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

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

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

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

The list of current Internet-Drafts can be accessed at http://www.ietf.org/1id-abstracts.html

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

Copyright Notice

Copyright (c) 2015 IETF Trust and the persons identified as the document authors. All rights reserved.

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

RFC 5575 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

Table of Contents

1. Introduction ................................................ 2
2. The Flow Specification encoding for NVO3 .................... 3
4. Security Considerations ...................................... 5
5. IANA Considerations ......................................... 5
  5.1. Normative References .................................... 5
  5.2. Informative References .................................. 6
6. Acknowledgments ............................................. 6

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.

RFC 5575 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

Hao, et al
Expires April 19, 2016
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

5.2. Informative References


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
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

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

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

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

The list of current Internet-Drafts can be accessed at http://www.ietf.org/1id-abstracts.html

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

Copyright Notice
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].

<table>
<thead>
<tr>
<th>type</th>
<th>extended community</th>
<th>RFC or Draft</th>
</tr>
</thead>
<tbody>
<tr>
<td>0x8006</td>
<td>traffic-rate</td>
<td>RFC5575</td>
</tr>
<tr>
<td>0x8007</td>
<td>traffic-action</td>
<td>RFC5575</td>
</tr>
<tr>
<td>0x8008</td>
<td>redirect</td>
<td>RFC5575</td>
</tr>
<tr>
<td>0x8009</td>
<td>traffic-marking</td>
<td>RFC5575</td>
</tr>
<tr>
<td>TBD</td>
<td>redirect to Tunnel</td>
<td>This draft</td>
</tr>
</tbody>
</table>

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


4.2. Informative References


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 Beijing Rd.
Beijing 100095
China
Email: lizhenbin@huawei.com

Lucy Yong
Huawei Technologies
Phone: +1-918-808-1918
Email: lucy.yong@huawei.com
Constrain Attribute announcement within BGP
draft-keyupate-idr-bgp-attribute-announcement-00.txt

Abstract

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

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

Status of This Memo

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

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

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

This Internet-Draft will expire on April 20, 2016.

Copyright Notice

Copyright (c) 2015 IETF Trust and the persons identified as the document authors. All rights reserved.

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

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

Table of Contents

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

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

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

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

1.1. Requirements Language

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

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

Path Attribute flags:

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

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

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

**P** Partial as defined in [RFC4271]

**E** Extended Length type as defined in [RFC4271]

**A** AS Wide Scope

**C** Member-AS in Confederation Scope

**M** Multi-AS Scope

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

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

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

3. Operation

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

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

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

When a BGP speaker receives or originates a route with an optional path attribute that has both, the AS Wide Scope bit set and the
Member-AS Scope bit set, it MUST announce it to all its EBGP peers within its administrative domain. Such an attribute MUST be filtered when the attribute is announced outside its administrative domain. The BGP peering boundaries for an administrative domain is a matter of a policy and is set by the operators.

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

4. IANA Considerations

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

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

<table>
<thead>
<tr>
<th>Bit Value (MSB)</th>
<th>Type</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Optional/Mandatory</td>
<td>RFC4271</td>
</tr>
<tr>
<td>2</td>
<td>Transitive/NonTransitive</td>
<td>RFC4271</td>
</tr>
<tr>
<td>3</td>
<td>Partial</td>
<td>RFC4271</td>
</tr>
<tr>
<td>4</td>
<td>Extended Length Type</td>
<td>RFC4271</td>
</tr>
<tr>
<td>5</td>
<td>AS Wide Scope</td>
<td>Current Draft</td>
</tr>
<tr>
<td>6</td>
<td>Member-AS in Confederation</td>
<td>Current Draft</td>
</tr>
</tbody>
</table>

5. Security Considerations

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

5.1. Acknowledgements

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

6. References
6.1. Normative References


6.2. Information References


Authors' Addresses

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

Email: keyupate@cisco.com
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].

Status of This Memo

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

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

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

This Internet-Draft will expire on April 19, 2016.
1. Introduction

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

- 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.
- 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.

- ACL: Access Control List
- BGP: Border Gateway Protocol
- FS: Flow Specification
- PBR: Policy-Based Routing
- RPD: Routing Policy Distribution
- 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:

- Business development or network failure introduces link congestion and overload.
- Network transmission quality decreased as the result of delay, loss and need to adjust traffic to other paths.
- 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 criteria directly.

