytes of the Path Attribute’s value field. This draft introduces three new Extended Path Attributes flags as follows:

Path Attribute flags:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 0 1 2 3 4 5 6 7 0 1 2 3 4 5 6 7 0 1 2 3 4 5 6 7
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                         R                                 |C|A|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

(R = MUST Be Zero)

A   AS Wide Scope

C   Member-AS in Confederation Scope

M   Multi-AS Scope

The first least 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 defined in BGP Path Attribute Flag field MUST be set. Otherwise a BGP speaker MUST consider an attribute as an error and malformed.

The second least 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 defined in BGP Path Attribute Flag field 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 first and the second most least significant bit together is defined as the Multiple AS Scope within a Single Administration. When both the first and the second 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.

4. Operation

When originating a well-known Path attribute, a BGP speaker MUST set both the AS Wide Scope and Member-AS Scope bit to 0. 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 any well-known Path attribute with either a AS Wide Scope bit set or a Member-AS Scope bit set then it SHOULD consider the attribute as malformed. The handling of malformed attributes SHOULD follow the procedures mentioned in [RFC7606].

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 admistrative domain. The BGP peering boundaries for an admistrative 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].

5. IANA Considerations

This draft define a new path attribute format for all undefined attribute types. We request IANA to record the use of new path attribute format for the following undefined attribute types (30-39, 41-127, 129-254).

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

Under "Border Gateway Protocol (BGP) Parameters" registry, "BGP Extended Path Attributes Flags" Reference: draft-keyupate-idr-bgp-attribute-annoucement-01 Registration Procedures as follows:

Bit Value (LSB)	Type	Reference
1	AS Wide Scope	Current Draft
2	Member-AS in Confederation	Current Draft

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 John Scudder, Jakob Heitz, Shyam Seturam, Juan Alcaide and Acee Lindem for the review and comments.

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>.

[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>.

7.2. 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

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

Email: uttaro@att.com

Bruno Decraene
Orange

Email: bruno.decraene@orange.com

Wim Henderickx
Alcatel Lucent
Copernicuslaan 50
Antwerp 2018
Belgium

Email: wim.henderickx@alcatel-lucent.com

Jeff Haas
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
USA

Email: jhaas@juniper.net

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

Z. Li
Huawei
L. Ou
Y. Luo
China Telcom Co., Ltd.
S. Lu
Tencent
S. Zhuang
N. Wu
Huawei

October 17, 2015

BGP FlowSpec Extensions for Routing Policy Distribution (RPD)
draft-li-idr-flowspec-rpd-01

Abstract

This document describes a mechanism to use BGP Flowspec address family as routing-policy distribution protocol. This mechanism is called BGP FlowSpec Extensions for Routing Policy Distribution (BGP-FS RPD).

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].

Copyright Notice

Table of Contents

1. Introduction

Some difficulties exist when optimize traffic paths on a traditional IP network:

- o Traffic can only be adjusted device by device. All routers that the traffic traverses need to be configured. The configuration workload is heavy. The operation is not only time consuming but also prone to misconfiguration for Service Providers.

- o The routing policies used to control network routes are complex, posing difficulties to subsequent maintenance, high maintenance skills are required.

Hence, an automatic mechanism for setting up routing policies is desirable which can simplify the complexity of routing policies configuration. This document describes a mechanism to use BGP Flowspec address family [RFC5575] as route-policy distribution protocol. This mechanism is called BGP FlowSpec Extensions for Routing Policy Distribution (BGP-FS RPD).

2. Definitions and Acronyms

BGP Flow Specification route: BGP Flow Specification routes are defined in RFC 5575. Each BGP Flow Specification route contains BGP Network Layer Reachability Information (NLRI) and Extended Community Attributes, which carry traffic filtering rules and actions to be taken on filtered traffic.

BGP Flow Specification peer relationship: A BGP Flow Specification peer relationship is established between the device that generates BGP Flow Specification routes and each network ingress that will transmit the BGP Flow Specification routes. After receiving the BGP Flow Specification routes, the peer delivers preferred BGP Flow Specification routes to the forwarding plane. The routes are then converted into traffic policies that control attack traffic.

- o ACL:Access Control List

- o BGP: Border Gateway Protocol

- o FS: Flow Specification

- o PBR:Policy-Based Routing

- o RPD: Routing Policy Distribution

- o VPN: Virtual Private Network

3. Problem Statements

It is obvious that providers have the requirements to adjust their business traffic from time to time because:

o Business development or network failure introduces link congestion and overload.

o Network transmission quality decreased as the result of delay, loss and need to adjust traffic to other paths.

o To control OPEX and CPEX, prefer the transit provider with lower price.

3.1. Inbound Traffic Control

In the scenario below, for reasons above, the provider of AS100 saying P may wish the inbound traffic from AS200 enters AS100 through link L3 instead of others. Since P doesn't have administration over AS200, so there is no way for P to modify the route selection critera directly.

```
                  Traffic from PE1 to Prefix1
             ----------------------------------->

   +-----------------+           +-------------------------+
   |     +---------+ |        L1 | +-----+     +-----------+|
   |     |Speaker1 | +-----------+ |IGW1|      |policy     ||
   |     +---------+ |**      L2**| +-----+    |controller||
   |                 |  **    **  |            +-----------+|
   | +---+           |    ****    |                         |
   | |PE1|           |    ****    |                         |
   | +---+           |  **    **  |                         |
   |     +---------+ |**      L3**| +-----+                 |
   |     |Speaker2 | +-----------+ |IGW2|      AS100        |
   |     +---------+ |        L4  | +-----+                 |
   |                 |            |                         |
   |    AS200        |            |                         |
   |                 |            |  ...                    |
   |                 |            |                         |
   |     +---------+ |            | +-----+    +--------+   |
   |     |Speakern | |            | |IGWn|     |Prefix1|    |
   |     +---------+ |            | +-----+    +--------+   |
   +-----------------+           +-------------------------+

               Prefix advertise from AS100 to AS200
             <-----------------------------------------
           Figure : Inbound Traffic Control case
```

3.2. Outbound Traffic Control

In this scenario, the provider of AS100 saying P wishes to prefer link L3 for the traffic to the destination Prefix2 among multiple exits and links. This preference can be dynamic and change frequently because of the reasons above. So the provider P expects an efficient and convenient solution.

```
                  Traffic from PE2 to Prefix2
             ------------------------------------>

   +-------------------------+            +-----------------+
   |+----------+      +----+ |L1          | +---------+     |
   ||policy    |      |IGW1| +------------+ |Speaker1 |     |
   ||controller|      +----+ |**        **| +---------+     |
   |+----------+             |L2**    **  |        +-------+|
   |                         |    ****    |        |Prefix2||
   |                         |    ****    |        +-------+|
   |                         |L3**    **  |                 |
   |      AS100       +----+ |**        **| +---------+     |
   |                  |IGW2| +------------+ |Speaker2 |     |
   |                  +----+ |L4          | +---------+     |
   |                         |            |                 |
   |+---+                    |            |    AS200        |
   ||PE2|              ...   |            |                 |
   |+---+                    |            |                 |
   |                  +----+ |            | +---------+     |
   |                  |IGWn| |            | |Speakern |     |
   |                  +----+ |            | +---------+     |
   +-------------------------+            +-----------------+

               Prefix advertise from AS200 to AS100
             <-----------------------------------------
```

4. Proposed Solution

BGP FlowSpec [RFC5575] leverages the BGP control plane to simplify the distribution of filter rules. New filter rules can be injected to all BGP peers simultaneously without changing router configuration. Though the typical application of is for DDOS mitigation, it doesn’t mean BGP Flowspec only takes effect on the forwarding plane.

This document introduces a mechanism that uses BGP Flowspec as a route-policy distribution protocol. It can be the same powerful as the device-based route-policy while still has the efficiency and convenience of BGP Flowspec.

This draft will use the term BGP-FS RPD as the abbreviation of FlowSpec Extensions for Routing Policy Distribution.

5. Protocol Extensions

5.1. FlowSpec Traffic Actions for Routing Policy Distribution

The traffic-action extended community consists of 6 bytes of which only the 2 least significant bits of the 6th byte (from left to right) are currently defined in [RFC5575]. Terminal Action (bit 47) and Sample (bit 46) defines in [RFC5575], this document defines Route Policy Distribution Flag(Bit 45).

The Flow Specification Traffic Actions for Routing Policy Distribution:

```
  40  41  42  43  44  45  46  47
+---+---+---+---+---+---+---+---+
| reserved          | R | S | T |
+---+---+---+---+---+---+---+---+
```

Figure 1: FlowSpec Traffic-action

Route Policy Distribution Flag(Bit 45): When this bit is set, the corresponding filtering rules will be used as Route Policy.

5.2. BGP Policy Attribute

This document defines and uses a new BGP attribute called the "BGP Policy attribute". This is an optional BGP attribute. The format of this attribute is defined as follows:

```
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                               |
|   Match fields (Variable)     |
|                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                               |
|   Action fields (Variable)    |
|                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2: BGP Policy Attribute

Match fields: Match Fields define the matching criteria for the BGP Policy Attribute.

Action fields: Action fields define the action being applied to the target route.

5.2.1. Match Fields Format

Match Fields define the matching criteria for the BGP Policy Attribute.

```
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   Match Type (2 octets)         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Number of Sub-TLVs (2 octets)   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                 |
|   Sub-TLVs (Variable)           |
|                                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 3: Match Fields Format

Match Type:

0: Permit, specifies the permit mode of a match rule. If a route matches the matching criteria of the BGP Policy Attribute, the actions defined by the Action fields of the BGP Policy Attribute are performed. If a route does not match the matching criteria for the BGP Policy Attribute, then nothing needs to do with this route.

1: Deny, specifies the deny mode of a match rule. In the deny mode, If a route does not match the matching criteria of the BGP Policy Attribute, the actions defined by the Action fields of the BGP Policy Attribute are performed. If a route matches the matching criteria of the BGP Policy Attribute, then nothing needs to do with this route.

Number of Sub-TLVs: The number of Sub-TLVs contain in Match fields.

The contents of Match fields are encoded as Sub-TLVs, where each TLV has the following format:

```
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|       Type (2 octets)           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|       Length (2 octets)         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                 |
|       Values (Variable)         |
|                                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 4: Sub-TLVs Format

Type: The Type field contains a value of 1-65534. The values 0 and 65535 are reserved for future use.

Length: The Length field represents the total length of a given TLV’s value field in octets.

Values: The Value field contains the TLV value.

Supported format of the TLVs can be:

Type 1: IPv4 Neighbor

Type 2: IPv6 Neighbor

Type 3: ASN List

...

To be added in later versions.

5.2.2. Action Fields Format

Action fields define the action being applied to the targeted route.

```
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   Action Type (2 octets)        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   Action Length (2 octets)      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                 |
|   Action Values (Variable)      |
|                                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 5: Action Fields Format

Action Type: The Action Type field contains a value of 1-65534. The values 0 and 65535 are reserved for future use.

Action Length: The Action Length field represents the total length of the Action Values in octets.

Action Values: The Action Values field contain parameters of the action.

Supported format of the TLVs can be:

Type 1: Route-Preference

Type 2: Route-Prepend-AS

...

To be added in later versions.

5.2.3. Operation Examples

5.2.3.1. Inbound Traffic Control

The traffic destined for Prefix1 needs to be scheduled to link Speaker1 -> IGW2 for transmission.

The Policy Controller constructs a BGP-FS RPD route and pushes it to all the IGW routers, the route carries:

1. Prefix1 in the Destination Prefix component of the BGP-FS NLRI;

2. Flow Specification Traffic Action Extended Community with the Route Policy Distribution Flag(Bit 45) set. When this bit is set, the corresponding filtering rules will be used as Routing Policies.

3. BGP Policy Attribute:

   * Match Type: 2, Deny

   * IPv4 Neighbor Sub-TLV: Local BGP Speaker IGW2, Remote BGP Peer Speaker1

   * Action Type: Route-Prepend-AS

   * Action Value: Prepend-AS times is 5

IGW1 processes the received BGP-FS RPD route as follows:

1. IGW1 gets the target prefix Prefix1 from the Destination Prefix component in the BGP FS NLRI of the BGP FS RPD route;

2. IGW1 identifies the Route Policy Distribution Flag carrying in the Flow Specification Traffic Action Extended Community, then IGW1 knows that the corresponding filtering rules will be used as Routing Policies.

3. IGW1 uses the target prefix Prefix1 to choose the matching routes, in this case, IGW1 will choose the current best route of Prefix1;

4. IGW1 gets the matching criteria from the BGP Policy Attribute: Local BGP Speaker IGW2, Remote BGP Speaker1;

5. IGW1 gets the action from the BGP Policy Attribute: Route-Prepend-AS, 5 times;

IGW1 checks the matching criteria and finds that it doesn’t hits the matching criteria: Local BGP Speaker IGW2, Remote BGP Speaker1, at the same time the Match Type is "Deny" mode, so IGW1 sends the best route of Prefix1 to Speaker1 with performing the Action instructions from the BGP-FS RPD route: Prepend Local AS 5 times.

IGW2 processes the received BGP FS RPD route as follows:

1. IGW2 gets the target prefix Prefix1 from the Destination Prefix component in the BGP-FS NLRI of the BGP FS RPD route;

2. IGW2 identifies the Route Policy Distribution Flag carrying in the Flow Specification Traffic Action Extended Community, then IGW2 knows that the corresponding filtering rules will be used as Routing Policies.

3. IGW2 uses the target prefix Prefix1 to choose the matching routes, in this case, IGW2 will choose the current best route of Prefix1;

4. IGW2 gets the matching criteria from the BGP Policy Attribute: Local BGP Speaker IGW2, Remote BGP Speaker1;

5. IGW2 gets the action from the BGP Policy Attribute: Route-Prepend-AS, 5 times;

IGW2 checks the matching criteria and finds that it hits the matching criteria: Local BGP Speaker IGW2, Remote BGP Peer Speaker1, but the Match Type is "Deny" mode, so IGW2 sends the best route of Prefix1 to Speaker1, without performing the Action instructions from the BGP-FS RPD route.

In the similar manner, other IGWs will perform the same Action instructions as IGW1. Then the traffic destined for Prefix1 has been be scheduled to link L3 for transmission.

5.2.3.2. Outbound Traffic Control

In this scenario, if the bandwidth usage of a link exceeds the specified threshold, the Policy Controller automatically identifies which traffic needs to be scheduled and the Policy Controller automatically calculates traffic control paths based on network topology and traffic information.

For example, the outbound traffic destined for Prefix2 needs to be scheduled to link IGW2 -> Speaker1 for transmission.

The Policy Controller sends a BGP-FS RPD route to IGW2, the route carries:

1. Prefix2 in the Destination Prefix component of the BGP-FS NLRI;

2. Flow Specification Traffic Action Extended Community with the Route Policy Distribution Flag(Bit 45) set. When this bit is set, the corresponding filtering rules will be used as Routing Policies.

3. BGP Policy Attribute:

   * Match Type: 1, Permit

   * IPv4 Neighbor Sub-TLV: Local BGP Speaker IGW2, Remote BGP Peer Speaker1

   * Action Type: Route-Preference

IGW2 processes the received BGP FS RPD route as follows:

1. IGW2 gets the target prefix Prefix2 from the Destination Prefix component in the BGP-FS NLRI of the BGP FS RPD route;

2. IGW2 identifies the Route Policy Distribution Flag carrying in the Flow Specification Traffic Action Extended Community, then IGW2 knows that the corresponding filtering rules will be used as Routing Policies.

3. IGW2 uses the target prefix Prefix2 to choose the matching routes, in this case, the prefix Prefix2 has two more routes:

Prefix	Next-Hop	Local BGP Speaker	Remote BGP Peer
Prefix2	Speaker1	IGW2	Speaker1
Prefix2	Speaker2	IGW2	Speaker2
...			

4. IGW2 gets the matching criteria from the BGP Policy Attribute: Local BGP Speaker IGW2, Remote BGP Peer Speaker1;

5. IGW2 gets the action from the BGP Policy Attribute: Route-Preference;

So IGW2 chooses the BGP route received from Speaker1 instead of Speaker2 as the best route and the outbound traffic destined for Prefix2 can be scheduled to link L3 for transmission.

5.3. BGP Wide Community

This section describes the option 2 for protocol extensions, which is completely different from section 5.2 by reusing BGP Wide Community introduced in [I-D.ietf-idr-wide-bgp-communities].

5.3.1. New Wide Community Atoms

New wide community atoms have to be introduced since the entrance and exit of traffic need to be designated precisely.

```
 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             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                         Value (variable)                      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 6: Wide Community Atoms

Supported format of the TLVs can be:

- o Type 1: Autonomous System number list
- o Type 2: IPv4 prefix (1 octet prefix length + prefix) list
- o Type 3: IPv6 prefix (1 octet prefix length + prefix) list
- o Type 4: Integer list
- o Type 5: IEEE Floating Point Number list
- o Type 6: Neighbor Class list
- o Type 7: User-defined Class list7
- o Type 8: UTF-8 String
- o Type TBD: BGP IPv4 neighbor --- Newly introduced in this draft
- o Type TBD: BGP IPv6 neighbor --- Newly introduced in this draft

5.3.2. Encoding examples

5.3.2.1. Inbound Traffic Control

As required in the case, traffic from PE1 to Prefix1 need to enter through L3, so IGWs except IGW2 should prepend ASN list to Prefix1 when populating to AS100. As shown in figure below, community "PREPEND N TIMES TO AS" and "Exclude Target(s) TLV" are be used.

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Container Type 1 (1)      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|1 0 0 0 0 0 0 0|
+-+-+-+-+-+-+-+-+
| Hop Count: 0  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Length:                    36 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Community: PREPEND N TIMES TO AS                           18 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Own ASN                                                   100 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Context ASN#                                              100 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|ExcTargetTLV(2)|   Length:                  11 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  IPv4Neig(TBD)|   Length:                   8 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Local Speaker                                           #IGW2 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Remote Speaker                                      #Speaker1 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Param TLV (3) |   Length:                   7 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Integer  (4) |   Length:                   4 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Prepend #                                                   5 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 7: Example encoding for Inbound Traffic Control case

5.3.2.2. Outbound Traffic Control

As required in the case, traffic from PE2 to Prefix2 need to exit through L3, so IGWs should perfer the route from IGW2 to Speaker1. As shown in figure below, community "LOCAL PREFERENCE" and "Target(s) TLV" are be used.

