Dissemination of Flow Specification Rules for NVO3
draft-hao-idr-flowspec-nvo3-00.txt

Abstract

This document defines BGP flow-spec extension for NVO3. A Flag in BGP Path Attribute is introduced to indicate the Flow-spec rules imposing on NVO3 outer or inner layer. A new subset of NVO3 specific component types and extended community also are defined.

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
1. Introduction

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

RFC5575 defines a new BGP Network Layer Reachability Information (NLRI) format used to distribute traffic flow specification rules. NLRI (AFI=1, SAFI=133) is for IPv4 unicast filtering. NLRI (AFI=1, SAFI=134) is for BGP/MPLS VPN filtering. The Flow specification match part only includes single layer IP information like source/destination prefix, protocol, ports, and etc, it can’t be used for overlay network like NVO3 directly.

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. GENEVE [draft-ietf-nvo3-geneve-00], GUE [draft-ietf-nvo3-gue-01] and GPE [draft-ietf-nvo3-vxlan-gpe-00] are three emerging NVO3 encapsulations in progress.

In NVO3 network, we also have requirements to deploy BGP Flow-spec for DDoS attack traffic mitigation or for traffic steering. The Flow specification rules in NVO3 network can be based on inner layer 2 Ethernet information, inner layer 3 IP information, outer layer 2 Ethernet information, outer layer 3 IP information, and/or NVO3 header information. However, current flow specification match part only includes single layer Ethernet or IP information, it can’t be imposed on the NVO3 traffic directly due to lack of Flag to indicate which layer should be filtered, current flow specification match part also doesn’t include the NVO3 header information.

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

2. Use cases

2.1. Flow-spec in Data Center

BGP Flow-spec can be used for DDOS mitigation and traffic steering for service chaining in a data center. A centralized BGP Flowspec speaker should be deployed in the data center.

When BGP Flow-spec is used for DDOS mitigation, the centralized speaker also acts as a traffic analyzer. When the analyzer detects abnormal traffic, it will automatically generate Flow-spec rules and distribute it to remote BGP peers, i.e., the remote NVEs connecting TSs or outside networks. Inner layer 2/3 information and VN ID should be included as the match part. Traffic-rate and traffic-marking are useful for simple attacks. However, for complicated attacks, traffic-redirect should be used for DDOS traffic cleaning. Currently traffic-redirect method only includes redirecting traffic into a VRF by specifying the VPN RT value and redirecting to IP. If the traffic cleaner is also attached to the NVO3 network, redirecting traffic into a NVO3 tunnel will be more straightforward to steer the abnormal traffic to the cleaner device attached to a remote NVE.

When BGP Flow-spec is used for service chaining, the classification rules is determined on the centralized speaker and is distributed to each classifier to steer the traffic to a service function path
(SFP). A SFP can be multiple segmented NVO3 tunnel stitched together
to forward the specific traffic through an ordered SFs. In this case,
redirecting traffic to a specified SFP should be enforced on the
classifier.

2.2. Flow-spec for Data Center Interconnection

```
  +-----+  +-----+  +-----+  +-----+
  | CE  |  | PE   |  | NVE1 |  \--|GW1|   | GW3|   | NVE3|
  +-----+  +-----+  +-----+  +-----+
  |       |  |       |  |       |  |       |
  | NVO-1 |  | MPLS  |  | NVO-2 |  |       |
  +-----+  +-----+  +-----+  +-----+  +-----+
  |       |  |       |  |       |  |       |
  | NVE2 |  |       |  |       |  |       |
  +-----+  +-----+  +-----+  +-----+  +-----+
  +-----+  +-----+  +-----+  +-----+
```

Figure 1 NVO3 data center interconnection

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

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

2. Segmented NVO3 tunnels across different data centers. NVE1 doesn’t
   perform end to end NVO3 encapsulation to NVE3 for DCI
interconnection. The GW performs NVO3 tunnel encapsulation, and
then transmits the inner original traffic through MPLS network to
peer data center GW. The peer data center GW terminates MPLS
encapsulation in WAN network, and then uses another NVO3
encapsulation to transmit the traffic to local NVE3.
In the first solution, to differentiate bandwidth and QoS among different tenants or applications, different TE tunnels in the WAN network will be used to carry the end to end VNO3 encapsulation traffic using VN ID, VNO3 outer header DSCP and etc as traffic classification match part. BGP Flow-spec can be used to set the traffic classification.