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

```
+-----------+      +---------+      +-----------+
| Speaker1 |      | L1      |      | IGW1 | policy|
|-----------+      +---------+      +-----------+
| **       |      **       | L2** |      | **       |
| PE1       |      ****     |       |      | controller|
+-----------+      +---------+      +-----------+
| **       |      **       |
| Speaker2 |      **       | L3** |      | IGW2 | AS100|
|-----------+      +---------+      +-----------+
| 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||
| | |                         |    ****    |        +-------+|
| | | AS100       +----+ |**        **| +---------+     |
| | |                  |IGW2| +------------+ |Speaker2 |     |
| | |                  +----+ |L4          | +---------+     |
| | +---+                    |            |    AS200        |
| | |PE2|              ...   |            |                 |
| | |+---+                    |            |                 |
| +-------------------------+            +-----------------+

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:

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

...
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:

<table>
<thead>
<tr>
<th>Prefix</th>
<th>Next-Hop</th>
<th>Local BGP Speaker</th>
<th>Remote BGP Peer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prefix2</td>
<td>Speaker1</td>
<td>IGW2</td>
<td>Speaker1</td>
</tr>
<tr>
<td>Prefix2</td>
<td>Speaker2</td>
<td>IGW2</td>
<td>Speaker2</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 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:

- Type 1: Autonomous System number list
- Type 2: IPv4 prefix (1 octet prefix length + prefix) list
- Type 3: IPv6 prefix (1 octet prefix length + prefix) list
- Type 4: Integer list
- Type 5: IEEE Floating Point Number list
- Type 6: Neighbor Class list
- Type 7: User-defined Class list
- Type 8: UTF-8 String
- Type TBD: BGP IPv4 neighbor --- Newly introduced in this draft
- 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Container Type 1 (1)      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|0 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 prefer 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Container Type 1 (1)      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|1 0 0 0 0 0 0 0 0|           |
+-+-+-+-+-+-+-+-+
| Hop Count: 0 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Length: 36 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Community: LOCAL PREFERENCE                                         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
5.4. 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:
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:

- Control route advertising: Only routes that match the rules specified in a policy are advertised.
- Control route receiving: Only the required and valid routes are received. This reduces the size of the routing table and improves network security.
- 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.
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.

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
11.2. Informative References

[I-D.ietf-idr-registered-wide-bgp-communities]

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
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

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

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

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

This Internet-Draft will expire on April 19, 2016.

Copyright Notice

Copyright (c) 2015 IETF Trust and the persons identified as the document authors. All rights reserved.
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.
3.2. Use cases

3.2.1. Deterministic ECMP

Entropy Label [RFC6790] is introduced to improve the ECMP capability by encapsulating 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.
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 next hop 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.
[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, EVPnP, MVPN, etc., it does 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:

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

Figure 1: NLRI Definition in RFC3107

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:

<table>
<thead>
<tr>
<th>Value</th>
<th>Tunnel Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>TBD</td>
<td>LDP LSP</td>
</tr>
<tr>
<td>TBD</td>
<td>RSVP-TE LSP</td>
</tr>
<tr>
<td>TBD</td>
<td>MPLS-based Segment Routing Best-effort Path</td>
</tr>
<tr>
<td>TBD</td>
<td>MPLS-based Segment Routing Traffic Engineering Path</td>
</tr>
</tbody>
</table>

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:

```
+--------------------------------------------------+
<table>
<thead>
<tr>
<th>Flags (1 octet)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tunnel Type (1 octets)</td>
</tr>
<tr>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>Tunnel Identifier (variable)</td>
</tr>
<tr>
<td>--------------------------------------------------</td>
</tr>
</tbody>
</table>
```

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:

<table>
<thead>
<tr>
<th>0 1 2 3 4 5 6 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
</tr>
</tbody>
</table>

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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|               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:

