Traffic classification in end-system IP VPNs.
draft-marques-sdnp-flow-spec-01

Abstract

   When IP VPNs are used to interconnect end-systems [I-D.marques-l3vpn-
   end-system] it may be desirable to introduce traffic control rules at
   a finer level of granularity than an IP destination address.

   This document extends the end-system IP VPN specification with
   support for fine grain traffic classification, filtering and
   redirection rules. It applies the existing BGP IP VPN flow
   specification dissemination mechanism [RFC5575] to end-system IP VPNs
   in order to provide the ability to control IP packets that match a
   specific pattern, which may include fields other than the IP
   destination address.

Status of this Memo

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

   Internet-Drafts are working documents of the Internet Engineering
   Task Force (IETF). Note that other groups may also distribute
   working documents as Internet-Drafts. The list of current Internet-
   Drafts is at http://datatracker.ietf.org/drafts/current/.

   Internet-Drafts are draft documents valid for a maximum of six months
   and may be updated, replaced, or obsoleted by other documents at any
   time. It is inappropriate to use Internet-Drafts as reference
   material or to cite them other than as "work in progress."

   This Internet-Draft will expire on October 01, 2012.
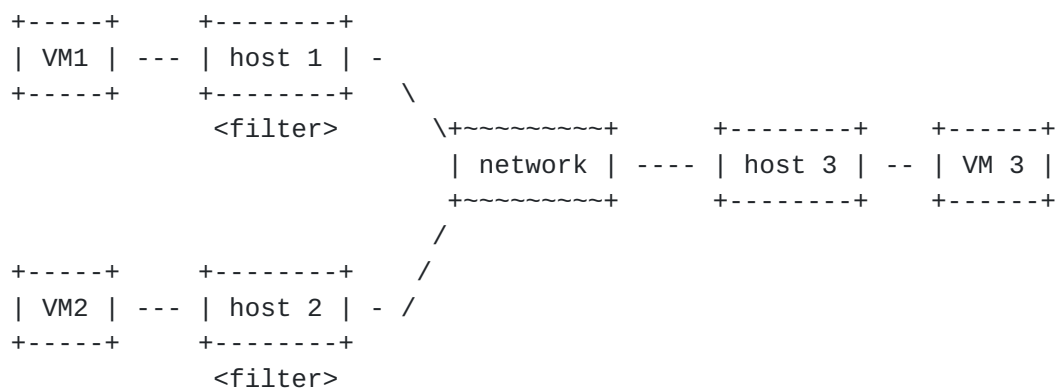
Copyright Notice

Table of Contents

1. Introduction

When end-system IP VPNs [I-D.marques-l3vpn-end-system] are used to
interconnect Virtual Machines or other multi-tenant applications it
may be desirable to control the flow of traffic between sender(s) and
receiver at a finer level of granularity than an IP destination host
prefix.

In the IP protocol model, ingress points map traffic into forwarding
equivalence classes (FECs) which are then given consistent treatment
through a transport network. This document defines a signaling
protocol that conveys traffic classification rules. These rules can
be applied by ingress points into an end-system IP VPN in order to
define FECs that depend on both the destination IP address of the
traffic as well as additional fiels such as the the transport
protocol and ports.

One example where this may be desirable is in scenarios where
different VPNs may exchange traffic directly. For instance, a VPN
that provides a common service to multiple tenants. In this case,
the owner of the destination address may wish to inject a traffic
rule that limits traffic to TCP packets to and from a specific port.
Another example is an application that request specific diffserv
[RFC2474] markings for certain types of traffic. In other
situations, network administrators may wish to inject specific rules
that temporarily redirect traffic.

This document uses a point-to-multipoint model for traffic filtering
rules where the traffic egress requests all the ingresses to perform
a given traffic classification action. The entity that advertises
the destination address of the traffic, or a proxy in its behalf,
injects a flow-based route advertisement into the signaling
infrastructure. This flow-based route is propagated according to VPN
policies to all the ingress points of the VPN, the end-systems which
contain VMs allowed to access the destination.

The traffic filtering rules are then applied at all the ingress
points of the VPN. The egress MAY also choose to apply the same rules
in cases where they are equivalent at both locations.

```
+-----+      +--------+
| VM1 | --- | host 1 | -
+-----+      +--------+   \
         <filter>      \+~~~~~~~~~+      +--------+    +------+
                        | network | ---- | host 3 | -- | VM 3 |
                        +~~~~~~~~~+      +--------+    +------+
                       /
+-----+      +--------+   /
| VM2 | --- | host 2 | - /
+-----+      +--------+
         <filter>
```

The figure above contains an example topology in which a given VM (VM
3) provides a common infrastructure service. VM1 and VM2 belong to
different tenants and are in VPNs which are allowed to access the
service in VM3.