In the second solution, a centralized BGP speaker can be deployed for DDOS mitigation in the WAN network. When the analyzer detects abnormal traffic, it will automatically generate Flow-spec rules and distribute it to each GW, the match part should include inner or outer L2/L3 layer or VNO3 header.

2.3. Requirements Summary

1. The match part should include inner L2/L3 header information and VNO3 header.

2. The Traffic Filtering Actions supports redirect to TE tunnel or VNO3 tunnel.

3. The Flow Specification encoding for VNO3

The NLRI format for this address family consists of a fixed-length Route Distinguisher field (8 bytes) followed by a flow specification, following the encoding defined in this document. 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. Flow rules are accepted by default, when received from remote PE routers.

Because VNO3 encapsulation includes outer layer 2/3 header, inner layer 2/3 header and VNO3 header, inner and outer layer Flag is introduced to indicate the Flow-spec rules imposing on outer or inner layer. The component types defined in [RFC5575],[draft-ietf-idr-flow-spec-v6-06] and [draft-ietf-idr-flowspec-l2vpn-01] combined with the Flag indicator can be used for the VNO3 traffic filtering. A bit in ‘traffic action’ is applied as the inner and outer layer indicator.
For NVO3 header part, the following additional component types are introduced.

Type TBD1 - VN ID

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

Defines a list of (operation, value) pairs used to match 24-bit VN ID fields. VN ID is tenant identification in NVO3 network. Values are encoded as 1- to 3-byte quantities.

Type TBD2 - NVO3 Proto Type

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

Defines a list of (operation, value) pairs used to match 8-bit Protocol Type fields. Values are encoded as 1-byte quantity.

Other types:

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


Layer Flag(Bit 45): When this bit is set, the corresponding filtering rules will be applied on the NVO3 inner layer. If not set, the corresponding filtering rules will be applied on the NVO3 outer layer.

<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>

Besides to support the above extended communities per RFC5575, similar to ’redirect to IP’ [REDIRECTIP], a new extended community
of ‘redirect to Tunnel’ is requested for service chaining and traffic steering. The sub-type value [to be assigned by IANA] indicates that the global administrator and local administrator fields encode a flow-spec ‘redirect to tunnel’ action. In the new extended communities 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           |C|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2: Local Administrator

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.

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.

A set of tunnels can be specified in the BGP Path Attribute. In [draft-rosen-idr-tunnel-encaps-00], the tunnel type and encapsulation information of VXLAN, NVGRE, VXLAN-GPE and etc are defined using Tunnel Encapsulation Attribute Sub-TLVs. In [draft-li-idr-mpls-path-programming-01], the tunnel type and encapsulation information of RSVP-TE, LDP, Segment Routing Path and etc are defined using Extended Unicast Tunnel Attributes.

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 or more ‘redirect to Tunnel’ extended communities, the BGP speaker SHOULD load-share the redirected packets across all the Tunnels.

5. Security Considerations

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

A bit in ’Traffic Action’ is requested as the inner and outer layer indicator for Flow-spec rules.

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

"Flow spec NVO3 Component Types”:

<table>
<thead>
<tr>
<th>Type</th>
<th>VN ID</th>
<th>NVO3 Proto Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>TBD1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TBD2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

IANA is requested to update the reference for the following assignment in the "BGP Extended Communities Type - extended, transitive" registry:

<table>
<thead>
<tr>
<th>Type value</th>
<th>Name</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>TBD</td>
<td>Redirect to Tunnel</td>
<td>[this document]</td>
</tr>
</tbody>
</table>

6.1. Normative References


6.2. Informative References


7. Acknowledgments

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