```
Li, et al. Expires April 19, 2016 [Page 10]
```
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]

[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.

11.2. Informative References

[I-D.ietf-pce-pce-initiated-lsp]


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
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 identify 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].

Status of This Memo

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

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

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

This Internet-Draft will expire on April 1, 2016.
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.
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 rule1 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].

<table>
<thead>
<tr>
<th>type</th>
<th>extended community</th>
<th>encoding</th>
</tr>
</thead>
<tbody>
<tr>
<td>TBD1</td>
<td>label-action</td>
<td>MPLS tag</td>
</tr>
<tr>
<td>--------</td>
<td>--------------------</td>
<td>----------</td>
</tr>
</tbody>
</table>
Label-action is described below:

<table>
<thead>
<tr>
<th>OpCode</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Push the MPLS tag</td>
</tr>
<tr>
<td>1</td>
<td>Pop the outermost MPLS tag in the packet</td>
</tr>
<tr>
<td>2</td>
<td>Swap the MPLS tag with the outermost MPLS tag in the packet</td>
</tr>
<tr>
<td>3-15</td>
<td>Reserved</td>
</tr>
</tbody>
</table>

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]


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
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].

Status of This Memo

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

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

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

This Internet-Draft will expire on April 20, 2016.
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.
3.1. Flow Description sub-TLV

The Flow Description sub-TLV is encoded as below:

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 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:
<table>
<thead>
<tr>
<th>Type Code</th>
<th>Function Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Immediate validation</td>
</tr>
<tr>
<td>1</td>
<td>Delayed validation</td>
</tr>
<tr>
<td>2</td>
<td>Timing validation</td>
</tr>
<tr>
<td>else codes</td>
<td>Reserved</td>
</tr>
</tbody>
</table>

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

Duration Type:
<table>
<thead>
<tr>
<th>Type Code</th>
<th>Function Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Permanent validation</td>
</tr>
<tr>
<td>1</td>
<td>Hard invalidation</td>
</tr>
<tr>
<td>2</td>
<td>Idle invalidation</td>
</tr>
<tr>
<td>else codes</td>
<td>Reserved</td>
</tr>
</tbody>
</table>

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:

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

7.2. Informative References

[I-D.eddy-idr-flowspec-exp]

[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]

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
Applying BGP flowspec rules on a specific interface set

draft-litkowski-idr-flowspec-interfaceset-02

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].
1. Use case

1.1. Specific filtering for DDoS

The figure 1 above displays a typical service provider Internet network owing Customers, Peers and Transit. To protect proactively 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 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
    |                     PE ------ (ebgp) - Peer2
|                     PE ------ (ebgp) - Cust4_INT
Cust3_VPN -- (ebgp) --- PE
    \                     PE ------ (ebgp) - Transit2
Peer1 ------ (ebgp) --

Figure 2

The figure 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 can imagine two cases:

- Maintaining complete ACLs using flowspec: in this case all the ingress ACL are maintained and deployed using BGPFlowspec. See section 7 for more details on security aspects.

- 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 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 permits to apply traffic action one step before and so may permit to prevent impact like congestions.

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 destined to MyAS. 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 an identification of interfaces within BGP Flowspec. All interfaces may be associated to one or more group-identifiers and a BGP Flowspec rule may also be associated with one or more group-identifiers including a filtering direction (input/output/both), so the FlowSpec rule will be applied only on interfaces belonging the the group identifier included in the BGP FlowSpec update.

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.

4. Interface-set extended community

This document proposes a new BGP extended community called "flow spec interface-set". This new BGP extended community is part of TRANSITIVE FOUR-OCTET AS-SPECIFIC EXTENDED COMMUNITY and has subtype TBD.

The Global Administrator field of this community MUST be set to the ASN of the originating router. The Local Administrator field is encoded as follows:
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.

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 with 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 behavior. In order to keep a deterministic and well known behaviour, an implementation SHOULD process the BGP FlowSpec ACL as follows:

- 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.

- 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.

```
<table>
<thead>
<tr>
<th>Inbound filters</th>
<th>Outbound filter</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Permanent -&gt; Dynamic -&gt; Forwarding -&gt; Dynamic -&gt; Permanent</td>
</tr>
</tbody>
</table>
```

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:

