INTERNET-DRAFT
Intended Status: Proposed Standard

D. Eastlake
Futurewei Technologies
W. Hao
S. Zhuang
Z. Li
Huawei Technologies
R. Gu
China Mobil

Expires: May 3, 2020
November 4, 2019

BGP Dissemination of Flow Specification Rules for Tunneled Traffic
draft-ietf-idr-flowspec-nvo3-07

Abstract

This draft specifies a Border Gateway Protocol Network Layer Reachability Information (BGP NLRI) encoding format for flow specifications (RFC 5575bis) that can match on a variety of tunneled traffic. In addition, flow specification components are specified for certain tunneling header fields.

Status of This Document

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

Distribution of this document is unlimited. Comments should be sent to the authors or the IDR Working Group mailing list <idr@ietf.org>.

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.

Table of Contents

1. Introduction

BGP Flow-spec [RFC5575bis] is an extension to BGP that supports the dissemination of traffic flow specification rules. It uses the BGP control plane to simplify the distribution of Access Control Lists (ACLs) and allows new filter rules to be injected to all BGP peers simultaneously without changing router configuration. A typical application of BGP Flow-spec is to automate the distribution of traffic filter lists to routers for Distributed Denial of Service (DDOS) mitigation.

BGP Flow-spec defines a BGP Network Layer Reachability Information (NLRI) format used to distribute traffic flow specification rules. AFI=1/SAFI=133 is for IPv4 unicast filtering. AFI=1/SAFI=134 is for IPv4 BGP/MPLS VPN filtering. [FlowSpecV6] and [Layer2- FlowSpec] extend the flow-spec rules for IPv6 and layer 2 Ethernet packets respectively. All these previous flow specifications match only a single level of IP/Ethernet information fields such as source/destination IP prefix, protocol type, source/destination MAC, ports, EtherType and the like.

In the cloud computing era, multi-tenancy has become a core requirement for data centers. It is increasingly common to see tunneled traffic with a field to distinguish tenants. An example is the Network Virtualization Over Layer 3 (NVO3 [RFC8014]) overlay technology that can satisfy multi-tenancy key requirements. VXLAN [RFC7348] and NVGRE [RFC7637] are two typical NVO3 encapsulations. Other encapsulations such as IP-in-IP or GRE may be encountered. Because these tunnel / overlay technologies involving an additional level of encapsulation, flow specification that can match on the inner header as well as the outer header are needed.

In summary, the Flow specifications should be able to include inner nested header information as well as fields specific to the type of tunneling in use such as virtual network / tenant ID. This draft specifies methods for accomplishing this using SAFI=TBD1 and a new NLRI encoding.

1.1 Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

The reader is assumed to be familiar with BGP terminology. The following terms and acronyms are used in this document with the meaning indicated:

ACL - Access Control List

DDOS - Distributed Denial of Service (Attack)

DSCP - Differentiated Services Code Point

GRE - Generic Router Encapsulation [RFC2890]

L2TPv3 - Layer Two Tunneling Protocol - Version 3 [RFC3931]

NLRI - Network Layer Reachability Information

NVGRE - Network Virtualization Using Generic Routing Encapsulation [RFC7637]

NVO3 - Network Virtual Overlay Layer 3 [RFC8014]

VN - virtual network

VXLAN - Virtual eXtensible Local Area Network [RFC7348]

2. Tunneled Traffic Flow Specification NLRI

The Flow-spec rules in [RFC5575bis], [FlowSpecV6], and [FlowSpecL2] can only recognize flows based on one level of header in a data packet. To enable flow specification of tunneled traffic, a new SAFI (TBD1) and NLRI encoding are introduced. This encoding, shown in Figure 1, enables flow specification of more than one layer of header when needed.

```
      +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
      | Length                          2 bytes       |
      +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
      | Tunnel Type                     2 bytes       |
      +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
    Flags:
      +--+--+--+--+--+--+--+--+
      | D| I| Reserved        |         1 byte
      +--+--+--+--+--+--+--+--+
    Optional Routing Discriminator:
      +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
      |                                               |
      +                                               +
      |                                               |
      + Routing Discriminator           8 bytes       +
      |                                               |
      +                                               +
      |                                               |
      +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
    Outer Flow-spec:
      +--+--+--+--+--+--+--+--+
      | Outer Flowspec Length :         1 or 2 bytes
      +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
      | Outer Flowspec                  variable      :
      +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
    Optional Inner Flow-spec:
      +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
      | Inner AFI                       2 bytes       |
      +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
      | Inner Flowspec Length :         1 or 2 bytes
      +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
      | Inner Flowspec                  variable      :
      +--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
```

