Inter Domain considerations for Constrained Route distribution
draft-litkowski-idr-rtc-interas-01

Abstract

   [RFC4684] defines Multi-Protocol BGP (MP-BGP) procedures that allow
   BGP speakers to exchange Route Target reachability information in
   order to limit the propagation of Virtual Private Networks (VPN)
   Network Layer Reachability Information (NLRI).

   [RFC4684] addresses both intra domain and inter domain distributions.
   Based on operational deployments, the current distribution model
   defined in [RFC4684] may cause some issue in specific scenarios.

   This document refines the route distribution rules for inter domain
   NLRIs in order to address these specific scenarios.

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 [RFC2119].

Copyright Notice

Table of Contents

1. External NLRI propagation

   [RFC4684] Section 3.1 and 3.2 describes propagation of Route Target
   NLRI between ASes and inside an AS and distinguish two types of NLRIs
   :

   o Locally originated NLRI where origin-as field of the NLRI is equal
      to the local AS number.

   o External NLRI where origin-as field of the NLRI is different from
      the local AS number.

   The global idea of inter AS propagation, is to propagate only VPN
   routes on shortest path towards the peer ASes using pruning of some
   branches of the distribution tree.

   Based on current implementations of RFC4684, we can see two flavors
   of pruning for interAS that are both compatible with RFC4684 text.

   o Pruning based on peering type : pruning rule is applied when RT
      membership path are learned from eBGP peers only. No pruning is
      applied when path is iBGP.

   o Pruning based on NLRI type : pruning rule is applied to external
      RT membership NLRIs (source AS different from local AS). This
      pruning rule applies both to eBGP and iBGP.

1.1. Peering type based pruning

```
  AS 400               AS 500
                          |
       ASBR1 --- (mpebgp vpnv4+rtc)___
             |                      \
             |                       \
       ASBR2 --- (mpebgp vpnv4+rtc) -- PE1      \
             |                           \
             |                      (mpibgp vpnv4+rtc)
             |          |                                              \
             |                              RR ----------- PE3
             |                             /
             |                      (mpibgp vpnv4+rtc)
             |                           /
       ASBR3 --- (mpebgp vpnv4+rtc) -- PE2
             |
             |
```
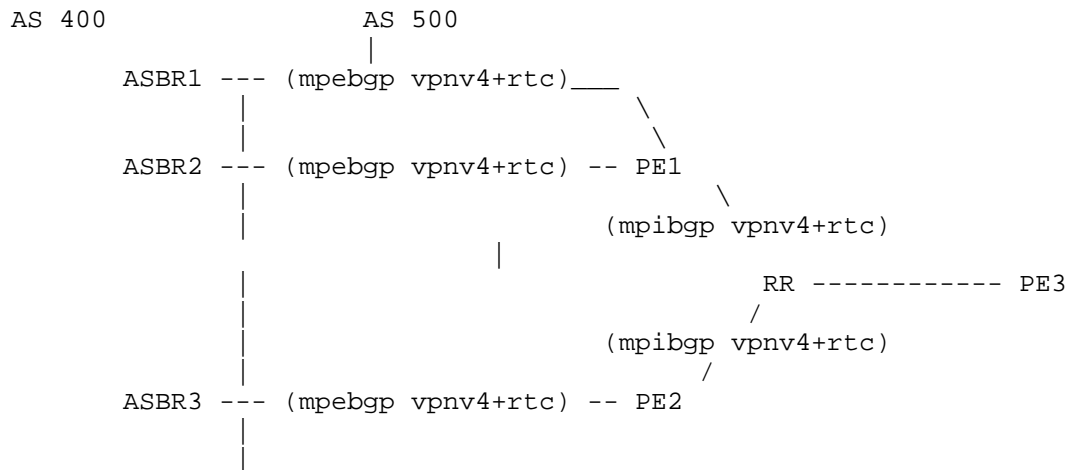
                             Figure 1