```
<table>
<thead>
<tr>
<th>Permanent ACL entry action</th>
<th>FlowSpec ACL entry action</th>
<th>Result action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drop</td>
<td>Drop</td>
<td>Drop</td>
</tr>
<tr>
<td>Drop</td>
<td>Accept</td>
<td>Drop</td>
</tr>
<tr>
<td>Accept</td>
<td>Drop</td>
<td>Drop</td>
</tr>
<tr>
<td>Accept</td>
<td>Accept</td>
<td>Accept</td>
</tr>
</tbody>
</table>
```
Example:

- **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.

<table>
<thead>
<tr>
<th>Inbound Flow</th>
<th>BGP Flow collection</th>
<th>BGP Flow</th>
<th>Outbound Flow collection</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Permanent</td>
</tr>
</tbody>
</table>

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. 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.

8. Acknowledgements

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

9. IANA Considerations

This document requests a new sub-type from the "TRANSITIVE FOUR-OCTET AS-SPECIFIC EXTENDED COMMUNITY SUB-TYPES" extended community registry. The sub-type name shall be ‘Flow spec interface-set’.
10. Normative References


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
Abstract

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

This document defines a new redirect-to-Path-id (32-bit or 128-bit) flow-spec action to provide advanced redirection capabilities. When activated, the flowspec signalled Path-id is used to identify the next-hop redirect information within a localized to the router Path-id redirect table. The Path-id redirect table is a table providing a list of Path-id’s and router local redirect information. This allows a Flowspec signalled redirect towards a next-hop IP address, next-hop local interface or next-hop (traffic engineered) tunnel interface.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [1].

Status of This Memo

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

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

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

Every flow-spec route is effectively a rule, consisting of a matching part (encoded in the NLRI field) and an action part (encoded in one or more BGP extended community). The flow-spec standard RFC5575 [2] defines widely-used filter actions such as discard and rate limit; it
also defines a redirect-to-VRF action for policy-based forwarding. Using the redirect-to-VRF action for redirecting traffic towards an alternate destination is useful for DDoS mitigation but using this technology can be cumbersome when there is need to redirect the traffic onto an engineered traffic path.

This draft proposes a new redirect-to-Path-id flow-spec action facilitating an anchor point for policy-based forwarding onto an engineered path. The router consuming and utilizing the flowspec rule makes a local mapping between the flowspec signalled redirect Path-id and locally available redirection information referenced by the Path-id. This locally available redirection information is derived from out-of-band programming or signalling.

The redirect-to-Path-id is encoded in a newly defined BGP extended Path-id community.

The construction of the Path-id redirect table and the technology used to create an engineered path are out-of-scope of this document.

2. Redirect to Path-id Communities

This document defines a new BGP extended community. The extended communities have a type indicating they are transitive and IPv4-address-specific or IPv6-address-specific, depending on whether the redirection Path-id is 32-bit or 128-bit. The sub-type value [to be assigned by IANA] indicates that the global administrator and local administrator fields encode a flow-spec ‘redirect to Path-id’ action. In the new extended community the 4-byte or 16-byte global administrator field encodes the 32-bit or 128-bit Path-id’s providing the redirection Path-id to allow a local to the router mapping reference to an engineered Path. The 2-byte local administrator field is formatted as shown in Figure 1.

```
0                   1
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|          Reserved       |TID|C|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 1

In the local administrator field the least-significant bit is defined as the ‘C’ (or copy) bit. When the ‘C’ bit is set the redirection
applies to copies of the matching packets and not to the original traffic stream.

The ‘TID’ field identifies a 2 bit Table-id field. This field is used to provide the router consuming the flowspec rule an indication how and where to use the Path-id when redirecting traffic.

All bits other than the ‘C’ bit in the local administrator field MUST be set to 0 by the originating BGP speaker and ignored by receiving BGP speakers.

3. Redirect using localized Path-id Router Mapping

When a BGP speaker receives a flow-spec route with a ‘redirect to Path-id’ extended community and this route represents the one and only best path, it installs a traffic filtering rule that matches the packets described by the NLRI field and redirects them (C=0) or copies them (C=1) towards the locally engineered Paths referenced by the extended community’s global administrator field. The BGP speaker is expected to do a local Path-id table lookup and identify inside this table a redirect path referenced by the Flowspec Path-id community field.

The router local Path-id table contains a list of Path-id’s mapped to redirect information. The redirect information can be a next-hop IP address, a local next-hop Interface or a next-hop tunnel.

- When the redirect information is a Next-hop IP address, then a recursive routing lookup to this destination address is performed and the traffic matching the flowspec rule is redirected to this next-hop IP address.

- In case of redirection to a local next-hop interface, the traffic matching the flowspec rule is redirected to the local next-hop interface.

