Constrained VPN route distribution

draft-marques-ppvpn-rt-constrain-01.txt

Status of this Memo

This document is an Internet-Draft and is in full conformance with all provisions of Section 10 of RFC2026.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

Abstract

This document defines MP-BGP procedures that allow BGP speakers to exchange Route Target reachability information. This information can be used to build a route distribution graph in order to limit the propagation of VPN NLRI (such as VPN-IPv4, VPN-IPv6 or L2-VPN NLRI) between different autonomous systems or distinct clusters of the same autonomous system.
1. Introduction

In RFC2547bis, PE routers use Route Target (RT) extended communities to control the distribution of routes into VRFs. Within a given iBGP mesh, PE routers need only to hold routes marked with Route Targets pertaining to VRFs that have local CE attachments.

It is common, however, for an autonomous system use route reflection [BGP-RR] in order to simplify the process of bringing up a new PE router in the network and to limit the size of the iBGP peering mesh.

In such a scenario, as well as when VPNs may have members in more than one autonomous system, the number of routes carried by the inter-cluster or inter-as distribution routers is an important consideration.

In order to limit the VPN routing information that is maintained at a given RR, RFC2547bis [RFC2547bis] suggests, in section 4.3.3,, the usage of "Cooperative Route Filtering" [ORF] between route reflectors.

As currently defined, "Cooperative Route Filtering" has a fundamental limitation in that it can only distribute information in a point-to-point fashion. As such, it doesn’t lend itself to be used to control the propagation of VPN NLRI information, either in an hierarchical way within an autonomous system, or between autonomous systems.

This limitation conditions the effectiveness of the suggestions presented in section 4.3.3. of RFC2547bis [RFC2547bis] in terms of their ability to limit the number of VPN routes known to the RRs. Of these, option 2 proposes that route reflectors build their inter-cluster Route Target filter based on the routes received from client PE routers. This assumes a symmetric model in which a VPN uses the same Route Target value for both Import and Export targets. An asymmetric model, such as an hub-and-spoke scenario, would not be supported by this suggestion. This proposal addresses this issue by basing itself on the Import Targets that define the VPN NLRI to VRF mapping.

While it would be possible to extend the encoding currently defined for extended-community ORF in order to achieve this purpose, BGP itself already has all the necessary machinery for dissemination of arbitrary information in a loop free fashion, both within a single autonomous system, as well as across multiple autonomous systems.

This document builds on the model described in RFC2547bis and on concept of cooperative route filtering by adding the ability to propagate Route Target information between iBGP meshes.
By using MP-BGP UPDATE messages to propagate Route Target information it is possible to reuse all this machinery including route reflection, confederations and inter-as information loop detection.

Received Route Target information can then be used to restrict advertisement of VPN NLRI to peers that have advertised their respective Route Targets, effectively building a route distribution graph. In this model, VPN NLRI routing information flows in the inverse direction of Route Target information.

This mechanism is applicable to any BGP NLRI that controls the distribution of routing information based on Route Targets, such as BGP L2VPNs [L2VPN] and VPLS [VPLS].

Throughout this document, the term NLRI, which originally expands to "Network Layer Reachability Information" is used to describe routing information carried via MP-BGP updates without any assumption of semantics.

2. Inter-AS VPN route distribution.

In order to better understand the problem at hand, it is helpful to divide it in its inter-AS and intra-AS components. Figure 1 represents an arbitrary graph of autonomous systems (a through j) interconnected in an ad-hoc fashion. The following discussion ignores the complexity of intra-AS route distribution.

```
+----------------------------------+
| +---+    +---+    +---+          |
| | a | -- | b | -- | c |          |
| +---+    +---+    +---+          |
|   |        |                     |
|   |        |                     |
| +---+    +---+    +---+    +---+ |
| | d | -- | e | -- | f | -- | j | |
| +---+    +---+    +---+    +---+ |
|       /            |            |
|       /             |            |
| +---+    +---+    +---+          |
| | g | -- | h | -- | i |          |
| +---+    +---+    +---+          |
+----------------------------------+
```

Figure 1.

Lets consider the simple case of a VPN with CE attachments in ASes a and i using a single Route Target to control VPN route distribution. Ideally we would like to build a flooding graph for the respective
VPN routes that would not include nodes \(\{c, g, h, j\}\).

