IDR Working Group S. Hares
Internet-Draft Hickory Hill Consulting
Intended status: Standards Track D. Eastlake
Expires: March 25, 2022 Futurewei
September 26, 2021

BGP Flow Specification Version 2
draft-hares-idr-flowspec-v2-03

Abstract

BGP flow specification version 1 (FSv1) defined in RFC 8955, RFC 8956, and RFC 9117 describes the distribution of traffic filter policy (traffic filters and actions) distributed via BGP. Multiple applications have used BGP FSv1 to distribute traffic filter policy. These applications include the following: mitigation of denial of service (DoS), enabling traffic filtering in BGP/MPLS VPNs, centralized traffic control of router firewall functions, and SFC traffic insertion.

During the deployment of BGP FSv1 a number of issues were detected due to lack of consistent TLV encoding for rules for flow specifications, lack of user ordering of filter rules and/or actions, and lack of clear definition of interaction with BGP peers not supporting FSv1. Version 2 of the BGP flow specification (FSv2) protocol addresses these features. In order to provide a clear demarcation between FSv1 and FSv2, a different NLRI encapsulates FSv2.

Status of This Memo

Table of Contents

1. Introduction

Modern IP routers have the capability to forward traffic and to classify, shape, rate limit, filter, or redirect packets based on administratively

defined policies. These traffic policy mechanisms allow the operator to define match rules that operate on multiple fields within header of an IP data packet. Upon a match, the traffic policy allows actions to be associated with each match rule. These rules can be more widely defined as “event-condition-action” (ECA) rules where the event is always the reception of a packet.

BGP ([RFC4271]) flow specification as defined by [RFC8955], [RFC8956], and [RFC9117] specifies the distribution of traffic filter policy (traffic filters and actions) via BGP to a mesh of BGP peers (IBGP and EBGP peers). The traffic filter policy is applied when packets are received on a router with the flow specification function turned on. The flow specification protocol defined in [RFC8955], [RFC8956], and [RFC9117] will be called BGP flow specification version 1 (BGP FSv1) in this draft.

Some modern IP routers also include the abilities of firewalls which can match on a sequence of packet events based on administrative policy. These firewall capabilities allow for user ordering of match rules and user ordering of actions per match.

Multiple deployed applications currently use BGP FSv1 to distribute traffic filter policy. These applications include: 1) mitigation of Denial of Service (DoS), 2) traffic filtering in BGP/MPLS VPNS, and 3) centralized traffic control for networks utilizing SDN control of router firewall functions, 4) classifiers for insertion in an SFC, and 5) filters for SRv6 routing.

During the deployment of BGP FSv1, the following issues were detected:

- lack of consistent TLV encoding prevented extension of encodings,
- inability to allow user defined order for filtering rules,
- inability to order actions to provide deterministic interactions or to allow users to define order for actions,
- no clearly defined mechanisms for BGP peers which do not support flow specification v1.

Networks currently cope with some of these issues by limiting the type of traffic filter policy sent in BGP. Current Networks does not have a good workaround/solution for applications that receive but do not understand FSv1 policies.

This document defines version 2 of the BGP flow specification protocol to address these shortcomings in BGP FSv1. Version 2 of BGP flow specification will be denoted as BGP FSv2.

BGP FSv1 as defined in [RFC8955], [RFC8955], and [RFC9117] specified 2 SAFIs (133, 134) to be used with IPv4 AFI (AFI = 1) and IPv6 AFI (AFI=2)

This document specifies 2 new SAFIs (TBD1, TBD2) for FSv2 to be used with 5 AFIs (1, 2, 6, 25, and 31) to allow user-ordered lists of traffic match filters for user-ordered traffic match actions encoded in Communities (Wide or Extended) or a SubTLV of the FSv2 NRLI.
FSv1 and FSv2 use different AFI/SAFIs to send flow specification filters. Since BGP AFI/SAFIs perform route selection per AFI/SAFI, this approach can be termed “ships in the night” based on AFI/SAFI.

FSv1 is a critical component of deployed applications. Therefore, this specification defines how FSv2 will interact with BGP peers that support either FSv2 or FSv1 or BGP peers that do not support either FSv1 or FSv2. It is expected that a transition to FSv2 will occur over time as new applications require FSv2 extensibility and user-defined ordering for rules and actions or network operators tire of the restrictions of FSv1 (error handling issues and restricted topologies.

This section contains a short review of FSv1 and an overview of FSv2.

Section 3 contains the definition of flow specification v2. Section 4 contains the encoding rules for FSv2 and user-based encoding sent via BGP, and section 5 describes how to validate FSv2 NLRI. Section 6 discusses how to combine FSv2 user-ordered match rules and FSv1 rules. Section 6 also discusses how to combine user-ordered actions, FSv1 actions, and default actions. Sections 7-10 address an alternate security mechanism, considerations for IANA, security in deployments, and manageability.

1.1. Flow Specification v1 (FSv1) Review

The FSv1 NLRI defined in [RFC8955] and [RFC8956] for this policy include 13 match conditions encoded for the following AFI/SAFIs

- IPv4 traffic: AFI:1, SAFI:133
- IPv6 Traffic: AFI:2 SAFI:133
- BGP/MPLS IPv4 VPN: AFI:1, SAFI: 134
- BGP/MPLS IPv6 VPN: AFI:2, SAFI: 134

If one considers the reception of the packet as an event, then BGP flow specification describes a set of Event-Condition-Action (ECA) policies where:

- event is the reception of a packet,
- condition stands for “match conditions” defined in the BGP NLRI as an n-tuple of component filters, and
- action taken is either: the default action (accept the packet for normal packet flow), or a set of actions (1 or more) defined in BGP Extended Community.

The flow specification conditions and actions combine to make up FSv1 specification rules. Each FSv1 NLRI must have a type 1 component (destination prefix) and Extended Communities with FSv1 actions can be attached to a single NLRI or multiple NLRIs in a BGP packet.

Within an AFI/SAFI pair, FSv1 rules are ordered based on the components in the packet (types 1-13) ordered from left-most to right-most and within the component types by value of the component. Rules are inserted in the rule list by component type where a FSv1 rule with existing component type has higher precedence than one missing a specific component type.

Since FSv1 specifications ([RFC8955],[RFC8956], and [RFC9117]) specify that the FSv1 NLRI MUST have a destination prefix (as component type 1) embedded in the flow specification, the FSv1 rules with destination components are ordered by IP Prefix comparison rules for IPv4 [RFC8955] and IPv6 [RFC8956]. [RFC8955] specifies that more specific prefixes (aka longest match) have higher precedence than that of less specific prefixes AND for prefixes of the same length the lower IP number is selected (lowest IP value). [RFC8956] specifies that if the offsets within component 1 are the same,

then the longest match and lowest IP comparison rules from [RFC8955] apply. If the offsets are different, then the lower offset has precedence.

These rules work to provide a set of FSv1 rules ordered by IP Destination Prefix by longest match and lowest IP address. [RFC8955] also states that the requirement for a destination prefix component “MAY be relaxed by explicit configuration” [RFC8955]. Since the rule insertions are based on comparing component types between two rules in order, this means the rules without destination prefixes are inserted after all rules which contain destination prefix component.

The actions specified by FSv1 have the following actions: accept packet (default), traffic flow limitation by bytes (6), traffic-action (7), redirect IPv4 (8), mark traffic (9), and traffic flow limitation by packets (12). The traffic action specifies two functions: either a termination of FSv1 filters and enable sampling/logging. Multiple actions may be associated with a FSv1 rule and may conflict. Implementers of FSv1 actions SHOULD document the interactions between FSv1 actions in their implementation.

Figure 1 shows a diagram of a FSv1 logical data structures with 5 rules. If FSv1 rules 1-4 have destination prefix components (type=1) and FSv1 rule 5 does not have a destination prefix, then FSv1 rule 5 will be inserted in the policy after rules 1-4.

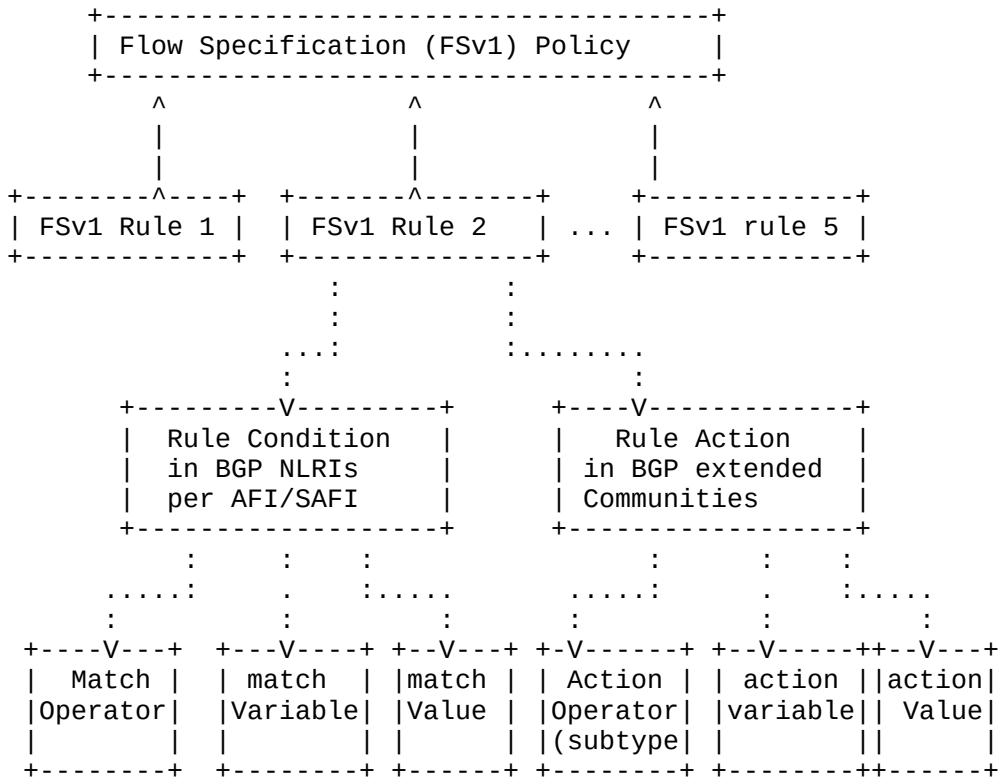


Figure 1: BGP Flow Specification Policy

1.2. Ordering for Flow Specification v2 (FSv2)

Flow Specification v2 allows the user to order the flow specification rules and the actions associated with a rule. Each FSv2 rule may have one or more match conditions and one or more associated actions.

