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

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

June 17, 2016

# BGP FlowSpec Extensions for Routing Policy Distribution (RPD)
## draft-li-idr-flowspec-rpd-02

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 December 19, 2016.

Copyright Notice

Copyright (c) 2016 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

## 1. Introduction

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

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

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

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

## 2. Definitions and Acronyms

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

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

- o ACL:Access Control List

- o BGP: Border Gateway Protocol

- o FS: Flow Specification

- o PBR:Policy-Based Routing

- o RPD: Routing Policy Distribution

- o VPN: Virtual Private Network

## 3. Problem Statements

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

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

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

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

### 3.1. Inbound Traffic Control

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

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

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

               Prefix1 advertise from AS100 to AS200
             <----------------------------------------
               Figure 1: Inbound Traffic Control case
```

### 3.2. Outbound Traffic Control

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

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

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

Figure 2: Outbound Traffic Control case

## 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 it is for DDOS mitigation, it doesn't mean BGP Flowspec only takes effect on the forwarding plane.

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

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

## 5. Protocol Extensions

### 5.1. FlowSpec Traffic Actions for Routing Policy Distribution

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

The Flow Specification Traffic Actions for Routing Policy Distribution:

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

Figure 3: 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. Option 1: 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 4: 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 5: Match Fields Format

Match Type:

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

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

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

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

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

Figure 6: Sub-TLVs Format

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

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

Values: The Value field contains the TLV value.

Supported format of the TLVs can be:

Type 1: IPv4 Neighbor

Type 2: IPv6 Neighbor

Type 3: ASN List

...

To be added in later versions.

#### 5.2.2. Action Fields Format

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

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

Figure 7: Action Fields Format

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

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

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

Supported format of the TLVs can be:

Type 1: Route-Preference

Type 2: Route-Prepend-AS

...

To be added in later versions.

### 5.2.3. Operation Examples

#### 5.2.3.1. Inbound Traffic Control

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

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

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

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

3. NO_ADVERTISE Community [RFC1997]

4. 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 and Speaker2 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 there is a speaker which 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. At the same time, IGW2 sends the best route of Prefix1 to Speaker2 with performing the Action instructions from the BGP-FS RPD route: Prepend Local AS 5 times.

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. NO_ADVERTISE Community [RFC1997]

4. BGP Policy Attribute:

    * Match Type: 1, Permit

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

    * Action Type: Route-Preference

    * Action Value: none

IGW2 processes the received BGP FS RPD route as follows:

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

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

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

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

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

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

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

### 5.3. Option 2: 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].

BGP Wide Community Attribute is a very useful tool that it can be used to convey different kinds of routing policies.

#### 5.3.1. New Wide Community Atoms

Wide Community Atoms define in [I-D.ietf-idr-wide-bgp-communities] , in that draft it defines Type 1 to Type 8.

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

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+
|     Type      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|            Length             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                         Value (variable)                      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 8: Wide Community Atoms

Supported format of the TLVs can be:

- o Type 1: Autonomous System number list
- o Type 2: IPv4 prefix (1 octet prefix length + prefix) list
- o Type 3: IPv6 prefix (1 octet prefix length + prefix) list
- o Type 4: Integer list
- o Type 5: IEEE Floating Point Number list
- o Type 6: Neighbor Class list

o Type 7: User-defined Class list7

o Type 8: UTF-8 String

o Type TBD: BGP IPv4 neighbor --- Newly introduced in this draft, which contains the BGP session IPv4 local address and the BGP session IPv4 peer address.

o Type TBD: BGP IPv6 neighbor --- Newly introduced in this draft, which contains the BGP session IPv6 local address and the BGP session IPv6 peer address.

### 5.3.2. Operation 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 BY AS" and "Exclude Target(s) TLV" are be used.

The encoding example using BGP Wide Community:

```
    0                   1                   2                   3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |     Container Type 1 (1)      |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |1 0 0 0 0 0 0 0|
   +-+-+-+-+-+-+-+-+
   | Hop Count: 0  |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   | Length:                    36 |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   | Community: PREPEND N TIMES BY AS                           17 |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   | 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 9: Example encoding for Inbound Traffic Control case

"PREPEND N TIMES BY AS" Wide Community has been defined in [I-D.ietf-idr-registered-wide-bgp-communities].

