Network Working Group
Request for Comments: 5575
Category: Standards Track

P. Marques
Cisco Systems
N. Sheth
Juniper Networks
R. Raszuk
Cisco Systems
B. Greene
Juniper Networks
J. Mauch
NTT America
D. McPherson
Arbor Networks

August 2009

Dissemination of Flow Specification Rules

Abstract

This document defines a new Border Gateway Protocol Network Layer Reachability Information (BGP NLRI) encoding format that can be used to distribute traffic flow specifications. This allows the routing system to propagate information regarding more specific components of the traffic aggregate defined by an IP destination prefix.

Additionally, it defines two applications of that encoding format: one that can be used to automate inter-domain coordination of traffic filtering, such as what is required in order to mitigate (distributed) denial-of-service attacks, and a second application to provide traffic filtering in the context of a BGP/MPLS VPN service.

The information is carried via the BGP, thereby reusing protocol algorithms, operational experience, and administrative processes such as inter-provider peering agreements.

Status of This Memo

This document specifies an Internet standards track protocol for the Internet community, and requests discussion and suggestions for improvements. Please refer to the current edition of the "Internet Official Protocol Standards" (STD 1) for the standardization state and status of this protocol. Distribution of this memo is unlimited.

Copyright Notice

Table of Contents

1. Introduction

Modern IP routers contain both the capability to forward traffic according to IP prefixes as well as to classify, shape, rate limit, filter, or redirect packets based on administratively defined policies.

These traffic policy mechanisms allow the router to define match rules that operate on multiple fields of the packet header. Actions such as the ones described above can be associated with each rule.

The n-tuple consisting of the matching criteria defines an aggregate traffic flow specification. The matching criteria can include elements such as source and destination address prefixes, IP protocol, and transport protocol port numbers.

This document defines a general procedure to encode flow specification rules for aggregated traffic flows so that they can be distributed as a BGP [RFC4271] NLRI. Additionally, we define the required mechanisms to utilize this definition to the problem of immediate concern to the authors: intra- and inter-provider distribution of traffic filtering rules to filter (distributed) denial-of-service (DoS) attacks.

By expanding routing information with flow specifications, the routing system can take advantage of the ACL (Access Control List) or firewall capabilities in the router’s forwarding path. Flow specifications can be seen as more specific routing entries to a unicast prefix and are expected to depend upon the existing unicast data information.

A flow specification received from an external autonomous system will need to be validated against unicast routing before being accepted. If the aggregate traffic flow defined by the unicast destination prefix is forwarded to a given BGP peer, then the local system can safely install more specific flow rules that may result in different forwarding behavior, as requested by this system.

The key technology components required to address the class of problems targeted by this document are:

1. Efficient point-to-multipoint distribution of control plane information.

2. Inter-domain capabilities and routing policy support.

3. Tight integration with unicast routing, for verification purposes.

Items 1 and 2 have already been addressed using BGP for other types of control plane information. Close integration with BGP also makes it feasible to specify a mechanism to automatically verify flow information against unicast routing. These factors are behind the choice of BGP as the carrier of flow specification information.

As with previous extensions to BGP, this specification makes it possible to add additional information to Internet routers. These are limited in terms of the maximum number of data elements they can hold as well as the number of events they are able to process in a given unit of time. The authors believe that, as with previous extensions, service providers will be careful to keep information levels below the maximum capacity of their devices.

It is also expected that, in many initial deployments, flow specification information will replace existing host length route advertisements rather than add additional information.

Experience with previous BGP extensions has also shown that the maximum capacity of BGP speakers has been gradually increased according to expected loads. Taking into account Internet unicast routing as well as additional applications as they gain popularity.

From an operational perspective, the utilization of BGP as the carrier for this information allows a network service provider to reuse both internal route distribution infrastructure (e.g., route reflector or confederation design) and existing external relationships (e.g., inter-domain BGP sessions to a customer network).