This FSv2 specification supports the components and actions for the following:

- IPv4 (AFI=1, SAFI: TBD1),
- IPv6 (AFI=2, SAFI: TBD2),
- L2 (AFI=6, SAFI: TDB1)
- BGP/MPLS IPv4 VPN: (AFI=1, SAFI: TBD2)
- BGP/MPLS IPv6 VPN: (AFI=2, SAFI: TBD2)
- BGP/MPLS L2VPN (AFI=25, SAFI: TDB2)
- SFC: (AFI=31, SAFI: TBD1)
- SFC VPN (AFI=31, SAFI: TBD2)

The FSv2 specification for tunnel traffic is outside the scope of this specification. The FSv1 specification for tunneled traffic is in [draft-ietf-idr-flowspec-nv03].

The basic principles regarding ordering of rules are simple:

1) Rule-0 (zero) is defined to be 0/0 with the “permit-all” action

2) FSv2 rules are ordered based on user-specified order. The user-specified order is carried FSv2 NLRI with the sentence that the numerical lower value takes precedence over the numerically higher value. For rules received with the same order value, the FSv1 rules apply (order by component type and then by value of the components).

3) FSv2 rules are added starting with Rule 1 and FSv1 rules are added after FSv2 rules.

> For this example, BGP Peer A has FSv2 data base with 10 FSv2 rules (1-10) and 10 FSv1 rules (301-310).

4) An FSv2 peer may receive BGP NLRI routes from a FSv1 peer or a BGP peer that does not support FSv1 or FSv2. The capabilities sent by a BGP peer indicate whether the AFI/SAFI can be received (FSv1 NLRI or FSv2 NLRI).

> Suppose an FSv2 peer (BGP Peer A) has the capability to send either FSv1 or FSv2. BGP Peer A peers with BGP Peers B, C, D and E.
>
> BGP Peer B can only send FSv1 routes (NLRI + Extended Community). BGP Peer C can send FSv2 routes (NLRI + path attributes (wide community or extended community or none)). BGP Peer D cannot send any FS routes. BGP E can send FSv2 and FSv1 routes.
>
> BGP Peer A sends FSv1 routes in its databases to BGP B. Since the FSv2 NLRI cannot be sent to the FSv1 peer, only the FSv1 NLRI is sent.

BGP Peer A sends to BGP C the FSv2 routes in its database (configured or received).

BGP peer A would not send the FSv1 NLRI or FSv2 NLRI to BGP Peer D. The BGP Peer D does not support for these NLRI.

BGP Peer A sends the NLRI for both FSv1 and FSv2 to BGP Peer E.

5) Associate a chain of actions to rules based on user-defined action number

FSv2 allows actions to be associated by the following: a) actions in an Extended Community, b) actions in a wide community, or c) actions within the FSv2 NRLI associated as a SubTLV.

Action user-order value zero is reserved.

An action associated with FSv2 NLRI using in a SubTLV always has a user-defined order.

The precedence between two actions with user-defined order (Wide community) is discussed in detail in section 6.2.

Examples of Action ordering:

An action associated with FSv2 NLRI using in a SubTLV always has a user-defined order. If two actions have the same user-defined order and the same action type, the ordering of the actions within the same type is defined by the action type (see section 4.2).

The use case for the Action which always associated with an NRLI is the DDOS match case that always drops the packet in order to kill off a widespread DDoS attack. The idea is easy, but the deployment issues may be more complex. An example may help illustrate this point.

Suppose BGP Peer A has a configured value for FSv2ExtComActionStart of 10. Suppose BGP Peer A receives the following attributes associated with the same FSv2 NLRI to form an action chain:

- a Wide Community action with user-defined order 10 from AS 2020 that limits packet-based rate limit of 600 packets per second,
- an action SubTLV with a user order of 11 from AS 10 that limits the packet base rate to zero packets per second,
- a Wide Community with a user order of 11 from AS 2021 that limits the packet-based rate limit of 50.

The FSv2 data base would store the following action chain:

- at user-defined action order 10:

1. a user action type of 12 (packet-based rate limit), with values of AS 2020 and float value of 1000.

- at user-defined action order 11 in order:
  1. A user action of type 12 with values of AS 10 and float value of zero,
  2. A user action of type 12 with value of AS 2021 and float value of 50.

When does the action chain stop?

The default process for the action chain is to stop on failure.

If setting the packet-based rate limit of 1000 works, the action chain would go on to set the value of zero. If this works, it would go on to set the value to 50. This set of actions may not be what the user wanted if this is a DDoS attack.

BGP FSv2 rules are ephemeral by default (just as BGP routes are ephemeral) upon a restart of a BGP session or a router.

After FSv2 NLRI are checked for errors in syntax, those with valid syntax are checked for the same validation procedure FSv1 NLRI uses [RFC8955] and [RFC9117]. See section 5 for a detailed discussion of validation and error handling.

Names may be associated with rules or actions in order for network management protocols (NETCONF/RESTCONF) to be able to provide detailed reports in the BGP Yang models.

Figure 2 provides a logical diagram of the FSv2 structure.

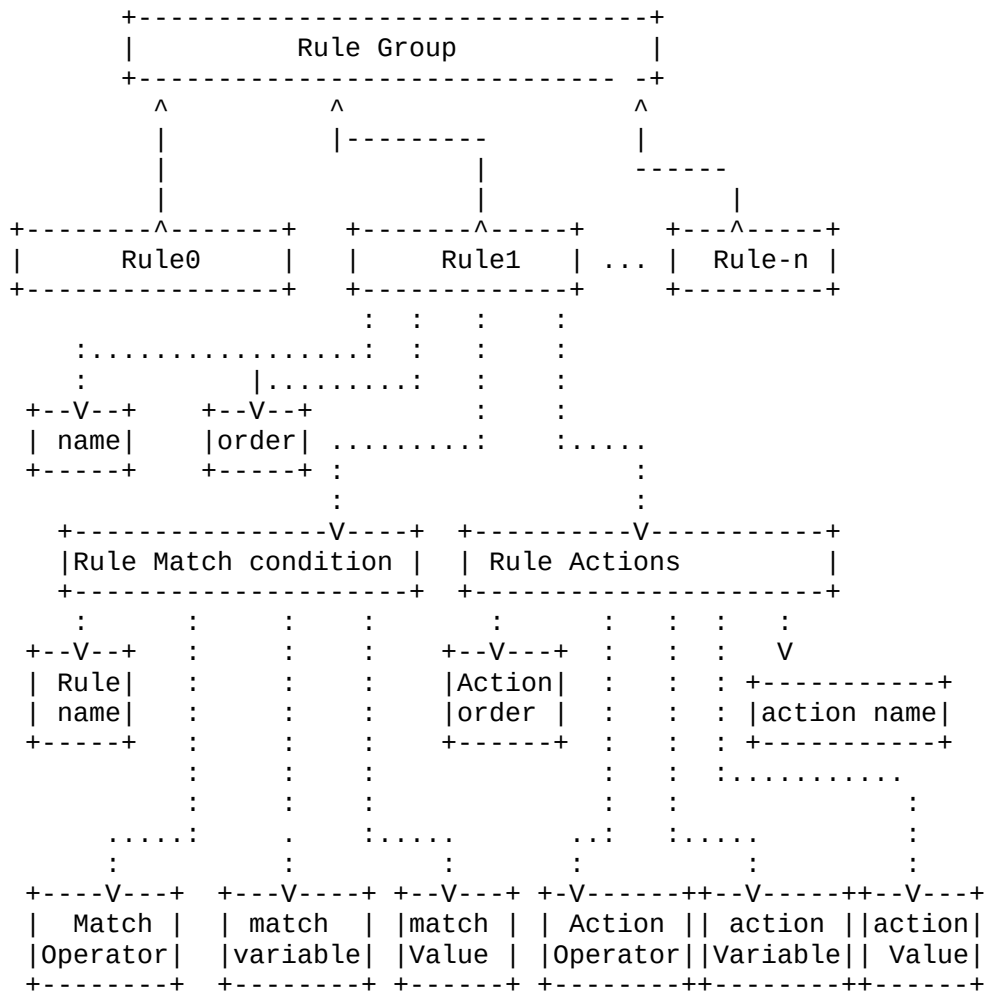


Figure 2: Order Flow Specification Data storage

2. Terminology

2.1. Definitions and Acronyms

AFI – Address Family Identifier

BGPSEC - secure BGP [RFC8205] updated by [RFC8206]

BGP Session ephemeral state - state which does not survive the loss of BGP peer,

Ephemeral state - state which does not survive the reboot of a software module, or a hardware reboot. Ephemeral state can be ephemeral configuration state or operational state.

configuration state - state which persist across a reboot of software module within a routing system or a reboot of a hardware routing device.

NETCONF - The Network Configuration Protocol [RFC6241].

RESTCONF - The RESTCONF configuration Protocol [RFC8040]

ROA - Route Origin Authentication [RFC6482]

SAFI – Subsequent Address Family Identifier

2.2. 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 BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals as shown here.

3. Flow Specification

A BGP Flow Specification is an n-tuple containing several match criteria that can be applied to IP traffic, traffic encapsulated in IP traffic or traffic associated with IP traffic. The following traffic filters are examples of traffic associated with IP traffic: IP packet or an IP packet inside a L2 packet (Ethernet), an MPLS packet, and SFC flow.

A given Flow Specification NLRI may be associated with a set of path attributes depending on the particular application, and attributes within that set may or may not include reachability information (e.g., NEXT_HOP). Extended Community or Wide Community attributes (well-known or AS-specific) MAY be used to encode a set of pre-determined actions.

A particular application is identified by a specific AFI/SAFI (Address Family Identifier/Subsequent Address Family Identifier) and responds to a distinct set of RIBs. Those RIBs should be treated independently of each other in order to assure noninterference between distinct applications.