```
    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
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |     Container Type 1 (1)      |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |1 0 0 0 0 0 0 0|
   +-+-+-+-+-+-+-+-+
   | Hop Count: 0  |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   | Length:                    36 |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   | Community: PREPEND N TIMES TO AS                           18 |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   | Own ASN                                                   100 |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   | Context ASN#                                              100 |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   | TargetTLV(1)  |   Length:                 11 |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |  IPv4Neig(TBD)|   Length:                  8 |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   | Local Speaker                                            #IGW2 |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   | Remote Speaker                                       #Speaker1 |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   | Param TLV (3) |   Length:                  7 |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |  Integer  (4) |   Length:                  4 |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   | Increment #                                               100 |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 8: Example encoding for Outbound Traffic Control case

5.4. Capability Negotiation

It is necessary to negotiate the capability to support BGP FlowSpec Extensions for Route Policy Distribution (RPD). The BGP FS RPD Capability is a new BGP capability [RFC5492]. The Capability Code for this capability is to be specified by the IANA. 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)     |
+-----------------------------------------------------+
|  Send/Receive (1 octet)                             |
+-----------------------------------------------------+
```

Figure 6: BGP FS RPD Capability

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].

Send/Receive: This field indicates whether the sender is (a) willing to receive Route Policies via BGP FLowSpec from its peer (value 1), (b) would like to send Route Policies via BGP FLowSpec to its peer (value 2), or (c) both (value 3) for the <AFI, SAFI>.

6. Consideration

6.1. Route-Policy

Routing policies are used to filter routes and control how routes are received and advertised. If route attributes, such as reachability, are changed, the path along which network traffic passes changes accordingly.

When advertising, receiving, and importing routes, the router implements certain policies based on actual networking requirements to filter routes and change the attributes of the routes. Routing policies serve the following purposes:

- o Control route advertising: Only routes that match the rules specified in a policy are advertised.
- o Control route receiving: Only the required and valid routes are received. This reduces the size of the routing table and improves network security.
- o Filter and control imported routes: A routing protocol may import routes discovered by other routing protocols. Only routes that satisfy certain conditions are imported to meet the requirements of the protocol.

o Modify attributes of specified routes Attributes of the routes: that are filtered by a routing policy are modified to meet the requirements of the local device.

o Configure fast reroute (FRR): If a backup next hop and a backup outbound interface are configured for the routes that match a routing policy, IP FRR, VPN FRR, and IP+VPN FRR can be implemented.

Routing policies are implemented using the following procedures:

1. Define rules: Define features of routes to which routing policies are applied. Users define a set of matching rules based on different attributes of routes, such as the destination address and the address of the router that advertises the routes.

2. Implement the rules: Apply the matching rules to routing policies for advertising, receiving, and importing routes.

7. Contributors

The following people have substantially contributed to the definition of the BGP-FS RPD and to the editing of this document:

Peng Zhou
Huawei
Email: Jewpon.zhou@huawei.com

8. IANA Considerations

TBD.

9. Security Considerations

TBD.

10. Acknowledgements

The authors would like to thank Acee Lindem, Jeff Haas for their comments to this work.

11. References

11.1. Normative References

[I-D.ietf-idr-wide-bgp-communities]
Raszuk, R., Haas, J., Lange, A., Amante, S., Decraene, B., Jakma, P., and R. Steenbergen, "Wide BGP Communities Attribute", draft-ietf-idr-wide-bgp-communities-00 (work in progress), June 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>.

[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>.

11.2. Informative References

[I-D.ietf-idr-registered-wide-bgp-communities]
Raszuk, R. and J. Haas, "Registered Wide BGP Community Values", draft-ietf-idr-registered-wide-bgp-communities-00 (work in progress), June 2015.

Authors' Addresses

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

Email: lizhenbin@huawei.com

Liang Ou
China Telcom Co., Ltd.
109 West Zhongshan Ave,Tianhe District
Guangzhou 510630
China

Email: oul@gsta.com

Yujia Luo
China Telcom Co., Ltd.
109 West Zhongshan Ave,Tianhe District
Guangzhou 510630
China

Email: luoyuj@gsta.com

Sujian Lu
Tencent
Tengyun Building,Tower A ,No. 397 Tianlin Road
Shanghai, Xuhui District 200233
China

Email: jasonlu@tencent.com

Shunwan Zhuang
Huawei
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: zhuangshunwan@huawei.com

Nan Wu
Huawei
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: eric.wu@huawei.com

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

Z. Li
Huawei
L. Ou
Y. Luo
China Telcom Co., Ltd.
S. Lu
Tencent
H. Chen
Futurewei
S. Zhuang
H. Wang
Huawei

July 7, 2019

BGP Extensions for Routing Policy Distribution (RPD)
draft-li-idr-flowspec-rpd-05

Abstract

It is hard to adjust traffic and optimize traffic paths on a traditional IP network from time to time through manual configurations. It is desirable to have an automatic mechanism for setting up routing policies, which adjust traffic and optimize traffic paths automatically. This document describes BGP Extensions for Routing Policy Distribution (BGP RPD) to support this.

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].

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

Copyright Notice

Table of Contents

1. Introduction

It is difficult to optimize traffic paths on a traditional IP network because of:

o Heavy configuration and error prone. Traffic can only be adjusted device by device. All routers that the traffic traverses need to be configured. The configuration workload is heavy. The operation is not only time consuming but also prone to misconfiguration for Service Providers.

- o Complex. The routing policies used to control network routes are complex, posing difficulties to subsequent maintenance, high maintenance skills are required.

It is desirable to have an automatic mechanism for setting up routing policies, which can simplify the routing policies configuration. This document describes extensions to BGP for Routing Policy Distribution to resolve these issues.

2. Terminology

The following terminology is used in this document.

- o ACL:Access Control List
- o BGP: Border Gateway Protocol
- o FS: Flow Specification
- o PBR:Policy-Based Routing
- o RPD: Routing Policy Distribution
- o VPN: Virtual Private Network

3. Problem Statements

It is obvious that providers have the requirements to adjust their business traffic from time to time because:

- o Business development or network failure introduces link congestion and overload.
- o Network transmission quality is decreased as the result of delay, loss and they need to adjust traffic to other paths.
- o To control OPEX and CPEX, prefer the transit provider with lower price.

3.1. Inbound Traffic Control

In the scenario below, for the reasons above, the provider of AS100 saying P may wish the inbound traffic from AS200 enters AS100 through link L3 instead of the others. Since P doesn't have any administration over AS200, so there is no way for P to modify the route selection criteria directly.

```
               Traffic from PE1 to Prefix1
          ------------------------------------>

+-----------------+            +-------------------------+
|     +---------+ |        L1  | +----+      +----------+|
|     |Speaker1 | +------------+ |IGW1|      |policy    ||
|     +---------+ |**      L2**| +----+      |controller||
|                 |  **    **  |             +----------+|
| +---+           |    ****    |                         |
| |PE1|           |    ****    |                         |
| +---+           |  **    **  |                         |
|     +---------+ |**      L3**| +----+                  |
|     |Speaker2 | +------------+ |IGW2|      AS100       |
|     +---------+ |        L4  | +----+                  |
|                 |            |                         |
|    AS200        |            |                         |
|                 |            |  ...                    |
|                 |            |                         |
|     +---------+ |            | +----+      +-------+   |
|     |Speakern | |            | |IGWn|      |Prefix1|   |
|     +---------+ |            | +----+      +-------+   |
+-----------------+            +-------------------------+

            Prefix1 advertised from AS100 to AS200
          <-----------------------------------------
```

Inbound Traffic Control case

3.2. Outbound Traffic Control

In the scenario below, the provider of AS100 saying P prefers link L3 for the traffic to the destination Prefix2 among multiple exits and links. This preference can be dynamic and changed frequently because of the reasons above. So the provider P expects an efficient and convenient solution.

```
                  Traffic from PE2 to Prefix2
             ----------------------------------->
   +-------------------------+            +-----------------+
   |+----------+      +----+ |L1          | +---------+     |
   ||policy    |      |IGW1| +------------+ |Speaker1 |     |
   ||controller|      +----+ |**        **| +---------+     |
   |+----------+             |L2**    **  |        +-------+|
   |                         |    ****    |        |Prefix2||
   |                         |    ****    |        +-------+|
   |                         |L3**    **  |                 |
   |      AS100       +----+ |**        **| +---------+     |
   |                  |IGW2| +------------+ |Speaker2 |     |
   |                  +----+ |L4          | +---------+     |
   |                         |            |                 |
   |+---+                    |            |    AS200        |
   ||PE2|              ...   |            |                 |
   |+---+                    |            |                 |
   |                  +----+ |            | +---------+     |
   |                  |IGWn| |            | |Speakern |     |
   |                  +----+ |            | +---------+     |
   +-------------------------+            +-----------------+

               Prefix2 advertised from AS200 to AS100
             <----------------------------------------
```

Outbound Traffic Control case

4. Protocol Extensions

A solution is proposed to use a new AFI and SAFI with the BGP Wide Community for encoding a routing policy.

4.1. Using a New AFI and SAFI

A new AFI and SAFI are defined: the Routing Policy AFI whose codepoint TBD1 is to be assigned by IANA, and SAFI whose codepoint TBD2 is to be assigned by IANA.

The AFI and SAFI pair uses a new NLRI, which is defined 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
+-+-+-+-+-+-+-+-+
|  NLRI Length  |
+-+-+-+-+-+-+-+-+
|  Policy Type  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Distinguisher (4 octets)                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Peer IP (4/16 octets)                     ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Where:

NLRI Length: 1 octet represents the length of NLRI.

Policy Type: 1 octet indicates the type of a policy. 1 is for export policy. 2 is for import policy.

Distinguisher: 4 octet value uniquely identifies the policy in the peer.

Peer IP: 4/16 octet value indicates an IPv4/IPv6 peer.

The NLRI containing the Routing Policy is carried in a BGP UPDATE message, which MUST contain the BGP mandatory attributes and MAY also contain some BGP optional attributes.

When receiving a BGP UPDATE message, a BGP speaker processes it only if the peer IP address in the NLRI is the IP address of the BGP speaker or 0.

The content of the Routing Policy is encoded in a BGP Wide Community.

4.2. BGP Wide Community

The BGP wide community is defined in [I-D.ietf-idr-wide-bgp-communities]. It can be used to facilitate the delivery of new network services, and be extended easily for distributing different kinds of routing policies.

4.2.1. New Wide Community Atoms

A wide community Atom is a TLV (or sub-TLV), which may be included in a BGP wide community container (or BGP wide community for short) containing some BGP Wide Community TLVs. Three BGP Wide Community TLVs are defined in [I-D.ietf-idr-wide-bgp-communities], which are BGP Wide Community Target(s) TLV, Exclude Target(s) TLV, and

Parameter(s) TLV. Each of these TLVs comprises a series of Atoms, each of which is a TLV (or sub-TLV). A new wide community Atom is defined for BGP Wide Community Target(s) TLV and a few new Atoms are defined for BGP Wide Community Parameter(s) TLV. For your reference, the format of the TLV is illustrated 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      |              Length             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                         Value (variable)                      ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Format of Wide Community Atom TLV

A RouteAttr Atom TLV (or RouteAttr TLV/sub-TLV for short) is defined and may be included in a Target TLV. It 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 (TBD1)  |       Length (variable)          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            sub-TLVs                           ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Format of RouteAttr Atom TLV

The Type for RouteAttr is TBD1 (suggested value 48) to be assigned by IANA. In RouteAttr TLV, three sub-TLVs are defined: IP Prefix, AS-Path and Community sub-TLV.

An IP prefix sub-TLV gives matching criteria on IPv4 prefixes. Its format is illustrated 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 (TBD2)  |          Length (N x 8)       |M-Type | Flags |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                          IPv4 Address                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Mask      |    GeMask     |    LeMask     |M-Type | Flags |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~       . . .
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                          IPv4 Address                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Mask      |    GeMask     |    LeMask     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Format of IPv4 Prefix sub-TLV

Type: TBD2 (suggested value 1) for IPv4 Prefix is to be assigned by IANA.

Length: N x 8, where N is the number of tuples <M-Type, Flags, IPv4 Address, Mask, GeMask, LeMask>.

M-Type: 4 bits for match types, four of which are defined:

M-Type = 0: Exact match.

M-Type = 1: Match prefix greater and equal to the given masks.

M-Type = 2: Match prefix less and equal to the given masks.

M-Type = 3: Match prefix within the range of the given masks.

Flags: 4 bits. No flags are currently defined.

IPv4 Address: 4 octets for an IPv4 address.

Mask: 1 octet for the mask length.

GeMask: 1 octet for match range, must be less than Mask or be 0.

LeMask: 1 octet for match range, must be greater than Mask or be 0.

For example, tuple <M-Type=0, Flags=0, IPv4 Address = 1.1.0.0, Mask = 22, GeMask = 0, LeMask = 0> represents an exact IP prefix match for 1.1.0.0/22.

<M-Type=1, Flags=0, IPv4 Address = 16.1.0.0, Mask = 24, GeMask = 24, LeMask = 0> represents match IP prefix 1.1.0.0/24 greater-equal 24.

<M-Type=2, Flags=0, IPv4 Address = 17.1.0.0, Mask = 24, GeMask = 0, LeMask = 26> represents match IP prefix 17.1.0.0/24 less-equal 26.

<M-Type=3, Flags=0, IPv4 Address = 18.1.0.0, Mask = 24, GeMask = 24, LeMask = 32> represents match IP prefix 18.1.0.0/24 greater-equal to 24 and less-equal 32.

Similarly, an IPv6 Prefix sub-TLV represents match criteria on IPv6 prefixes. Its format is illustrated 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(TBD3)  |          Length (N x 20)      |M-Type | Flags |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      IPv6 Address (16 octets)                 ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Mask      |    GeMask     |     LeMask    |M-Type | Flags |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~       . . .
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       IPv6 Address (16 octets                 ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Mask      |    GeMask     |     LeMask    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Format of IPv6 Prefix sub-TLV

An AS-Path sub-TLV represents a match criteria in a regular expression string. Its format is illustrated 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 (TBD4)   |      Length (Variable)        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     AS-Path Regex String                      |
:                                                               :
|                                                               ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Format of AS Path sub-TLV

Type: TBD4 (suggested value 2) for AS-Path is to be assigned by IANA.

Length: Variable, maximum is 1024.

AS-Path Regex String: AS-Path regular expression string.

A community sub-TLV represents a list of communities to be matched all. Its format is illustrated 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 (TBD5)  |         Length (N x 4 + 1)      |     Flags    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      Community 1 Value                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                              . . .                             ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      Community N Value                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Format of Community sub-TLV

Type: TBD5 (suggested value 3) for Community is to be assigned by IANA.

Length: N x 4 + 1, where N is the number of communities.

Flags: 1 octet. No flags are currently defined.

In Parameter(s) TLV, two action sub-TLVs are defined: MED change sub-TLV and AS-Path change sub-TLV. When the community in the container is MATCH AND SET ATTR, the Parameter(s) TLV includes some of these sub-TLVs. When the community is MATCH AND NOT ADVERTISE, the Parameter(s) TLV's value is empty.

A MED change sub-TLV indicates an action to change the MED. Its format is illustrated 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 (TBD6)  |           Length (5)          |       OP       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            Value                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Format of MED Change sub-TLV

Type: TBD6 (suggested value 1) for MED Change is to be assigned by IANA.

Length: 5.

OP: 1 octet. Three are defined:

OP = 0: assign the Value to the existing MED.

OP = 1: add the Value to the existing MED. If the sum is greater than the maximum value for MED, assign the maximum value to MED.

OP = 2: subtract the Value from the existing MED. If the existing MED minus the Value is less than 0, assign 0 to MED.

Value: 4 octets.

An AS-Path change sub-TLV indicates an action to change the AS-Path. Its format is illustrated 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 (TBD7)  |        Length (n x 5)         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                              AS1                              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Count1     |
+-+-+-+-+-+-+-+-+
~       . . .
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                              ASn                              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Countn     |
+-+-+-+-+-+-+-+-+
```

Format of AS-Path Change sub-TLV

Type: TBD7 (suggested value 2) for AS-Path Change is to be assigned by IANA.

Length: n x 5.

ASi: 4 octet. An AS number.

Counti: 1 octet. ASi repeats Counti times.

The sequence of AS numbers are added to the existing AS Path.