Figure 1. Tunneled Traffic Flow-spec NLRI

Length - The NLRI Length encoded as an unsigned integer including the Tunnel Type.

Tunnel Type - The type of tunnel using a value from the IANA BGP Tunnel Encapsulation Attribute Tunnel Types registry.

Flags: D bit - Indicates the presence of the Routing Discriminator (see below).

Flags: I bit - Indicates the presence of an inner AFI and Flow-spec.

Flags: Reserved - Six bits that MUST be sent as zero and ignored on receipt.

Routing Discriminator - If the outer layer 3 address belongs to a BGP/MPLS VPN, the routing discriminator can be included to support traffic filtering within that VPN. Because NVO3 outer layer addresses normally belong to a public network, a Route Distinguisher field is normally not needed for NVO3.

Outer Flowspec / Length - The flow specification for the outer header. The length is encoded as provided in Section 4.1 of [RFC5575bis]. The AFI for the outer flowspec is that AFI at the beginning of the BGP multiprotocol MP_REACH_NLRI or MP_UNREACH_NLRI containing the tunneled traffic flow specification NLRI.

Inner AFI - Depending on the Tunnel Type, there may be an inner AFI that indicates the address family for the inner flow specification. There is no need for a SAFI as it is automatically TBD1, the SAFI for a tunneled traffic flow specification.

Inner Flowspec / Length - Depending on the Tunnel Type, there may be an inner flow specification for the header level encapsulated within the outer header. The length is encoded as provided in Section 4.1 of [RFC5575bis].

2.1 SAFI Code Point

Use of the tunneled traffic flow specification NLRI format is indicated by SAFI=TBD1. This is used in conjunction with the AFI for the outer layer 3 header, that is AFI=1 for IPv4 and AFI=2 for IPv6.

2.2 Component Code Points

For flow specification based on certain tunnel header fields, the component types below are added. These are associated with the Tunnel Type and MAY appear in the outer flow specification or, if it is present, in the inner flow specification.

Type TBD2 - VN ID
Encoding: <type (1 octet), length (1 octet), [op, value]+>.

Defines a list of {operation, value} pairs used to match the 24-bit VN ID that is used as the tenant identification in some tunneling headers. For VXLAN encapsulation, the VN ID is the VNI. For NVGRE encapsulation, the VN ID is the VSID. op is encoded as specified in Section 4.2.3 of [RFC5575bis]. Values are encoded as 1- to 3-byte quantities.

Type TBD3 - Flow ID
Encoding: <type (1 octet), length (1 octet), [op, value]+>

Defines a list of {operation, value} pairs used to match 8-bit Flow ID fields which are currently only useful for NVGRE encapsulation. op is encoded as specified in Section 4.2.3 of [RFC5575bis]. Values are encoded as 1-byte quantity.

Type TBD4 - Session
Encoding: <type (1 octet), length (1 octet), [op, value]+>

Defines a list of {operation, value} pairs used to match a 32-bit Session field. This field is called Key in GRE [RFC2890] encapsulation and Session ID in L2TPv3 encapsulation. op is encoded as specified in Section 4.2.3 of [RFC5575bis]. Values are encoded as a 1, 2, or 4 byte quantity.

Type TBD5 - Cookie
Encoding: <type (1 octet), length (1 octet), [op, value]+>

Defines a list of {operation, value} pairs used to match a variable length Cookie field. This is only useful in L2TPv3 encapsulation. op is encoded as specified in Section 4.2.3 of [RFC5575bis]. Values are encoded as a 1, 2, 4, or 8 byte quantity. If the Cookie does not fit exactly into the value length, it is left justified, that is, padded with following bytes the MUST be sent as zero and ignored on receipt.