BGP processing treats the NLRI as a key to entries in AFI/SAFI BGP databases. Entries that are placed in the Loc-RIB are then associated with a given set of semantics which are application dependent. Standard BGP mechanisms such as update filtering by NLRI or by attributes such as AS_PATH or large communities apply to the BGP Flow Specification defined NLRI-types.

Network operators can control the propagation of BGP routes by enabling or disabling the exchange of routes for a particular AFI/SAFI pair on a particular peering session. As such, the Flow Specification may be distributed to only a portion of the BGP infrastructure.

4. Distribution of Flow Specification Information

The BGP Flow Specification version 2 (FSv2) uses an NRLI with the format for AFIs for IPv4 (AFI=1), IPv6 (AFI=2), L2 (AFI=6), L2VPN (AFI=25), and SFC (AFI=31) with 2 SAFI (TBD1, TBD2) to support transmission of the flow

specification which support user ordering of traffic filters and actions for IP traffic and IP VPN traffic.

This NLRI information is encoded using MP_REACH_NLRI and MP_UNREACH_NLRI attributes defined in [RFC4760]. When advertising FSv2 NLRIs, the length of the Next-Hop Network Address MUST be set to 0. Upon reception, the Network Address of the Next-Hop field MUST be ignored.

Implementations wishing to exchange flow specification rules MUST use BGP's Capability Advertisement facility to exchange the Multiprotocol Extension Capability Code (Code 1) as defined in [RFC4760] and indicate a capability for flow specification v2 (Code TBD4).

The AFI/SAFI NLRI for BGP Flow Specification has the format:

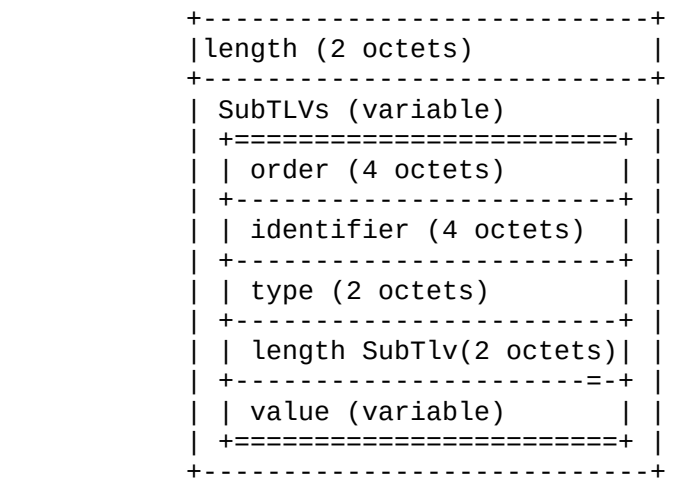


Figure 3: Flow Specification v2 format

where:

length: length of field including all SubTLVs in octets. The combined lengths of any FSv2 NLRI in the MP_REACH_NLRI or MP_UNREACH_NLRI plus the BGP path attributes, the BGP NLRI length and the BGP header must be less than the packet size.

order: flow-specification global rule order number (4 octets).

identifier: identifier for the rule (used for NM/Logging) (4 octets)

type: contains a type for the TLV format of the NRLI (2 octets).

The TLV types are valid FSv2 Filter-Action Types:

0 - reserved
1 - FSv2 IP Header traffic rules
2 - FSv2 Actions
3 - FSv2 L2 traffic rules
4 – FSv2 SFC filter rules

Table 1 – Valid FSv2 Filter Types

- length-tlv - is the length of the value part of the TLV
- value – depends on the TLV type

4.1 IP header SubTLV (type = 1)

The format of the IP header TLV value field is shown in figure 4. The AFI/SAFI field includes the AFI (2 octets), SAFI (1 octet).

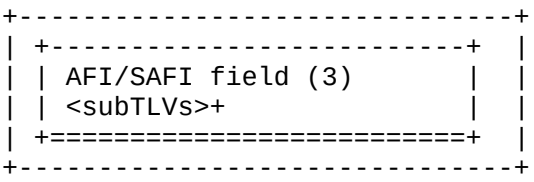


Figure 4 – IP Header TLV

Each SubTLV has the format:

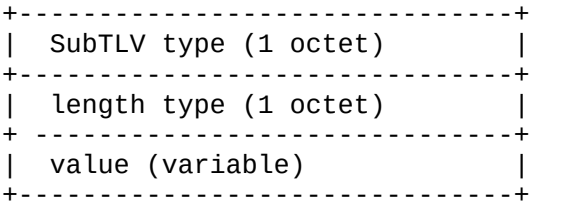


Figure 5 – IP header SubTLV format

Where:

SubTLV type: are listed by component values defined in the “Flow Specification Component types” registry for IPv4 and IPv6 by [RFC8955],[RFC8956], and [draft-li-idr-flowspec-srv6).

The FSv2 valid component types are:

1 – IP Destination prefix
2 – IP Source prefix
3 – IPv4 Protocol / IPv6 Upper Layer Protocol
4 – Port
5 – Destination Port
6 – Source Port
7 – ICMPv4 type / ICMPv6 type
8 – ICMPv4 code / ICPv6 code
9 – TCP Flags
10 – Packet length
11 – DSCP (Diffserv Code Point)
12 – Fragment
13 – Flow Label
14 – Portions of SID

Table 2 – Valid FSv2 component types

Length: length of SubTLV (varies depending on SubTLV type).

Value: For descriptions of components 1-13 see [RF8955] and[RFC8956].
For components 14-15 see [draft-li-idr-flowspec-srv6].

Some of the components use the numeric operator [numeric_op] and [bitmask_op] defined in [RFC8955].

Ordering within the TLV in FSv2: The transmission of SubTLVs within a flow specification rule must be sent ascending order by SubTLV type. If the subTLV types are the same, then the value field between the SubTLV is compared using mechanisms defined in [RFC8955] and [RFC8956]. NLRIs having TLVs which do not follow the above ordering rules MUST be considered as malformed by a BGP FSv2 propagator. This rule prevents any ambiguities that arises from the multiple copies of the same NLRI from multiple BGP FSv2 propagators. A BGP implementation SHOULD treat such malformed NLRIs as ‘Treat-as-withdraw’.

See [RFC8955], [RFC8956], and [draft-li-flowspec-srv6] for specific ordering rules.

4.1.1 IP Destination Prefix (type = 1)

IPv4 Name: IP Destination Prefix (reference: [RFC8955])
IPv6 Name: IPv6 Destination prefix (reference: [RFC8956])

IPv4 length: Prefix length
IPv4 Value: IPv4 Prefix (variable length)

IPv6 length: length of value
IPv6 Value: [offset (1 octet)] [pattern (variable)][padding(variable)]

If IPv6 length = 0 and offset = 0, then component matches every address. Otherwise, length must be offset <length < 129 or component is malformed.

4.1.2 IP Source Prefix (type = 2)

IPv4 Name: IP Source Prefix (reference: [RFC8955])
IPv6 Name: IPv6 Source prefix (reference: [RFC8956])

IPv4 length: Prefix length
IPv4 Value: Source IPv4 Prefix (variable length)

IPv6 length: length of value
IPv6 value: [offset (1 octet)] [pattern (variable)][padding(variable)]

If IPv6 length = 0 and offset = 0, then component matches every address. Otherwise, length must be offset <length < 129 or component is malformed.

4.1.3 IP Protocol (type = 3)

IPv4 Name: IP Protocol (reference: [RFC8955])
IPv6 Name: IPv6 Upper-Layer Protocol: (reference: [RFC8956])

IPv4 length: variable
IPv4 Component Value format: [numeric_op, value]+

IPv6 length: variable
IPv6 Component Value format: [numeric_op, value}+

Where: value is a single octet.

4.1.4 Port (type = 4)

IPv4/IPv6 Name: Port (reference: [RFC8955][RFC8956])
Filter defines: a set of port values to match either destination port or source port.

IPv4 length: variable
IPv4 Component Value format: [numeric_op, value]+

IPv6 length: variable
IPv6 Component Value format: [numeric_op, value}+

Where: value in this component is a single octet.

Note1: in the presence of the port (destination or source port), only a TCP (port 6) or UDP (port 17) packet can match the entire flow specification. If the packet is fragmented and this is not the first fragment then the system will may not be able to find the header. At this point, the FSv2 filter may fail to detect the correct flow. Similarly, if other IP options or the encapsulating security payload (ESP) is present, the node may not be able to describe the transport header. Again, the FSv2 filter may fail to detect the flow.

This problem comes from the inheritance of the FSv1 filter component for port. If more detail is desired, a new FSv2 filter should be defined.

Note2: Although IPv6 allows for more than one Next Header field in the packet, the main goal of the Type 3 FSv2 component is to match the first upper layer protocol value.

4.1.5 Destination Port (type = 5)

IPv4/IPv6 Name: Destination Port (reference: [RFC8955][RFC8956])
Defines: a list of match filters for destination port for TCP or UDP within a received packet

Length: variable
Component Value format: [numeric_op, value]+

4.1.6 Source Port (type = 6)

IPv4/IPv6 Name: Source Port (reference: [RFC8955][RFC8956])
Defines: a list of match filters for source port for TCP or UDP within a received packet

Length: variable
Component Value format: [numeric_op, value]+

4.1.7 ICMP Type (type = 7)

IPv4: ICMP Type (reference: RFC8955])
Defines: a list of match criteria for ICMPv4 type

IPv6: ICMPv6 Type (reference: RFC8955])
Defines: a list of match criteria for ICMPv6 type

Length: variable
Component Value format: [numeric_op, value]+

4.1.8 ICMP Code (type = 8)

IPv4: ICMP Code (reference: RFC8955])
Defines: a list of match criteria for ICMPv4 type

IPv6: ICMPv6 Code (reference: RFC8955])
Defines: a list of match criteria for ICMPv6 type

Length: variable
Component Value format: [numeric_op, value]+
.