4.3. Capability Negotiation

It is necessary to negotiate the capability to support BGP Extensions for Routing Policy Distribution (RPD). The BGP RPD Capability is a new BGP capability [RFC5492]. The Capability Code for this capability is to be specified by the IANA. 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)  |
+--------------------------------------------------+
|  Send/Receive (1 octet)                          |
+--------------------------------------------------+
```

BGP RPD Capability

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].

Send/Receive: This field indicates whether the sender is (a) willing to receive Routing Policies from its peer (value 1), (b) would like to send Routing Policies to its peer (value 2), or (c) both (value 3) for the <AFI, SAFI>.

5. Consideration

5.1. Route-Policy

Routing policies are used to filter routes and control how routes are received and advertised. If route attributes, such as reachability, are changed, the path along which network traffic passes changes accordingly.

When advertising, receiving, and importing routes, the router implements certain policies based on actual networking requirements to filter routes and change the attributes of the routes. Routing policies serve the following purposes:

- o Control route advertising: Only routes that match the rules specified in a policy are advertised.

- o Control route receiving: Only the required and valid routes are received. This reduces the size of the routing table and improves network security.

- o Filter and control imported routes: A routing protocol may import routes discovered by other routing protocols. Only routes that satisfy certain conditions are imported to meet the requirements of the protocol.

- o Modify attributes of specified routes Attributes of the routes: that are filtered by a routing policy are modified to meet the requirements of the local device.

- o Configure fast reroute (FRR): If a backup next hop and a backup outbound interface are configured for the routes that match a routing policy, IP FRR, VPN FRR, and IP+VPN FRR can be implemented.

Routing policies are implemented using the following procedures:

1. Define rules: Define features of routes to which routing policies are applied. Users define a set of matching rules based on different attributes of routes, such as the destination address and the address of the router that advertises the routes.

2. Implement the rules: Apply the matching rules to routing policies for advertising, receiving, and importing routes.

6. Contributors

The following people have substantially contributed to the definition of the BGP-FS RPD and to the editing of this document:

Peng Zhou
Huawei
Email: Jewpon.zhou@huawei.com

7. Security Considerations

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

8. Acknowledgements

The authors would like to thank Acee Lindem, Jeff Haas, Jie Dong, Lucy Yong, Qiandeng Liang, Zhenqiang Li for their comments to this work.

9. IANA Considerations

This document requests assigning a new AFI in the registry "Address Family Numbers" as follows:

Code Point	Description	Reference
TBD (36879 suggested)	Routing Policy AFI	This document

This document requests assigning a new SAFI in the registry "Subsequent Address Family Identifiers (SAFI) Parameters" as follows:

Code Point	Description	Reference
TBD(179 suggested)	Routing Policy SAFI	This document

This document defines a new registry called "Routing Policy NLRI". The allocation policy of this registry is "First Come First Served (FCFS)" according to [RFC8126].

Following code points are defined:

Code Point	Description	Reference
1	Export Policy	This document
2	Import Policy	This document

This document requests assigning a code-point from the registry "BGP Community Container Atom Types" as follows:

TLV Code Point	Description	Reference
TBD1 (48 suggested)	RouteAttr Atom	This document

This document defines a new registry called "Route Attributes Sub-TLV" under RouteAttr Atom TLV. The allocation policy of this registry is "First Come First Served (FCFS)" according to [RFC8126].

Following Sub-TLV code points are defined:

Code Point	Description	Reference
0	Reserved	
1	IP Prefix Sub-TLV	This document
2	AS-Path Sub-TLV	This document
3	Community Sub-TLV	This document
4 - 255	To be assigned in FCFS	

This document defines a new registry called "Attribute Change Sub-TLV" under Parameter(s) TLV. The allocation policy of this registry is "First Come First Served (FCFS)" according to [RFC8126].

Following Sub-TLV code points are defined:

Code Point	Description	Reference
0	Reserved	
1	MED Change Sub-TLV	This document
2	AS-Path Change Sub-TLV	This document
3 - 255	To be assigned in FCFS	

10. References

10.1. Normative References

[I-D.ietf-idr-wide-bgp-communities]
Raszuk, R., Haas, J., Lange, A., Decraene, B., Amante, S., and P. Jakma, "BGP Community Container Attribute", draft-ietf-idr-wide-bgp-communities-05 (work in progress), July 2018.

[RFC1997] Chandra, R., Traina, P., and T. Li, "BGP Communities Attribute", RFC 1997, DOI 10.17487/RFC1997, August 1996, <https://www.rfc-editor.org/info/rfc1997>.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <https://www.rfc-editor.org/info/rfc4271>.

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

[RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, DOI 10.17487/RFC5492, February 2009, <https://www.rfc-editor.org/info/rfc5492>.

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

[RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 8126, DOI 10.17487/RFC8126, June 2017, <https://www.rfc-editor.org/info/rfc8126>.

10.2. Informative References

[I-D.ietf-idr-registered-wide-bgp-communities] Raszuk, R. and J. Haas, "Registered Wide BGP Community Values", draft-ietf-idr-registered-wide-bgp-communities-02 (work in progress), May 2016.

Authors' Addresses

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

Email: lizhenbin@huawei.com

Liang Ou
China Telcom Co., Ltd.
109 West Zhongshan Ave,Tianhe District
Guangzhou 510630
China

Email: oul@gsta.com

Yujia Luo
China Telcom Co., Ltd.
109 West Zhongshan Ave,Tianhe District
Guangzhou 510630
China

Email: luoyuj@gsta.com

Sujian Lu
Tencent
Tengyun Building,Tower A ,No. 397 Tianlin Road
Shanghai, Xuhui District 200233
China

Email: jasonlu@tencent.com

Huaimo Chen
Futurewei
Boston, MA
USA

Email: Huaimo.chen@futurewei.com

Shunwan Zhuang
Huawei
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: zhuangshunwan@huawei.com

Haibo Wang
Huawei
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: rainsword.wang@huawei.com

Network Working Group Z. Li
Internet-Draft Z. Zhuang
Intended status: Standards Track Huawei Technologies
Expires: April 19, 2016 S. Lu
Tencent
October 17, 2015

BGP Extensions for Service-Oriented MPLS Path Programming (MPP)
draft-li-idr-mpls-path-programming-02

Abstract

Service-oriented MPLS programming (SoMPP) is to provide customized service process based on flexible label combinations. BGP will play an important role for MPLS path programming to download programmed MPLS path and map the service path to the transport path. This document defines BGP extensions to support Service-oriented MPLS path programming.

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].

Table of Contents

1. Introduction

The label stack capability of MPLS would have been utilized well to implement flexible path programming to satisfy all kinds of service requirements. But in the distributed environment, the flexible programming capability is difficult to implement and always confined to reachability. As the introducing of central control in the network, the flexible MPLS programming capability becomes possible owing to two factors: 1. It becomes easier to allocate label for more purposes than reachability; 2. It is easy to calculate the MPLS path in a global network view. Moreover, the MPLS path programming capability can be utilized to satisfy more requirements of service bearing in the service layer which is defined as Service-oriented MPLS path programming. BGP will play an important role for MPLS path programming to download programmed MPLS path and map the service path to the transport path. This document defines BGP extensions to support Service-oriented MPLS path programming.

2. Terminology

BGP: Border Gateway Protocol

EVPN: Ethernet VPN

L2VPN: Layer 2 VPN

L3VPN: Layer 3 VPN

MPP: MPLS Path Programming

MVPN: Multicast VPN

RR: Route Reflector

SR-Path: Segment Routing Path

NLRI: Network Layer Reachability Information

3. Architecture and Usecases of SoMPP

3.1. Architecture

The architecture of BGP-based MPLS path programming is shown in the Figure 1. Central control plays an important role in MPLS path programming. It can extend the MPLS path programming capability easily. The central controller can calculate path in a global network view and implement the MPLS path programming to satisfy different requirements of services. The result of MPLS path programming can be advertised from the central controller to the client nodes through BGP extensions to the ingress PEs. When client nodes receives the result of MPLS path programming, it will install the MPLS forwarding entry for the specified BGP prefix to implement the service process.

```
                  +--------------------+
                  |      Central       |
                  |     Controller     |
       |----------|(Path Calculation   |---------|
       |          | /Path Programming)|         |
       |          +--------------------+         |
       |                                         |
    MPLS Path                                 MPLS Path
       |                                         |
       |                                         |
       |                                         |
     +---------+        +---------+        +---------+
     | CLIENT  |        | CLIENT  |        | CLIENT  |
     |         | ...... |         | ...... |         |
     |  (PE)   |        |   (P)   |        |  (PE)   |
     |         |        |         |        |         |
     +---------+        +---------+        +---------+
```

Figure 1 BGP-based MPLS Path Programming

3.2. Usecases

3.2.1. Deterministic ECMP

Entropy Label[RFC6790] is introduced to improve the ECMP capability by encapsulate the entropy label in the MPLS label stack. The existing implementation is always to calculate the entropy label based on the header of packets by specific hash algorithm in the ingress node. That is, the entropy label is determined locally by the ingress node. The method can improve the hash of packets in the network for load-sharing. But since the ingress node lacks the knowledge of the global traffic pattern of the network and calculates the entropy label by itself it may be not able to improve the ECMP capability accurately and in some cases it may deteriorate the imbalance of load-sharing.

With the central controlled MPLS path programming, the central controller can collect the global traffic pattern information of the network and based on the information deterministically calculate the entropy label for specific flows to help improve the load-sharing of the network. Then the central controller can download the label stack information with the deterministic entropy label to the ingress PEs for the specific BGP prefix. The ingress node can install the MPLS forwarding entry shown in the following figure to help optimize the ECMP of the flow specified by the BGP prefix, then optimize the ECMP of the whole network.

```
+----------+      +----------+----------+
|   BGP    | ---> |  Entropy |BGP Prefix| ---> Transport
|  Prefix  |      |   Label  |   Label  |        Tunnel
+----------+      +----------+----------+
```

3.2.2. Centralized Mapping of Service to Tunnels

In the network there can be multiple tunnels to one specific destination which satisfy different constraints. In the traditional way, the tunnel is set up by the distributed forwarding nodes. As the PCE-initiated LSP setup [I-D.ietf-pce-pce-initiated-lsp]is introduced, the tunnel with different constraints can be set up in the central controlled way. In order to satisfy different service requirements, it is necessary to provide the capability to flexibly map the service to different tunnels which constraints can satisfy the required service requirement. Since the central controller has enough information of the whole network view, it can be an effective way to map the service (such as L3VPN and L2VPN) to the tunnel by the central controller and advertise the mapping information to the ingress PE of the service to guide the mapping in the forwarding node.

There can be two types of behaviors to map service to the tunnel:

1. Specify the tunnel type: with the method BGP will carry the tunnel type information for the BGP prefix. When the ingress PE receives the information, it will use the tunnel type and the nexthop address (or other specified target IP address) to search the corresponding tunnels to bear the flow specified by the BGP prefix. If there are more than one tunnels, the ingress PE will load share the traffic across all the tunnels.

2. Specify the specific tunnel: For MPLS TE/SR-TE tunnel, there can be multiple MPLS TE tunnels from one ingress PE to a specific destination with different constraints. BGP can carry the tunnel identifier information for the BGP prefix from the controller to the ingress node. When the ingress PE receives the information, it will use the tunnel identifier information to search the corresponding tunnels to bear the flow specified by the BGP prefix. If there are multiple tunnel identifiers, the ingress PE will load share the traffic across all the tunnels.

4. Download of MPLS Path

According to the service requirements, the central controller can combine MPLS labels flexibly. Then it can download the service label combination for specific prefix.BGP extensions are necessary to advertise label stacks for the prefix in NLRI field.

```
+--------------------------+
|   Length (1 octet)       |
+--------------------------+
|   Label (3 octets)       |
+--------------------------+
............................
+--------------------------+
|   Prefix (variable)      |
+--------------------------+
```

Figure 1: NLRI Definition in RFC3107

[RFC3107] defines above NLRI to advertise label binding for specific prefix. The label field can carry one or more labels. Each label is encoded as 3 octets, where the high-order 20 bits contain the label value, and the low order bit contains "Bottom of Stack". But for other AFI/SAFIs using label binding such as VPNv4, VPNv6, EVPN, MVPN, etc., it dose not support the capability to carry more labels for the specific prefix. Moreover for the AFI/SAFIs which do not support label binding capability originally, but may possibly adopt MPLS path programming now, there is no label field in the NLRI. In order to support flexible MPLS path programming, this document defines and uses a new BGP attribute called the "Extended Label attribute". This is an optional transitive BGP attribute. The format of this attribute is defined as follows:

```
+--------------------------+
|   Label 1 (3 octets)     |
+--------------------------+
|   Label 2 (3 octets)     |
+--------------------------+
............................
+--------------------------+
|   Label n (3 octets)     |
+--------------------------+
```

Figure 2: Extended Label Attribute

The Label field carries one or more labels (that corresponds to the stack of labels [[RFC3032]]). Each label is encoded as 3 octets, where the high-order 20 bits contain the label value, and the low order bit contains "Bottom of Stack" (as defined in [[RFC3032]]).

The central controller for MPLS path programming could build a route with Extended Label attribute and send it to the ingress routers.

Upon receiving such a route from the central Controller, the ingress router SHOULD select such a route as the best path. If a packet comes into the ingress router and uses such a path, the ingress router will encapsulate the stack of labels which is derived from the Extended Label Attribute of the route into the packet and forward the packet along the path.

The "Extended Label attribute" can be used for various BGP address families. Before using this attribute, firstly, it is necessary to negotiate the capability between two nodes to support MPLS path programming for a specific BGP address family. If negotiation fails, a node MUST NOT send this attribute and MUST discard this attribute when it receives.

5. Download of Mapping of Service Path to Transport Path

5.1. Specify Tunnel Type

[I-D.ietf-idr-tunnel-encaps] proposes the Tunnel Encapsulation Attribute which can be used without BGP Encapsulation SAFI to specify a set of tunnels. It defines a series of Encapsulation Sub-TLVs for particular tunnel types. It also defines the Remote Endpoint Attributes Sub-TLV to specify the remote tunnel endpoint address for each tunnel which can be different the BGP nexthop. The Tunnel Encapsulation Attributes can be reused for the MPLS path programming to specify the tunnel types, the encapsulation and the remote tunnel endpoint address which can determine a set of tunnels which the service can map to. Now the limited MPLS tunnel types are defined for the Tunnel Encapsulation Attributes. In order to support MPLS path programming, the following MPLS tunnel types are to be defined:

Value	Tunnel Type
TBD	LDP LSP
TBD	RSVP-TE LSP
TBD	MPLS-based Segment Routing Best-effort Path
TBD	MPLS-based Segment Routing Traffic Engineering Path

5.2. Specify Specific Tunnel

Besides specifying the tunnel types to determine the set of tunnels which the service traffic can map to, the specific tunnels can be specified directly by the tunnel identifiers when map the service traffic to the path. BGP extensions is necessary that through the community attribute of BGP the identifier of the transport path can be carried when advertise the specific prefix.

In order to support the application, this document defines a new BGP attribute called the "Extended Unicast Tunnel attribute". This is an optional transitive BGP attribute. The format of this attribute is defined as follows:

```
+-------------------------------------------------+
| Flags (1 octet)                                 |
+-------------------------------------------------+
| Tunnel Type (1 octets)                          |
+-------------------------------------------------+
| Tunnel Identifier (variable)                    |
+-------------------------------------------------+
```

The Flags is reserved and must be set as zero. The Tunnel Type identifies the type of the tunneling technology used for the unicast service path. The tunnel type determines the syntax and semantics of the Tunnel Identifier field. This document defines following Tunnel Types:

+ 0 - No tunnel information present
+ 1 - RSVP-TE LSP
+ 2 - MPLS-based Segment Routing Traffic Engineering Path

Tunnel Specific Attributes contains the attributes of the tunnel. The field is optional. The value depends on the tunnel type. It will be defined in the future versions.

When the Tunnel Type is set to "No tunnel information present", the Tunnel attribute carries no tunnel information (no Tunnel Identifier). when the type is used, the tunnel used for the service path is determined by the ingress router.

When the Tunnel Type is set to RSVP - Traffic Engineering (RSVP-TE) Label Switched Path (LSP), the Tunnel Identifier is <C-Type, Tunnel Sender Address, Tunnel ID, Tunnel End-point Address> as specified in [RFC3209] If C-Type = 7, Tunnel Sender Address and Tunnel End-point Address are IPv4 address in 4 octets. If C-Type = 8, Tunnel Sender Address and Tunnel End-point Address are IPv6 address in 16 octets. The other fields in the RSVP-TE LSP Identifier are the same as specified in [RFC3209].

When the Tunnel Type is set to MPLS-based Segment Routing Traffic Engineering Path, the Tunnel Identifier is <C-Type, Tunnel Sender Address, Tunnel ID, Tunnel End-point Address>. If C-Type = 7, Tunnel Sender Address and Tunnel End-point Address are IPv4 address in 4 octets. If C-Type = 8, Tunnel Sender Address and Tunnel End-point Address are IPv6 address in 16 octets. The tunnel identifier is similar as that of RSVP-TE LSP.

BGP can carry multiple Extended Unicast Tunnel Attributes for specific prefix. If there are multiple tunnel identifiers, the ingress PE will load share the traffic across all the specified tunnels for the service traffic determined by the specific BGP prefix.

6. Route Flag Extended Community

In order to make the MPLS path programming to take effect, the route advertised by the central controller after the MPLS Path Programming should be selected by the ingress PE over other routes for the same BGP prefix. There are two options of BGP extensions for the purpose:

Option 1: A new BGP Extended Community called as the "Route Flag Extended Community" can be introduced. The Type value is to be assigned by IANA.

The Route Flag Extended Community is used to carry the flag appointed by the BGP central controller.

The format of this extended community is defined as follows:

```
   0     1     2     3     4     5     6     7
 +-----+-----+-----+-----+-----+-----+-----+-----+
 |    Type   |  Reserved                   |Flag |
 +-----+-----+-----+-----+-----+-----+-----+-----+
```

Flag = 0, Treat as normal route
Flag = 1, Treat as best route

When a router receives a BGP route with a Route Flag Extended Community and the Flag set to "1", it SHOULD use the route as the best route when select the route from multiple routes for a specific prefix.

Option 2: [I-D.ietf-idr-custom-decision] defines a new Extended Community, called the Cost Community, which can be used in tie breaking during the best path selection process. The Cost Community can be reused by the MPLS path programming to set the "Point of Insertion" as 128 to make the route advertised by the central controller to be chosen.

7. Destination Node Attribute

This document defines and uses a new BGP attribute called as the "Destination Node attribute" which Type value is to be assigned by

IANA. The Destination Node attribute is an optional non-transitive attribute that can be applied to any address family.

The Destination Node attribute is used to carry a list of node addresses, which are intended to be used to determine the nodes where the route with such attribute SHOULD be considered. If a node receives a BGP route with a Destination Node attribute, it MUST check the node address list. If one address of the list belongs to this node, the route MUST be used in this node. Otherwise the route MUST be ignored silently.

The format of this attribute is defined 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|               AFI             |      SAFI     |   Reserved    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                                                               ~
~               Destination Node Address List                   ~
~                                                               ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

AFI: Address Family Identifier (16 bits).

SAFI: Subsequent Address Family Identifier (8 bits).

Reserved: One octet reserved for special flags

Destination Node Address List: The list of IPv4 (AFI=1) or IPv6 (AFI=2) address.

8. Capability Negotiation

It is necessary to negotiate the capability to support MPLS path programming. The MPLS-Path-Programming Capability is a new BGP capability [RFC5492]. The Capability Code for this capability is to be specified by the IANA. 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)     |
+-----------------------------------------------------+
|  Send/Receive (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].

Send/Receive: This field indicates whether the sender is (a) willing to receive programming MPLS paths from its peer (value 1), (b) would like to send programming MPLS paths to its peer (value 2), or (c) both (value 3) for the <AFI, SAFI>.

9. IANA Considerations

TBD.

10. Security Considerations

TBD.

11. References

11.1. Normative References

[I-D.ietf-idr-custom-decision]
Retana, A. and R. White, "BGP Custom Decision Process", draft-ietf-idr-custom-decision-06 (work in progress), April 2015.

[I-D.ietf-idr-tunnel-encaps]
Rosen, E., Patel, K., and G. Velde, "Using the BGP Tunnel Encapsulation Attribute without the BGP Encapsulation SAFI", draft-ietf-idr-tunnel-encaps-00 (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>.

[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>.

[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>.

[RFC5036] Andersson, L., Ed., Minei, I., Ed., and B. Thomas, Ed., "LDP Specification", RFC 5036, DOI 10.17487/RFC5036, October 2007, <http://www.rfc-editor.org/info/rfc5036>.

[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>.

11.2. Informative References

[I-D.ietf-pce-pce-initiated-lsp]
Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE Model", draft-ietf-pce-pce-initiated-lsp-04 (work in progress), April 2015.

[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>.

[RFC6790] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", RFC 6790, DOI 10.17487/RFC6790, November 2012, <http://www.rfc-editor.org/info/rfc6790>.

Authors' Addresses

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

Email: lizhenbin@huawei.com

Shunwan Zhuang
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: zhuangshunwan@huawei.com

Sujian Lu
Tencent
Tengyun Building,Tower A ,No. 397 Tianlin Road,Xuhui District
Shanghai 200233
China

Email: jasonlu@tencent.com

Network Working Group Z. Li
Internet-Draft S. Zhuang
Intended status: Standards Track Huawei Technologies
Expires: May 3, 2017 S. Lu
Tencent
October 30, 2016

BGP Extensions for Service-Oriented MPLS Path Programming (MPP)
draft-li-idr-mpls-path-programming-04

Abstract

Service-oriented MPLS programming (SoMPP) is to provide customized service process based on flexible label combinations. BGP will play an important role for MPLS path programming to download programmed MPLS path and map the service path to the transport path. This document defines BGP extensions to support service-oriented MPLS path programming.

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

Copyright Notice

Table of Contents

1. Introduction

The label stack capability of MPLS would have been utilized well to implement flexible path programming to satisfy all kinds of service requirements. But in the distributed environment, the flexible programming capability is difficult to implement and always confined to reachability. As the introducing of central control in the network, the flexible MPLS programming capability becomes possible owing to two factors: 1. It becomes easier to allocate label for more purposes than reachability; 2. It is easy to calculate the MPLS path in a global network view. Moreover, the MPLS path programming capability can be utilized to satisfy more requirements of service bearing in the service layer which is defined as Service-oriented MPLS path programming. BGP will play an important role for MPLS path programming to download programmed MPLS path and map the service path to the transport path. This document defines BGP extensions to support Service-oriented MPLS path programming.

2. Terminology

BGP: Border Gateway Protocol

EVPN: Ethernet VPN

L2VPN: Layer 2 VPN

L3VPN: Layer 3 VPN

MPP: MPLS Path Programming

MVPN: Multicast VPN

RR: Route Reflector

SR-Path: Segment Routing Path

NLRI: Network Layer Reachability Information

3. Architecture and Usecases of SoMPP

3.1. Architecture

The architecture of BGP-based MPLS path programming is shown in the Figure 1. Central control plays an important role in MPLS path programming. It can extend the MPLS path programming capability easily. The central controller can calculate path in a global network view and implement the MPLS path programming to satisfy different requirements of services. The result of MPLS path programming can be advertised from the central controller to the client nodes through BGP extensions to the ingress PEs. When client nodes receives the result of MPLS path programming, it will install the MPLS forwarding entry for the specified BGP prefix to implement the service process.

```
                  +-------------------+
                  |      Central      |
                  |     Controller    |
       |----------|(Path Calculation  |--------|
       |          | /Path Programming)|        |
       |          +-------------------+        |
       |                                       |
    MPLS Path                               MPLS Path
       |                                       |
       |                                       |
       |                                       |
     +--------+         +--------+         +--------+
     | CLIENT |         | CLIENT |         | CLIENT |
     |        | ......  |        | ......  |        |
     |  (PE)  |         |  (P)   |         |  (PE)  |
     |        |         |        |         |        |
     +--------+         +--------+         +--------+
```

Figure 1 BGP-based MPLS Path Programming

3.2. Usecases

3.2.1. Deterministic ECMP

Entropy Label[RFC6790] is introduced to improve the ECMP capability by encapsulate the entropy label in the MPLS label stack. The existing implementation is always to calculate the entropy label based on the header of packets by specific hash algorithm in the ingress node. That is, the entropy label is determined locally by the ingress node. The method can improve the hash of packets in the network for load-sharing. But since the ingress node lacks the knowledge of the global traffic pattern of the network and calculates the entropy label by itself it may be not able to improve the ECMP capability accurately and in some cases it may deteriorate the imbalance of load-sharing.

With the central controlled MPLS path programming, the central controller can collect the global traffic pattern information of the network and based on the information deterministically calculate the entropy label for specific flows to help improve the load-sharing of the network. Then the central controller can download the label stack information with the deterministic entropy label to the ingress PEs for the specific BGP prefix. The ingress node can install the MPLS forwarding entry shown in the following figure to help optimize the ECMP of the flow specified by the BGP prefix, then optimize the ECMP of the whole network.

```
+----------+      +----------+---------+
|   BGP    | ---> |  Entropy |BGP Prefix| ---> Transport
|  Prefix  |      |   Label  |   Label  |        Tunnel
+----------+      +----------+---------+
```

3.2.2. Centralized Mapping of Service to Tunnels

In the network there can be multiple tunnels to one specific destination which satisfy different constraints. In the traditional way, the tunnel is set up by the distributed forwarding nodes. As the PCE-initiated LSP setup [I-D.ietf-pce-pce-initiated-lsp] is introduced, the tunnel with different constraints can be set up in the central controlled way. In order to satisfy different service requirements, it is necessary to provide the capability to flexibly map the service to different tunnels which constraints can satisfy the required service requirement. Since the central controller has enough information of the whole network view, it can be an effective way to map the service (such as L3VPN and L2VPN) to the tunnel by the central controller and advertise the mapping information to the ingress PE of the service to guide the mapping in the forwarding node.

There can be two types of behaviors to map service to the tunnel:

1. Specify the tunnel type: with the method BGP will carry the tunnel type information for the BGP prefix. When the ingress PE receives the information, it will use the tunnel type and the nexthop address (or other specified target IP address) to search the corresponding tunnels to bear the flow specified by the BGP prefix. If there are more than one tunnels, the ingress PE will load share the traffic across all the tunnels.

2. Specify the specific tunnel: For MPLS TE/SR-TE tunnel, there can be multiple MPLS TE tunnels from one ingress PE to a specific destination with different constraints. BGP can carry the tunnel identifier information for the BGP prefix from the controller to the ingress node. When the ingress PE receives the information, it will use the tunnel identifier information to search the corresponding tunnels to bear the flow specified by the BGP prefix. If there are multiple tunnel identifiers, the ingress PE will load share the traffic across all the tunnels.

4. Advertising Label Stacks in BGP

According to the service requirements, the central controller can combine MPLS labels flexibly. Then it can download the service label combination for specific prefix. BGP extensions are necessary to advertise label stacks for the prefix in NLRI field.

```
+--------------------------+
|   Length (1 octet)       |
+--------------------------+
|   Label (3 octets)       |
+--------------------------+
............................
+--------------------------+
|   Prefix (variable)      |
+--------------------------+
```

Figure 2: NLRI Definition in RFC3107

[RFC3107] defines above NLRI to advertise label binding for specific prefix. The label field can carry one or more labels. Each label is encoded as 3 octets, where the high-order 20 bits contain the label value, and the low order bit contains "Bottom of Stack". But for the other AFI/SAFIs using label binding such as IPv4 Flowspec, IPv6 Flowspec, VPNv4, VPNv6, EVPN, MVPN, etc., it dose not support the capability to carry more labels for the specific prefix. Moreover for the AFI/SAFIs which do not support label binding capability originally, but may possibly adopt MPLS path programming now, there is no label field in the NLRI. In order to support flexible MPLS path programming, this document defines and uses a new BGP attribute called the "Extended Label attribute". This is an optional transitive BGP attribute. The attribute type code is (TBA by IANA), the value field of this attribute is defined as follows:

```
+--------------------------+
|   Label 1 (3 octets)     |
+--------------------------+
|   Label 2 (3 octets)     |
+--------------------------+
............................
+--------------------------+
|   Label n (3 octets)     |
+--------------------------+
```

Figure 3: Extended Label Attribute

The Label field carries one or more labels (that corresponds to the stack of labels [[RFC3032]]). Each label is encoded as 3 octets, where the high-order 20 bits contain the label value, and the low order bit contains "Bottom of Stack" (as defined in [[RFC3032]]). In the last label, the S bit MUST be "1"; in the other labels, the S bit MUST be "0".

The "Extended Label attribute" can be used for various BGP address families. Before using this attribute, firstly, it is necessary to negotiate the capability between two nodes to support MPLS path programming for a specific BGP address family. If negotiation fails, a node MUST NOT send this attribute and MUST discard this attribute when it receives.

4.1. Download of MPLS Path

The Central Controller for MPLS path programming could build a route with Extended Label attribute and send it to the ingress routers.

Upon receiving such a route from the Central Controller, the ingress router SHOULD select such a route as the best path. If a packet comes into the ingress router and uses such a path, the ingress router will encapsulate the stack of labels which is derived from the Extended Label Attribute of the route into the packet and forward the packet along the path.

4.2. Mapping Traffic to MPLS Path

The Extended Label attribute can be used for BGP Flowspec address families. BGP advertises the Flowspec with the Extended Label attribute, so the flow packets can be redirected to the MPLS Path which is derived from the Extended Label Attribute.

5. Download of Mapping of Service Path to Transport Path

5.1. Specify Tunnel Type

[I-D.ietf-idr-tunnel-encaps] proposes the Tunnel Encapsulation Attribute which can be used without BGP Encapsulation SAFI to specify a set of tunnels. It defines a series of Encapsulation Sub-TLVs for particular tunnel types. It also defines the Remote Endpoint Attributes Sub-TLV to specify the remote tunnel endpoint address for each tunnel which can be different the BGP nexthop. The Tunnel Encapsulation Attributes can be reused for the MPLS path programming to specify the tunnel types, the encapsulation and the remote tunnel endpoint address which can determine a set of tunnels which the service can map to. Now the limited MPLS tunnel types are defined for the Tunnel Encapsulation Attributes. In order to support MPLS path programming, the following MPLS tunnel types are to be defined:

Value	Tunnel Type
TBD	LDP LSP
TBD	RSVP-TE LSP
TBD	MPLS-based Segment Routing Best-effort Path
TBD	MPLS-based Segment Routing Traffic Engineering Path

5.2. Specify Specific Tunnel

Besides specifying the tunnel types to determine the set of tunnels which the service traffic can map to, the specific tunnels can be specified directly by the tunnel identifiers when map the service traffic to the path. BGP extensions is necessary that through the community attribute of BGP the identifier of the transport path can be carried when advertise the specific prefix.

In order to support the application, this document defines a new BGP attribute called the "Extended Unicast Tunnel attribute". This is an optional transitive BGP attribute. The attribute type code is (TBA by IANA), the value field of this attribute is defined as follows:

```
+---------------------------------------------------+
| First Tunnel entry (variable)                     |
+---------------------------------------------------+
| Second Tunnel entry (variable)                    |
+---------------------------------------------------+
| ...                                               |
+---------------------------------------------------+
| N-th Tunnel entry (variable)                      |
+---------------------------------------------------+
```

The Tunnel entry is defined as follows:

```
+-------------------------------------------------+
|  Flags (1 octet)                                |
+-------------------------------------------------+
|  Tunnel Type (1 octets)                         |
+-------------------------------------------------+
|  Tunnel Identifier (variable)                   |
+-------------------------------------------------+
| Tunnel Specific Attributes (Variable)(Optional)|
+-------------------------------------------------+
```

The Flags is reserved and must be set as zero. The Tunnel Type identifies the type of the tunneling technology used for the unicast service path. The tunnel type determines the syntax and semantics of the Tunnel Identifier field. This document defines following Tunnel Types:

+ 0 - No tunnel information present

+ 1 - RSVP-TE LSP

+ 2 - MPLS-based Segment Routing Traffic Engineering Path

Tunnel Specific Attributes contains the attributes of the tunnel. The field is optional. The value depends on the tunnel type. It will be defined in the future versions.

When the Tunnel Type is set to "No tunnel information present", the Tunnel attribute carries no tunnel information (no Tunnel Identifier). when the type is used, the tunnel used for the service path is determined by the ingress router.

When the Tunnel Type is set to RSVP - Traffic Engineering (RSVP-TE) Label Switched Path (LSP), the Tunnel Identifier is <C-Type, Tunnel Sender Address, Tunnel ID, Tunnel End-point Address> as specified in [RFC3209] If C-Type = 7, Tunnel Sender Address and Tunnel End-point Address are IPv4 address in 4 octets. If C-Type = 8, Tunnel Sender Address and Tunnel End-point Address are IPv6 address in 16 octets. The other fields in the RSVP-TE LSP Identifier are the same as specified in [RFC3209].

When the Tunnel Type is set to MPLS-based Segment Routing Traffic Engineering Path, the Tunnel Identifier is <C-Type, Tunnel Sender Address, Tunnel ID, Tunnel End-point Address>. If C-Type = 7, Tunnel Sender Address and Tunnel End-point Address are IPv4 address in 4 octets. If C-Type = 8, Tunnel Sender Address and Tunnel End-point Address are IPv6 address in 16 octets. The tunnel identifier is similar as that of RSVP-TE LSP.

BGP can carry multiple Tunnel entries in one Extended Unicast Tunnel attribute for specific prefix. If there are multiple tunnel entries, the ingress PE can load share the traffic across all the specified tunnels for the service traffic determined by the specific BGP prefix, or selects the primary / Backup tunnels from the multiple tunnel entries.

The "Redirect-to-Tunnel Action" for BGP Flowspec has been described in[I-D.hao-idr-flowspec-redirect-tunnel]. This document reuses the tunnel identifier and defines it in the Extended Unicast Tunnel attribute which can be used for "Redirect-to-Tunnel Action".

6. Route Flag Extended Community

In order to make the MPLS path programming to take effect, the route advertised by the central controller after the MPLS Path Programming should be selected by the ingress PE over other routes for the same BGP prefix. There are two options of BGP extensions for the purpose:

Option 1: A new BGP Extended Community called as the "Route Flag Extended Community" can be introduced. The Type value is to be assigned by IANA.

The Route Flag Extended Community is used to carry the flag appointed by the BGP central controller.

The format of this extended community is defined as follows:

```
    0     1     2     3     4     5     6     7
 +-----+-----+-----+-----+-----+-----+-----+-----+
 |    Type   |  Reserved                   |Flag |
 +-----+-----+-----+-----+-----+-----+-----+-----+