While it is certainly possible to address this problem using other mechanisms, the authors believe that this solution offers the substantial advantage of being an incremental addition to already deployed mechanisms.

In current deployments, the information distributed by the flow-spec extension is originated both manually as well as automatically. The latter by systems that are able to detect malicious flows. When automated systems are used, care should be taken to ensure their correctness as well as to limit the number and advertisement rate of flow routes.

This specification defines required protocol extensions to address most common applications of IPv4 unicast and VPNv4 unicast filtering. The same mechanism can be reused and new match criteria added to address similar filtering needs for other BGP address families (for example, IPv6 unicast). The authors believe that those would be best to be addressed in a separate document.

2. Definitions of Terms Used in This Memo

NLRI - Network Layer Reachability Information

RIB - Routing Information Base

Loc-RIB - Local RIB

AS - Autonomous System number

VRF - Virtual Routing and Forwarding instance

PE - Provider Edge router

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

3. Flow Specifications

A flow specification is an n-tuple consisting of several matching criteria that can be applied to IP traffic. A given IP packet is said to match the defined flow if it matches all the specified criteria.

A given flow may be associated with a set of attributes, depending on the particular application; such attributes may or may not include reachability information (i.e., NEXT_HOP). Well-known or AS-specific community attributes can be used to encode a set of predetermined actions.

A particular application is identified by a specific (Address Family Identifier, Subsequent Address Family Identifier (AFI, SAFI)) pair [RFC4760] and corresponds to a distinct set of RIBs. Those RIBs should be treated independently from each other in order to assure non-interference between distinct applications.

BGP itself treats the NLRI as an opaque key to an entry in its databases. Entries that are placed in the Loc-RIB are then associated with a given set of semantics, which is application dependent. This is consistent with existing BGP applications. For instance, IP unicast routing (AFI=1, SAFI=1) and IP multicast reverse-path information (AFI=1, SAFI=2) are handled by BGP without any particular semantics being associated with them until installed in the Loc-RIB.

Standard BGP policy mechanisms, such as UPDATE filtering by NLRI prefix and community matching, SHOULD apply to the newly defined NLRI-type. Network operators can also control propagation of such routing updates by enabling or disabling the exchange of a particular (AFI, SAFI) pair on a given BGP peering session.

4. Dissemination of Information

We define a "Flow Specification" NLRI type that may include several components such as destination prefix, source prefix, protocol, ports, etc. This NLRI is treated as an opaque bit string prefix by BGP. Each bit string identifies a key to a database entry with which a set of attributes can be associated.

This NLRI information is encoded using MP_REACH_NLRI and MP_UNREACH_NLRI attributes as defined in RFC 4760 [RFC4760]. Whenever the corresponding application does not require Next-Hop information, this shall be encoded as a 0-octet length Next Hop in the MP_REACH_NLRI attribute and ignored on receipt.

The NLRI field of the MP_REACH_NLRI and MP_UNREACH_NLRI is encoded as a 1- or 2-octet NLRI length field followed by a variable-length NLRI value. The NLRI length is expressed in octets.

```
                      +-------------------------------+
                      |    length (0xnn or 0xfn nn)  |
                      +-------------------------------+
                      |    NLRI value  (variable)    |
                      +-------------------------------+
```

flow-spec NLRI

If the NLRI length value is smaller than 240 (0xf0 hex), the length field can be encoded as a single octet. Otherwise, it is encoded as an extended-length 2-octet value in which the most significant nibble of the first byte is all ones.

In the figure above, values less-than 240 are encoded using two hex digits (0xnn). Values above 240 are encoded using 3 hex digits (0xfnnn). The highest value that can be represented with this encoding is 4095. The value 241 is encoded as 0xf0f1.

The Flow specification NLRI-type consists of several optional subcomponents. A specific packet is considered to match the flow specification when it matches the intersection (AND) of all the components present in the specification.