In order to achieve this we will rely on \(\text{AS}a\) and \(\text{AS}i\) generating a NLRI consisting of <as#, route-target>. Receipt of such an advertisement by one of the \(\text{AS}\)es in the network will signal the need to distribute VPN routes containing this Route Target community to the peer that advertised this route.

Using routes that include both route-target and originator as#, allows BGP speakers to use standard path selection rules concerning as-path length (and other policy mechanisms) to prune duplicate paths in the flooding graph, while maintaining the information required to reach all autonomous systems advertising the Route Target.

In the example above, \(\text{AS}e\) needs to maintain a path to \(\text{AS}a\) in order to flood VPN routing information originating from \(\text{AS}i\) and vice-versa. It should however prune less preferred paths such as the longer path to \(\text{AS}i\) with as-path (\(g \ h \ i\)).

Extending the example above to include \(\text{AS}j\) as a member of the VPN distribution graph would cause \(\text{AS}f\) to advertise 2 Route Target routes to \(\text{e}\), one containing origin \(\text{AS}i\) and one containing origin \(\text{AS}j\). While advertising a single path, lets assume (\(f \ j\)) is selected, would be sufficient to guarantee that VPN information flows to all VPN member \(\text{AS}\)es, the information concerning the path (\(f \ i\)) is necessary to prune the arc (\(g \ h \ i\)) from the route distribution graph.

As with other approaches for building distribution graphs, the benefits of this mechanism are directly proportional to how "sparse" is the VPN membership. Standard RFC2547 inter-AS behavior can be seen as a dense-mode approach, to make the analogy with multicast routing protocols.

3. Intra-AS VPN route distribution

As indicated above, the inter-AS VPN route distribution graph, for a given route-target, is constructed by creating a directed arc on the inverse direction of received Route Target UPDATEs containing an NLRI of the form <as#, route-target>.

Inside the BGP topology of a given autonomous-system, as far as external routes are concerned (route-targets where the as# is not the local as), it is easy to see that standard BGP route selection and advertisement rules [BGP-BASE] will allow a transit AS to create the necessary flooding state.

Consider a IPv4 NLRI prefix, sourced by a single AS, which
distributed via BGP within a given transit AS. BGP protocol rules guarantee that BGP speaker has a valid route that can be used for forwarding of data packets for that destination prefix, in the inverse path of received routing updates.

By the same token, and given that a <as#, route-target> key provides uniqueness between several ASes that may be sourcing this route-target, BGP route selection and advertisement procedures guarantee that a valid VPN route distribution path exists to the origin of the Route Target advertisement.

Route Target routes that are originated within the autonomous-system however require more careful examination. Several PE routers within a given autonomous-system may source the the same NLRI <as#, route-target>, thus default route advertisement rules are no longer sufficient to guarantee that within the given AS each node in the distribution graph has selected a feasible path to each of the PEs that import the given route-target.

When processing Route Target routes for which the as# is equal to the local autonomous system, it is necessary to consider all available iBGP paths for a given RT prefix when performing outbound route filtering, not just the best path.

In addition, when advertising Route Target NLRI information sourced by the local autonomous system to an iBGP peer, a BGP speaker shall modify its procedure to calculate the BGP attributes such that:

- When advertising a route to a route-reflector client, the Originator attribute shall be set to the router-id of the advertiser and the Next-hop attribute shall be set of the local address for that session.

- When advertising a route to a non client peer, if the best path as selected by path selection procedure described in section 9.1 of [BGP-BASE], is a route received from a non-client peer, and there is an alternative path to the same destination from a client, the attributes of the client path are advertised to the peer.

The first of these route advertisement rules is designed such that the originator of a route does not drop a route which is reflected back to it, thus allowing the route reflector to use this route in order to signal the client that it should distribute VPN routes with the specific target towards the reflector.

The second rule makes is such that any BGP speaker present in an iBGP mesh can signal the interest of its route reflection clients in receiving VPN routes for that target.
An alternative solution to the procedure given above would have been to source different routes per PE, such as NLRI of the form <originator-id, route-target>, and aggregate them at the edge of the network. The solution adopted is considered to be advantageous over the former given that it requires less routing-information within a given AS.