Type TBD6 - VXLAN-GPE Flags
Encoding: <type (1 octet), length (1 octet), [op, bitmask]+>

Defines a list of {operation, value} pairs used to match against the VSLAN-GPE flags field. op is encoded as in Section 4.2.9 of [RFC5575bis]. bitmask is encoded as 1 byte.

2.3 Specific Tunnel Types

The following subsections describe how to handle flow specification for several specific tunnel types.

2.3.1 VXLAN

The headers on a VXLAN [RFC7348] data packet are an outer Ethernet header, an outer IP header, a UDP header, the VXLAN header, and an inner Ethernet header. This inner Ethernet header is frequently, but not always, followed by an inner IP header. If the tunnel type is VXLAN, the I flag MUST be set.

The version (IPv4 or IPv6) of the outer IP header is indicated by the AFI at the beginning of the multiprotocol MP_REACH_NLRI or MP_UNREACH_NLRI containing the tunneled traffic flow specification NLRI. The outer flowspec is used to filter the outer headers and the UDP header.

The inner flowspec is used on the Inner Ethernet header [FlowSpecL2]. If the inner AFI is 25, then whether or not an IP header follows the inner Ethernet header is ignored and the inner flowspec SHOULD NOT contain and IPv4 or IPv6 flowspec components. If the inner AFI is 1 or 2, to match the flowspec the Inner Ethernet header must be followed by an IPv4 or IPv6 header, respectively, and the inner flowspec is also used to filter that inner IP header.

A component filtering on the VXLAN header VN ID (VNI) can appear in either the outer or inner flowspec. The inner MAC/IP address is associated with a VN ID. In the NVO3 terminating into a VPN scenario, if multiple access VN IDs map to one VPN instance, one shared VN ID can be carried in the Flow-Spec rule to enforce the rule on the entire VPN instance and the shared VN ID and VPN correspondence should be configured on each VPN PE beforehand. In this case, the function of the layer 3 VN ID is the same as a Route Discriminator: it acts as the identification of the VPN instance.

2.3.2 VXLAN-GPE

VXLAN-GPE [GPE] is similar to VXLAN and the VXLAN-GPE header is the same size as the VXLAN header but has been extended from the VXLAN header by specifying a number of bits that are reserved in the VXLAN header. In particular, a number of additional flag bits are specified and a Next Protocol field is added that is valid if the P flag bit is set. These flags bits can be tested using the VXLAN-GPE Flags component defined above. VXLAN and VXLAN-GPE are distinguished by the port number in the UDP header the precedes the VXLAN or VXLAN-GPE headers.

If the VXLAN-GPE header P flag is zero, then the header is followed by the same sequence as for VXLAN and the same flow-spec choices apply (see Section 2.3.1).

If the VXLAN-GPE header P flag is one and that header's next protocol field is 1, then the VXLAN-GPE header is followed by an IPv4 header. The inner AFI/flowspec match only if the inner AFI is 1 and the inner flowspec matches.

If the VXLAN-GPE header P flag is one and that header's next protocol field is 2, then the VXLAN-GPE header is followed by an IPv6 header. The inner AFI/flowspec match only if the inner AFI is 2 and the inner flowspec matches.

2.3.3 NVGRE

NVGRE [RFC7637] is very similar to VXLAN except that the UDP header and VXLAN header immediately after the outer IP header are replaced by a GRE (Generic Router Encapsulation) header. The GRE header as used in NVGRE has no Checksum or Reserved1 field as shown in [RFC2890] but there are Virtual Subnet ID and FlowID fields in place of what is labeled in [RFC2890] as the Key field. Processing and restrictions for NVGRE are as in Section 2.3.1 eliminating references to a UDP header and replacing references to the VXLAN header and its VN ID with references to the GRE header and its VN ID (VSID) and Flow ID.