The following component types are defined:

Type 1 - Destination Prefix

Encoding: <type (1 octet), prefix length (1 octet), prefix>

Defines the destination prefix to match. Prefixes are encoded as in BGP UPDATE messages, a length in bits is followed by enough octets to contain the prefix information.

Type 2 - Source Prefix

Encoding: <type (1 octet), prefix-length (1 octet), prefix>

Defines the source prefix to match.

Type 3 - IP Protocol

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

Contains a set of {operator, value} pairs that are used to match the IP protocol value byte in IP packets.

The operator byte is encoded as:

```
  0   1   2   3   4   5   6   7
+---+---+---+---+---+---+---+---+
| e | a |  len  | 0 |lt |gt |eq |
+---+---+---+---+---+---+---+---+
```

Numeric operator

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

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.

len - The length of the value field for this operand is given as (1 << len).

lt - less than comparison between data and value.

gt - greater than comparison between data and value.

eq - equality between data and value.

The bits lt, gt, and eq can be combined to produce "less or equal", "greater or equal", and inequality values.

Type 4 - Port

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

Defines a list of {operation, value} pairs that matches source OR destination TCP/UDP ports. This list is encoded using the numeric operand format defined above. Values are encoded as 1- or 2-byte quantities.

Port, source port, and destination port components evaluate to FALSE if the IP protocol field of the packet has a value other than TCP or UDP, if the packet is fragmented and this is not the first fragment, or if the system in unable to locate the transport header. Different implementations may or may not be able to decode the transport header in the presence of IP options or Encapsulating Security Payload (ESP) NULL [RFC4303] encryption.

Type 5 - Destination port

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

Defines a list of {operation, value} pairs used to match the destination port of a TCP or UDP packet. Values are encoded as 1- or 2-byte quantities.

Type 6 - Source port

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

Defines a list of {operation, value} pairs used to match the source port of a TCP or UDP packet. Values are encoded as 1- or 2-byte quantities.

Type 7 - ICMP type

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

Defines a list of {operation, value} pairs used to match the type field of an ICMP packet. Values are encoded using a single byte.

The ICMP type and code specifiers evaluate to FALSE whenever the protocol value is not ICMP.

Type 8 - ICMP code

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

Defines a list of {operation, value} pairs used to match the code field of an ICMP packet. Values are encoded using a single byte.

Type 9 - TCP flags

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

Bitmask values can be encoded as a 1- or 2-byte bitmask. When a single byte is specified, it matches byte 13 of the TCP header [RFC0793], which contains bits 8 though 15 of the 4th 32-bit word. When a 2-byte encoding is used, it matches bytes 12 and 13 of the TCP header with the data offset field having a "don’t care" value.

As with port specifiers, this component evaluates to FALSE for packets that are not TCP packets.

This type uses the bitmask operand format, which differs from the numeric operator format in the lower nibble.

```
  0   1   2   3   4   5   6   7
+---+---+---+---+---+---+---+---+
| e | a |  len  | 0 | 0 |not| m |
+---+---+---+---+---+---+---+---+
```

e, a, len - Most significant nibble: (end-of-list bit, AND bit, and length field), as defined for in the numeric operator format.

not - NOT bit. If set, logical negation of operation.

m - Match bit. If set, this is a bitwise match operation defined as "(data & value) == value"; if unset, (data & value) evaluates to TRUE if any of the bits in the value mask are set in the data.

Type 10 - Packet length

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

Match on the total IP packet length (excluding Layer 2 but including IP header). Values are encoded using 1- or 2-byte quantities.

Type 11 - DSCP (Diffserv Code Point)

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

Defines a list of {operation, value} pairs used to match the 6-bit DSCP field [RFC2474]. Values are encoded using a single byte, where the two most significant bits are zero and the six least significant bits contain the DSCP value.