This specification allows VM3 to advertise a traffic filtering rule,
as a flow-spec route, requesting the VPN Forwarders for hosts 1 and 2
to limit any traffic flow to VM3's destination IP address such that,
for instance, only packets for a specific TCP destination port are
allowed.

It is important to note that traffic filtering does not avoid the
need for application level authorization and authentication.

When a flow-spec route is advertised, the number of possible ingress
points it not known in advance. There is no mechanism to generate a
positive or negative acknowledgement from the ingress points. This
is in contrast to the more traditional network management operation
in which the management station is aware of all the agents that must
be controlled.

As with the base end-system IP VPN specification, the forwarding and
signaling networks are distinct. Flow-spec routes are advertised by

the egress end-system or by a proxy in its behalf. The routes are
injected into one or more XMPP signaling gateways and propagated
using the BGP flow-spec address family [RFC5575].

Using the same vrf-import and export policies that define the IP VPN, the flow-spec routes are then imported from BGP into a vpn-specific database and advertised to all the ingress end-system, which apply them.

This document limits itself to "stateless" traffic classification rules that classify a given IP packet independently of any previous data traffic.

2. VPN Forwarder functionality

In order to implement the functionality described in this document a VPN Forwarder MUST support stateless traffic classification rules that are capable of matching the TCP/IP protocol fields defined in [RFC5575].

This document assumes that this traffic filtering functionality can be associated with a particular Virtual Routing and Forwarding (VRF) table, either directly or through the virtual interfaces associated with the VRF. Conceptually, the traffic classification rules described in this document are applied at the VRF level.

The BGP Flow Specification [RFC5575] document lists a set of TCP/IP packet header fields and match operations that are though to be a minimum common set of supported functionality among implementations.

The defined packet header fields are:

o IPv4 destination address.

o IPv4 source address.

o IP protocol identifier.

o Transport Ports: Source, Destination or Either.

o ICMP Type and Code.

o TCP flags.

o Packet length.

o Diffserv Code Point.

o IPv4 fragmentation flags.

When numeric values are specified (i.e. fields other than IP addresses), the match operator can specify a list of values with

inequality operators. Note that this may result in one logical rule, as defined by this specification to be implemented as multiple classification rules on the underlying implementation.

The match operator is defined via the following BNF grammar:

<match> ::= <terms>

<terms> ::= <term>

    | <term> "||" <terms>

    | <term> "&&" <terms>

<term> ::= <operator> value

<operator> ::= "<" | "<=" | "=" | "!=" | ">=" | ">"

As an example, a value range is expressed as: ">= begin && <= end".

The result of a flow-spec rule is one of the following actions:

o allow

o deny

o rate-limit

o redirect

o copy

o log

o set-dscp

The redirect and copy actions have as a target an FEC which should contain an unique UUID [RFC4122] identifier as well as information regarding the IP next-hop address and label used for forwarding.

The copy action instructs the system to generate a copy of the original packet and forward to the specified FEC. Both copy and log actions have an additional parameter which controls whether all matching packets or a sample is subject to the specified treatment.

The 'set-dscp' action specifies the DSCP value to be assigned to the outer IP header of the packet, when a packet is encapsulated.

3. XML schema

In the end-system IP VPN [I-D.marques-l3vpn-end-system]
specification, IP reachability information is encoded as XMPP "item"
information belonging to collection nodes where each collection is
the IP reachability information for a given VPN. End-systems can
publish and receive notifications for these nodes.

This document uses the same approach. It uses a collection with the
name of "<vpn-customer-name>/ip4-flow-spec" to publish and receive
updates corresponding to IPv4 flow-spec routes. When an end-system
published a node into such a collection it must generate a node name
that is unique among the nodes that it publishes. It then associates
that node with the collection.

XML encoding used by flow-spec items:

```
<item>
  <entry xmlns='http://ietf.org/protocol/bgpvpn/ip4-flow-spec'>
    <ip4-destination>10.0.1/24</ip4-destination>
    <ip4-source>20.0.128/20</ip4-source>
    <ip4-protocol>=6 || =17</ip4-protocol>
    <port>=80</port>
    <destination-port>=80</destination-port>
    <source-port>=80</source-port>
    <icmp-type>=1</icmp-type>
    <icmp-code>=1</icmp-code>
    <tcp-flags>=(syn|rst|ack|fin)</tcp-flags>
    <ip-length>>40</ip-length>
    <dscp>=0</dscp>
    <ip4-fragment>=(df|first|more|last)</ip4-fragment>
    <action>
      <accept/>
      <deny/>
      <rate-limit rate='10pps'/>
      <redirect>
        <fec uuid='550e8400-e29b-41d4-a716-446655440000'>
      <nlri af='1'>'infrastructure-ip-address'</nlri>
      <label>1</label>
  </fec>
      </redirect>
      <copy>
        <fec>...</fec>
  <sample/>
      </copy>
      <log/>
      <set-dscp>128</set-dscp>
    </action>
  </entry>
</item>
```