 Flag = 0, Treat as normal route
 Flag = 1, Treat as best route
```

When a router receives a BGP route with a Route Flag Extended Community and the Flag set to "1", it SHOULD use the route as the best route when select the route from multiple routes for a specific prefix.

Option 2: [I-D.ietf-idr-custom-decision] defines a new Extended Community, called the Cost Community, which can be used in tie breaking during the best path selection process. The Cost Community can be reused by the MPLS path programming to set the "Point of Insertion" as 128 to make the route advertised by the central controller to be chosen.

7. Destination Node Attribute

This document defines and uses a new BGP attribute called as the "Destination Node attribute" which Type value is to be assigned by IANA. The Destination Node attribute is an optional non-transitive attribute that can be applied to any address family.

The Destination Node attribute is used to carry a list of node addresses, which are intended to be used to determine the nodes where the route with such attribute SHOULD be considered. If a node receives a BGP route with a Destination Node attribute, it MUST check the node address list. If one address of the list belongs to this node, the route MUST be used in this node. Otherwise the route MUST be ignored silently.

The format of this attribute is defined 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|              AFI              |      SAFI     |    Reserved   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                                                               ~
~               Destination Node Address List                   ~
~                                                               ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

AFI: Address Family Identifier (16 bits).

SAFI: Subsequent Address Family Identifier (8 bits).

Reserved: One octet reserved for special flags

Destination Node Address List: The list of IPv4 (AFI=1) or IPv6 (AFI=2) address.

8. Capability Negotiation

It is necessary to negotiate the capability to support MPLS path programming. The MPLS-Path-Programming Capability is a new BGP capability [RFC5492]. The Capability Code for this capability is to be specified by the IANA. 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)  |
+--------------------------------------------------+
|  Send/Receive (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].

Send/Receive: This field indicates whether the sender is (a) willing to receive programming MPLS paths from its peer (value 1), (b) would like to send programming MPLS paths to its peer (value 2), or (c) both (value 3) for the <AFI, SAFI>.

9. Acknowledgments

The authors of this document would like to thank Lucy Yong, Susan Hares, Eric Wu, Weiguo Hao, Pingan Li, Zhengqiang Li and Jie Dong for their reviews and comments of this document.

10. IANA Considerations

TBD.

11. Security Considerations

The security considerations of [RFC4271] and [RFC5575] are applicable.

12. References

12.1. Normative 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-01 (work in progress), March 2016.

[I-D.ietf-idr-custom-decision]
Retana, A. and R. White, "BGP Custom Decision Process", draft-ietf-idr-custom-decision-07 (work in progress), November 2015.

[I-D.ietf-idr-tunnel-encaps]
Rosen, E., Patel, K., and G. Velde, "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-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>.

[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>.

[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>.

[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>.

12.2. Informative References

[I-D.ietf-pce-pce-initiated-lsp] Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE Model", draft-ietf-pce-pce-initiated-lsp-07 (work in progress), July 2016.

[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>.

[RFC6790] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", RFC 6790, DOI 10.17487/RFC6790, November 2012, <http://www.rfc-editor.org/info/rfc6790>.

Authors' Addresses

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

Email: lizhenbin@huawei.com

Shunwan Zhuang
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: zhuangshunwan@huawei.com

Sujian Lu
Tencent
Tengyun Building,Tower A ,No. 397 Tianlin Road
Shanghai, Xuhui District 200233
China

Email: jasonlu@tencent.com

Idr Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 1, 2016

Q. Liang
J. You
Huawei
R. Raszuk
Nozomi
D. Ma
Cisco Systems

September 29, 2015

Carrying Label Information for BGP FlowSpec
draft-liang-idr-bgp-flowspec-label-01

Abstract

This document specifies a method in which the label mapping information for a particular FlowSpec rule is piggybacked in the same Border Gateway Protocol (BGP) Update message that is used to distribute the FlowSpec rule. Based on the proposed method, the Label Switching Routers (LSRs) (except the ingress LSR) on the Label Switched Path (LSP) can use label to indentify the traffic matching a particular FlowSpec rule; this facilitates monitoring and traffic statistics for FlowSpec rules.

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].

Copyright Notice

Table of Contents

1. Introduction

[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 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.

In BGP VPN/MPLS networks, when FlowSpec rules on multiple forwarding devices in the network bound with labels form one or more LSPs, only the ingress LSR (Label Switching Router) needs to identify a particular traffic flow based on the matching criteria and then 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.

Though the FlowSpec rule could use the label(s) bound with the best-match unicast route for the destination prefix embedded in the FlowSpec rule or the best-match route to the target IP in the 'redirect to IP' action, this way means that if two or more FlowSpec rules have the same best-match unicast route for the embedded destination prefix or the same best-match route to target IP in the 'redirect to IP' action; they would be mapped to the same label. This would affect monitoring and traffic statistics facilities, because each FlowSpec rule requires an independent statistic and log data, which is described in Section 9 [RFC5575]. The LSRs (except the ingress LSR) on the LSP can use label to indentify the traffic matching a particular FlowSpec rule; this facilitates monitoring and traffic statistics for FlowSpec rules.

So this document proposes that the BGP router supports to allocate a label to one or more FlowSpec rule(s), the forwarding path is still decided by the best-match unicast route for the embedded destination prefix or the best-match route to target IP in the 'redirect to IP' action. Figure 1 gives an example that FlowSpec rule bound with a label is disseminated in the network.

```
               Option-B inter-AS connection
               |<------AS1----->|    |<------AS2----->|
               +-----+    +-----+    +-----+    +-----+
    VPN 1,IP1..| PE1 |====|ASBR1|----|ASBR2|====| PE2 |..VPN1,IP2
               +-----+    +-----+    +-----+    +-----+
                  | LDP LSP1 |          | LDP LSP2 |
                  | -------> |          | -------> |
                  |-------BGP VPN Flowspec LSP---->|
              (Label1)    (Label2)   (Label3)   (Label4)
                 Figure 1: Usage of FlowSpec with Label
```

```
FlowSpec rule1 (injected in PE2):
     Filters:
        destination ip prefix:IP2/32
        source ip prefix:IP1/32
     Actions:
        traffic-marking: 1

Labels allocated for FlowSpec1:
     Label4 allocated by PE2
     Label3 allocated by ASBR2
     Label2 allocated by ASBR1
     Label1 allocated by PE1

PE2 disseminates the FlowSpec1 bound with Label4 to ASBR2.
ASBR2 disseminates the FlowSpec1 bound with Label3 to ASBR1.
ASBR1 disseminates the FlowSpec1 bound with Label2 to PE1.

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(--)
```

So ASBR1 can do traffic statistics for FlowSpec rule 1 based on Label2; ASBR2 can do it based on Label3; and PE2 can do it based on Label4.

2. Terminology

This section contains definitions of terms used in this document.

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.

3. Protocol Extensions

In this document, BGP is used to distribute the FlowSpec rule bound with label(s). 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

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 |         Reserved      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+ Label
|                Label                  | Exp |S|       TTL     | Stack
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+ Entry
```

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

Type: the same as defined in [RFC4360]

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

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.

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.

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.

Label: the same as defined in [RFC3032].

A FlowSpec rule MAY include one or more ordering label-action(s). The arrival order of the label-actions decides the action order.

If the BGP router allocates a label for a FlowSpec rule and disseminates the labeled FlowSpec rule to the upstream peers, it can use the label to match the traffic identified by the FlowSpec rule in the forwarding plane.

4. IANA Considerations

For the purpose of this work, IANA should allocate value for the type of label-action:

TBD1 for label-action

5. Security considerations

This extension to BGP does not change the underlying security issues inherent in the existing BGP.

6. Acknowledgement

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

7. Normative 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.

[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>.

[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>.

[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>.

Authors' Addresses

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

Email: liangqiandeng@huawei.com

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

Email: youjianjie@huawei.com

Robert Raszuk
Nozomi

Email: robert@raszuk.net

Dan Ma
Cisco Systems

Email: danma@cisco.com

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

Q. Liang
S. Hares
J. You
Huawei
R. Raszuk
Nozomi
D. Ma
Cisco Systems

March 21, 2016

Carrying Label Information for BGP FlowSpec
draft-liang-idr-bgp-flowspec-label-02

Abstract

This document specifies a method in which the label mapping information for a particular FlowSpec rule is piggybacked in the same Border Gateway Protocol (BGP) Update message that is used to distribute the FlowSpec rule. Based on the proposed method, the Label Switching Routers (LSRs) (except the ingress LSR) on the Label Switched Path (LSP) can use label to indentify the traffic matching a particular FlowSpec rule; this facilitates monitoring and traffic statistics for FlowSpec rules.

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].

Copyright Notice

Table of Contents

1. Introduction

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

This section contains definitions of terms used in this document.

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.

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).

[I-D.yong-idr-flowspec-mpls-match] provides a match BGP-FS that may be used with this action to match and direct MPLS packets.

Example of Use:

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
```

```
           |<------AS1----->|    |<------AS2----->|
           +-----+    +-----+    +-----+    +-----+
VPN 1,IP1..| PE1 |====|ASBR1|----|ASBR2|====| PE2 |..VPN1,IP2
           +-----+    +-----+    +-----+    +-----+
             | LDP LSP1 |          | LDP LSP2 |
             | -------> |          | -------> |
             |-------BGP VPN Flowspec LSP---->|
          (Label1)    (Label2)   (Label3)   (Label4)
```

Figure 1: Usage of FlowSpec with Label

BGP-FS rule1 (locally configured):

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

Actions: Extended Communities
   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]
```

```
   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 (NRLI + 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]
```

4. Protocol Extensions

In this document, BGP is used to distribute the FlowSpec rule bound with label(s). 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

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
 |                Label                  | Exp |S|       TTL     | Stack
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+ Entry
```

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

Type: the same as defined in [RFC4360]

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

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.

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: A FlowSpec rule MAY include one or more ordering label-action(s). If multiple label action extended communities are associated with a BGP-FS Rule, this gives the order of this in the list. The Last 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. IANA Considerations

For the purpose of this work, IANA should allocate the following Extended community:

TBD1 for label-action

6. Security considerations

This extension to BGP does not change the underlying security issues inherent in the existing BGP.

7. Acknowledgement

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

8. References

8.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <http://www.rfc-editor.org/info/rfc2119>.

[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>.

[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>.

[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>.

8.2. 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.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., Liang , Q., and J. You , "Dissemination of Flow Specification Rules 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
Nozomi

Email: robert@raszuk.net

Dan Ma
Cisco Systems

Email: danma@cisco.com

IDR Working Group Q. Liang
Internet-Draft J. You
Intended status: Standards Track S. Zhuang
Expires: April 20, 2016 Huawei Technologies
October 18, 2015

BGP FlowSpec with Time Constraints
draft-liang-idr-bgp-flowspec-time-00

Abstract

The BGP flow specification (FlowSpec) is an additional tool to mitigate the effects of Distributed Denial of Service (DDoS) attacks. Since DDoS attacks are dynamic, filtering of a flow may only be necessary for some specified time, and be undesirable at other times. This document proposes a new BGP path attribute called "Flow Extended Attribute", which carries expected valid period information for a FlowSpec rule. So network administrators can control certain types of traffic in a specified period.

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].

Table of Contents

1. Introduction

The BGP flow specification (FlowSpec) defined in [RFC5575] is an n-tuple consisting of several matching criteria, which gives network operators an additional tool to mitigate the effects of Distributed Denial of Service (DDoS) attacks on their networks. 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 flow actions, such as rate limit, redirect, and marking, associated with each flow specification rule. A Border Gateway Protocol 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 specification rules.

Since DDoS attacks are dynamic, redirection or filtering of a flow may only be necessary for some specified time, and be undesirable at other times [I-D.eddy-idr-flowspec-exp]. Thus, network administrators may only need to control certain types of traffic in a specified period; they can configure or inject a FlowSpec rule with a valid period, which determines when the said FlowSpec rule is effective. There's another use case for this usage, for example, the network administrator may need to ensure reliable transmission for high priority services (e.g. video traffic) for VIP and limit the bandwidth for low priority services (e.g. web browsing) for common users during peak network utilization periods.

The current BGP FlowSpec protocol cannot support to control the valid period of a FlowSpec rule precisely in the network. For example, the network administrator may want to validate a FlowSpec rule on different BGP routers simultaneously; firstly the rule should be disseminated to those BGP routers. But since those BGP routers would receive this FlowSpec rule with different delay, the FlowSpec rule may be valid at different time slightly. Therefore the BGP router can specify a time parameter as the valid period when installing a FlowSpec rule.

This document proposes a new BGP path attribute called "Flow Extended Attribute", which carries expected valid period information for a FlowSpec rule. Besides, in order to make the FlowSpec rule more readable in diagnosing and logging, the "Flow Extended Attribute" can also carry the flow description information for the FlowSpec rule.

2. Terminology

This section contains definitions of terms used in this document.

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

3. Protocol Extensions

In this document, BGP is used to distribute FlowSpec rules bound with a "Flow Extended Attribute". This "Flow Extended Attribute" is an optional transitive attribute that is composed of a set of Type-Length-Value (TLV) encodings, including Flow Description sub-TLV and Flow Validation Period sub-TLV.

```
        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 Octets)         |        Length (2 Octets)      |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |                                                               |
       |                             Value                             |
       |                                                               |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 1:TLV Format

3.1. Flow Description sub-TLV

The Flow Description sub-TLV is encoded as 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
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |      1 (Flow Description)     |        Length (2 Octets)      |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |                                                               |
       ~                Flow Description (variable Octets)             ~
       |                                                               |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2:Flow Description sub-TLV Format

Type: Flow Description (Type Code: 1)

Length: the size of the value field (typically in bytes)

Flow Description: This field is an ASCII string, padded on the right with null bytes (\0). It is usually used as a flow name or flow function description. The length of this field SHOULD be no more than 256 octets.

3.2. Flow Validity Period sub-TLV

The Flow Validity Period sub-TLV is encoded as 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
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |   2 (Flow Validity Period)    |        Length (2 Octets)      |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       | Starting Time Type (2 Octets) |  Duration Type (2 Octets)     |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |                  Starting Time (seconds)                      |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |                Starting Time (microseconds)                   |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |                   Duration (seconds)                          |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |                 Duration (microseconds)                       |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |                   Delay Time (seconds)                        |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |                 Delay Time (microseconds)                     |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |                  Periodic Time (seconds)                      |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |                Periodic Time (microseconds)                   |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 3:Flow Validity Period sub-TLV Format

Type:Flow Validity Period (Type Code: 2)

Length: the size of the value field (typically in bytes)

Starting Time Type:

Type Code	Function Description
0	Immediate validation
1	Delayed validation
2	Timing validation
else codes	Reserved

When the "Starting Time Type" is set to 2, the BGP Speaker should be clock synchronized [I-D.litkowski-idr-bgp-timestamp].

Duration Type:

Type Code	Function Description
0	Permanent validation
1	Hard invalidation
2	Idle invalidation
else codes	Reserved

When the "Duration Type" is set to 0, the corresponding FlowSpec rule is always valid until it is withdrawn by BGP signaling. When the "Duration Type" is set to 1, the corresponding FlowSpec rule is only valid in a specified duration defined by the "Duration" field. When the "Duration Type" is set to 2, the corresponding FlowSpec rule is valid until no traffic has matched it for a duration defined by the "Duration" field.

Starting Time: Expressed in seconds and microseconds since midnight (zero hour), January 1, 1970 (UTC). Precision of the "Starting Time" is implementation-dependent. If the "Starting Time Type" is set to 0, this field is invalid.

Duration: if the "Duration Type" is set to 0, this field is invalid.

Delay Time: Only when the "Starting Time Type" is set to 1, this field is valid. If the "Starting Time" is set to a valid value,the first valid period of the FlowSpec rule bound with this "Flow Extended Attribute" is [Starting Time + Delay, Starting Time + Delay + Duration]; if not, and assuming that the current time of the BGP router is T1, then the first valid period of the FlowSpec rule bound with this "Flow Extended Attribute" is [T1 + Delay, T1 + Delay + Duration].

Periodic Time: If zero, the value is unavailable. The FlowSpec rule bound with this "Flow Extended Attribute" would be valid periodically. The "Periodic Time" MUST be not less than the "Duration", otherwise this sub-TLV is invalid.

The BGP router may not actively withdraw a FlowSpec rule, which has been invalid. However, it should withdraw a FlowSpec rule according to the BGP signaling normally.

4. IANA Considerations

For the purpose of this work, IANA should allocate a new code from the "BGP Path Attributes" Registry to "BGP Flow Extended Attribute".

IANA is requested to change the registration policy of the "BGP Flow Extended Attribute Sub-TLVs" registry to the following:

- o The values 0 and 255 are reserved.
- o The values in the range 1-127 are to be allocated using the "Standards Action" registration procedure.
- o The values in the range 128-251 are to be allocated using the "First Come, First Served" registration procedure.
- o The values in the range 252-254 are reserved for experimental use;

IANA shall not allocate values from this range.

IANA is requested to assign a code point from the "BGP Flow Extended Attribute Sub-TLVs" registry for "Flow Description", with this document being the reference.

IANA is requested to assign a code point from the "BGP Flow Extended Attribute Sub-TLVs" registry for "Flow Validity Period", with this document being the reference.

5. Security Considerations

This extension to BGP does not change the underlying security issues inherent in the existing BGP.

6. Acknowledgements

The authors would like to thank Zhenbin Li and Weiguo Hao for their comments.

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>.

[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. Informative References

[I-D.eddy-idr-flowspec-exp]
Eddy, W., Dailey, J., and G. Clark, "Experimental BGP Flow Specification Extensions", draft-eddy-idr-flowspec-exp-00 (work in progress), August 2015.

[I-D.ietf-idr-tunnel-encaps]
Rosen, E., Patel, K., and G. Velde, "Using the BGP Tunnel Encapsulation Attribute without the BGP Encapsulation SAFI", draft-ietf-idr-tunnel-encaps-00 (work in progress), August 2015.

[I-D.litkowski-idr-bgp-timestamp]
Litkowski, S., Patel, K., and J. Haas, "Timestamp support for BGP paths", draft-litkowski-idr-bgp-timestamp-02 (work in progress), March 2015.

Authors' Addresses

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

Email: liangqiandeng@huawei.com

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

Email: youjianjie@huawei.com

Shunwan Zhuang
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: zhuangshunwan@huawei.com

Routing Area Working Group S. Litkowski
Internet-Draft Orange
Intended status: Standards Track A. Simpson
Expires: June 10, 2016 Alcatel Lucent
K. Patel
Cisco
J. Haas
Juniper Networks
December 8, 2015

Applying BGP flowspec rules on a specific interface set
draft-litkowski-idr-flowspec-interfaceset-03

Abstract

BGP Flow-spec is an extension to BGP that allows for the dissemination of traffic flow specification rules. The primary application of this extension is DDoS mitigation where the flowspec rules are applied in most cases to all peering routers of the network.

This document will present another use case of BGP Flow-spec where flow specifications are used to maintain some access control lists at network boundary. BGP Flowspec is a very efficient distributing machinery that can help in saving OPEX while deploying/updating ACLs. This new application requires flow specification rules to be applied only on a specific subset of interfaces and in a specific direction.

The current specification of BGP Flow-spec does not detail where the flow specification rules need to be applied.

This document presents a new interface-set flowspec action that will be used in complement of other actions (marking, rate-limiting ...). The purpose of this extension is to inform remote routers on where to apply the flow specification.

This extension can also be used in a DDoS mitigation context where a provider wants to apply the filtering only on specific peers.

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

Copyright Notice

Table of Contents

1. Use case

1.1. Specific filtering for DDoS

```
                       ----------------    --- (ebgp) - Peer3 (BW 10G)
                      /                  \/
                     |                   /|
                     |                PE --- (ebgp) - Transit1(BW 4x10G)
 Cust1 --- (ebgp) --- PE                  |
                     |                PE ---- (ebgp) - Peer2 (BW 4*10G)
                     |                   \|
 Cust2 --- (ebgp) --- PE                  |----- (ebgp) - Customer3
                    /|                    |
 Peer1(BW10G)-(ebgp) |                PE --- (ebgp) - Transit2(BW 4x10G)
                     |                    |
                      \                  /
                       -----------------
```

Figure 1

The figure 1 above displays a typical service provider Internet network owing Customers, Peers and Transit. To protect pro actively against some attacks (e.g. DNS, NTP ...), the service provider may want to deploy some rate-limiting of some flows on peers and transit links. But depending on link bandwidth, the provider may want to apply different rate-limiting values.