Type 12 - Fragment

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

Uses bitmask operand format defined above.

```
  0   1   2   3   4   5   6   7
+---+---+---+---+---+---+---+---+
|   Reserved    |LF |FF |IsF|DF |
+---+---+---+---+---+---+---+---+
```

Bitmask values:

+ Bit 7 - Don’t fragment (DF)

+ Bit 6 - Is a fragment (IsF)

+ Bit 5 - First fragment (FF)

+ Bit 4 - Last fragment (LF)

Flow specification components must follow strict type ordering. A given component type may or may not be present in the specification, but if present, it MUST precede any component of higher numeric type value.

If a given component type within a prefix in unknown, the prefix in question cannot be used for traffic filtering purposes by the receiver. Since a flow specification has the semantics of a logical AND of all components, if a component is FALSE, by definition it cannot be applied. However, for the purposes of BGP route propagation, this prefix should still be transmitted since BGP route distribution is independent on NLRI semantics.

The <type, value> encoding is chosen in order to account for future extensibility.

An example of a flow specification encoding for: "all packets to 10.0.1/24 and TCP port 25".

destination	proto	port
0x01 18 0a 00 01	03 81 06	04 81 19

Decode for protocol:

Value		
0x03	type	
0x81	operator	end-of-list, value size=1, =
0x06	value	

An example of a flow specification encoding for: "all packets to 10.0.1/24 from 192/8 and port {range [137, 139] or 8080}".

destination	source	port
0x01 18 0a 01 01	02 08 c0	04 03 89 45 8b 91 1f 90

Decode for port:

Value		
0x04	type	
0x03	operator	size=1, >=
0x89	value	137
0x45	operator	&, value size=1, <=
0x8b	value	139
0x91	operator	end-of-list, value-size=2, =
0x1f90	value	8080

This constitutes an NLRI with an NLRI length of 16 octets.

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 RFC 4760 [RFC4760]. The (AFI, SAFI) pair carried in the Multiprotocol Extension Capability MUST be the same as the one used to identify a particular application that uses this NLRI-type.

5. Traffic Filtering

Traffic filtering policies have been traditionally considered to be relatively static.

The popularity of traffic-based, denial-of-service (DoS) attacks, which often requires the network operator to be able to use traffic filters for detection and mitigation, brings with it requirements that are not fully satisfied by existing tools.

Increasingly, DoS mitigation requires coordination among several service providers in order to be able to identify traffic source(s) and because the volumes of traffic may be such that they will otherwise significantly affect the performance of the network.

Several techniques are currently used to control traffic filtering of DoS attacks. Among those, one of the most common is to inject unicast route advertisements corresponding to a destination prefix being attacked. One variant of this technique marks such route advertisements with a community that gets translated into a discard Next-Hop by the receiving router. Other variants attract traffic to a particular node that serves as a deterministic drop point.

Using unicast routing advertisements to distribute traffic filtering information has the advantage of using the existing infrastructure and inter-AS communication channels. This can allow, for instance, a service provider to accept filtering requests from customers for address space they own.

There are several drawbacks, however. An issue that is immediately apparent is the granularity of filtering control: only destination prefixes may be specified. Another area of concern is the fact that filtering information is intermingled with routing information.

The mechanism defined in this document is designed to address these limitations. We use the flow specification NLRI defined above to convey information about traffic filtering rules for traffic that should be discarded.

This mechanism is primarily designed to allow an upstream autonomous system to perform inbound filtering in their ingress routers of traffic that a given downstream AS wishes to drop.

In order to achieve this goal, we define an application-specific NLRI identifier (AFI=1, SAFI=133) along with specific semantic rules.

BGP routing updates containing this identifier use the flow specification NLRI encoding to convey particular aggregated flows that require special treatment.

Flow routing information received via this (AFI, SAFI) pair is subject to the validation procedure detailed below.

5.1. Order of Traffic Filtering Rules

With traffic filtering rules, more than one rule may match a particular traffic flow. Thus, it is necessary to define the order at which rules get matched and applied to a particular traffic flow. This ordering function must be such that it must not depend on the arrival order of the flow specification’s rules and must be constant in the network.