4. Route Target advertisements

Route Target routing information is advertised in BGP UPDATE messages using the MP_REACH_NLRI and MP_UNREACH_NLRI attributes [BGP-MP]. The <AFI, SAFI> value pair used to identify this NLRI is (AFI=1, SAFI=132).

The Next Hop field of MP_REACH_NLRI attribute shall be interpreted as an IPv4 address, whenever the length of NextHop address is 4 octets, and as a IPv6 address, whenever the length of the NextHop address is 16 octets.

The NLRI field in the MP_REACH_NLRI and MP_UNREACH_NLRI is a prefix of 0 to 96 bits encoded as defined in section 4 of [BGP-MP].

This prefix is structured as follows:

```
+-------------------------------+
| origin as        (4 octects) |
+-------------------------------+
| route target     (8 octects) |
+-------------------------------+
```

Except for the default route target, which is encoded as a 0 length prefix, the minimum prefix length is 32 bits. Thus, the origin AS must be set on a prefix.

Route targets can then be expressed as prefixes, where for instance, a prefix would encompass all route target extended communities assigned by a given Global Administrator [BGP-EXTCOMM].

The default route target can be used to indicate to a peer the willingness to receive all VPN route advertisements such as, for instance, the case of route reflector speaking to one of its PE router clients.
5. Capability Advertisement

A BGP speaker that wishes to exchange Route Target information must use the the Multiprotocol Extensions Capability Code as defined in [BGP-MP], to advertise the corresponding (AFI, SAFI) pair.

A BGP speaker MAY participate in the distribution of Route Target information while not using the learned information for purposes of VPN NLRI route filtering, although the latter is discouraged.

6. Operation

A VPN NLRI route should be advertised to a peer that participates in the exchange of Route Target information if that peer has advertised either the default Route Target or any of the targets contained in the extended communities attribute of the VPN route in question.

When a BGP speaker receives a BGP UPDATE that advertises or withdraws a given Route Target, it should examine the RIB-OUTs of VPN NLRIs and reevaluate the advertisement status of routes that match the Route Target in question.

A BGP speaker should generate the minimum set of BGP VPN route updates necessary to transition between the previous and current state of the route distribution graph that is derived from Route Target information.

7. Deployment considerations

One method that may be used to limit VPN route distribution is to partition routes into different planes. In this scenario, a PE router typically has a redundant connection to each route distribution plane.

A simple example of this methodology is an autonomous-system where routes containing "odd" Route Targets are handled by plane A and "even" Route Targets by plane B. Each plane consisting of at least two RRs, so that PE routers can have a redundant connection to both A and B.

This proposal is orthogonal to the technique described above. If a given routing distribution plane contains more than one cluster, then RT-based VPN route filtering can limit the number of routes that needs to be maintained in each cluster.

The effectiveness of RT-based filtering depends on how sparse the VPN
For instance, in the inter-as case, it is likely that a given VPN is connected to only a subset of all participating ASes. The only current mechanism to limit the scope of VPN route flooding is through manual filtering on the EBGP border routers. With the current proposal such filtering will be performed based on the dynamic RT-route information.

In some inter-as deployments not all RTs used for a given VPN have external significance. For example, a VPN can use an hub RT and a spoke RT internally to an autonomous-system. The spoke RT does not have meaning outside this AS and so it may be stripped at an external border router. The same policy rules that result in extended community filtering can be applied to RT-route filtering in order to avoid advertising an RT-route for the spoke-RT in the example above.

Throughout this document, we assume that autonomous-systems agree on an RT assignment convention. RT translation at the extern border router boundary, is considered to be a local implementation decision, as it should not affect inter-operability.

8. Security considerations

This document does not alter the security properties of BGP-based VPNs.

9. Acknowledgments

This proposal is based on the extended community route filtering mechanism defined in [ORF].

Ahmed Guetari was instrumental in defining requirements for this proposal.

The authors would also like to thank Yakov Rekhter, Dan Tappan, Dave Ward, John Scudder, and Jerry Ash for their comments and suggestions.
10. References


[RFC2547bis] "BGP/MPLS VPNs", Rosen et. al., draft-ietf-ppvpn-rfc2547bis-03.txt, 10/02.


11. Authors’ Addresses