2.3.4 L2TPv3

The headers on an L2TPv3 [RFC3931] packets are an outer Ethernet header, an outer IP header, the L2TPv3 header, an inner Ethernet header, and possibly an inner IP header if indicated by the inner Ethernet header EtherType. The outer flowspec operates on the outer headers that precede the GRE header. The version of IP is specified by the outer AFI at the beginning of the MP_REACH_NLRI or MP_UNREACH_NLRI.

The L2TPv3 header consists of a 32-bit Session ID followed by a variable length Cookie (maximum length 8 bytes). The Session ID and Cookie can be filtered for by using the Session and Cookie flowspec components. To filter on Cookie or even be able to bypass Cookie and parse the remainder of the L2TPv3 packet, the node implementing flowspec needs to know the length and/or value of the Cookie fields of interest. This is negotiated at L2TPv3 session establishment and it is out of scope for this document how the node would learn this information. Of course, if flowspec is being used for DDOS mitigation and the Cookie has a fixed length and/or value in the DDOS traffic, this could be learned by inspecting that traffic.

If the I flag bit is zero, then no filtering is done on data beyond the L2TPv3 header. If the I flag is one, indicating the presence of an inner flowspec, and the node implementing flowspec does not know the length of the L2TPv3 header Cookie, the match fails. If that node does know the length of that Cookie, the inner flowspec if matched against the headers at the beginning of that data using the inner AFI. If the inner AFI is 1 or 2, then an inner IP header is required and filtering can be done on the Ethernet header immediately after the L2TPv3 header and the following IPv4 or IPv6 headers respectively. If the inner AFI is 25, filtering SHOULD only be done on the inner Ethernet header [FlowSpecL2].

2.3.5 GRE

Generic Router Encapsulation (GRE [RFC2890]) is a common type of encapsulation. The outer flowspec operates on the outer headers that precede the GRE header. The version of IP is specified by the outer AFI at the beginning of the MP_REACH_NLRI or MP_UNREACH_NLRI.

If the I flag bit is zero, no filtering is done on data after the GRE header. If the I flag bit is one, then there is an inner AFI and flowspec and the Protocol Type field of the GRE header must match the inner AFI as follows for the flowspec to match:

GRE Protocol Type	Inner AFI
0x0800 (IPv4)	1
0x86DD (IPv6)	2
0x6558	25

With the I flag a one and the inner AFI and GRE Protocol Type fields match, the inner flowspec is used to filter the inner Ethernet header (AFI=25) or the inner IP and Ethernet headers (AFI=1 or 2).

2.3.6 IP-in-IP

IP-in-IP encapsulation is shown when the outer IP header indicates an inner IP IPv4 or IPv6 header by the value of the outer IP header's Protocol (IPv4) or Next Protocol (IPv6) field. If the Tunnel Type is

IP-in-IP, the I flag MUST be set.

The version of the outer IP header (IPv4 or IPv6) matched is indicated by the AFI at the beginning of the MP_REACH_NLRI or MP_UNREACH_NLRI. The version of the inner IP header is indicated by the inner AFI. The outer flowspec applies to the outer IP header and the inner flowspec applies to the inner IP header.

2.4 Tunneled Traffic Actions

The previously specified traffic filtering actions are used for tunneled traffic [RFC5575bis] [FlowSpecL2]. For Traffic Marking in NVO3, only the DSCP in the outer header can be modified.

3. Order of Traffic Filtering Rules

In comparing an applicable tunneled traffic flow specification with a non-tunneled flow specification, the tunneled specification has precedence.

If comparing two tunneled traffic flow specifications, if both are applicable, the tunnel types will be the same. If only one has a Routing Discriminator, it has precedence. If both have a Routing Discriminator, those discriminators are compared as unsigned integers and the one with the smaller magnitude Routing Discriminator has precedence.

If neither has a Routing Discriminator or they have equal Routing Discriminators, the order of precedence is determined by comparing the outer flowspec.

If the outer flowspecs are equal and the tunnel type calls for an inner flowspec, then the precedence is determined by comparing inner AFI as an unsigned integer with the inner AFI having the smaller magnitude having precedence.