The relative order of two flow specification rules is determined by comparing their respective components. The algorithm starts by comparing the left-most components of the rules. If the types differ, the rule with lowest numeric type value has higher precedence (and thus will match before) than the rule that doesn’t contain that component type. If the component types are the same, then a type-specific comparison is performed.

For IP prefix values (IP destination and source prefix) precedence is given to the lowest IP value of the common prefix length; if the common prefix is equal, then the most specific prefix has precedence.

For all other component types, unless otherwise specified, the comparison is performed by comparing the component data as a binary string using the memcmp() function as defined by the ISO C standard. For strings of different lengths, the common prefix is compared. If equal, the longest string is considered to have higher precedence than the shorter one.

Pseudocode:

```
flow_rule_cmp (a, b)
{
    comp1 = next_component(a);
    comp2 = next_component(b);
    while (comp1 || comp2) {
        // component_type returns infinity on end-of-list
        if (component_type(comp1) < component_type(comp2)) {
            return A_HAS_PRECEDENCE;
        }
        if (component_type(comp1) > component_type(comp2)) {
            return B_HAS_PRECEDENCE;
        }

        if (component_type(comp1) == IP_DESTINATION || IP_SOURCE) {
            common = MIN(prefix_length(comp1), prefix_length(comp2));
            cmp = prefix_compare(comp1, comp2, common);
            // not equal, lowest value has precedence
            // equal, longest match has precedence
        } else {
            common =
               MIN(component_length(comp1), component_length(comp2));
            cmp = memcmp(data(comp1), data(comp2), common);
            // not equal, lowest value has precedence
            // equal, longest string has precedence
        }
    }

    return EQUAL;
}
```

6. Validation Procedure

Flow specifications received from a BGP peer and that are accepted in the respective Adj-RIB-In are used as input to the route selection process. Although the forwarding attributes of two routes for the same flow specification 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 non-feasible. 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 flow specification NLRI, to allow other validation procedures.

A flow specification NLRI must be validated such that it is considered feasible if and only if:

a) The originator of the flow specification matches the originator of the best-match unicast route for the destination prefix embedded in the flow specification.

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

By originator of a BGP route, we mean either the BGP originator path attribute, as used by route reflection, or the transport address of the BGP peer, if this path attribute is not present.

The underlying concept is that the neighboring AS that advertises the best unicast route for a destination is allowed to advertise flow-spec information that conveys a more or equally specific destination prefix. Thus, as long as there are no more specific unicast routes, received from a different neighboring AS, which would be affected by that filtering rule.

The neighboring AS is the immediate destination of the traffic described by the flow specification. If it requests these flows to be dropped, that request can be honored without concern that it represents a denial of service in itself. Supposedly, the traffic is being dropped by the downstream autonomous system, and there is no added value in carrying the traffic to it.

BGP implementations MUST also enforce that the AS_PATH attribute of a route received via the External Border Gateway Protocol (eBGP) contains the neighboring AS in the left-most position of the AS_PATH attribute. While this rule is optional in the BGP specification, it becomes necessary to enforce it for security reasons.

7. Traffic Filtering Actions

This specification defines a minimum set of filtering actions that it standardizes as BGP extended community values [RFC4360]. This is not meant to be an inclusive list of all the possible actions, but only a subset that can be interpreted consistently across the network.

Implementations should provide mechanisms that map an arbitrary BGP community value (normal or extended) to filtering actions that require different mappings in different systems in the network. For instance, providing packets with a worse-than-best-effort, per-hop behavior is a functionality that is likely to be implemented differently in different systems and for which no standard behavior is currently known. Rather than attempting to define it here, this can be accomplished by mapping a user-defined community value to platform-/network-specific behavior via user configuration.

The default action for a traffic filtering flow specification is to accept IP traffic that matches that particular rule.