4.1.9 TCP Flags (type = 9)

IPv4/IPv6: TCP Flags Code (reference: [RFC8955][RFC8956])
Defines: a list of match criteria for TCP Control bits

Length: variable
Component Value format: [bitmask_op, value]+

Note: 2 octets bitmask match is always used

4.1.10 Packet length (type = 10)

IPv4/IPv6: TCP Flags Code (reference: [RFC8955][RFC8956])
Defines: a list of match criteria for length of packet which excludes L2 header but includes IP header.

Length: variable
Component Value format: [numeric_op, value]+

4.1.11 DSCP (DiffServe Code Point) (type = 11)

IPv4/IPv6: DSCP Code (reference: [RFC8955][RFC8956])

Defines: a list of match criteria for DSCP code values to match the 6-bit DSCP field.

Length: variable
Component Value format: [numeric_op, value]+

4.1.12 Fragment (type = 12)

IPv6: TCP Flags Code (reference: [RFC8955][RFC8956])
Defines: a list of match criteria for specific IP fragments.

Length: variable
Component Value format: [bitmask_op, value]+
Bitmask values are:

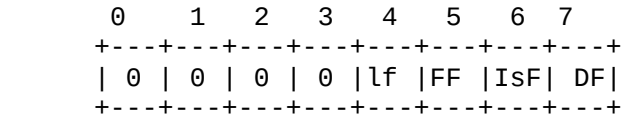


Figure 6

DF (don't fragment) (match If IP header flags bit 1 (DF) is 1.
IsF(is a fragment other than first) –
match if IP header fragment offset is not 0.
FF (First Fragment): Match if [RFC0791] IP Header Fragment offset
Is zero and Flags Bit-2 (MF) is 1.
LF (last Fragment): Match if [RFC7091] IP header Fragment is not 0
And Flags bit-2 (MF) is 0.
0 – must be sent in NLRI encoding as 0, and must be ignored during
Reception.

4.1.13: Flow Label (type = 13)

IPv6: Flow Label (reference: [RFC8956])
Defines: a list of match criteria for 20-bit Flow Label in the
IPv6 header field.

Length: variable
Component Value format: [numeric_op, value]+

4.1.14: Some Parts of SID (Service Identifier) (type = 14)

IPv6: Service Identifier Matches
(reference: draft-li-idr-flowspec-srv6-07.txt)

Defines: a list of match bit match criteria for some combinations of LOC, FUNCT and ARG in SID
or whole SID.

Length: variable
Component Value format:
[type, LOC-Len, FUNCT-Len, ARG-Len, [op, value]+]

- type (1 octet): This indicates the new component type (TBD1, which is to be assigned by IANA).
- LOC-Len (1 octet): This indicates the length in bits of LOC in SID.
- FUNCT-Len (1 octet): This indicates the length in bits of FUNCT in SID.
- ARG-Len (1 octet): This indicates the length in bits of ARG in SID.
- [op, value]+: This contains a list of {operator, value} pairs that are used to match some parts of SID.

The total of three lengths (i.e., LOC length + FUNCT length + ARG length) MUST NOT be greater than 128. If it is greater than 128, an error occurs and Error Handling is applied according to [RFC7606] and [RFC4760].

The operator (op) byte is encoded as:

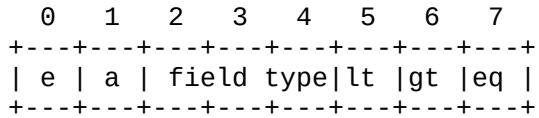


Figure 7

where the behavior of each operator bit has clear symmetry with that of [RFC8955]'s Numeric Operator field.

e - end-of-list bit. Set in the last {op, value} pair in the sequence.

a - AND bit. If unset, the previous term is logically ORed with the current one. If set, the operation is a logical AND. It should be unset in the first operator byte of a sequence. The AND operator has higher priority than OR for the purposes of evaluating logical expressions.

field type:

000: SID's LOC

001: SID's FUNCT

010: SID's ARG

011: SID's LOC:FUNCT

100: SID's FUNCT:ARG

101: SID's LOC:FUNCT:ARG

For an unknown type, Error Handling is applied according to [RFC7606] and [RFC4760].

lt - less than comparison between data' and value'.

gt - greater than comparison between data' and value'.

eq - equality between data' and value'.

The data' and value' used in lt, gt and eq are indicated by the field type in a operator and the value field following the operator.

The value field depends on the field type and has the value of SID's some parts rounding up to bytes (refer to the table below).

Field Type	Value
SID's LOC	value of LOC bits
SID's FUNCT	value of FUNCT bits
SID's ARG	value of ARG bits
SID's LOC:FUNCT	value of LOC:FUNCT bits
SID's FUNCT:ARG	value of FUNCT:ARG bits
SID's LOC:FUNCT:ARG	value of LOC:FUNCT:ARG bits

Figure 8

4.2 Encoding of Actions

The FSv2 actions may be sent in an extended community, a wide community or an NLRI.

The extended community encodes the Flow Specification value in the extended community format [RFC4360].

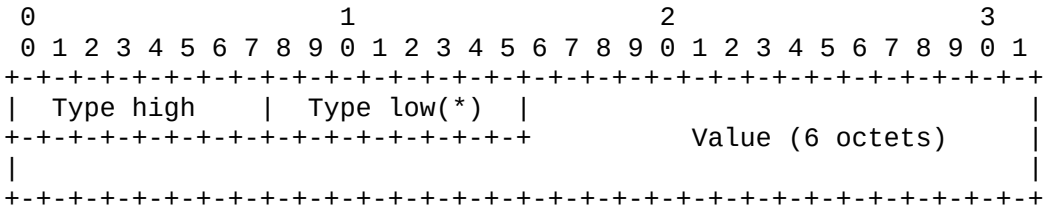


Figure 9

The Wide community definition for Flow Specification v2 is as follows:

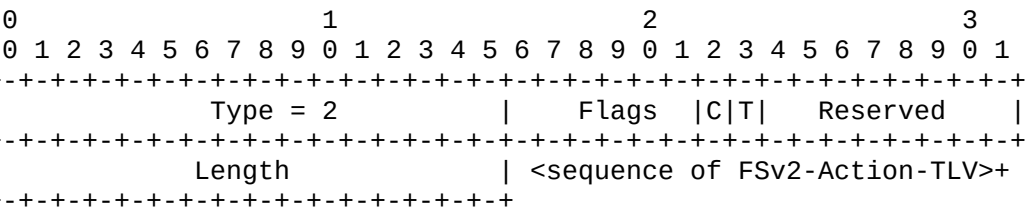


Figure 10

Where:

FSv2-Action-TLV

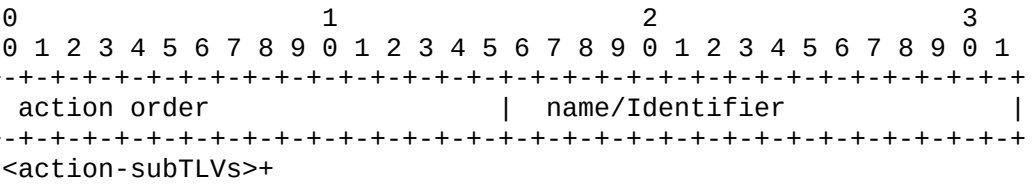


Figure 11

FS-SubTLVs

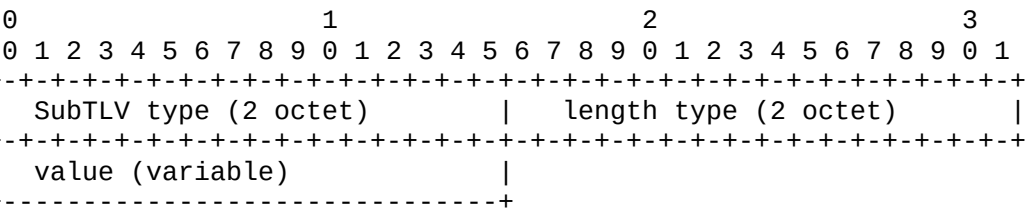


Figure 12

The FSv2 Action TLV may be included in the NLRI to be associated with a specific NLRI. (Note inclusion with the FSv2 NLRI does not have good scaling properties.)

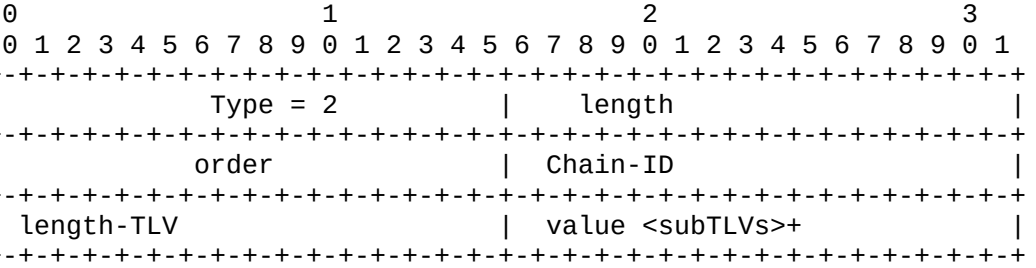


Figure 13

Where
Action-order – is the user defined action within the list
Chain ID – is a 2-byte identifier for action chain
Actions – are a sequence of action SubTLVs.

Each Action SubTLV has the format:

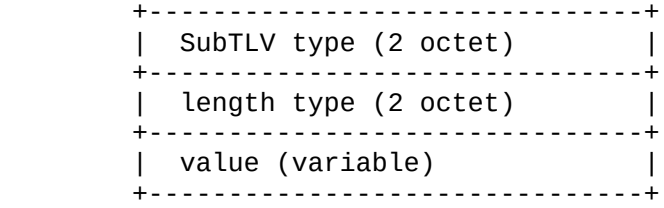


Figure 14

Where
- SubTLV type values are action type values shown in the table below.
- Length – is the length of the action subtlv
- Value is specific to the sub-tlv

Action	Description
00	reserved
01	Action Chain Operation
02	traffic actions per interface group
06	traffic rate limited by bytes
07	traffic action (terminal/sample)
08	redirect IPv4
09	mark DSCP value
10	associate L2 Information
11	associate E-Tree Information
12	traffic rate limited by packets
13	redirect to IPv6
14	SFC Classifier Info (moved from OD to OE)
15	redirect to Indirection-id (move from 0x00)
15-21	TBA (to be assigned)
22	VLAN-Action (0x16)[draft-ietf-idr-flowspec-l2vpn-17]
23	TPID-Action (0x17) [draft-ietf-idr-flowspec-l2vpn-17]
24-254	TBA (to be assigned)
255	reserved

Table 3 – FSv2 Action types

Ordering of Actions within a rule:

The actions are first stored in user-defined order. If multiple actions exist for a single action order value, then the actions will be ordered by action component type followed by value.