- In case of a next-hop tunnel, the traffic matching the flowspec rule is redirected to the next-hop tunnel. This tunnel could be instantiated through various means (i.e. manual configuration, PCEP, RSVP-TE, WAN Controller, Segment Routing, etc...).

4. 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 Path-id’ extended community, as it is to all types of flow-spec routes. This means that a flow-spec route with a destination prefix subcomponent SHOULD NOT be accepted from an EGBP
peer unless that peer also advertised the best path for the matching unicast route.

TBC (add what to check regarding Path-id and redirect-ip Next-hop usage)

5. Localized Path-id Table

Each router participating in the Path-id based Flowspec redirect has a localized Path-id indexed table. The exact nature on how this table is populated is out of scope of this document. The Path-id localized table provides a list of Path-id’s, each followed by a set of Labels or encapsulation information to push for the traffic matching the flowspec rule. If the flowspec rule signals multiple Path-id communities, then it is a localized decision on the Flowspec consuming device how the set of Path-id’s will be pushed upon Flowspec matching traffic.

6. Security Considerations

A system using ‘redirect to Path-id’ extended community can cause during the redirect mitigation of a DDoS attack an overflow of traffic being received by the mitigation infrastructure.

7. Acknowledgements

This document has been contributed by Adam Simpson, Mustapha Aissaoui, Jan Mertens.

8. IANA Considerations

This document requests a new sub-type from the "Transitive IPv4-Address-Specific" extended community registry. The sub-type name shall be 'Flow-spec Redirect to 32-bit Path-id'.

This document requests a new sub-type from the "Transitive IPv6-Address-Specific" extended community registry. The sub-type name shall be 'Flow-spec Redirect to 128-bit Path-id'.

9. References

9.1. Normative References

9.2. Informative References


Authors’ Addresses

Gunter Van de Velde  
Alcatel-Lucent  
Antwerp  
BE  
Email: gunter.van_de_velde@alcatel-lucent.com

Wim Henderickx  
Alcatel-Lucent  
Antwerp  
BE  
Email: wim.henderickx@alcatel-lucent.com

Keyur Patel  
Cisco Systems  
170 W. Tasman Drive  
San Jose, CA  95134  
USA  
Email: keyupate@cisco.com
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

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

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

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

This Internet-Draft will expire on April 21, 2016.

Copyright Notice

Copyright (c) 2015 IETF Trust and the persons identified as the document authors. All rights reserved.

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

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

Where:

- Protocol-ID set to the new Protocol-ID: BGP-Segment-Allocation
- Node Descriptors and Link Descriptors defined in [I-D.ietf-idr-ls-distribution] can be reused.
Following TLV will be used in Link Attribute:

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

```
+--------+
|  Protocol-ID |
+--------+
| Identifier (64 bits) |
+--------+
```

Where:

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

Following TLV will be used in Prefix Attribute:

- BGP-LS-SR TLV 1034: SR Capabilities
- BGP-LS-SR TLV 1035: SR Algorithm
- 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.

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:

- Protocol-ID set to the new Protocol-ID: BGP-Segment-Allocation
- Local Node Descriptors contains
  - BGP Router-ID: 7.7.7.7
  - Local ASN: AS7
  - BGP-LS Identifier: 70000
- Prefix Descriptors
  - 7.7.7.7/32
- 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 an 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:

- Protocol-ID set to the new Protocol-ID: BGP-Segment-Allocation
- 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.
The Node-SIDs and their corresponding label value mapping could be like this:

<table>
<thead>
<tr>
<th>Prefix</th>
<th>Index Value</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.1.1.1/32</td>
<td>1001</td>
<td>10</td>
</tr>
<tr>
<td>10.1.1.20/32</td>
<td>1020</td>
<td>1</td>
</tr>
<tr>
<td>20.1.1.1/32</td>
<td>2001</td>
<td>5</td>
</tr>
</tbody>
</table>

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

- Protocol-ID set to the new Protocol-ID: BGP-Segment-Allocation
- Local Node Descriptors contains
  * Mapping Server’s node descriptor
- 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.

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]

[I-D.ietf-isis-segment-routing-extensions]

[I-D.ietf-rtgwg-bgp-routing-large-dc]

[I-D.ietf-spring-segment-routing]

[I-D.ietf-spring-segment-routing-ldp-interop]
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