   In the figure above, ASBR1,ASBR2 and ASBR3 are MPLS VPN nodes part of
   the AS 400. We consider that all these ASBRs are importing the same
   RT : 400:1, which is also exported by PE3. All ASBRs will generate
   the same RT membership NLRI 400:400:1/96 towards their PE. PE2 will
   send its path for this RT membership to RR. As PE1 has two ebgp
   paths for the same RT membership NLRI, it will apply pruning (as per
   peering type based pruning policy), if we consider that path from
   ASBR1 is the best path, RT distribution tree will only have a branch
   to ASBR1, and so ASBR2 will not receive any VPN route for RT 400:1
   from PE1. PE1 will also send the RT membership NLRI to RR. RR will
   so have two paths for NLRI 400:400:1/96. As both path are iBGP, no
   pruning will be applied (as per peering type based pruning policy),
   and RR will create tree branches for 400:1 to both PE1 and PE2. As a
   result, VPN routes originated by PE3 with RT 400:1 will be sent by RR
   to PE1 and PE2. PE1 will propagate the routes only to ASBR1. PE2
   will propagate the routes to ASBR3. AS 400 will have knowledge from
   PE3 routes only from ASBR1 and ASBR2.

1.2. NLRI type based pruning

   We consider the same setup as in Figure 1. All ASBRs will generate
   the same RT membership NLRI 400:400:1/96 towards their PE. PE2 will
   send its path for this RT membership to RR. As PE1 has two ebgp
   paths for the same external RT membership NLRI, it will apply pruning
   (as per NLRI type based pruning policy, pruning is applied because
   NLRI is external), if we consider that path from ASBR1 is the best
   path, RT distribution tree will only have a branch to ASBR1, and so
   ASBR2 will not receive any VPN route for RT 400:1 from PE1. PE1 will
   also send the RT membership NLRI to RR. RR will so have two paths
   for NLRI 400:400:1/96. As the NLRI is external, pruning will be
   applied : if we consider that path from PE1 is the best one, a single
   branch of distribution tree will be added towards PE1. As a result,
   VPN routes originated by PE3 with RT 400:1 will be sent by RR to PE1
   only. PE1 will propagate the routes only to ASBR1. AS 400 will have
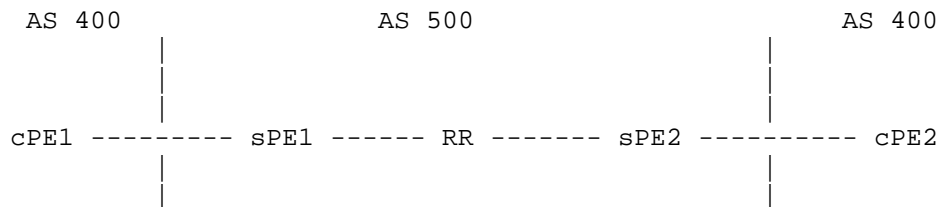   knowledge from PE3 routes only from ASBR1.

```
    AS 400                   AS 500                        AS 400
       |                        |                            |
       |                        |                            |
       |                        |                            |
   cPE1 --------- sPE1 ------ RR ------- sPE2 ---------- cPE2
       |                                                     |
       |                                                     |
       |                    Figure 2
```

   Figure 2 presents at typical case where an AS (AS400) uses another AS
   (AS500) as transit to build VPN services. If cPE1 and cPE2 shares a
   common VPN using RT 400:1, in case of NLRI type based pruning in
   AS500, RR in AS500 will perform pruning of VPN routes for NLRI
   400:400:1/96. Considering that path from sPE1 is considered as best
   path, sPE2 will be pruned and cPE2 will never receive VPN routes from
   cPE1. This issue is discussed further in Section 2.

1.3. Analysis of both approaches

   Both pruning approaches have pros and cons. Service Provider will
   need to be aware of this pros/cons while deploying inter AS RTC.

   o NLRI type based pruning helps in saving BGP paths in network
      nodes, inter AS distribution tree is only established on shortest
      path (at AS boundary and within the AS). In figure 1, PE2 does
      not receive VPN routes for RT 400:1 because these routes are
      already advertised through another path. This approach prevents
      hot potatoe routing and transit for disjoint ASes.

   o Peering type based pruning is based on the fact that the local AS
      does not know the precise location of the VPNs in the peer AS, so
      there is no reason for a route reflector to perform blind pruning
      that may lead to suboptimal routing. In figure 1, if we consider
      that ASBR3 is located in New York City, and ASBR1/2 are located in
      San Francisco. Considering that PE3 is located in Washington,
      performing NLRI type based pruning will prevent ASBR3 to receive
      PE3 routes, so routing from Washington to New York City will
      transit through San Francisco. We must note that in case of ASBR1
      and ASBR2 being in two far cities, peering type based pruning will
      also suffer from suboptimal routing. The other point in favor of
      peering type pruning is faster convergence. In figure 1, when PE1
      fails, backup routes are already available in AS400 through ASBR3.