Action specifications must include descriptions of order comparison for the values within the action.

4.2.1 Action 1 – Action Chain operation (ACO)

SubTLV: 0x01
Length: variable
Value
AC-failure-type – byte that determines the action failure
AC-failure-value – variable depending on action chain type

Actions may successfully complete or fail and an Action chain must deal with it. The default value stored for an action change that does not have this action chain is “stop on failure”.

AC-Failure types:

0x00 – default – stop on failure
0x01 – continue on failure (best effort on actions)
0x02 – conditional stop on failure – depending on AC-Failure-value
0x03 – rollback – do all or nothing - depending in AC-Failure-value

AC-Failure values: TBD

Interactions: With all other Actions
Ordering within Action type: By AC-Failure type

4.2.2 Traffic Actions per interface set (0x02)

SubTLV: 2
Length: 8 octets (6 in extended community)
Value field:
[4-octet-AS] [GroupID 2-octet] [action 2-octet]

Where:
Group-ID: identifier for group in 2 octets (14 lower bits)

Note: Extended Community format will have 2 bits for action.

Action:
Outbound(0x1): FSv2 rule MUST be applied in outbound Direction to interface set identified by Group-id

Inbound (0x2): FSv2 rule must be applied in inbound Direction to interface set identified by Group-ID

Value ordering: AS, then Group ID, then Action bytes.

Reference: [draft-ietf-idr-flowspec-interface-set]
Conflict: with any bi-direction action such as
1) traffic rate limited by bytes, or
2) traffic rate limited by packets.

4.2.3 Traffic rate limited by bytes (0x6)

SubTLV: 6 (0x6)
Length: 8 octets
Value field: [4-octet-AS] [float (4 bytes)]

4-octet-AS - 4 byte AS number.
If FSv1 passes the lower 2 bytes of 4 byte AS number,
Use [TBD-TBD] as higher 2 bytes to identify.

Float – maximum byte rate in IEEE floating point [IEEE.754.1985] format
in units per second. A value of 0 should result in all traffic
for the particular flow to be discarded.

Value ordering: AS then float value
Reference: [RFC8955]
Action Conflict: traffic-rate-packets.

4.2.4 Traffic Action (0x7)

SubTLV: 7
Length: 1
Value field: [1-octet action]

Where traffic actions values are:
1 = Terminal flow specification action
2 – Sample – enables sampling and logging
3 = Terminal action + sample

Value ordering: by traffic action values

Reference: [RFC8955]
Conflicts: none
Copy in path redirect (01), Redirect to IPv4 (8),
And Redirect to IPv6 may also duplicate packets.

Authors note: Recommend FSv2 centralize copies

4.2.5 Redirect to IPv4 [0x8)

SubTLV: 0x8
Length: 12 octets
Value field: [4-byte-AS] [IPv4 address (4 octet] [ID (4octet)]
[Flag (1 octet)]

Where
4-octet-AS – is a 4-byte AS in a Route Target
IPv4 address - is a 4-byte IP Address in RT
ID – the 4-octet value set by user

Flag (1 octet) with the following definition

```
           0- reserved
           1- copy and redirect copy

  Value ordering:  AS, then IP address, the ID (lowest to highest)
      No AS specified uses AS Value of zero.
      No IP specified uses IP value of zero.
      No ID specified uses ID value of zero.

  Reference [RFC8955] [draft-ietf-idr-flowspec-ip-02.txt]
  Conflicts: Redirect to indirection ID, traffic action sampling
```

4.2.6 Traffic marking (0x9)

```
   SubTLV: 9
   Length: 1 octet
   Value: DSCP field with the 2 left most bits zero

       0 1 2 3 4 5 6 7
      +---------------+
      |R R DSCP bits  |
      +---------------+
        Figure 15


     Where:
           RR – reserved bits (set to zero to send,
                 Ignored upon reception and set to 0)

           DSCP – 6 bits of DSCP values

     Ordering within Value:  based on DSCP value
     Reference: [RFC8955]
     Conflicts: none
```

4.2.7 Traffic rate limited by packets (12/0xC)

```
    SubTLV= 12 (0xC)
    Length: 8
    Value field:  [4-octet-AS] [float (4 octet)]

    Where:
      4-octet AS – is the AS setting this value
      Float – specifies maximum rate [IEEE.754.185] format
              in packets per second.  A traffic rate of zero should result
              in all packets being discard.  The traffic rate should not
              be negative

      Ordering within value: AS, then float
      Reference: [RFC8955]
```

4.2.8 Traffic redirect to IPv6 (13, 0xD)

```
    SubTLV = 13 (0xD)
```

```
    Length = 24 octets
    Value field: [4-octet-as] [IPv6-address (16 octets)]
                 [local administrator (2 octets]
                 [Flag (1 octets)]

   Where: 4-octet-AS – is AS requesting action
          IPv6-address – is redirection address
          Local administrator – 2 bytes assigned
          Flag (1 octet) with the following definition

            0- reserved
            1- copy and redirect copy

   Ordering within value: AS, IPv6 address, flag (low to high)
   Reference: [RFC8956][draft-ietf-idr-redirect-ip-02.txt]
```

4.2.9 Traffic insertion in SFC (14, OXE)

```
    Function: Redirect traffic to a specific point in the SFC
              for insertion in the SFC forwarding.

    SubTLV = (0xE)
       Note: replace IANA 0xD FSv1 with FSv2 OxE.

    Length = 6 octets
    Value field:  [SPI (3 octets)][SI (1 octet)][SFT (2 octet)]

    Where:
      SPI – is the service path identifier
      SI – is the service index
      SFT – is the service function type.

   Value ordering: SPI, then, SI, then SFT (lowest to highest)
   Reference: [RFC9015]
```

4.2.10 Flow Specification Redirect to Indirection-ID (0x0F)

```
  SubTLV: 0x01 (note: current value is 0x00 for FSv1)
  Length: 6 octets
  Value field:
   [Flags (1 octet)] [ID-Type (1 octet)][Generalized-ID (4 octets)]

  Where:
    Flags –
      [S-ID]- sequence number for indirection IDs (3 bits)
              Value of zero means sequence is not set and all other
              S-ID values should be ignored
      [C] – copy packets matching this ID

    ID-Type: type of indirection ID
      0 – localized ID
      1 – Node with SID/index in MPLS SR
      2 – Node with SID/label in MPLS SR
      3 – Node with Binding Segment ID with SID/Index
      4 – Node with Binding Segment ID with SID/Label
```

5 - Tunnel ID
Generalized-ID (G-ID): indirection value

Action ordering: first indirection ID, then Generalized ID.
Action Value ordering: ID-Type

Reference: [draft-ietf-idr-flowspec-path-redirect]
Conflicts with: rt-redirect (0x08)

4.2.11 VLAN action (23,[type 0x16])

Function: Rewrite inner or outer VLAN header
SubTLV: 23 (0x16)
Length:
Value: [Rewrite-actions (2 octets)]
[vlan-PCP-DE-1 (2 octets)]
[vlan-PCP-DE-2 [2 octets)]

Where: Rewrite-actions – is a bit mask of push/pop actions
Value ordering: rewrite-actions, VLAN1, VLAN2, PCP-DE1, PCP-DE2

Reference: [draft-ietf-idr-flowspec-l2vpn]

4.2.12 TPID action (23,[type 0x17])

Function: Replace Inner or outer TP
SubTLV: 23 (0x17)
Length:
Value: [Rewrite-actions (2 octets)]
[TP-ID-1 (2 octets)]
[TP-ID-2 (2 octets)]

Where: rewrite-actions are bitmask (2 octets)
With 2 actions

Value Ordering: rewrite-actions, TP-ID-1, TP-ID-2

4.2.13 Comparison of the Action storage

Table 2 provides a comparison between Action SubTLV and Extended community format.

Table 1 – Actions in 3 forms

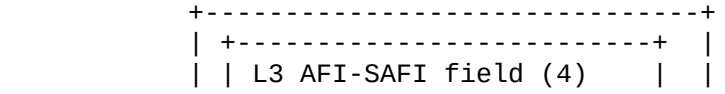
Subtlv Type	Action Name	Action SubTLV form (Wide community + NLRI)	Extended Community form

Subtlv Type	Action Name	Action SubTLV form (Wide community + NLRI)	Extended Community form
1	Action Chain operation (ACO)	Type: 01 Length: variable Value: Type (1 octet] Value (variable based on type)	Not applicable
2	Traffic actions per interface group	Type = 02 Length: 7 [4-octet-as] [group-3-octet] [flags-1-octet] [this document]	0x0702 or 0x4702 Length=6 octets [4-octet-AS] [Flags-group] (2 octets] [draft-ietf-idr-interface-set]
3-5	Reserved		
6	traffic rate limited by bytes	type=06 length=8 octets [4-byte-AS] [float(4O)] [this document]	0x8006 Length=6 octets [2-octet-AS] [float (4O)] Ref: [RFC8966]
7	traffic action	type=7 length=1 Flags (1 octet) [This document]	0x8007 Length=6 octets Flags (6 octets) Ref: [RFC8966]
8	Redirect IP	type=8 length=12 [4-byte-AS] [IPv4-address] [4-byte-ID]	0x8008 Length=6 octets [AS-2-octets] [IPv4 address] 0x8108 Length=6 octets [IPv4-4-octets] [ID-2-octets] 0x8208 Length=6 octets [AS-4-octets] [ID-2-octets]
9	Traffic mark	Type=9 Length=1 [DSCP-1-octet]	Type=0x8009 Length=6 [DSCP-1-octet]
10	L2VPN	Type=10 Length=6 [encap-2-octet] [CTL-2-octet] [L2MTU-2-octet]	0x8010 Length=6 [encap-1-octet] [CTL-1-octet] [L2MTU-2-octet]
11	Reserved		
12	Traffic rate limited by packets	Type=0xc Length=8 octets [4-byte-AS] [4-octet-float]	Type=0xc Length=6 octets [2-byte-AS] [4-octet-float]