For 4*10G links peer/transit, it may want to apply a rate-limiting of DNS flows of 1G, while on 10G links, the rate-limiting would be set to 250Mbps. Customer interfaces must not be rate-limited.

BGP Flow-spec infrastructure may already be present on the network, and all PEs may have a BGP session running flowspec address family. The Flowspec infrastructure may be reused by the service provider to implement such rate-limiting in a very quick manner and being able to adjust values in future quickly without having to configure each node one by one. Using the current BGP flowspec specification, it would not be possible to implement different rate limiter on different interfaces of a same router. The flowspec rule is applied to all interfaces in all directions or on some interfaces where flowspec is activated but flowspec rule set would be the same among all interfaces.

Section Section 3 will detail a solution to address this use case using BGP Flowspec.

1.2. ACL maintenance

```
                              ---------------    --- (ebgp) - Cust4_VPN
                             /                 \/
    Cust1_INT -- (ebgp) --- PE                 /|
                            |              PE ------ (ebgp) - Transit1
    Cust3_VPN -- (ebgp) --- PE                  |
                            |              PE ------ (ebgp) - Peer2
                            |                  \|
    Cust2_INT -- (ebgp) --- PE                  |----- (ebgp) - Cust4_INT
                           /|                   |
    Peer1 ------ (ebgp) --  |              PE ------ (ebgp) - Transit2
                            |                   |
                             \                 /
                              ----------------
```

Figure 2

The figure 1 above displays a typical service provider multiservice network owing Customers, Peers and Transit for Internet, as well as VPN services. The service provider requires to ensure security of its infrastructure by applying ACLs at network boundary. Maintaining and deploying ACLs on hundreds/thousands of routers is really painful and time consuming and a service provider would be interested to deploy/updates ACLs using BGP Flowspec. In this scenario, depending on the interface type (Internet customer, VPN customer, Peer, Transit ...) the content of the ACL may be different.

We foresee two main cases :

o Maintaining complete ACLs using flowspec : in this case all the ingress ACL are maintained and deployed using BGPFlowspec. See section Section 8 for more details on security aspects.

o Requirement of a quick deployment of a new filtering term due to a security alert : new security alerts often requires a fast deployment of new ACL terms. Using traditional CLI and hop by hop provisioning, such deployment takes time and network is unprotected during this time window. Using BGP flowspec to deploy such rule, a service provider can protect its network in few seconds. Then the SP can decide to keep the rule permanently in BGP Flowspec or update its ACL or remove the entry (in case equipments are not vulnerable anymore).

Section Section 3 will detail a solution to address this use case using BGP Flowspec.

2. Collaborative filtering and managing filter direction

[RFC5575] states in Section 5. : "This mechanism is primarily designed to allow an upstream autonomous system to perform inbound filtering in their ingress routers of traffic that a given downstream AS wishes to drop.".

In case of networks collaborating in filtering, there is a use case for performing outbound filtering. Outbound filtering allows to apply traffic action one step before and so may allow to prevent impact like congestions.

```
  ----------------------
/                        \
|   Upstream provider     |
\                        /
  -----------------------
    |                |
    |P2              |P1
  ----------------------
/                        \
|        MyAS             |
\                        /
  -----------------------
```

Figure 3

In the figure above, MyAS is connected to an upstream provider. If a malicious traffic comes in from the upstream provider, it may congestion P1 or P2 links. If MyAS apply inbound filtering on P1/P2 using BGP Flowspec, the congestion issue will not be solved.

Using collaborative filtering, the upstream provider may propose to MyAS to filter malicious traffic sent to it. We propose to enhance [RFC5575] to make myAS able to send BGP FlowSpec updates (on eBGP sessions) to the upstream provider to request outbound filtering on peering interfaces towards MyAS. When the upstream provider will receive the BGP Flowspec update from MyAS, the BGP flowspec update will contain request for outbound filtering on a specific set of interfaces. The upstream provider will apply automatically the requested filter and congestion will be prevented.

3. Interface specific filtering using BGP flowspec

The use case detailed above requires application of different BGP Flowspec rules on different set of interfaces. The basic specification detailed in [RFC5575] does not address this and does not give any detail on where the FlowSpec filter need to be applied.

We propose to introduce, within BGP Flowspec, an identification of interfaces where a particular filter should apply on. Identification of interfaces within BGP Flowspec will be done through group identifiers. A group identifier marks a set of interfaces sharing a common administrative property. Like a BGP community, the group identifier itself does not have any significance. It is up to the network administrator to associate a particular meaning to a group identifier value (e.g. group ID#1 associated to Internet customer interfaces). The group identifier is a local interface property. Any interface may be associated with one or more group identifiers using manual configuration.

When a filtering rule advertised through BGP Flowspec must be applied only to particular sets of interfaces, the BGP Flowspec BGP update will contain the identifiers associated with the relevant sets of interfaces. In addition to the group identifiers, it will also contain the direction the filtering rule must be applied in (see Section 4).

Configuration of group identifiers associated to interfaces may change over time. An implementation MUST ensure that the filtering rules (learned from BGP Flowspec) applied to a particular interface are always updated when the group identifier mapping is changing.

Considering figure 2, we can imagine the following design :

- Internet customer interfaces are associated with group-identifier 1.
- VPN customer interfaces are associated with group-identifier 2.
- All customer interfaces are associated with group-identifier 3.
- Peer interfaces are associated with group-identifier 4.
- Transit interfaces are associated with group-identifier 5.
- All external provider interfaces are associated with group-identifier 6.
- All interfaces are associated with group-identifier 7.

If the service provider wants to deploy a specific inbound filtering on external provider interfaces only, the provider can send the BGP flow specification using group-identifier 6 and including inbound direction.

There are some cases where nodes are dedicated to specific functions (Internet peering, Internet Edge, VPN Edge, Service Edge ...), in this kind of scenario, there is an interest for a constrained distribution of filtering rules that are using the interface specific filtering. Without the constrained route distribution, all nodes will received all the filters even if they are not interested in those filters. Constrained route distribution of flowspec filters would allow for a more optimized distribution.

4. Interface-set extended community

This document proposes a new BGP Route Target extended community called "flowspec interface-set". This document so expands the definition of the Route Target extended community to allow a new value of high order octet (Type field) to be TBD (in addition to the values specified in [RFC4360]).

In order to ease intra-AS and inter-AS use cases, this document proposes to have a transitive as well as a non transitive version of this extended community.

This new BGP Route Target extended community is encoded as follows :

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Type (TBD)   |     0x02      |   Autonomous System Number    :
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
:     AS Number (cont.)         |O|I|      Group Identifier     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

The flags are :

- O : if set, the flow specification rule MUST be applied in outbound direction to the interface set referenced by the following group-identifier.

- I : if set, the flow specification rule MUST be applied in input direction to the interface set referenced by the following group-identifier.

Both flags can be set at the same time in the interface-set extended community leading to flow rule to be applied in both directions. An interface-set extended community with both flags set to zero MUST be treated as an error and as consequence, the FlowSpec update MUST be discarded.

The Group Identifier is coded as a 14-bit number (values goes from 0 to 16383).

Multiple instances of the interface-set community may be present in a BGP update. This may appear if the flow rule need to be applied to multiple set of interfaces.

Multiple instances of the community in a BGP update MUST be interpreted as a "OR" operation : if a BGP update contains two interface-set communities with group ID 1 and group ID 2, the filter would need to be installed on interfaces belonging to Group ID 1 or Group ID 2.

As using a Route Target, route distribution of flowspec NLRI with interface-set may be subject to constrained distribution as defined in [RFC4684]. Constrained route distribution for flowspec routes using interface-set requires discussion and will be addressed in a future revision of the document.

5. Interaction with permanent traffic actions

[RFC5575] states that BGP Flowspec is primarily designed to allow upstream AS to perform inbound filtering in their ingress routers. This specification does not precise where this ingress filtering should happen in the packet processing pipe.

This proposal enhances [RFC5575] in order to add action on traffic coming from or going to specific interfaces. Based on this enhancement, some new requirements come to implementations.

An implementation SHOULD apply input actions (I bit set) within the input packet processing pipe. An implementation SHOULD apply output actions (O bit set) within the output packet processing pipe.

As input and output processing pipes may also involve already present static/permanent features that will manipulate the packet, the next sections will try to clarify how the static behaviors should interact will BGP flowspec actions.

5.1. Interaction with interface ACLs

Deploying interface specific filters using BGP FlowSpec (dynamic entries) may interfere with existing permanent interface ACL (static entries). The content of the existing permanent ACL MUST NOT be altered by dynamic entries coming from BGP FlowSpec. Permanent ACLs are using a specific ordering which is not compatible with the ordering of FS rules and misordering of ACL may lead to undesirable behaviour. In order, to keep a deterministic and well known behaviour, an implementation SHOULD process the BGP FlowSpec ACL as follows :

- o In inbound direction, the permanent ACL action is applied first followed by FlowSpec action. This gives the primary action to the permanent ACL as it is done today.

- o In outbound direction, FlowSpec action action is applied first followed by permanent ACL. This gives the final action to the permanent ACL as it is done today.

```
                         Inbound filters                         Outbound filter
s
           ---------     -------     ----------     -------     ---------
          |Permanent| -> |Dynamic| -> |Forwarding| -> |Dynamic| -> |Permanent|
```

In order for a flow to be accepted, the flow must be accepted by the two ACLs and a flow is rejected when one of the ACL rejects it as described in the table below :

Permanent ACL entry action	FlowSpec ACL entry action	Result action
Drop	Drop	Drop
Drop	Accept	Drop
Accept	Drop	Drop
Accept	Accept	Accept

Example :

- o ACL permanent IN :

  - * Entry 1 : permit udp from 10/8 to 11/8 port 53

  - * Entry 2 : permit tcp from 10/8 to 11/8 port 22

   * Entry 3 : deny ip from 10/8 to 11/8

- ACL dynamic FlowSpec IN :

   * Entry 1 : deny udp from 10.0.0.1/32 to 11/8 port 53

   * Entry 2 : permit tcp from 10/8 to 11/8 port 80

In the example above :

- a UDP flow from 10.0.0.1 to 11.0.0.2 on port 53 will be rejected because the dynamic ACL rejects it.

- a UDP flow from 10.0.0.2 to 11.0.0.2 on port 53 will be accepted because both ACLs accept it.

- a TCP flow from 10.0.0.2 to 11.0.0.2 on port 80 will be rejected because permanent ACL rejects it.

5.2. Interaction with flow collection

A router may activate flow collection features (used in collaboration with Netflow export). Flow collection can be done at input side or output side. As for ACL, an implementation SHOULD process :

- BGP FS rules after the inbound flow collection : in case of DDoS protection, it is important to keep monitoring of attack flows and so performing action, after collection.

- BGP FS rules before the outbound flow collection : purpose of outbound flow collection is really to track flows that are exiting the interface. BGP FS rules should not interfere in this.

```
        Inbound                                                Outbound
        Flow          BGP                            BGP         Flow
                collection    FS                             FS           collecti
on
           ---------      -------      ----------      -------      ---------
          |Permanent| -> |Dynamic| -> |Forwarding| -> |Dynamic| -> |Permanent|
```

6. Scaling of per interface rules

Creating rules that are applied on specific interfaces may create forwarding rules that may be harder to share.

An implementation SHOULD take care about trying to keep sharing forwarding structures as much as possible in order to limit the scaling impact. How the implementation would do so is out of scope of the document.

7. Deployment considerations

There are some cases where a particular BGP Flowspec NLRI may be advertised to different interface groups with a different action. For example, a service provider may want to discard all ICMP traffic from customer interfaces to infrastructure addresses and want to rate-limit the same traffic when it comes from some internal platforms. These particular cases require ADD-PATH to be deployed in order to ensure that all paths (NLRI+interface group+actions) are propagated within the BGP control plane. Without ADD-PATH, only a single "NLRI+interface group+actions" will be propagated, so some filtering rules will never be applied.

8. Security Considerations

Managing permanent Access Control List by using BGP Flowspec as described in Section 1.2 helps in saving roll out time of such ACL. However some ACL especially at network boundary are critical for the network security and loosing the ACL configuration may lead to network open for attackers.

By design, BGP flowspec rules are ephemeral : the flow rule exists in the router while the BGP session is UP and the BGP path for the rule is valid. We can imagine a scenario where a Service Provider is managing the network boundary ACLs by using only FlowSpec. In this scenario, if , for example, an attacker succeed to make the internal BGP session of a router to be down , it can open all boundary ACLs on the node, as flowspec rules will disappear due to the BGP session down.

In reality, the chance for such attack to occur is low, as boundary ACLs should protect the BGP session from being attacked.

In order to complement the BGP flowspec solution is such deployment scenario and provides security against such attack, a service provider may activate Long lived Graceful Restart [I-D.uttaro-idr-bgp-persistence] on the BGP session owning Flowspec address family. So in case of BGP session to be down, the BGP paths of Flowspec rules would be retained and the flowspec action will be retained.

9. Acknowledgements

Authors would like to thanks Wim Hendrickx for his valuable comments.

10. IANA Considerations

This document requests a new type from the "BGP Transitive Extended Community Types" extended community registry. This type name shall be 'FlowSpec'.

This document requests a new type from the "BGP Non-Transitive Extended Community Types" extended community registry. This type name shall be 'FlowSpec'.

This document requests creation of a new registry called "FlowSpec Extended Community Sub-Types". This registry contains values of the second octet (the "Sub-Type" field) of an extended community when the value of the first octet (the "Type" field) is to one of those allocated in this document.

Within this new registry, this document requests a new subtype (suggested value 0x02), this sub-type shall be named "interface-set".

11. References

11.1. Normative References

[I-D.ietf-idr-rtc-no-rt]
Rosen, E., Patel, K., Haas, J., and R. Raszuk, "Route Target Constrained Distribution of Routes with no Route Targets", draft-ietf-idr-rtc-no-rt-04 (work in progress), November 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>.

[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>.

[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>.

11.2. Informative References

[I-D.uttaro-idr-bgp-persistence]
Uttaro, J., Chen, E., Decraene, B., and J. Scudder, "Support for Long-lived BGP Graceful Restart", draft-uttaro-idr-bgp-persistence-03 (work in progress), November 2013.

Authors’ Addresses

Stephane Litkowski
Orange

Email: stephane.litkowski@orange.com

Adam Simpson
Alcatel Lucent

Email: adam.simpson@alcatel-lucent.com

Keyur Patel
Cisco

Email: keyupate@cisco.com

Jeff Haas
Juniper Networks

Email: jhaas@juniper.net

IDR Working Group G. Van de Velde, Ed.
Internet-Draft W. Henderickx
Intended status: Standards Track Nokia
Expires: January 9, 2017 K. Patel
A. Sreekantiah
Cisco Systems
Z. Li
S. Zhuang
N. Wu
Huawei Technologies
July 8, 2016

Flowspec Indirection-id Redirect
draft-vandevelde-idr-flowspec-path-redirect-03

Abstract

Flowspec 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, particularly if the redirected traffic needs to be steered into an engineered path or into a service plane.

This document defines a new extended community known as redirect-to-indirection-id (32-bit) flowspec action to provide advanced redirection capabilities on flowspec clients. When activated, the flowspec extended community is used by a flowspec client to find the correct next-hop entry within a localised indirection-id mapping table.

The functionality present in this draft allows a network controller to decouple flowspec functionality from the creation and maintainance of the network's service plane itself including the setup of tunnels and other service constructs that could be managed by other network devices.

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

Copyright Notice

Table of Contents

1. Introduction

Flowspec RFC5575 [2] is an extension to BGP that allows for the dissemination of traffic flow specification rules. This has many possible applications, however the primary one for many network operators is the distribution of traffic filtering actions for DDoS mitigation.

Every flowspec policy 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 communities). 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 to steer traffic towards an alternate destination is useful for DDoS mitigation but using this technology can be cumbersome when there is need to steer the traffic onto an engineered traffic path.

This draft proposes a new redirect-to-indirection-id flowspec action facilitating an anchor point for policy-based forwarding onto an engineered path or into a service plane. The flowspec client consuming and utilizing the new flowspec indirection-id extended-community finds the redirection information within a localised indirection-id mapping table. The localised mapping table is a table construct providing at one side the table key and at the other side next-hop information. The table key consists out the combination of indirection-id type and indirection-id 32-bit value.

The redirect-to-indirection-id flowspec action is encoded in a newly defined BGP extended community. In addition, the type of redirection can be configured as an extended community indirection-id type field.

This draft defines the indirection-id extended-community and the wellknown indirection-id types. The specific solution to construct the localised indirection-id mapping table are out-of-scope of this document.

2. indirection-id and indirection-id table

An indirection-id is an abstract number (32-bit value) used as identifier for a localised indirection decision. The indirection-id will allow a flowspec client to redirect traffic into a service plane or onto an engineered traffic path. e.g. When a BGP flowspec controller signals a flowspec client the indirection-id extended community, then the flowspec client uses the indirection-id to make a recursive lookup to retrieve next-hop information found in a localised indirection mapping table.

The indirection-id table is a router localised table. The indirection-id table is constructed out of table keys mapped to flowspec client localised redirection information. The table key is created by the combination of the indirection-id type and the indirection-id 32-bit value. Each entry in the indirection-table key maps to sufficient information (parameters regarding encapsulation, interface, QoS, etc...) to successfully redirect traffic.

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 to send a single flowspec policy route (i.e. flowspec_route#1) to many BGP flowspec clients. This flowspec route signals the Flowspec clients to redirect traffic onto a tunnel towards a single IP destination address.