The sequence of XML elements in an item SHOULD follow the "flow specification" NLRI type order as the example above. IP source and destination prefixes are encoded in their standard textual representation of <dotted notation>"/"<prefix-length>. Protocol and Port elements are expressed using the match operator syntax documented above. "<port>" and "<destination-port>" or "<source-port>" SHOULD be mutually exclusive. The icmp type and code fields as well as ip-length and dscp are again encoded using the value match operator. The "<tcp-flags>" element uses either an equality or match operation of the TCP header flags. A binary match is expressed as "m /(syn|rst|ack|fin)/". The "<ip4-fragment>" element may also use a binary match operation.

4. Signaling gateway functionality

As with IP reachabilty information, signaling gateways create a routing database for each 'vpn-customer-name'. An XMPP client (a VPN Forwarder) can publish and subscribe to multiple of these databases. Each "virtual interface" on the end-system is associated with a virtual routing table on the gateway.

From a signaling perspective, the gateway functions as a IP VPN PE as described in section 8 of [RFC5575]. As with IP reachability, this document uses the XMPP interface to delegate the forwarding functionality to the VPN Forwarder, separating it from the signaling node.

5. Applications

This specification provides a mechanism to distribute traffic classification rules to many enforcement points. This may of interest in applications where it is desirable to avoid the standard approach of a centralized enforcement point. Typically in situations where the volume of traffic or the nature of the problem make it more cost effective to do so.

One such application is the enforcement of stateless traffic forwarding rules for infrastructure services. An application level services, such as a storage server may need to support multiple data-center tenants. In this scenario the storage VPN advertises a given address prefix, which contains both the anycast IP address of the load-balancers as the addresses of individual servers. Using VPN import policies, the data-center management solution allows the tenant specific VPNs to see these routes. The tenant VPN addresses must also be reachable on the storage VPN, in this example.

This specification allows the storage service to block out traffic that does not match the specific transport protocols used to provide

this service. It also allows confirming traffic to be marked with
the appropriate diffserv classification. The network administrator
case also use this mechanism for diagnostic purposes.

6. **Security Considerations**

There are two independent areas that are worth examining when it comes to security. The integrity of the control plane information and the forwarding actions.

This document assumes that all signaling interactions use mutual authentication, where all communication channels are authenticated.

For traffic filtering and redirection this mechanism assumes a "best-effort" model. The ingress points will strive to perform the actions specified by the egress. However there are no strict guarantees that the actions can be applied successfully on an ingress points or that the order of operations is such that no non-conforming traffic is ever presented to the egress.

For traffic filtering rules, the egress point can choose to apply the rules also in order to provide stronger guarantees.

Applications should themselves authenticate its communication peers my methods that do not depend on the IP addresses used at the network layer.

7. References

[I-D.marques-l3vpn-end-system]
          Marques, P., Fang, L., Pan, P., Shukla, A., Napierala, M.
          and N. Bitar, "BGP-signaled end-system IP/VPNs.",
          Internet-Draft draft-marques-l3vpn-end-system-05, March
          2012.

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

[RFC4122] Leach, P., Mealling, M. and R. Salz, "A Universally Unique
          IDentifier (UUID) URN Namespace", RFC 4122, July 2005.

[RFC5575] Marques, P., Sheth, N., Raszuk, R., Greene, B., Mauch, J.
          and D. McPherson, "Dissemination of Flow Specification
          Rules", RFC 5575, August 2009.

Authors' Addresses

   Pedro Marques
   Contrail Systems
   440 N Wolfe Rd
   Sunnyvale, CA 94085

   Email: roque@contrailsystems.com

Luyuan Fang
Cisco Systems
111 Wood Avenue South
Iselin, NJ 08830

Email: lufang@cisco.com

Ping Pan
Infinera Corp
140 Caspian Ct.
Sunnyvale, CA 94089

Email: ppan@infinera.com

Amit Shukla
Juniper Networks
1194 N. Mathilda Av.
Sunnyvale, CA 94089

Email: amit@juniper.net

Maria Napierala
AT&T Labs
200 Laurel Avenue
Middletown, NJ 07748

Email: mnapierala@att.com