Inter-Domain Routing   H. Gredler, Ed.
Internet-Draft   K. Vairavakkalai
Intended status: Informational   C. Ramachandran
Expires: July 21, 2015   B. Rajagopalan
                                                  Juniper Networks, Inc.
                                                                L. Fang
                                                              Microsoft
                                                               E. Aries
                                                               Facebook
                                                       January 17, 2015

Egress Peer Engineering using BGP-LU
draft-gredler-idr-bgplu-epe-02

Abstract

   The MPLS source routing paradigm provides path control for both
   intra- and inter- Autonomous System (AS) traffic. For Intra-AS path
   control, protocols like RSVP-TE and CR-LDP are utilized. This
   documents outlines how MPLS routers may use the BGP labeled unicast
   protocol (BGP-LU) for doing traffic-engineering on inter-AS links as
   outlined in [I-D.filsfils-spring-segment-routing-central-epe].

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].

Status of This Memo

Table of Contents

1. Introduction

   Today BGP-LU [RFC3107] is used both as an intra-AS
   [I-D.ietf-mpls-seamless-mpls] and inter-AS routing protocol. BGP-LU
   may advertise a MPLS transport path between IGP regions and
   Autonomous Systems. Those paths may span one or more router hops.

This document describes advertisement and use of one-hop MPLS label-switched paths (LSP) for traffic-engineering the links between Autonomous Systems.

Consider Figure Figure 1: an ASBR router (R2) advertises a labeled host route for the remote-end IP address of its link (IP3). The BGP next-hop gets set to R2s loopback IP address. For the advertised Label <N> a forwarding action of 'POP and forward' to next-hop (IP3) is installed in R2's MPLS forwarding table. Now consider if R2 had several links and R2 would advertise labels for all of its inter-AS links. By pushing the corresponding MPLS label <N> on the label-stack an ingress router R1 may control the egress peer selection.

```
          AS1               :          AS2
                            :
+----+   iBGP   +----+  :  eBGP   +----+
| R1 |----------| R2 |-IP2----IP3-| R3 |
+----+          +----+  :         +----+
                            :
   -----------traffic-flow---------->
   <------------route-flow-----------
```

                  Figure 1: single-hop LSPs

Of course, since R1 and R2 may not be directly connected to each other, if the interior routers within AS1 do not maintain routes to external destinations, carrying traffic to such destinations would require a tunnel from R1 to R2. Such tunnel could be realized as either a MPLS Label Switch Path (LSP), or by GRE.

2. Motivation, Rationale and Applicability

BGP-LU is often just seen as a 'stitching' protocol for connecting Autonomous Systems. BGP-LU is often not viewed as a viable protocol for solving the Inter-domain traffic-engineering problem.

With this document the authors want to clarify the use of BGP-LU for Egress Peering traffic-engineering purposes and encourage both implementers and network operators to use a widely deployed and operationally well understood protocol, rather than inventing new protocols or new extensions to the existing protocols.

3. Sample Topology

The following topology (Figure 2) and IP addresses shall be used throughout the Egress Peering Engineering advertisement examples.

```
                              :              :
              AS 1            :     AS 2      :    AS 4
                              :              :
                              :      +-----+ :
                          /IP2--:-IP3--|ASBR3|   :
   +-----+          +-----+-IP4--:-IP5--+-----+-----------+-----+
   | R1   +-------------+ASBR1|    :                       |ASBR6|
   +--+--+             +--+--+-IP6--:-IP7--+-----+-----------+-----+
      |                   |  \   :     |ASBR4|    :   /
      |                   |   \  :     +-----+    :  /
      |                   |     IP8-                ---
      |                   |      \ ............... /
      |                   |        IP9-          ---
      |                   |      :    \      /   :
      |                   |      :     \    /    :
   +--+--+             +--+--+   :     +--+--+   :
   | R2   +-------------+ASBR2|-IP10-:-IP11-|ASBR5|   :
   +-----+             +-----+   :     +-----+   :
                              :              :
                              :     AS3      :
                              :              :
```
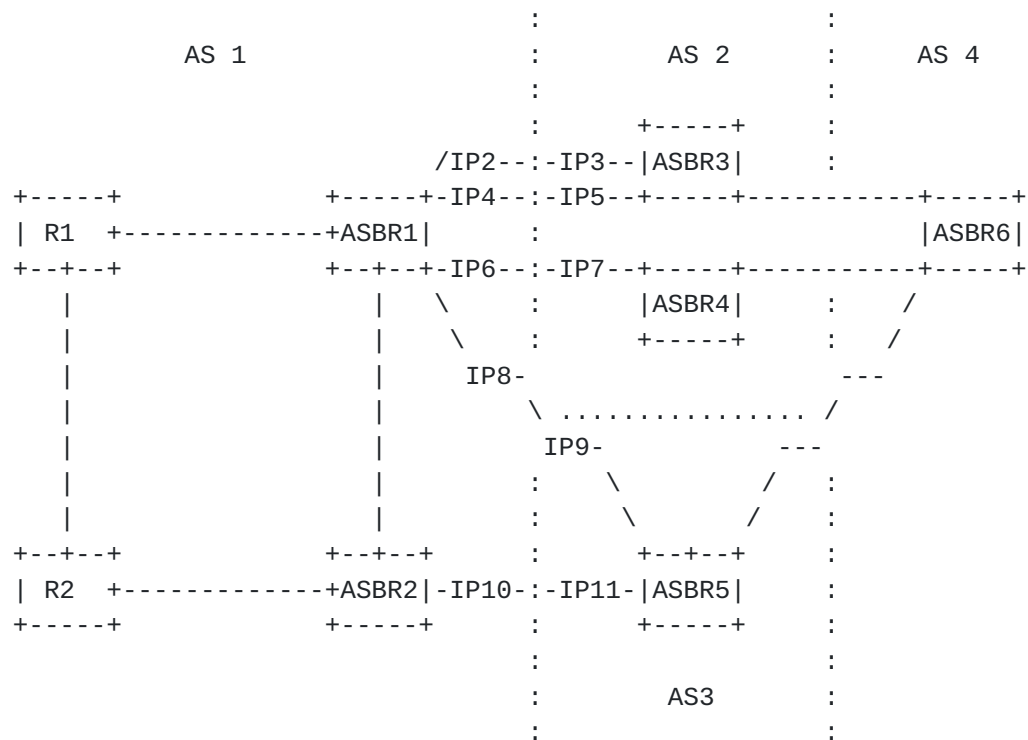
Figure 2: Sample Topology