For this first use-case scenario, the flowspec client receives from the flowspec controller a flowspec route (i.e. flowspec_route#1) including the redirect-to-indirection-id extended community. The redirect-to-indirection-id extended community contains the key (indirection-id type + indirection-id 32-bit value) to select the corresponding next-hop information from the flowpsec client localised indirection-id table. The resulting next-hop information for this use-case is a remote tunnel end-point IP address with accordingly sufficient tunnel encapsulation information to forward the packet accordingly.

For redirect to shortest path tunnel it is required that the tunnel MUST be operational and allow packets to be exchanged between tunnel head- and tail-end.

3.2. Redirection to path-engineered tunnels

For a second use-case, it is expected that the flowspec client redirect traffic matches the flowspec rule, onto a path engineered tunnel. The path engineered tunnel on the flowspec client SHOULD be created by out-of-band mechanisms. Each path engineered tunnel deployed for flowspec redirection, has a unque key as an identifier. consequently, the key (=indirection-id type and indirection-id 32-bit value) uniquely identifies a single path engineered tunnel on the flowspec client. The localised indirection-id mapping table is the collection of all keys corresponding all path engineered tunnels on the flowspec client.

For this second use-case scenario, the flowspec controller sends a flowspec route (i.e. flowspec_route#2) to the flowspec clients. The flowspec clients, respectively receive the flowspec route. The redirect-to-indirection-id extended community contains the key (indirection type + indirection-id 32-bit value) to select the corresponding next-hop information from the flowpsec client localised indirection-id table. The resulting next-hop information for this use-case is path engineered tunnel information and has sufficient tunnel encapsulation information to forward the packet according the expectations of the flowspec controller.

A concrete example of this use-case 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, while the indirection-id type references the Bindging SID semantic. 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 an 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-id field now encoding the associated value for the binding SID.

3.3. Redirection to Next-next-hop tunnels

A Third use-case is when a BGP Flowspec controller sends a single flowspec policy route to flowpsec clients to signal redirection towards next-next-hop tunnels. In this use-case The flowspec rule is instructing the Flowspec client to redirect traffic using a sequence of indirection-id extended communities. The sequence of indirection-ids is managed using Tunnel IDs (TID). i.e. a classic example would be DDoS mitigation towards Segment Routing Central Egress Path Engineering [4]. To steer DDoS traffic towards egress peer engineering paths, a first indirection-id will steer traffic onto a tunnel to an egress router, while a second indirection-id is used steer the traffic at this egress router onto a particular interface or towards a peer. The flowspec client will for this use-case dynamically append all segment routing segments to steer the DDoS traffic through the EPE path.

To achieve this type of redirection to next-next-hop tunnels, multiple indirection-ids, each using a unique Tunnel ID are imposed upon a the flowspec policy rule. The Tunnel ID will allow the flowspec client to sequence the indirection-ids for correct next-next-hop tunnel constructs.

4. Redirect to indirection-id Community

This document defines a new BGP extended community known as a Redirect-to-indirection-id extended community. This extended community is a new transitive extended community with the Type and the Sub-Type field to be assigned by IANA. The format of this extended community is show in Figure 1.

```
 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          | Sub-Type      | Flags(1 octet)| Indirection ID|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                  Generalized indirection_id                   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 1

The meaning of the extended community fields are as follows:

Type: 1 octet to be assigned by IANA.

Sub-Type: 1 octet to be assigned by IANA.

Flags: 1 octet field. Following Flags are defined.

```
 0             1
 0 1 2 3 4 5 6 7
+-+-+-+-+-+-+-+-+
| RES |  TID  |C|
+-+-+-+-+-+-+-+-+
```

Figure 2

The least-significant Flag 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 4 bit Table-id field. This field is used to provide the flowspec client an indication how and where to sequence the received indirection-ids to redirecting traffic. TID value 0 indicates that Table-id field is NOT set and SHOULD be ignored.

All bits other than the 'C' and 'TID' bits MUST be set to 0 by the originating BGP speaker and ignored by receiving BGP speakers.

Indirection ID: 1 octet value. This draft defines following indirection_id Types:

0 - Localised ID

1 - Node ID

2 - Agency ID

3 - AS (Autonomous System) ID

4 - Anycast ID

5 - Multicast ID

6 - Tunnel ID (Tunnel Binding ID)

7 - VPN ID

8 - OAM ID

9 - ECMP (Equal Cost Multi-Path) ID

10 - QoS ID

11 - Bandwidth-Guarantee ID

12 - Security ID

13 - Multi-Topology ID

5. Redirect using localised indirection-id mapping table

When a BGP speaker receives a flowspec policy route with a 'redirect to indirection-id' extended community and this route represents the one and only best path or an equal cost multipath, 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. To construct the local recursed path, the flowspec client does a local indirection-id mapping table lookup using the key comprised of the indirection-id 32-bit value and indirection-id type to retrieve the correct redirection information.

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 semenatics 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 emposed. 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 respective '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 overflow of traffic 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 type and sub-type for the Redirect to indirection-id Extended community from the "Transitive Extended community" registry. The Type name shall be "Redirect to indirection-id Extended Community" and the Sub-type name shall be 'Flow-spec Redirect to 32-bit Path-id'.

In addition, this document requests IANA to create a new registry for Redirect to indirection-id Extended Community INDIRECTION-IDs as follows:

Under "Transitive Extended Community:"

Registry: "Redirect Extended Community indirection_id"

Reference: [RFC-To-Be]

Registration Procedure(s): First Come, First Served

Registry: "Redirect Extended Community indirection_id"

Value	Code	Reference
0	Localised ID	[RFC-To-Be]
1	Node ID	[RFC-To-Be]
2	Agency ID	[RFC-To-Be]
3	AS (Autonomous System) ID	[RFC-To-Be]
4	Anycast ID	[RFC-To-Be]
5	Multicast ID	[RFC-To-Be]
6	Tunnel ID (Tunnel Binding ID)	[RFC-To-Be]
7	VPN ID	[RFC-To-Be]
8	OAM ID	[RFC-To-Be]
9	ECMP (Equal Cost Multi-Path) ID	[RFC-To-Be]
10	QoS ID	[RFC-To-Be]
11	Bandwidth-Guarantee ID	[RFC-To-Be]
12	Security ID	[RFC-To-Be]
13	Multi-Topology ID	[RFC-To-Be]

Figure 3

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)
Nokia
Antwerp
BE

Email: gunter.van_de_velde@nokia.com

Wim Henderickx
Nokia
Antwerp
BE

Email: wim.henderickx@nokia.com

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

Email: keyupate@cisco.com

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

Email: asreekan@cisco.com

Zhenbin Li
Huawei Technologies
Huawei Bld., No. 156 Beiquing Rd
Beijing 100095
China

Email: lizhenbin@huawei.com

Shunwan Zhuang
Huawei Technologies
Huawei Bld., No. 156 Beiquing Rd
Beijing 100095
China

Email: zhuangshunwan@huawei.com

Nan Wu
Huawei Technologies
Huawei Bld., No. 156 Beiquing Rd
Beijing 100095
China

Email: eric.wu@huawei.com

Network Working Group N. Wu
Internet-Draft Z. Zhuang
Intended status: Standards Track Huawei
Expires: April 21, 2016 October 19, 2015

BGP Extensions for Segment Allocation
draft-wu-idr-bgp-segment-allocation-ext-00

Abstract

This document defines extensions to the BGP-LS to distribute/push the segment information to its administrative SR domain and describes some use cases.

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

Copyright Notice

Table of Contents

1. Introduction

In those networks with a central controller, it may be beneficial to allocate and manage SIDs for the network since the controller has the whole link-state database in mind. This document proposes BGP extensions to allocate SIDs in a centralized manner instead of distribution way.

2. Terminology

- MPP: MPLS Path Programming
- RR: Route Reflector
- SID: Segment Identifier
- SR-Path: Segment Routing Path

3. Motivation

3.1. Allocating Segment in BGP Networks

It is possible that BGP may be the only routing protocol in some networks, such as the one described in [I-D.ietf-rtgwg-bgp-routing-large-dc]. If Segment Routing [I-D.ietf-spring-segment-routing] is going to be used for in the dataplane, it will be better to allocate SIDs in a centralized manner since no IGP flooding mechanism to advertise now.

In order to allocating SIDs, the centralized allocator SHOULD collect BGP network topology database ahead, which at least consists of BGP speakers, prefixes and adjacencies among them. No concrete technique for collecting this database has been specified in this document.

3.2. Allocating Segment in IGP Networks

In the scenario SR & LDP interoperation described in [I-D.ietf-spring-segment-routing-ldp-interop], if mapping entries are allocated in a centralized manner, e.g. a controller, it is possible that Binding SIDs will be populated to a designated SRMS through a protocol instead of IGP, no matter whether the SRMS is a dedicated server or function module.

4. Protocol Extensions

This section defines a new Protocol-ID called as BGP-Segment-Allocation (TBA) in the BGP-LS specification. The use of a new Protocol-ID allows separation and differentiation between the NLRIs carrying Segment Allocation information from the NLRIs carrying IGP link-state information as defined in [I-D.ietf-idr-ls-distribution].

4.1. Node NLRI for Segment Allocation

This section describes the Node NLRI used for allocating the Node-SID. The Node NLRI Type uses descriptors and attributes already defined in [I-D.ietf-idr-ls-distribution]. The format of the Node 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                                 //
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Where:

o Protocol-ID set to the new Protocol-ID: BGP-Segment-Allocation

o Node Descriptors defined in [I-D.ietf-idr-ls-distribution] can be reused

This NLRI MAY contain BGP-LS-SR TLV 1033 (SID/Label Binding) as its attribute.

4.2. Link NLRI for Segment Allocation

This section describes the Link NLRI used for allocating the Adj-SID. The format of the 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                                       //
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Where:

o Protocol-ID set to the new Protocol-ID: BGP-Segment-Allocation

o Node Descriptors and Link Descriptors defined in [I-D.ietf-idr-ls-distribution] can be reused.

Following TLV will be used in Link Attribute:

- o BGP-LS-SR TLV 1034: SR Capabilities
- o BGP-LS-SR TLV 1035: SR Algorithm
- o BGP-LS-SR TLV 1099: Adj-SID
- o BGP-LS-SR TLV 1036: Peer-SID
- o BGP-LS-SR TLV 1037: Peer-Set-SID

4.3. Prefix NLRI for Segment Allocation

This section describes the Prefix NLRI used for Allocating the Prefix-SID. The format of the 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                                 //
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
//      Prefix Descriptors                                     //
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Where:

- o Protocol-ID set to the new Protocol-ID: BGP-Segment-Allocation
- o Node Descriptors and Prefix Descriptors defined in [I-D.ietf-idr-ls-distribution] can be reused.

Following TLV will be used in Prefix Attribute:

- o BGP-LS-SR TLV 1034: SR Capabilities
- o BGP-LS-SR TLV 1035: SR Algorithm
- o BGP-LS-SR TLV 1158: Prefix SID

5. Applications

5.1. Allocating Segments for BGP Networks

As shown below, we assume:

Each node is its own AS (Node X has AS X). The loopback of Node X is 1.1.1.x/32.

Each node peers with its neighbors via BGP session.

Each node peers with Controller via BGP session.

Local BGP-LS Identifier in Node X is set to X0000.

```
                           /----------\
                          |            |
                          | Controller |
                          |            |
                           \----------/
                                 |
         +-----------------------+-------------------------------+
         |                       |                               |
         |              +------+ | +------+                      |
         |              |  N7  |-+-|  N8  |           Tier-1     |
         |              +------+   +------+                      |
         |                |  |       |  |                        |
         |      +---------+  |       |  +----------+             |
         |      | +-------+--+-------+--+-------+  |             |
         |      | |       |  |       |  |       |  |             |
         |    +----+     +----+     +----+     +----+            |
         |    | N3 |-----| N4 |     | N5 |-----| N6 | Tier-2     |
         |    +----+     +----+     +----+     +----+            |
         |       |         |           |         |               |
         |       | +-----+ |           | +-----+ |               |
         |       +-| N1  |-+           +-| N2  |-+    Tier-3     |
         |         +-----+               +-----+                 |
         |          | | |                 | | |                  |
         |      <- Servers ->         <- Servers ->              |
         +-------------------------------------------------------+
```

When the controller has collected the topology information of this BGP network, it can start segment allocation to the network.

5.1.1. Node-SID Distribution via a Prefix NLRI

A Node-SID represents a Node and has a global significance, something like a loopback of a router. Like an operator assigns a loopback's to their routers, it's expected that the Node-SID value will be assigned to every node. The assigned value can be an absolute or Index value and must be globally unique. In order to push a Node-SID for a router(e.g., N7), Controller advertise a Prefix NLRI to all the routers of the BGP-SR Network, where:

- o Protocol-ID set to the new Protocol-ID: BGP-Segment-Allocation
- o Local Node Descriptors contains
  - * BGP Router-ID: 7.7.7.7
  - * Local ASN: AS7
  - * BGP-LS Identifier: 70000
- o Prefix Descriptors
  - * 7.7.7.7/32
- o Prefix Attribute contains
  - * BGP-LS-SR TLV 1034: SR Capabilities
  - * BGP-LS-SR TLV 1035: SR Algorithm
  - * BGP-LS-SR TLV 1158: Prefix SID, With the N-flag (node-SID flag) set.
  - * Other Prefix Attributes.

5.1.2. Adj-SID Distribution via a Link NLRI

In order to push a Adj-SID for a router(e.g., N7 connects to N8), Controller advertise a Link NLRI to all the routers of the BGP-SR Network, where:

- o Protocol-ID set to the new Protocol-ID: BGP-Segment-Allocation
- o Local Node Descriptors contains
  - * BGP Router-ID: 7.7.7.7
  - * Local ASN: AS7

      * BGP-LS Identifier: 70000

   o Remote Node Descriptors contains

      * BGP Router-ID: 8.8.8.8

      * Local ASN: AS8

      * BGP-LS Identifier: 80000

   o Link Descriptors

      * BGP session IPv4 local address: 7.7.7.7

      * BGP session IPv4 peer address: 8.8.8.8

   o Link Attribute contains

      * BGP-LS-SR TLV 1034: SR Capabilities

      * BGP-LS-SR TLV 1035: SR Algorithm

      * BGP-LS-EPE TLV 1036: Peer-Node-SID

      * Other Prefix Attributes.

In the similar way, the controller can distribute Peer-Adj-SID and Peer-Set-SID.

5.2. Allocating Segments for IGP Networks

In IGP networks deployed with SR, the method defined in [I-D.ietf-idr-ls-distribution] to populate the topology database and the SRGB to the controller.

A controller may use the extensions defined in this document to populate mapping entries to the SRMS. Then the SRMS will advertise this mapping to all the SR Nodes via IGP.

In the following figure, LSR1-10 and LSR20 are only running LDP and R21-to-R25 Routers are SR capable Routers. R21 and R22 will be running both SR and LDP as they are on the border between SR and LDP. The whole network is running single IGP let’s say IS-IS.

```
                                      /----------\
                                     |            |
                                     | Controller |
                                     |            |
                                      \----------/
                                          /   \
                                         /     \
+----------------------------+----------/-------\---------------+
|     LDP Domain             |         /SR Domain\              |
|                            |        /           \             |
| ------                  -------    /   ------    \ ---------  |
| |LSR1|\                 | R21 |---+----| R23|\----\|Mapping|  |
| ------ \               /-------\      / ------\    |Server |  |
| 10.1.1.1\             / 20.1.1.1\    / 20.1.1.3\   --------+  |
|          \           /     |     \  /           \  30.1.1.1   |
| ------    \ ------- /      |      \/             \ ------     |
| |LSR2|-----\|LSR20|/       |      /\              \| R25|     |
| ------    / -------\       |     /  \             /------     |
| 10.1.1.2 / 10.1.1.20\      |    /    \           / 20.1.1.5   |
| ...     /            \     |   /      \         /             |
| ------ /              \-------/        \------ /              |
||LSR10|/                | R22 |----------|R24 |/               |
| ------                 -------          ------                |
| 10.1.1.10              20.1.1.2         20.1.1.4              |
|                            |                                  |
|                            |                                  |
+----------------------------+----------------------------------+
```

The Node-SIDs and their corresponding label value mapping could be like this:

Prefix	Index Value	Range
10.1.1.1/32	1001	10
10.1.1.20/32	1020	1
20.1.1.1/32	2001	5

The controller will advertise a node NLRI to Mapping Server, where:

o Protocol-ID set to the new Protocol-ID: BGP-Segment-Allocation

o Local Node Descriptors contains

   * Mapping Server’s node descriptor

o Node Attribute contains

* BGP-LS-SR TLV-1033: SID/Label Binding TLV

* Other Prefix Attributes

Mapping Server will convert BGP-LS-SR TLV-1033 to IS-IS TLV-149, and advertise this mapping to all the SR Nodes via IS-IS.

```
 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    |0|0|           |    Weight     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|            Range = 10         |      /32      |      10       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|      .1       |       .1      |      .1       |Prefix-SID Type|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| sub-TLV Length|     Flags     |   Algorithm   |               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                             1 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Type      |     Length    |0|0|           |    Weight     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|            Range = 1          |      /32      |       10      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|      .1       |       .1      |      .20      |Prefix-SID Type|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| sub-TLV Length|     Flags     |   Algorithm   |               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                            20 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Type      |     Length    |0|0|           |    Weight     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|            Range = 5          |      /32      |      20       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|      .1       |       .1      |      .1       |Prefix-SID Type|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| sub-TLV Length|     Flags     |   Algorithm   |               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                             1 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

A node receiving a MS entry for a prefix MUST check the existence of such prefix in its link-state database prior to consider and use the associated SID. This has been defined in [I-D.ietf-isis-segment-routing-extensions].

6. IANA Considerations

TBD.

7. Security Considerations

TBD.

8. Acknowledgements

TBD.

9. References

9.1. 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-isis-segment-routing-extensions]
Previdi, S., Filsfils, C., Bashandy, A., Gredler, H., Litkowski, S., Decraene, B., and J. Tantsura, "IS-IS Extensions for Segment Routing", draft-ietf-isis-segment-routing-extensions-05 (work in progress), June 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.

[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-06 (work in progress), October 2015.