Subtlv Type	Action Name	Action SubTLV form (Wide community + NLRI)	Extended Community form
13	Redirect to IPv6	Type=Oxd Length=22 [4-byte-AS] [IPv6-address (16 octets)] [Local-Administrator (2 octets)]	Type=0x000c Length=18 [IPv6 address (16 octets)) [Local-Administrator [2 octets)
14	SFC insertion	Type:0xE [SPI (3 octets)] [SI (1)] [SFT (2 octet)]	Type: 0xD (FSv1) Type: OXE (FSv2) [SPI (3 octets)] [SI (1 octet)][SFT (2 octet)]
15	Redirect to indirection ID	Type: 01 Length:8 octets [flags (1) ID type (1) G-ID [1]	0x0900 Length:6 octets
16-22	Served		
23	VLAN action	Type:0x16 [Rewrite-action (2 octets)] [vlan-PCP-DE-1 (2 octets)] [vlan-PCP-DE-2 [2 octets)]	Type: TBD [Rewrite-action (2 octets)] [vlan-PCP-DE-1 (2 octets)] [vlan-PCP-DE-2 [2 octets)]
24	TPID action	Type:0x17 [Rewrite-actions (2 octets)] [TP-ID-1 (2 octets)] [TP-ID-2 (2 octets)]	Type: TBD [Rewrite-actions (2 octets)] [TP-ID-1 (2 octets)] [TP-ID-2 (2 octets)]

4.3 L2 and L2VPN FSv2 Filters

The FSv2 filters for L2 flow and L2VPN flows may be sent in an extended community, a wide community or in the action SubTLV in the NLRI. This section describes the encoding of the value field for the Action TLV.



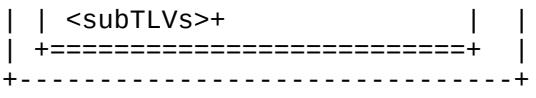


Figure 15

Where the SubTLVs have the following components shown in Table 4.

Component Types Table

Component type	Description
1	EtherType
2	Source MAC
3	Destination MAC
4	DSAP (destination service access point)
5	SSAP (source service access point)
6	control field in LLC
7	SNAP
8	VLAN ID
9	VPAN PCP
10	Inner VLAN ID
11	Inner VLAN PCP
12	VLAN DEI
13	VLAN DEI
14	Source MAC special bits
15	Destination MAC special bits

Table 4 – L2 VPN components

Reference: [draft-ietf-idr-flowspec-l2vpn]

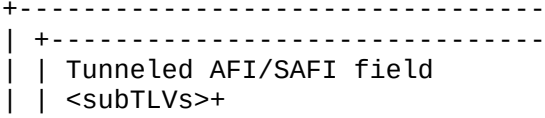
See [draft-ietf-idr-flowspec-l2vpn] for the details on the format and value fields for each component.

Ordering of L2 FSv2 rules will be by user-defined order of the rule. For FSv2 filters within the same rule, the ordering will be by component number and then by value within the component. See [draft-ietf-idr-flowspec-l2vpn] for the ordering of values.

4.4 FSv2 SFC NLRI Traffic Filters

The FSv2 filters allow for filtering of the SFC NLRI family of routes. The traffic NLRIs filtered are from SFC AFI/SAFI (AFI = 31, SAFI=9).

The FSv2 filters provide this filtering with SFC AFI (AFI=31) and SAFI for FSv2 filters (SAFI = TB1).



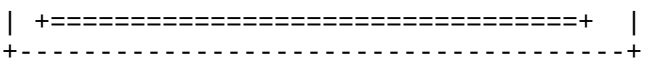


Figure 20

Each SubTLV has the format:

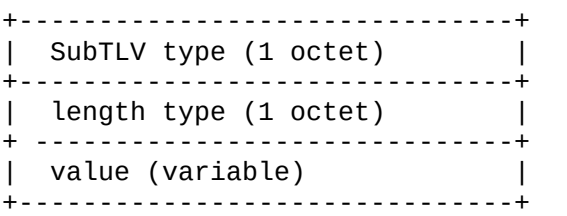


Figure 21 – Tunneled SubTLV format

The components listed are:
1 = SFIR RD Type (types 1, 2, 3)
2 = SFIR RD Value
3 = SFIR Pool ID
4 = SFIR MPLS context/label
5 = SFPR SPI
6 = SPF attribute fields

Table 6 – SFC Filter types

Ordering is by: User-defined rule order, component number, and then value within component.

Reference: [RFC9015], TBD

5.0 Validation of FSv2 NLRI

The validation of FSv2 NLRI adheres to the combination of rules for general BGP FSv1 NLRI found in [RFC8955], [RFC8956], [RFC9112], and the specific additions made for SFC NLRI [RFC9015], L2VPN NLRI [draft-ietf-idr-flowspec-l2vpn], and tunneled NLRI [draft-ietf-idr-flowspec-nv03].

To provide clarity, the full validation process for flow specification routes (v1 or v2) is described in this section rather than simply refer to the portions of these RFCs. Validation only occurs after BGP UPDATE packet, the FSv2 NLRI and the path attributes relating to FSv2 (Extended community and Wide Community) have been determined to be well-formed. Any MALFORMED FSv2 NRLI is handled as a “TREAT as WITHDRAW”.

5.1 Validation of FS NLRI (FSv1 or FSv2)

Flow specification received from a BGP peer that are accepted in the respective Adj-RIB-In are used as input to the route selection process. Although the forwarding attributes of the two routes for same prefix may be the same, BGP is still required to perform its path selection algorithm in order to select the correct set of attributes to advertise.

The first step of the BGP Route selection procedure (section 9.1.2 of [RFC4271]) is to exclude from the selection procedure routes that are considered unfeasible. In the context of IP routing information, this step is used to validate that the NEXT_HOP Attribute of a given route is resolvable.

The concept can be extended in the case of the Flow Specification NLRI to allow other validation procedures.

The FSv2 validation process validates the FSv2 NLRI with following unicast routes received over the same AFI (1 or 2) but different SAFIs:

- Flow specification routes (FSv1 or FSv2) received over SAFI=133 will be validated against SAFI=1,
- Flow Specification routes (FSv1 or FSv2) received over SAFI=134 will be validated against SAFI=128, and
- Flow Specification routes (FSv1 or FSv2) [AFI =1, 2] received over SAFI=77 will be validated using only the Outer Flow Spec against SAFI = 133.

The FSv2 validates L2 FSv2 NLRI with the following L2 routes received over the same AFI (25), but a different SAFI:

- Flow specification routes (FSv1 or FSv2)received over SAFI=135 are validated against SAFI=128.

In the absence of explicit configuration, a Flow specification NLRI (FSv1 or FSv2) MUST be validated such that it is considered feasible if and only if all of the conditions are true:

a) A destination prefix component is embedded in the Flow Specification,
b) One of the following conditions must hold true:
    1. The originator of the Flow Specification matches the originator of the best-math unicast route for the destination prefix embedded in the flow specification (this is the unicast route with the longest possible prefix length covering the destination prefix embedded in the flow specification).
    2. The AS_PATH attribute of the flow specification is empty or contains only an AS_CONFED_SEQUENCE segment [RFC5065].
        1. This condition should be enabled by default.
        2. This condition may be disabled by explicit configuration on a BGP Speaker,
        3. As an extension to this rule, a given non-empty AS_PATH (besides AS_CONFED_SEQUENCE segments) MAY be permitted by policy].

c) There are no “more-specific” unicast routes when compared with the flow destination prefix that have been received from a different neighbor AS than the best-match unicast route, which has been determined in rule b.

However, rule a may be relaxed by explicit configuration, permitting Flow Specifications that include no destination prefix component. If such is the case, rules b and c are moot and MUST be disregarded.

By “originator” of a BGP route, we mean either the address of the originator in the ORIGINATOR_ID Attribute [RFC4456] or the source address of the BGP peer, if this path attribute is not present.

BGP implementation MUST enforce that the AS in the left-most position of the AS_PATH attribute of a Flow Specification Route (FSv1 or FSv2) received via the Exterior Border Gateway Protocol (eBGP) matches the AS in the left-most position of the AS_PATH attribute of the best-match unicast rout for the destination prefix embedded in the Flow Specification (FSv1 or FSv2) NLRI.

The best-match unicast route may change over time independently of the Flow Specification NLRI (FSv1 or FSv2). Therefore, a revalidation of the Flow Specification MUST be performed whenever unicast routes change. Revalidation is defined as retesting rules a to c as described above.

5.2 Validation of Flow Specification actions

Flow Specification actions may be mapped using Extended Communities, Wide Communities or a FSv2 NLRI TLV. The scaling of FSv2 actions implies that Extended Communities and wide communities which can associate an action to a large number of NLRIs will be most often used. Therefore, it is likely the FSv2 NLRI TLV for actions will be used for Internet Wide actions.

The ordering of precedence for these actions in the absence of user-defined ordering, is the following precedence (best to lowest) FSv2 NLRI action TLV, Wide Communities and Extended communities.

Actions may conflict, duplicate, or complementation other actions. An example of conflict is the packet rate limiting by byte and by packet. The example of duplicate is the request to copy or to sample a packet. An example of complementary actions is to redirect a copy of a packet. Specifications of new FSv2 action specification and implementations should document the interaction of FSv2 actions and FSv1 actions.

Well-formed syntactically correct actions should be linked to a filtering rule in order the actions should be enacted. If one action in the ordered list fails, the default procedure is for the action process for this rule to stop and flag the error via system management. By explicit configuration, the action processing may continue after errors.

Implementations MAY wish to log the actions taken by FS actions (FSv1 or FSv2).

5.3 Error handling and Validation

The following two error handling rules must be followed by all BGP speakers support the following two FSv2 error handling rules:

- FSv2 NLRI having TLVs which do not have the correct lengths or syntax must be considered MALFORMED.

- FSv2 NLRIs having TLVs which do not follow the above ordering rules described in 4.1 MUST be considered as malformed by a BGP FSv2 propagator.