If the inner AFIs are equal, precedence is determined by comparing the inner flow specifications.

4. Flow Spec Validation

Flow-specs received over AFI=1/SAFI=TBD1 or AFI=2/SFAI=TBD1 are validated, using only the outer Flow-spec, against routing reachability received over AFI=1/SAFI=133 and AFI=2/SAFI=133 respectively, as modified by [FlowSpecOID].

5. Security Considerations

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

6. IANA Considerations

IANA is requested to assign a new SAFI as follows:

Value	Description	Reference
TBD1	Tunneled traffic flow specification rules	[This document]

IANA is requested to assign two new values in the "Flow Spec Component Types" registry as follows:

Type	Name	Reference
TBD2	VN ID	[this document]
TBD3	Flow ID	[this document]
TBD4	Session	[this document]
TBD5	Cookie	[this document]
TBD6	VXLAN-GPE Flags	[this document]

Normative References

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

[RFC2890] - Dommety, G., "Key and Sequence Number Extensions to GRE", RFC 2890, DOI 10.17487/RFC2890, September 2000, <https://www.rfc-editor.org/info/rfc2890>.

[RFC3931] - Lau, J., Ed., Townsley, M., Ed., and I. Goyret, Ed., "Layer Two Tunneling Protocol - Version 3 (L2TPv3)", RFC 3931, DOI 10.17487/RFC3931, March 2005, <https://www.rfc-editor.org/info/rfc3931>.

[RFC7348] - Mahalingam, M., Dutt, D., Duda, K., Agarwal, P., Kreeger, L., Sridhar, T., Bursell, M., and C. Wright, "Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", RFC 7348, DOI 10.17487/RFC7348, August 2014, <https://www.rfc-editor.org/info/rfc7348>.

[RFC7637] - Garg, P., Ed., and Y. Wang, Ed., "NVGRE: Network Virtualization Using Generic Routing Encapsulation", RFC 7637, DOI 10.17487/RFC7637, September 2015, <https://www.rfc-editor.org/info/rfc7637>.

[RFC8174] - Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

[FlowSpecL2] - W. Hao, etc, "Dissemination of Flow Specification Rules for L2 VPN", draft-ietf-idr-flowspec-l2vpn, work in progress.

[FlowSpecOID] - J. Uttaro, J. Alcaide, C. Filsfils, D. Smith, P. Mohapatra, "Revised Validation Procedure for BGP Flow Specifications", draft-ietf-idr-bgp-flowspec-oid, work in progress.

[FlowSpecV6] - R. Raszuk, etc, "Dissemination of Flow Specification Rules for IPv6", draft-ietf-idr-flow-spec-v6, work in progress.

[RFC5575bis] - Hares, S., Loibl, C., Raszuk, R., McPherson, D., Bacher, M., "Dissemination of Flow Specification Rules", draft-ietf-idr-rfc5575bis-17, Work in progress, January 2019.

Informative References

[RFC8014] - Black, D., Hudson, J., Kreeger, L., Lasserre, M., and T. Narten, "An Architecture for Data-Center Network Virtualization over Layer 3 (NVO3)", RFC 8014, DOI 10.17487/RFC8014, December 2016, <https://www.rfc-editor.org/info/rfc8014>.

[GPE] - P. Quinn, etc, "Generic Protocol Extension for VXLAN", draft-ietf-nvo3-vxlan-gpe, work in progress.

Acknowledgments

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

Authors' Addresses

Donald Eastlake
Futurewei Technologies
2386 Panoramic Circle
Apopka, FL 32703 USA

Tel: +1-508-333-2270
Email: d3e3e3@gmail.com

Weiguo Hao
Huawei Technologies
101 Software Avenue,
Nanjing 210012 China

Email: haoweiguo@huawei.com

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

Email: zhuangshunwan@huawei.com

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

Email: lizhenbin@huawei.com

Rong Gu
China Mobile

Email: gurong_cmcc@outlook.com

Copyright, Disclaimer, and Additional IPR Provisions

Copyright (c) 2019 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.