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. NO_ADVERTISE Community [RFC1997]

4. Wide BGP Community Attribute:

PREPEND N TIMES BY AS:

```
    Type: 0x0001                S = src AS #
    F = 0x80                    C = 0x00000000
    H = 0                       T = none
    L = 36 octets               E = Type_TBD (BGP IPv4 neighbor)
    R = 17                      P = Type_4 (0x05)
```

Where "BGP IPv4 neighbor" Atom TLV contains:
The BGP session IPv4 local address: Local BGP Speaker IGW2
The BGP session IPv4 peer address: Remote BGP Peer Speaker1

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 action type from the Wide BGP Community Attribute: PREPEND N TIMES BY AS;

5. IGW1 gets the matching criteria from the Wide BGP Community Attribute: Exclude the BGP IPv4 neighbor: <Local BGP Speaker IGW2, Remote BGP Speaker1>;

6. IGW1 gets the parameter for "PREPEND N TIMES BY AS" from the Wide BGP Community Attribute: 5 times;

IGW1 checks the matching criteria and finds that it doesn't hits the exclude matching criteria: Local BGP Speaker IGW2, Remote BGP Speaker1, so IGW1 sends the best route of Prefix1 to Speaker1 and Speaker2 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 action type from the Wide BGP Community Attribute: PREPEND N TIMES BY AS;

5. IGW2 gets the matching criteria from the BGP Policy Attribute: Exclude the BGP IPv4 neighbor: <Local BGP Speaker IGW2, Remote BGP Speaker1>;

6. IGW2 gets the parameter for "PREPEND N TIMES BY AS" from the Wide BGP Community Attribute: 5 times;

IGW2 checks the matching criteria and finds that there is a speaker which hits the exclude matching criteria: Local BGP Speaker IGW2, Remote BGP Peer Speaker1, so IGW2 sends the best route of Prefix1 to Speaker1 without performing the Action instructions from the BGP-FS RPD route, at the same time, IGW2 sends the best route of Prefix1 to Speaker2 with performing the Action instructions from the BGP-FS RPD route: Prepend Local AS 5 times.

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.

#### 5.3.2.2. Outbound Traffic Control

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

The encoding example using BGP Wide Community:

```
  0                   1                   2                   3
  0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |     Container Type 1 (1)      |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |1 0 0 0 0 0 0 0|
 +-+-+-+-+-+-+-+-+
 | Hop Count: 0  |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 | Length:                    36 |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 | Community: LOCAL PREFERENCE                                 20 |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 | Own ASN                                                    100 |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 | Context ASN#                                               100 |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 | TargetTLV(1)  |   Length:                 11 |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |  IPv4Neig(TBD)|   Length:                  8 |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 | Local Speaker                                            #IGW2 |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 | Remote Speaker                                       #Speaker1 |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 | Param TLV (3) |   Length:                  7 |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |  Integer  (4) |   Length:                  4 |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 | Increment #                                                100 |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 10: Example encoding for Outbound Traffic Control case

"LOCAL PREFERENCE" Wide Community has been defined in [I-D.ietf-idr-registered-wide-bgp-communities]

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. NO_ADVERTISE Community [RFC1997]

4. Wide BGP Community Attribute:

```
LOCAL PREFERENCE:
       Type: 0x0001                 S = src AS #
       F = 0x80                     C = 0x00000000
       H = 0                        T = Type_TBD (BGP IPv4 neighbor)
       L = 36 octets                E = none
       R = 20                       P = Type_4 (0x64)
```

Where "BGP IPv4 neighbor" Atom TLV contains:
The BGP session IPv4 local address: Local BGP Speaker IGW2
The BGP session IPv4 peer address: Remote BGP Peer Speaker1

IGW2 processes the received BGP FS RPD route as follows:

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

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

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

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

4. IGW2 gets the action type from the Wide BGP Community Attribute: LOCAL PREFERENCE;

5. IGW2 gets the matching criteria from the Wide BGP Community Attribute: Local BGP Speaker IGW2, Remote BGP Peer Speaker1;

6. IGW2 gets the parameter for "LOCAL PREFERENCE" from the Wide BGP Community Attribute: increment 100;

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.

## 5.4. Capability Negotiation

It is necessary to negotiate the capability to support BGP FlowSpec Extensions for Route Policy Distribution (RPD). The BGP FS RPD Capability is a new BGP capability [RFC5492]. The Capability Code for this capability is to be specified by the IANA. The Capability Length field of this capability is variable. The Capability Value field consists of one or more of the following tuples:

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

Figure 11: BGP FS RPD Capability

The meaning and use of the fields are as follows:

Address Family Identifier (AFI): This field is the same as the one used in [RFC4760].

Subsequent Address Family Identifier (SAFI): This field is the same as the one used in [RFC4760].

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

# 6. Consideration

## 6.1. Route-Policy

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

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

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

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

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

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

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

Routing policies are implemented using the following procedures:

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

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

## 7. Contributors

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

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

## 8. IANA Considerations

TBD.

## 9. Security Considerations

TBD.

## 10. Acknowledgements

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

## 11. References

### 11.1. Normative References

[I-D.ietf-idr-wide-bgp-communities]
Raszuk, R., Haas, J., Lange, A., Amante, S., Decraene, B., Jakma, P., and R. Steenbergen, "Wide BGP Communities Attribute", draft-ietf-idr-wide-bgp-communities-02 (work in progress), May 2016.

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

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

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

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

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

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

## 11.2. Informative References

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

Authors' Addresses

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

Email: lizhenbin@huawei.com

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

Email: oul@gsta.com

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

Email: luoyuj@gsta.com

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

Email: jasonlu@tencent.com

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