The above two rules prevent any ambiguity that arises from the multiple copies of the same NLRI from multiple BGP FSv2 propagators.

A BGP implementation SHOULD treat such malformed NLRIs as ‘Treat-as-withdraw’.

An implementation for a BGP speaker supporting both FSv1 and FSv2 must support the error handling for both FSv1 and FSv2. The storage of the BGP FSv1 and FSv2 must support both the AFI/SAFI and the configuration which translates FSv1 NLRI into FSv2 NLRI for order storage.

6.0 Ordering for Flow Specification v2 (FS-v2)

Flow Specification v2 allows the user to order flow specification rules and the actions associated with a rule. Each FSv2 rule has one or more match conditions and one or more actions associated with each rule.

This section describes how to order FSv2 filters received from a peer prior to transmission to another peer. The same ordering should be used for the ordering of forwarding filtering installed based on only FSv2 filters.

Section 7.0 describes how a BGP peer that supports FSv1 and FSv2 should order the flow specification filters during the installation of these flow specification filters into FIBs or firewall engines in routers.

The BGP distribution of FSv1 NLRI and FSv2 NLRI and their associated path attributes for actions (Wide Communities and Extended Communities) is “ships-in-the-night” forwarding of different AFI/SAFI information. This recommended ordering provides for deterministic ordering of filters sent by the BGP distribution.

6.1 Ordering of FSv2 NLRI Filters

The basic principles regarding ordering of rules are simple:

1) Rule-0 (zero) is defined to be 0/0 with the “permit-all” action

BGP peers which do not support flow specification permit traffic for routes received. Rule-0 is defined to be “permit-all” for 0/0 which is the normal case for filtering for routes received by BGP.

By configuration option, the “permit-all” may be set to “deny-all” if traffic rules on routers used as BGP must have a “route” AND a firewall filter to allow traffic flow.

2) FSv2 rules are ordered based on the user-defined order numbers specified in the FSv2 NLRI (rules 1-n).

3) If multiple FSv2 NLRI have the same user-defined order, then the filters are ordered by type of FSv2 NRLI filters (see Table 1, section 4) with lowest numerical number have the best precedence.

For the same user-defined order and the same value for the FSv2 filters type, then the filters are ordered by FSv2 the component type for that FSv2 filter type (see Tables 3-6) with the lowest number having the best precedence.

For the same user-defined order, the same value of FSv2 Filter Type, and the same value for the component type, then the filters are ordered by value within the component type. Each component type defines value ordering.

For component types inherited from the FSv1 component types, there are the following two types of comparisons:

FSv1 component value comparison for the IP prefix values, compares the length of the two prefixes. If the length is different, the longer prefix has precedence. If the length is the same, the lower IP number has precedence.

For all other FSv1 component types, unless specified, the component data is compared using the memcmp() function defined by [ISO_IEC_9899]. For strings with the same length, the lowest string memcmp() value has precedence. For strings of different lengths, the common prefix is compared. If the common string prefix is not equal, then the string with the lowest string prefix has higher precedence. If the common prefix is equal, the longest string is considered to have higher precedence.

Notes:

- Since the user can define rules that re-order these value comparisons, this order is arbitrary and set to provide a deterministic default.
- All the ordering by type of FSv2 NLRI filter, component type, or component value is only done within the same order.

6.2 Ordering of the Actions

The FSv2 specification allows for actions to be associated by:
a) a FSv2 Action TLV in an FSv2 NLRI,
b) a Wide Community path attribute, or
c) an Extended Community path attribute.

Actions may be ordered by user-defined action order number from 1-n (where n is 2**16-2 and 2**16-1 is reserved).

Extended community actions are associated with order number 32768 [0x8000] or a specific configured value for the FSv2 domain.

Action user-order number zero is defined to have an Action type of “Set Action Chain operation” (value 0x01) that defines the default action chain process. For details on “set action chain operation” see section 4.2.0 and section 6.2.1.

If the user-defined action number for an action is the same, then the actions are ordered by FSv2 action types (see Table 3 for a list of action types). If the user-defined action number and the FSv2 action types are the same, then the order must be defined by the FSv2 action.

6.2.1 Action Chain Operation

The “Action Chain Operation” (ACO) changes the way the actions after this action in an action chain handles a failure. If no action chain operations are set, then the default action of “stop upon failure” (value 0x00) will be used for the chain.

An example may help illustrate how action ordering and the “operation of action” change

Suppose we have the following 4 actions defined for a match:

- Redirect to indirection ID (0x01)
  Sent by Wide Community, user-defined match 2
- Traffic rate limit by bytes (0x07)
  Sent by Wide Community, user-defined match 1
- Traffic sample (0x07)
  Sent by extended community
- SF classifier Info (0x0E)
  Sent by extended community

These 4 filters rate limit a potential DDoS attack by: a) redirect the packet to indirection ID (for slower speed processing), sample to local hardware, and forward the attack traffic via a SFC to a data collection box.

The FSv2 action list for the match would look like this

- Action 0: Operation of action chain (stop upon failure)
- Action 1: Traffic Rate limit by byte (0x07)
- Action 2: Redirect to Redirection ID (01)
- Action 32768 (0x8000): Traffic Sample (0x07)
- Action 32768 (0x8000): SFC Classifier (0x0E)

If the redirect to a redirection ID fails, then the other actions also do not occur. This may not be what is needed for a DDoS accounting.

Suppose the following 5 actions were defined for a FSv2 filter:

- Set Action Chain Operation, user defined match 2

(AC-Failure) 0x00 to 0x01 (continue upon failure)

- Redirect to indirection ID (0x01)
  Sent by Wide Community, user-defined match 2
- Traffic rate limit by bytes (0x07)
  Sent by Wide Community, user-defined match 1
- Traffic sample (0x07)
  Sent by extended community
- SFC classifier Info (0x0E)
  Sent by extended community

The FSv2 action list for the match would look like this

- Action 0: Operation of action chain (stop upon failure)
- Action 01: Traffic Rate limit by byte (0x07)
- Action 02: Set Action Chain Operation (0x00)
  (AC-failure) set 0x01 (continue on failure)
- Action 02: Redirect to Redirection ID (01)
- Action 32768 (0x8000): Traffic Sample (0x07)
- Action 32768 (0x8000): SFC Classifier (0x0E)

If the redirect to a redirection ID fails, the action chain will continue on to sample the data and enact SFC classifier actions.

Note: The scaling for associating actions is better with the Communities which can be associated with many FSv2 filters. The FSv2 action with the FSv2 NLRI should be used in rare cases such as the “Die Internet Worm case” where a particular filter match identifies a pernicious Internet worm that must die off and not be past. In such an example, installing the actions to stop the packets needs to stay with the filter.

6.2.2 Summary of FSv2 action ordering

Operators should use user-defined ordering to clearly specify the actions desired upon a match. The FSv2 actions default ordering is specified to provide deterministic order for actions which have the same user-defined order and same type.

Default Order User-order Match + Type match order

FS Action	Order (lowest value to highest)
0x01 – Action chain operation	Failure flag
0x02 – traffic actions per Interface group	AS, then Group-ID, then Action ID
0x03-0x05 to be assigned	TBD
0x06 – Traffic rate limit	AS, then float value
0x07 – Traffic Action	traffic Action value

0x08 – Redirect to IP	AS, then IP Address, then ID
0x09 – Traffic Marking	DSCP value (lowest to highest)
0x0C – Redirect to Indirect ID	AS, then Float value
0x0D – Traffic Redirect to IPv6	AS, IPv6 value, then local Admin
0x0E – Traffic insertion to SFC	SPI, then SI, the SFT
0xOF – Redirect to Indirection-ID	ID-type, then Generalized-ID
0x10-0x15 – to be assigned	TBD
0x16 – VLAN action	rewrite-actions, VALN1, VLAN2, PCP-DE1, PCP-DE2
0x17 – TPID action	rewrite actions, TP-ID-1, TP-ID-2

7. Ordering of FS filters for BGP Peers support FSv1 and FSv2

Flow Specification v2 allows the user to order flow specification rules and the actions associated with a rule. Each FSv2 rule has one or more match conditions and one or more actions associated with each rule.

Some BGP peers will support FSv1 and FSv2. This section describes the best practice for ordering the FSv1 and FSv2 filter rules.

One simple rule captures the best practice: Order the FSv1 filters after the FSv2 filter by placing the FSv1 filters after the FSv2 filters.

To operationally make this work, all flow specification filters should be included the same data base with the FSv1 filters being assigned a user-defined order beyond the normal size of FSv2 user-ordered values.

Suppose you might have 10,000 rules for the FSv2 filters. Assign all the FSv1 user defined rules to 10,001 (or better yet 20,000). The FSv1 rules will be ordered by the components and component values.

All FSv1 actions are defined ordered actions in FSv2. Translate your FSv1 actions into FSv2 ordered actions for storing in a common FSv1-FSv2 flow specification data base.

8. Manageability of FSv2

Operational issues drive the deployment of BGP flow specification as a quick and scalable way to distribute filters. The early operations accepted the fact validation of the distribution of filter needed to be done outside of the BGP distribution mechanism. Other mechanisms (NETCONF/RESTCONF or PCEP) have reply-request protocols.

These features within BGP have not changed. BGP still does not have action-reply feature.

NETCONF/RESTCONF latest enhancements provide action/response features which scale. The combination of a quick distribution of filters via BGP and a

long-term action in NETCONF/RESTCONF that ask for reporting of the installation of FSv2 filters may provide the best scalability.

The combination of NETCONF/RESTCONF NM protocols and BGP focuses each protocol on the strengths of scalability.

FSv2 will be deployed in webs of BGP peers which have some BGP peers passing FSv1, some BGP peers passing FSv2, some BGP peers passing FSv1 and FSv2, and some BGP peers not passing any routes.

The TLV encoding and deterministic behaviors of FSv2 do not get around the need for careful design of the distribution of flow specification filters in this mixed environment. The needs of networks for flow specification are different depending on the network topology and the deployment technology for BGP peers sending flow specification.

Suppose we have a centralized RR connected to DDoS processing sending out flow specification to a second tier of RR who distribute the information to targeted nodes. This type of distribution has one set of needs for FSv2 and the transition from FSv1 to FSv2.