The following extended community values can be used to specify particular actions.

type	extended community	encoding
0x8006	traffic-rate	2-byte as#, 4-byte float
0x8007	traffic-action	bitmask
0x8008	redirect	6-byte Route Target
0x8009	traffic-marking	DSCP value

Traffic-rate: The traffic-rate extended community is a non-transitive extended community across the autonomous-system boundary and uses following extended community encoding:

The first two octets carry the 2-octet id, which can be assigned from a 2-byte AS number. When a 4-byte AS number is locally present, the 2 least significant bytes of such an AS number can be used. This value is purely informational and should not be interpreted by the implementation.

The remaining 4 octets carry the rate information in IEEE floating point [IEEE.754.1985] format, units being bytes per second. A traffic-rate of 0 should result on all traffic for the particular flow to be discarded.

Traffic-action: The traffic-action extended community consists of 6 bytes of which only the 2 least significant bits of the 6th byte (from left to right) are currently defined.

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

- Terminal Action (bit 47): When this bit is set, the traffic filtering engine will apply any subsequent filtering rules (as defined by the ordering procedure). If not set, the evaluation of the traffic filter stops when this rule is applied.

- Sample (bit 46): Enables traffic sampling and logging for this flow specification.

Redirect: The redirect extended community allows the traffic to be redirected to a VRF routing instance that lists the specified route-target in its import policy. If several local instances match this criteria, the choice between them is a local matter (for example, the instance with the lowest Route Distinguisher value can be elected). This extended community uses the same encoding as the Route Target extended community [RFC4360].

Traffic Marking: The traffic marking extended community instructs a system to modify the DSCP bits of a transiting IP packet to the corresponding value. This extended community is encoded as a sequence of 5 zero bytes followed by the DSCP value encoded in the 6 least significant bits of 6th byte.

8. Traffic Filtering in BGP/MPLS VPN Networks

Provider-based Layer 3 VPN networks, such as the ones using a BGP/MPLS IP VPN [RFC4364] control plane, have different traffic filtering requirements than Internet service providers.

In these environments, the VPN customer network often has traffic filtering capabilities towards their external network connections (e.g., firewall facing public network connection). Less common is the presence of traffic filtering capabilities between different VPN attachment sites. In an any-to-any connectivity model, which is the default, this means that site-to-site traffic is unfiltered.

In circumstances where a security threat does get propagated inside the VPN customer network, there may not be readily available mechanisms to provide mitigation via traffic filter.

This document proposes an additional BGP NLRI type (AFI=1, SAFI=134) value, which can be used to propagate traffic filtering information in a BGP/MPLS VPN environment.

The NLRI format for this address family consists of a fixed-length Route Distinguisher field (8 bytes) followed by a flow specification, following the encoding defined in this document. The NLRI length field shall include both the 8 bytes of the Route Distinguisher as well as the subsequent flow specification.

Propagation of this NLRI is controlled by matching Route Target extended communities associated with the BGP path advertisement with the VRF import policy, using the same mechanism as described in "BGP/ MPLS IP VPNs" [RFC4364] .

Flow specification rules received via this NLRI apply only to traffic that belongs to the VRF(s) in which it is imported. By default, traffic received from a remote PE is switched via an MPLS forwarding decision and is not subject to filtering.

Contrary to the behavior specified for the non-VPN NLRI, flow rules are accepted by default, when received from remote PE routers.

9. Monitoring

Traffic filtering applications require monitoring and traffic statistics facilities. While this is an implementation-specific choice, implementations SHOULD provide:

- A mechanism to log the packet header of filtered traffic.

- A mechanism to count the number of matches for a given flow specification rule.

10. Security Considerations

Inter-provider routing is based on a web of trust. Neighboring autonomous systems are trusted to advertise valid reachability information. If this trust model is violated, a neighboring autonomous system may cause a denial-of-service attack by advertising reachability information for a given prefix for which it does not provide service.

As long as traffic filtering rules are restricted to match the corresponding unicast routing paths for the relevant prefixes, the security characteristics of this proposal are equivalent to the existing security properties of BGP unicast routing.

