Network Working Group   P. Mohapatra
Internet-Draft   Sproute Networks
Intended status: Standards Track   A. Sreekantiah
Expires: April 11, 2015   K. Patel
                                                          B. Pithawala
                                                         Cisco Systems
                                                                 A. Lo
                                                       Arista Networks
                                                       October 8, 2014

           Automatic Route Target Filtering for legacy PEs
                   draft-ietf-idr-legacy-rtc-04.txt

Abstract

   This document describes a simple procedure that allows "legacy" BGP
   speakers to exchange route target membership information in BGP
   without using mechanisms specified in [RFC4684]. The intention of
   the proposed technique is to help in partial deployment scenarios and
   is not meant to replace [RFC4684].

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

   [RFC4684] provides a powerful and general means for BGP speakers to
   exchange and propagate Route Target reachability information and
   constrain VPN route distribution to achieve high scale. However, it
   requires that all the BGP speakers in the network are upgraded to
   support this functionality. For example, in a network with route
   reflectors (RR), if one PE client in the cluster doesn't support
   constrained distribution, the cluster degenerates into storing and

processing all the VPN routes. The route reflectors need to request
and store all the network routes since they do not receive route
target membership information from the legacy PEs. The RR will also
generate all those routes to the legacy PEs and the legacy PEs will
end up filtering the routes and store the subset of VPN routes that
are of interest.

This document specifies a mechanism for such legacy PE devices using
existing configuration and toolset to provide similar benefits as
[RFC4684]. At the same time, it is backward-compatible with the
procedures defined in [RFC4684]. It also allows graceful upgrade of
the legacy router to be [RFC4684] capable.

1.1. 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 RFC 2119 [RFC2119].

2. Basic Idea

The basic idea is to make use of VPN unicast route exchange from the
legacy PEs to a new BGP speaker (e.g. an RR) to signal RT
membership. The legacy PEs announce a set of "special" routes with
mapped RTs to the RR along with a standard community (defined in this
document). The presence of the community triggers the RR to extract
the RTs and build RT membership information.

3. Detailed Operation

3.1. Legacy PE Behavior

The following simple steps are performed on the legacy PE device:

o Collect the "import route targets" of all the configured customer
   VRFs. Let's call this set 'IRTS'.

o Create a special "route-filter VRF" with a route distinguisher(RD)
   that's configured with the same value across the network for all
   legacy PE devices. Note: the equivalence of the RD value is for
   optimization - the operator may choose to use different values.

o Originate one or more routes in this VRF and attach a subset of
   'IRTS' as "translated route-target extended communities" with each
   route so as to evenly distribute the RTs (and to make sure they
   can fit into one BGP UPDATE message). Collectively, the union of
   the "translated route-target extended communities" of all these
   routes is equal to the set 'IRTS'. The translated RTs are

      attached as export route-targets for the routes originated in the
      route-filter VRF.

   o The translation of the IRTs is necessary in order to refrain from
      importing "route-filter" VRF routes into VPN VRFs that would
      import the same route-targets. The translation of the IRTS is
      done as follows. For a given IRT, the equivalent translated RT
      (TRT) is constructed by means of swapping the value of the high-
      order octet of the Type field for the IRT (as defined in
      [RFC4360]).

```
 0                   1            0                   1
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5  0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|      0x00     |     0x02      |  |      0x01     |     0x02      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|2B AS                          |  |2B AS => IP(high)              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+<=>+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Local Admin(high)              |  |Local Admin(high) => IP(low)   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Local Admin(low)               |  |Local Admin(low) => Local Admin|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

 0                   1            0                   1
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5  0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|      0x01     |     0x02      |  |      0x02     |     0x02      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|IP(high)                       |  |IP(high) => 4B AS(high)        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+<=>+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|IP(low)                        |  |IP(low) => 4B AS(low)          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Local Admin                    |  |Local Admin => Local Admin     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

 0                   1            0                   1
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5  0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|      0x02     |     0x02      |  |      0x00     |     0x02      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|4B AS(high)                    |  |4B AS(high) => 2B AS           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+<=>+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|4B AS(low)                     |  |4B AS(low) => Local Admin(high)|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Local Admin                    |  |Local Admin => Local Admin(low)|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

As an example, if IRT R= 65500:12244(hex: 0x0002ffdc00002fd4), equivalent route-filter TRT: 255.220.0.0:12244(hex: 0x0102ffdc00002fd4). One shortcoming of the translation mechanism is a possible collision between IRTs and TRTs if the network has been configured with RTs of multiple higher order octet types (2-byte AS, IP address, and 4-byte AS). It is expected that such a configuration is rare in practice.

o As an alternative to the translation of the IRTS, the subset of the 'IRTS' can be attached as-is (without swapping the type field as described earlier) as "export route-target extended communities" with each route so as to evenly distribute the RTs (and to make sure they can fit into one BGP UPDATE message). In this case, the IRT subsets can be attached in outbound policy to avoid the route-filter VRFs from being imported into VPN VRFs. Also in this case, the route-filter VRF routes must be tagged with a different special community (from that associated with the translated RTs) as described in Section 4 so that the receiving BGP speaker can distinguish the two cases.

o The routes are marked with NO_ADVERTISE and NO_EXPORT well-known communities as well as the appropriate new community that's defined in this document Section 4. Note that there is no specific provision made to disallow configuration of subsequent route policies that can potentially alter the set of communities attached to "route-filter" VRF routes. The protocol behavior in such a case is undefined and the use of those policy statements is discouraged.

3.2. RR Behavior

Upon receiving the "route-filter" routes, the BGP speaker does its usual processing to store them in its local RIB. It recognizes them as route-filter routes based on the association of the new standard community as defined in this document. If required (as indicated by the community value), it translates the attached route-target extended communities (TRT) to equivalent import route-targets (IRT). Finally it creates the route-target filter list for each legacy client by collecting the entire set of route targets. From this point onwards, the behavior is similar to that defined in [RFC4684]. The RR does not propagate the routes further because of their association with NO_ADVERTISE community. Also the VPN EoR that is sent by the legacy PE should also be used as an indication that the legacy PE is done sending the route-filter information as per the procedures defined in [RFC4684] for implementing a EoR mechanism to signal the completion of initial RT membership exchange.

3.2.1. **Generating Route Target Membership NLRIs for the legacy PE
        clients**

   The RR MAY also translate the received extended communities from
   legacy clients into route target membership NLRIs as if it had
   received those NLRIs from the client itself. This is useful for
   further propagation of the NLRIs to rest of the network to create RT
   membership flooding graph. When the route_filter routes are received
   with same RD (from all legacy PE speakers), processing of the paths
   to generate equivalent NLRIs becomes fairly easy.

4. **ROUTE_FILTER Community**

   This memo defines four BGP communities that are attached to BGP
   UPDATE messages at the legacy PE devices and processed by the route
   reflectors as defined above. They are as follows:


```
   +----------------------------+-------------------------------------+
   |         Community          | Meaning                             |
   +----------------------------+-------------------------------------+
   |       ROUTE_FILTER_v4      | RTs are attached as-is for VPNv4    |
   |                            | route filtering                     |
   |            ...             | ...                                 |
   |       ROUTE_FILTER_v6      | RTs are attached as-is for VPNv6    |
   |                            | route filtering                     |
   |            ...             | ...                                 |
   | ROUTE_FILTER_TRANSLATED_v4 | Translated RTs are attached for     |
   |                            | VPNv4 route filtering               |
   |            ...             | ...                                 |
   | ROUTE_FILTER_TRANSLATED_v6 | Translated RTs are attached for     |
   |                            | VPNv6 route filtering               |
   +----------------------------+-------------------------------------+
```


   In the absence of (or lack of support of) AF specific communities
   (ROUTE_FILTER_v6, ROUTE_FILTER_TRANSLATED_v6), the ROUTE_FILTER_v4 or
   ROUTE_FILTER_TRANSLATED_v4 MAY be treated by an implementation as a
   default VPN route-filter community to build a combination VPN filter
   for all VPN AFs (VPNv4, VPNv6) present on the RR. This is in
   accordance with the procedures in [RFC4684] to build combination
   route-filters for VPN AFs and AF specific route-filters defined in
   [I-D.keyur-bgp-af-specific-rt-constrain]. If this is the case, then
   subsequent receipt of any "route-filter" routes with AF specific
   communities (ROUTE_FILTER_v6, ROUTE_FILTER_TRANSLATED_v6) will
   override the default filters sent with ROUTE_FILTER_v4 or
   ROUTE_FILTER_TRANSLATED_v4 for the VPNv6 AFI when support for the AF
   specific communities exists.

5. Deployment Considerations

When both the legacy PE and the RR support extended community based
Outbound Route Filtering as in [I-D.chen-bgp-ext-community-orf] this
may be used as a alternate solution for the legacy PE to signal RT
membership information, in order to realize the same benefits as
[RFC4684]. Also extended community ORF can be used amongst the RRs
in lieu of [RFC4684] to realize similar benefits.

6. Contributors

Significant contributions were made by Stephane Litkowski, Luis M
Tomotaki and James Uttaro which the authors would like to
acknowledge.

7. Acknowledgments

The authors would like to thank Rob Shakir for his review and
comments.

8. IANA Considerations

IANA shall assign new code points from BGP first-come first-serve
communities for the four communities as listed in Section 4.

9. Security Considerations

There are no additional security risks introduced by this design.

10. References

10.1. Normative References

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

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

[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.

   [RFC4684]   Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk,
               R., Patel, K., and J. Guichard, "Constrained Route
               Distribution for Border Gateway Protocol/MultiProtocol
               Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual
               Private Networks (VPNs)", RFC 4684, November 2006.

10.2. Informational References

   [I-D.chen-bgp-ext-community-orf]
               Chen, E., Lo, A., and K. Patel, "Extended Community Based
               Outbound Route Filter for BGP-4", draft-chen-bgp-ext-
               community-orf-02 (work in progress), December 2011.

   [I-D.keyur-bgp-af-specific-rt-constrain]
               Patel, K., Raszuk, R., Djernaes, M., Dong, J., and M.
               Chen, "IPv6 AF Extensions for Route Target Distribution",
               draft-keyur-bgp-af-specific-rt-constrain-01 (work in
               progress), March 2011.

Authors' Addresses

   Pradosh Mohapatra
   Sproute Networks

   Email: mpradosh@yahoo.com

   Arjun Sreekantiah
   Cisco Systems
   170 W. Tasman Drive
   San Jose, CA 95134
   USA

   Email: asreekan@cisco.com

   Keyur Patel
   Cisco Systems
   170 W. Tasman Drive
   San Jose, CA 95134
   USA

   Email: keyupate@cisco.com

   Burjiz Pithawala
   Cisco Systems
   170 W. Tasman Drive
   San Jose, CA 95134
   USA

   Email: bpithaw@cisco.com

   Alton Lo
   Arista Networks
   5470 Great America Parkway
   Santa Clara, CA 95054
   USA

   Email: altonlo@aristanetworks.com