[I-D.ietf-spring-segment-routing-ldp-interop]
Filsfils, C., Previdi, S., Bashandy, A., Decraene, B., and S. Litkowski, "Segment Routing interoperability with LDP", draft-ietf-spring-segment-routing-ldp-interop-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>.

9.2. Informative References

[I-D.gredler-idr-bgp-ls-segment-routing-extension]
Gredler, H., Ray, S., Previdi, S., Filsfils, C., Chen, M., and J. Tantsura, "BGP Link-State extensions for Segment Routing", draft-gredler-idr-bgp-ls-segment-routing-extension-02 (work in progress), October 2014.

[I-D.ietf-idr-bgpls-segment-routing-epe]
Previdi, S., Filsfils, C., Ray, S., Patel, K., Dong, J., and M. Chen, "Segment Routing Egress Peer Engineering BGP-LS Extensions", draft-ietf-idr-bgpls-segment-routing-epe-00 (work in progress), June 2015.

Authors' Addresses

Nan Wu
Huawei
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: eric.wu@huawei.com

Shunwan Zhuang
Huawei
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: zhuangshunwan@huawei.com

H. Chen
Futurewei
Z. Li
Huawei
Z. Li
China Mobile
Y. Fan
Casa Systems
M. Toy
Verizon
L. Liu
Fujitsu

October 30, 2019

BGP Extensions for IDs Allocation
draft-wu-idr-bgp-segment-allocation-ext-04

Abstract

This document describes extensions to the BGP for IDs allocation. The IDs are SIDs for segment routing (SR), including SR for IPv6 (SRv6). They are distributed to their domains if needed.

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

Table of Contents

1. Introduction

In a network with a central controller, the controller has the link state information of the network, including the resource such as traffic enginerring and SIDs information. It is valuable for the controller to allocate and manage the resources including SIDs of the network in a centralized way, especially for the SIDs representing network resources [I-D.ietf-teas-enhanced-vpn].

When BGP as a controller allocates an ID, it is natural and beneficial to extend BGP to send it to its corresponding network elements.

PCE may be extended to send IDs to their corresponding network elements after the IDs are allocated by a controller. However, when BGP is already deployed in a network, using PCE for IDs will need to deploy an extra protocol PCE in the network. This will increase the CapEx and OpEx.

Yang may be extended to send IDs to their corresponding network elements after the IDs are allocated by a controller. However, Yang progress may be slow. Some people may not like this.

There may not be these issues when BGP is used to send IDs. In addition, BGP may be used to distribute IDs into their domains easily when needed. It is also fit for the dynamic and static allocation of IDs.

This document proposes extensions to the BGP for sending Segment Identifiers (SIDs) for segment routing (SR) including SRv6 to their corresponding network elements after SIDs are allocated by the controller. If needed, they will be distributed into their network domains.

2. Terminology

The following terminology is used in this document.

SR: Segment Routing.

SRv6: SR for IPv6

SID: Segment Identifier.

IID: Indirection Identifier.

SR-Path: Segment Routing Path.

SR-Tunnel: Segment Routing Tunnel.

RR: Route Reflector.

MPP: MPLS Path Programming.

NAI: Node or Adjacency Identifier.

TED: Traffic Engineering Database.

3. Protocol Extensions

A new AFI and SAFI are defined: the Identifier AFI and the SID SAFI whose codepoints are to be assigned by IANA. A few new NLRI TLVs are defined for the new AFI/SAFI, which are Node, Link and Prefix SID NLRI TLVs. When a SID for a node, link or prefix is allocated by the controller, it may be sent to a network element in a UPDATE message containing a MP_REACH NLRI with the new AFI/SAFI and the SID NLRI TLV. When the SID is withdrawn by the controller, a UPDATE message containing a MP_UNREACH NLRI with the new AFI/SAFI and the SID NLRI TLV may be sent to the network element.

3.1. Node SID NLRI TLV

The Node SID NLRI TLV is used to represent the IDs such as SID associated with a node. Its format is illustrated in the Figure below, which is similar to the corresponding one defined in [RFC7752].

```
 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 (TBDa for Node SID)   |            Length             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Protocol ID   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                           Identifier                          |
|                           (8 octets)                          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Peer IP (4/16 bytes for IPv4/IPv6 Address)                   ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                   Local Node Descriptors TLV                  ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                           Sub-TLVs                            ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Where:

Type (TBDa): It is to be assigned by IANA.

Length: It is the length of the value field in bytes.

Peer IP: 4/16 octet value indicates an IPv4/IPv6 peer. When receiving a UPDATE message, a BGP speaker processes it only if the peer IP is the IP address of the BGP speaker or 0.

Protocol-ID, Identifier, and Local Node Descriptor: defined in [RFC7752], can be reused.

Sub-TLVs may be some of the followings:

SR-Capabilities TLV (1034): It contains the Segment Routing Global Base (SRGB) range(s) allocated for the node.

SR Local Block TLV (1036): The SR Local Block (SRLB) TLV contains the range(s) of SIDs/labels allocated to the node for local SIDs.

SRv6 SID Node TLV (TBD1): A new TLV, called SRv6 Node SID TLV, contains an SRv6 SID and related information.

SRv6 Locator TLV (TBD2): A new TLV, called SRv6 Locator TLV, contains an SRv6 locator and related information.

The format of SRv6 SID Node TLV is illustrated 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)         |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Reserved    |     Flags    |    SRv6 Endpoint Function     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        SRv6 Identifier                        |
|                          (128 bits)                           |
|                                                               |
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
~                        Optional sub-TLVs                      ~
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

SRv6 Node SID TLV

Type: TBD1 for SRv6 Node SID TLV is to be assigned by IANA.

Length: Variable.

Flags: 1 octet. No flags are defined now.

SRv6 Endpoint Function: 2 octets. The function associated with SRv6 SID.

SRv6 Identifier: 16 octets. IPv6 address representing SRv6 SID.

Reserved: MUST be set to 0 while sending and ignored on receipt.

SRv6 node SID inherits the topology and algorithm from its locator.

The format of SRv6 locator TLV is illustrated 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 (TBD2)          |             Length            |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |R|R|R|R|          MT-ID        |   Algorithm   |     Flags     |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                            Metric                             |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    | Locator-Size  | Locator (variable)...
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                                                               |
    ~                       Optional sub-TLVs                       ~
    |                                                               |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

SRv6 Locator TLV

Type: TBD2 for SRv6 Locator TLV is to be assigned by IANA.

Length: Variable.

MT-ID: Multitopology Identifier as defined in [RFC5120].

Algorithm: 1 octet. Associated algorithm.

Flags: 1 octet. As described in [I-D.ietf-lsr-isis-srv6-extensions].

Metric: 4 octets. As described in [RFC5305].

Locator-Size: 1 octet. Number of bits in the Locator field (1 to 128).

Locator: 1 to 16 octets. SRv6 Locator encoded in the minimum number of octets for the given Locator-Size.

Reserved: MUST be set to 0 while sending and ignored on receipt.

3.2. Link SID NLRI TLV

The Link SID NLRI TLV is used to represent the IDs such as SID associated with a link. Its format is illustrated in the Figure below, which is similar to the corresponding one defined in [RFC7752].

```
     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 (TBDb for Link SID)   |            Length             |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    | Protocol ID   |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    ~                    Identifier (8 octets)                      ~
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |  Peer IP (4/16 bytes for IPv4/IPv6 Address)                   ~
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    ~                    Local Node Descriptors TLV                 ~
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    ~                    Remote Node Descriptors TLV                ~
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    ~                        Link Descriptors TLV                   ~
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    ~                            Sub-TLVs                           ~
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Where:

Type (TBDb): It is to be assigned by IANA.

Length: It is the length of the value field in bytes.

Peer IP: 4/16 octet value indicates an IPv4/IPv6 peer.

Protocol-ID, Identifier, Local Node Descriptors, Remote Node Descriptors and Link Descriptors:
defined in [RFC7752], can be reused.

The Sub-TLVs may be some of the followings:

Adj-SID TLV (1099): It contains the Segment Identifier (SID) allocated for the link/adjacency.

LAN Adj-SID TLV (1100): It contains the Segment Identifier (SID) allocated for the adjacency/link to a non-DR router on a broadcast, NBMA, or hybrid link.

SRv6 Adj-SID TLV (TBD3): A new TLV, called SRv6 Adj-SID TLV, contains an SRv6 Adj-SID and related information.

SRv6 LAN Adj-SID TLV (TBD4): A new TLV, called SRv6 LAN Adj-SID TLV, contains an SRv6 LAN Adj-SID and related information.

The format of an SRv6 Adj-SID TLV is illustrated 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 (TBD3)           |            Length           |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |     Weight    |   Algorithm    |B|S|P|            Flags       |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |            Reserved             |   SRv6 Endpoint Function    |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                         SRv6 Identifier                       |
    |                           (128 bits)                          |
    |                                                               |
    |                                                               |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                                                               |
    ~                         Optional sub-TLVs                     ~
    |                                                               |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

SRv6 Adj-SID TLV

Type: TBD3 for SRv6 Adj-SID TLV is to be assigned by IANA.

Length: Variable.

Weight: 1 octet. The value represents the weight of the SID for the purpose of load balancing.

Algorithm: 1 octet. Associated algorithm.

Flags: 2 octets. Three flags are defined in [I-D.ietf-lsr-isis-srv6-extensions].

SRv6 Endpoint Function: 2 octets. The function associated with SRv6 SID.

SRv6 Identifier: 16 octets. IPv6 address representing SRv6 SID.

Reserved: MUST be set to 0 while sending and ignored on receipt.

The format of an SRv6 LAN Adj-SID TLV is illustrated 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 (TBD4)          |            Length             |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |     Weight    |   Algorithm   |B|S|P|            Flags      |O|
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |           Reserved            |    SRv6 Endpoint Function     |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |    neighbor Router ID (4 octets) / System ID (6 octets)       ~
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                       SRv6 Identifier                         |
    |                         (128 bits)                            |
    |                                                               |
    |                                                               |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                                                               |
    ~                       Optional sub-TLVs                       ~
    |                                                               |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

SRv6 LAN Adj-SID TLV

Type: TBD4 for SRv6 LAN Adj-SID TLV is to be assigned by IANA.

Length: Variable.

Weight: 1 octet. The value represents the weight of the SID for the purpose of load balancing.

Algorithm: 1 octet. Associated algorithm.

Flags: 2 octets. Three flags B, S and P are defined in [I-D.ietf-lsr-isis-srv6-extensions]. Flag O set to 1 indicating OSPF neighbor Router ID of 4 octets, set to 0 indicating IS-IS neighbor System ID of 6 octets.

SRv6 Endpoint Function: 2 octets. The function associated with SRv6 SID.

SRv6 Identifier: 16 octets. IPv6 address representing SRv6 SID.

Reserved: MUST be set to 0 while sending and ignored on receipt.

3.3. Prefix SID NLRI TLV

The Prefix SID NLRI TLV is used to represent the IDs such as SID associated with a prefix. Its format is illustrated in the Figure below, which is similar to the corresponding one defined in [RFC7752].

```
     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 (TBDc for Prefix SID)  |            Length             |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    | Protocol ID   |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    ~                       Identifier (8 octets)                   ~
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |  Peer IP (4/16 bytes for IPv4/IPv6 Address)                   ~
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    ~                    Local Node Descriptors TLV                 ~
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    ~                        Prefix Descriptors TLV                 ~
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    ~                            Sub-TLVs                           ~
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Where:

Type (TBDc): It is to be assigned by IANA.

Length: It is the length of the value field in bytes.

Peer IP: 4/16 octet value indicates an IPv4/IPv6 peer.

Protocol-ID, Identifier, Local Node Descriptors and Prefix Descriptors:
defined in [RFC7752], can be reused.

Sub-TLVs may be some of the followings:

Prefix-SID TLV (1158): It contains the Segment Identifier (SID) allocated for the prefix.

Prefix Range TLV (1159): It contains a range of prefixes and the Segment Identifier (SID)s allocated for the prefixes.

3.4. Capability Negotiation

It is necessary to negotiate the capability to support BGP Extensions for sending and receiving Segment Identifiers (SIDs). The BGP SID Capability is a new BGP capability [RFC5492]. The Capability Code for this capability is to be specified by the IANA. 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)    |
+---------------------------------------------------+
| Send/Receive (1 octet)                            |
+---------------------------------------------------+
```

BGP SID Capability

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].

Send/Receive: This field indicates whether the sender is (a) willing to receive SID from its peer (value 1), (b) would like to send SID to its peer (value 2), or (c) both (value 3) for the <AFI, SAFI>.

4. IANA Considerations

This document requests assigning a new AFI in the registry "Address Family Numbers" as follows:

Code Point	Description	Reference
TBDx	Identifier AFI	This document

This document requests assigning a new SAFI in the registry "Subsequent Address Family Identifiers (SAFI) Parameters" as follows:

Code Point	Description	Reference
TBDy	SID SAFI	This document

This document defines a new registry called "SID NLRI TLVs". The allocation policy of this registry is "First Come First Served (FCFS)" according to [RFC8126].

Following TLV code points are defined:

Code Point	Description	Reference
1 (TBDa)	Node SID NLRI	This document
2 (TBDb)	Link SID NLRI	This document
3 (TBDc)	Prefix SID NLRI	This document

This document requests assigning a code-point from the registry "BGP-LS Node Descriptor, Link Descriptor, Prefix Descriptor, and Attribute TLVs" as follows:

TLV Code Point	Description	Reference
TBD1	SRv6 Node SID	This document
TBD2	SRv6 Allocator	This document
TBD3	SRv6 Adj-SID	This document
TBD4	SRv6 LAN Adj-SID	This document

5. Security Considerations

Protocol extensions defined in this document do not affect the BGP security other than those as discussed in the Security Considerations section of [RFC7752].

6. Acknowledgements

The authors would like to thank Eric Wu, Robert Raszuk, Zhengquiang Li, and Ketan Talaulikar for their valuable suggestions and comments on this draft.

7. References

7.1. Normative References

[I-D.ietf-idr-flowspec-path-redirect]
Velde, G., Patel, K., and Z. Li, "Flowspec Indirection-id Redirect", draft-ietf-idr-flowspec-path-redirect-10 (work in progress), October 2019.

[I-D.ietf-isis-segment-routing-extensions]
Previdi, S., Ginsberg, L., Filsfils, C., Bashandy, A., Gredler, H., and B. Decraene, "IS-IS Extensions for Segment Routing", draft-ietf-isis-segment-routing-extensions-25 (work in progress), May 2019.

[I-D.ietf-lsr-isis-srv6-extensions]
Psenak, P., Filsfils, C., Bashandy, A., Decraene, B., and Z. Hu, "IS-IS Extension to Support Segment Routing over IPv6 Dataplane", draft-ietf-lsr-isis-srv6-extensions-03 (work in progress), October 2019.

[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-11 (work in progress), June 2016.

[I-D.ietf-spring-segment-routing]
Filsfils, C., Previdi, S., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", draft-ietf-spring-segment-routing-15 (work in progress), January 2018.

[I-D.ietf-spring-segment-routing-ldp-interop]
Bashandy, A., Filsfils, C., Previdi, S., Decraene, B., and S. Litkowski, "Segment Routing interworking with LDP", draft-ietf-spring-segment-routing-ldp-interop-15 (work in progress), September 2018.

[I-D.li-ospf-ospfv3-srv6-extensions]
Li, Z., Hu, Z., Cheng, D., Talaulikar, K., and P. Psenak, "OSPFv3 Extensions for SRv6", draft-li-ospf-ospfv3-srv6-extensions-05 (work in progress), August 2019.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

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

[RFC5120] Przygienda, T., Shen, N., and N. Sheth, "M-ISIS: Multi Topology (MT) Routing in Intermediate System to Intermediate Systems (IS-ISs)", RFC 5120, DOI 10.17487/RFC5120, February 2008, <https://www.rfc-editor.org/info/rfc5120>.

[RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, DOI 10.17487/RFC5305, October 2008, <https://www.rfc-editor.org/info/rfc5305>.

[RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, DOI 10.17487/RFC5492, February 2009, <https://www.rfc-editor.org/info/rfc5492>.

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

[RFC7752] Gredler, H., Ed., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and Traffic Engineering (TE) Information Using BGP", RFC 7752, DOI 10.17487/RFC7752, March 2016, <https://www.rfc-editor.org/info/rfc7752>.

[RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 8126, DOI 10.17487/RFC8126, June 2017, <https://www.rfc-editor.org/info/rfc8126>.

7.2. Informative References

[I-D.gredler-idr-bgp-ls-segment-routing-extension]
Gredler, H., Ray, S., Previdi, S., Filsfils, C., Chen, M., and J. Tantsura, "BGP Link-State extensions for Segment Routing", draft-gredler-idr-bgp-ls-segment-routing-extension-02 (work in progress), October 2014.

[I-D.ietf-idr-bgpls-segment-routing-epe]
Previdi, S., Talaulikar, K., Filsfils, C., Patel, K., Ray, S., and J. Dong, "BGP-LS extensions for Segment Routing BGP Egress Peer Engineering", draft-ietf-idr-bgpls-segment-routing-epe-19 (work in progress), May 2019.

[I-D.ietf-teas-enhanced-vpn]
Dong, J., Bryant, S., Li, Z., Miyasaka, T., and Y. Lee, "A Framework for Enhanced Virtual Private Networks (VPN+) Service", draft-ietf-teas-enhanced-vpn-03 (work in progress), September 2019.

Authors' Addresses

Huaimo Chen
Futurewei
Boston, MA
USA

Email: Huaimo.chen@futurewei.com

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

Email: lizhenbin@huawei.com

Zhenqiang Li
China Mobile
No. 29 Finance Street, Xicheng District
Beijing 100029
P.R. China

Email: li_zhenqiang@hotmail.com

Yanhe Fan
Casa Systems
USA

Email: yfan@casa-systems.com

Mehmet Toy
Verizon
USA

Email: mehmet.toy@verizon.com

Lei Liu
Fujitsu
USA

Email: liulei.kddi@gmail.com