   As a summary, NLRI type based pruning helps in saving BGP paths in
   the transit networks, while peering type based pruning permits more
   optimal routing and faster convergence with the drawback of
   propagating additional routes. Peering type based pruning may also
   experience convergence or suboptimal routing case in case a single
   node is attached to multiple routers in the external AS.

2. Problem statement : disjoint peer AS

   The previous section described how inter AS route distribution works
   and pros and cons of the existing approaches. Apart of these pros/
   cons, pruning in both solutions may lead to some problematic
   situation where the remote AS is disjoint, as already shown in
   Section 1.2.

```
+-------+
| DC1   | -- CE1 -- (mpebgp vpnv4+rtc) -- PE1
+-------+                                      \
                                        (mpibgp vpnv4+rtc)


   \
                                            RR
                                           /
                                        (mpibgp vpnv4+rtc)
+-------+                             /
| DC2   | -- CE2 -- (mpebgp vpnv4+rtc) -- PE2
+-------+
```
                              Figure 3

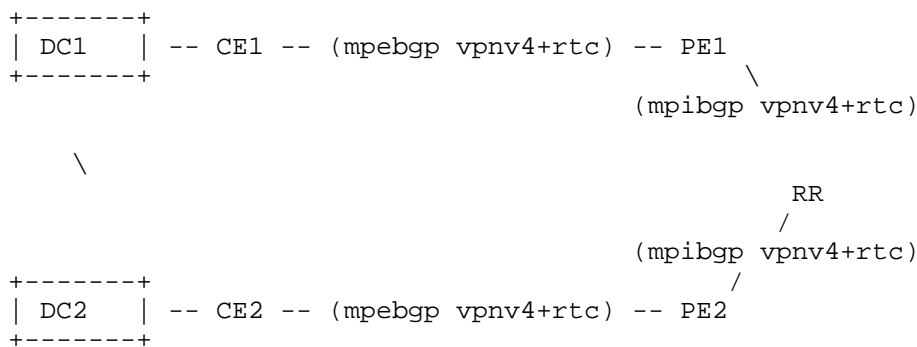   The figure above describes another typical service provider scenario
   where datacenters are connected through MPLS VPN interas option B
   with the Service Provider network. Route Target Constraint (RTC) is
   deployed on MPeBGP sessions as well as internally in the service
   provider network to ensure optimal distribution of VPN routes
   (required for scaling reason). In this scenario, both Datacenters

are using the same AS number, generally a private ASN (65000) like a
typical PE-CE connection. As we expect DCs to communicate between
each other, some features like "as-override" are deployed on PEs to
overcome ASPATH loop issue.

In the Figure 3, CE1 and CE2 are advertising the RT 1:1 respectively
to PE1 and PE2, the generated NLRI would be 65000:1:1/96. According
to procedures defined in [RFC4684] Section 3.2, both PEs are using
the standard BGP route selection and advertising rules. So both PEs
are advertising their path for 65000:1:1/96 to the route-reflector.
In case of NLRI type based pruning, route-reflector will establish
the distribution tree only to PE1 (considering PE1 is the best path).

Due to this behavior, VPN routes from DC1 would never to send to DC2
because PE2 is not part of the flooding tree and as DC1 and DC2 are
disjoint, even if they are using the same ASN, there is no
communication possible between them.