Suppose we have Data Center with a 3-tier backbone trying to distribute DDoS or other filters from the spine to combinational nodes, to the leaf BGP nodes. The BGP peers may use RR or normal BGP distribution.

Suppose we have a corporate network with a few AS sending DDoS filters using basic BGP from a variety of sites.

These examples are given to indicate that BGP FSv2 like so many BGP protocols needs to be carefully tuned to aid the mitigation services within the network. This protocol suite starts the migration toward better tools using FSv2, but it does not end it. With FSv2 TLVs and deterministic actions, new operational mechanisms can start to be understood and utilized.

This FSv2 specification is merely the start of a revolution of work – not the end.

9. Optional Security Additions

This section discusses the optional BGP Security additions for BGP FSv2 relating to BGPSEC [RFC8205] and ROA.

9.1. BGP FS v2 and BGPSEC

FSv1 RFCs ([RFC8955] and [RFC8956]) do not comment on how FSv1 Specification may be passed over BGPSEC [RFC8205].

FSv1 and FSv2 may be sent via BGPSEC.

9.2. BGP FS v2 with ROA

BGP Flow Specification v2 can utilize Route Origin Authentication (ROA [RFC6482]. If BGP-FS v2 is used with BGPSEC and ROA, the first thing is to

validate the route within BGPSEC and second to utilize BGP ROA to validate the route origin.

The BGP-FS peers using both ROA and BGP-FS validation determine that a BGP Flow specification is valid if and only if one of the following cases:

- If the BGP Flow Specification NLRI has a IPv4 or IPv6 address in a destination address match filter and the following is true:
  - o A BGP ROA has been received to validate the originator, and the route is the best-match unicast route for the destination prefix embedded in the match filter; or
- If a BGP ROA has not been received that matches the IPv4 or IPv6 destination address in the destination filter, the match filter must abide by the [RFC8955] and [RFC8956] validation rules as follows:
  - o The originator match of the flow specification matches the originator of the best-match unicast route for the destination prefix filter embedded in the flow specification, and
  - o No more specific unicast routes exist when compared with the flow destination prefix that have been received from a different neighboring AS than the best-match unicast route (see step a of the validation procedure described in section 5.1).

The best match is defined to be the longest-match NLRI with the highest preference.

10. IANA Considerations

This section complies with [RFC7153].

10.1. Flow Specification V2 SAFIs

IANA is requested to assign two SAFI Values from the registry at https://www.iana.org/assignments/safi-namespace from the Standard Action Range as follows:

Value	Description	Reference
TBD1	BGP-FS V2	[This document]
TBD2	BGP-FS V2 VPN	[this document]

10.2. BGP Capability Code

IANA is requested to assign a Capability Code from the registry at https://www.iana.org/assignments/capability-codes/ from the IETF Review range as follows:

Value	Description	Reference	Controller
TBD3	Flow Specification V2	[this document]	IETF

10.3. Filter IP component types

IANA is requested to indicate [this draft] as a reference on the following assignments in the Flow Specification Component Types Registry:

Value	Description	Reference
1	Destination filter	[RFC8955][RFC8956][this draft]
2	Source Prefix	[RFC8955][RFC8956][this draft]
3	IP Protocol	[RFC8955][RFC8956][this draft]
4	Port	[RFC8955][RFC8956][this draft]
5	Destination Port	[RFC8955][RFC8956][this draft]
6	Source Port	[RFC8955][RFC8956][this draft]
7	ICMP Type [v4 or v6]	[RFC8955][RFC8956][this draft]
8	ICMP Code [v4 or v6]	[RFC8955][RFC8956][this draft]
9	TCP Flags [v4]	[RFC8955][RFC8956][this draft]
10	Packet Length	[RFC8955][RFC8956][this draft]
11	DSCP marking	[RFC8955][RFC8956][this draft]
12	Fragment	[RFC8955][RFC8956][This draft]
13	Flow Label	[RFC8956] [This draft]
14	whole SID	[draft-li-idr-flowspec-srv6]
15	Partial SID	[draft-li-idr-flowspec-srv6]

10.4. Filter IP component types

IANA is requested to create the following two new registries on a new "Flow Specification v2 TLV types”.

Name: BGP-FS v2 TLV types
Reference: [this document]
Registration Procedures: 0x01-0x3FFF Standards Action.

Type	Use	Reference
0x00	Reserved	[this document]
0x01	IP traffic rules	[this document]
0x02	FSv2 Actions	[this document]
0x03	L2 traffic rules	[this document]
0x04	tunnel traffic rules	[this document]
0x05	SFC AFI filter rules	[this document]
0x06-0x3FFF	Unassigned	[this document]
0x4000-0x7FFF	Vendor specific	[this document]
0x8000-0xFFFFFFFF	Reserved	[this document]

Name: BGP-FS v2 Action types
Reference: [this document]
Registration Procedure: 0x01-0x3FFF Standards Action.

Type	Use	Reference

Type	Use	Reference
0x00	Reserved	[this document]
0x01	Redirect to Indirection-id	[this document]
0x02	Traffic actions per interface group	[this document]
0x03	Unassigned	[this document]
0x04	Unassigned	[this document]
0x05	Unassigned	[this document]
0x06	traffic rate limited by bytes	[this document]
0x07	traffic action (terminal/sample)	[this document]
0x08	redirect IPv4	[this document]
0x09	mark DSCP value	[this document]
0x0a	associate L2 Information	[this document]
0x0b	associate E-Tree Information	[this document]
0xoc	traffic rate limited by packets	[this document]
Type	Use	Reference
0x0d	redirect to IPv6	[this document]
0xoe	SFC Classifier Info (moved from OD to OE)	[this document]
0x0f to 0x15	unassigned	[this document]
0x16	VLAN-Action	[this document] [draft-ietf-idr-flowspec-l2vpn-17]
0x17	TPID-Action	[this document] [draft-ietf-idr-flowspec-l2vpn-17]
0x18- 0x3ff	Unassigned	[this document]
0x4000- 0x7FFF	Vendor specific	[this document]
0c8000- 0xFFFFFFFF	Reserved	[this document]

11. Security Considerations

The use of ROA [RFC6482] improves on [RFC8955] by checking if route origination is valid. This check can improve the validation sequence for a multiple-AS environment.

The use of BGPSEC [RFC8205] to secure the packet can increase security of BGP flow specification information sent in the packet.

The use of the reduced validation within an AS [RFC9112] can provide adequate validation for distribution of flow specification within an single autonomous system for prevention of DDOS.

Distribution of flow filters may provide insight into traffic being sent within an AS, but this information should be composite information that does not reveal the traffic patterns of individuals.

12. References

12.1. Normative References

[RFC9117]
Uttaro, J., Alcaide, J., Filsfils, C., Smith, D., and P. Mohapatra, "Revised Validation Procedure for BGP Flow Specifications", RFC9117, July 2021.

[I-D.ietf-idr-flowspec-l2vpn]
Hao, W., Eastlake, D. E., Litkowski, S., and S. Zhuang, "BGP Dissemination of L2 Flow Specification Rules", draft-ietf-idr-flowspec-l2vpn-17 (work in progress), May 2021.

[I-D.ietf-idr-flowspec-nvo3]
Eastlake, D., Weiguo, H., Zhuang, S., Li, Z., and R. Gu, "BGP Dissemination of Flow Specification Rules for Tunneled Traffic", draft-ietf-idr-flowspec-nvo3-13 (work in progress), February 2021.

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

[I-D.li-idr-flowspec-srv6-05.txt
Li, Z, Li, L, Chen, H. Loibl, C, Misrha, C., Fan, Y. Zhu, Y., Liu, L, Liu, X., draft-li-idr-flowspec-srv6-05 (work in progress), March 2021.

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

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

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, DOI 10.17487/RFC4360, February 2006, <https://www.rfc-editor.org/info/rfc4360>.

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

[RFC6482] Lepinski, M., Kent, S., and D. Kong, "A Profile for Route Origin Authorizations (ROAs)", RFC 6482, DOI 10.17487/RFC6482, February 2012, <https://www.rfc-editor.org/info/rfc6482>.

[RFC7153] Rosen, E. and Y. Rekhter, "IANA Registries for BGP Extended Communities", RFC 7153, DOI 10.17487/RFC7153,

March 2014, <https://www.rfc-editor.org/info/rfc7153>.

[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>.

[RFC8955] Loibl, C., Hares, S., Raszuk, R., McPherson, D., and M. Bacher, "Dissemination of Flow Specification Rules", RFC 8955, DOI 10.17487/RFC8955, December 2020, <https://www.rfc-editor.org/info/rfc8955>.

[RFC8956] Loibl, C., Ed., Raszuk, R., Ed., and S. Hares, Ed., "Dissemination of Flow Specification Rules for IPv6", RFC 8956, DOI 10.17487/RFC8956, December 2020, <https://www.rfc-editor.org/info/rfc8956>.

12.2. Informative References

[RFC6241] Enns, R., Ed., Bjorklund, M., Ed., Schoenwaelder, J., Ed., and A. Bierman, Ed., "Network Configuration Protocol (NETCONF)", RFC 6241, DOI 10.17487/RFC6241, June 2011, <https://www.rfc-editor.org/info/rfc6241>.

[RFC8040] Bierman, A., Bjorklund, M., and K. Watsen, "RESTCONF Protocol", RFC 8040, DOI 10.17487/RFC8040, January 2017, <https://www.rfc-editor.org/info/rfc8040>.

[RFC8205] Lepinski, M., Ed. and K. Sriram, Ed., "BGPsec Protocol Specification", RFC 8205, DOI 10.17487/RFC8205, September 2017, <https://www.rfc-editor.org/info/rfc8205>.

[RFC8206] George, W. and S. Murphy, "BGPsec Considerations for Autonomous System (AS) Migration", RFC 8206, DOI 10.17487/RFC8206, September 2017, <https://www.rfc-editor.org/info/rfc8206>.

Authors' Addresses

Susan Hares
Hickory Hill Consulting
7453 Hickory Hill
Saline, MI 48176
USA

Email: shares@ndzh.com

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

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