Where it is not the case, this would open the door to further denial-of-service attacks.

Enabling firewall-like capabilities in routers without centralized management could make certain failures harder to diagnose. For example, it is possible to allow TCP packets to pass between a pair of addresses but not ICMP packets. It is also possible to permit packets smaller than 900 or greater than 1000 bytes to pass between a pair of addresses, but not packets whose length is in the range 900-1000. Such behavior may be confusing and these capabilities should be used with care whether manually configured or coordinated through the protocol extensions described in this document.

11. IANA Considerations

A flow specification consists of a sequence of flow components, which are identified by a an 8-bit component type. Types must be assigned and interpreted uniquely. The current specification defines types 1 though 12, with the value 0 being reserved.

For the purpose of this work, IANA has allocated values for two SAFIs: SAFI 133 for IPv4 dissemination of flow specification rules and SAFI 134 for VPNv4 dissemination of flow specification rules.

The following traffic filtering flow specification rules have been allocated by IANA from the "BGP Extended Communities Type - Experimental Use" registry as follows:

0x8006 - Flow spec traffic-rate

0x8007 - Flow spec traffic-action

0x8008 - Flow spec redirect

0x8009 - Flow spec traffic-remarking

IANA created and maintains a new registry entitled: "Flow Spec Component Types". The following component types have been registered:

Type 1 - Destination Prefix

Type 2 - Source Prefix

Type 3 - IP Protocol

Type 4 - Port

Type 5 - Destination port

Type 6 - Source port

Type 7 - ICMP type

Type 8 - ICMP code

Type 9 - TCP flags

Type 10 - Packet length

Type 11 - DSCP

Type 12 - Fragment

In order to manage the limited number space and accommodate several usages, the following policies defined by RFC 5226 [RFC5226] are used:

Range	Policy
0	Invalid value
[1 .. 12]	Defined by this specification
[13 .. 127]	Specification Required
[128 .. 255]	First Come First Served

The specification of a particular "flow component type" must clearly identify what the criteria used to match packets forwarded by the router is. This criteria should be meaningful across router hops and not depend on values that change hop-by-hop such as TTL or Layer 2 encapsulation.

The "traffic-action" extended community defined in this document has 46 unused bits, which can be used to convey additional meaning. IANA created and maintains a new registry entitled: "Traffic Action Fields". These values should be assigned via IETF Review rules only. The following traffic-action fields have been allocated:

47 Terminal Action

46 Sample

0-45 Unassigned

12. Acknowledgments

The authors would like to thank Yakov Rekhter, Dennis Ferguson, Chris Morrow, Charlie Kaufman, and David Smith for their comments.

Chaitanya Kodeboyina helped design the flow validation procedure.

Steven Lin and Jim Washburn ironed out all the details necessary to produce a working implementation.

13. Normative References

[IEEE.754.1985] Institute of Electrical and Electronics Engineers, "Standard for Binary Floating-Point Arithmetic", IEEE Standard 754, August 1985.

[RFC0793] Postel, J., "Transmission Control Protocol", STD 7, RFC 793, September 1981.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC2474] Nichols, K., Blake, S., Baker, F., and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474, December 1998.

[RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

[RFC4303] Kent, S., "IP Encapsulating Security Payload (ESP)", RFC 4303, December 2005.

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, February 2006.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.

[RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.

Authors' Addresses

Pedro Marques
Cisco Systems
170 West Tasman Drive
San Jose, CA 95134
US
EMail: roque@cisco.com

Nischal Sheth
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US
EMail: nsheth@juniper.net

Robert Raszuk
Cisco Systems
170 West Tasman Drive
San Jose, CA 95134
US
EMail: raszuk@cisco.com

Barry Greene
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US
EMail: bgreene@juniper.net

Jared Mauch
NTT America
101 Park Ave
41st Floor
New York, NY 10178
US
EMail: jmauch@us.ntt.net

Danny McPherson
Arbor Networks
EMail: danny@arbor.net