The same issue may appear if two MPeBGP sites using the same ASN are
connected on the same PE like in figure 4. In this situation both
NLRI type based pruning and Peering type based pruning solutions are
impacted.
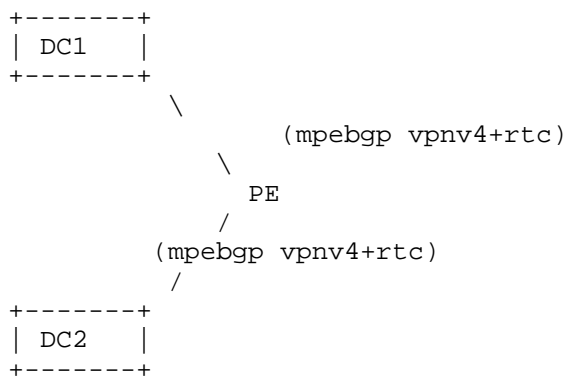
```
   +-------+
   | DC1   |
   +-------+
           \
                   (mpebgp vpnv4+rtc)
             \
               PE
             /
         (mpebgp vpnv4+rtc)
          /
   +-------+
   | DC2   |
   +-------+
           Figure 4
```

3. Proposal

   This document proposes to introduce some new behavior in complement
   of [RFC4684] to manage the disjoint AS case.

   In order to support our scenario, path pruning MAY be disabled by
   configuration for a given origin AS (different from the local AS).
   Implementations MAY also permit path pruning to be disabled for
   private AS numbers by default, but must make provision for it to be
   selectively enabled if such a feature is present.

   This modification in establishing route distribution tree may create
   unnecessary flooding states in the situations where a real AS is
   multihomed to a service provider network (as displayed in Figure 3).

```
    ASN 65000                                                  ASN 64000
   +-----------+                                              +------------+
   |   ASBR3   | -- (mpebgp vpnv4+rtc) -- ASBR1     PE1 ---- | CE1 --- DC1 |
   |   |       |                             \     /         +------------+
   |   |       |                           (mpibgp vpnv4+rtc)
   |(vpnv4+rtc)|                                                    \  /
   |   |       |                             RR
   |   |       |                           /  \
   |   |       |                        (mpibgp vpnv4+rtc)        ASN 64000
   |   |       |                          /    \          +------------+
   |   ASBR4   | -- (mpebgp vpnv4+rtc) -- ASBR2     PE2 ---- | CE2 --- DC2 |
   +-----------+                                              +------------+
```

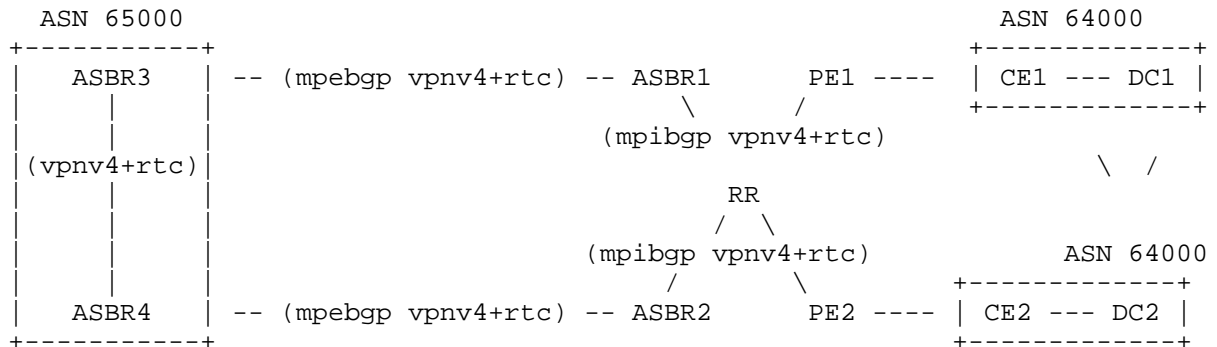   Figure 3

   In the figure above, disabling pruning is required for AS64000 but it
   may be interesting to keep it enabled for AS65000. Implementations
   may require support for such granularity as proposed previously.

4. Security considerations

   This document does not introduce any new security issue compared to
   [RFC4684].

5. Acknowledgements

6. IANA Considerations

   There is no IANA consideration.

7. Normative References

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

   [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.

Authors' Addresses

    Stephane Litkowski
    Orange Business Service

    Email: stephane.litkowski@orange.com

    Jeff Haas
    Juniper Networks

    Email: jhaas@juniper.net

    Keyur Patel
    Cisco Systems

    Email: keyupate@cisco.com