3.1. Loopback IP addresses and Router-IDs

o R1: 192.168.1.1/32

o R2: 192.168.1.2/32

o ASBR1: 192.168.1.11/32

o ASBR2: 192.168.1.12/32

o ASBR3: 192.168.1.13/32

o ASBR4: 192.168.1.14/32

o ASBR5: 192.168.1.15/32

o ASBR6: 192.168.1.16/32

3.2. Link IP addresses

o ASBR1 (10.0.0.2/31) to ASBR3 (10.0.0.3/31) link #1

o ASBR1 (10.0.0.4/31) to ASBR3 (10.0.0.5/31) link #2

o ASBR1 (10.0.0.6/31) to ASBR4 (10.0.0.7/31)

o ASBR1 (10.0.0.8/31) to ASBR5 (10.0.0.9/31)

o ASBR2 (10.0.0.10/31) to ASBR5 (10.0.0.11/31)

4. Service Route Advertisement

   In Figure 3 a simple network layout is shown. There are two classes
   of BGP speakers:

   1. Ingress Routers

   2. Controllers

   Ingress routers receive BGP-LU routes from the ASBRs. Each BGP-LU
   route corresponds to an egress link. Furthermore Ingress routers
   receive their service routes using the BGP protocol. The BGP Add-
   paths extension [I-D.ietf-idr-add-paths] ensures that multiple paths
   to a given service route may get advertised.

   Controllers receive BGP-LU routes from the ASBRs as well. However
   the service routes may be received either using the BGP protocol plus
   the BGP Add-paths extension [I-D.ietf-idr-add-paths] or alternatively
   The BGP Monitoring protocol [I-D.ietf-grow-bmp] (BMP). BMP has got
   support for advertising the RIB-In of a BGP router. As such it might
   be a suitable protocol for feeding all potential egress paths of a
   service-route from a ASBR into a controller.

5. Egress Next-hop Advertisement

   An ASBR assigns for each of its next-hops facing an eBGP peer, a
   distinct label and advertises it to its internal BGP mesh. The ASBR
   programs a forwarding action 'POP and forward' into the MPLS
   forwarding table. Note that the neighboring AS is not required to
   support exchanging NLRIs with the local AS using BGP-LU. It is the
   local ASBR (ASBR{1,2}) which generates the BGP-LU routes into its
   iBGP mesh or controller facing session(s). The forwarding next-hop
   for those routes points to the link-IP addresses of the remote ASBRs
   (ASBR{3,4,5}). Note that the generated BGP-LU routes always match
   the BGP next-hop that the remote ASBRs set their BGP service routes
   to, such that the software component doing route-resolution
   understands the association between the BGP service route and the
   BGP-LU forwarding route.

5.1. **iBGP meshing and BGP nexthop rewrite policy**

   Throughout this document we describe how the BGP next-hop of both BGP
   Service Routes and BGP-LU routes shall be rewritten. This may clash
   with existing network deployments and existing network configurations
   guidelines which e.g. may mandate to rewrite the BGP next-hop when
   an BGP update enters an AS.

   The Egress peering use case assumes a central controller like it is
   shown Figure 3. In order to support both existing BGP nexthop
   guidelines and the suggestion described in this document and
   implementation SHOULD support several internal BGP peer-groups:

   1. iBGP peer group for Ingress Routers

   2. iBGP peer group for Controllers

   The first peer group MAY be left unchanged and use any existing BGP
   nexthop rewrite policy. The second peer group MUST use the BGP
   rewrite policy described in this document for both service and BGP-LU
   routes.

   Of course a common iBGP peer group and a common rewrite policy may be
   used as the proposed policy is compatible with existing routing
   software implementations of BGP next-hop route resolution.

```
    +-----------+
    | Ingress   |
    |  Router   |
    +-----------+
         ^
         |
    +-----------+
    |    BGP    |                +------------+
    |   Route   |-------------->| Controller |
    | Reflector |                +------------+
    +-----------+                      ^
        ^   ^                          |
        |   |                          |
        |   +------------------+       |
        |                      |   |
        v                      v   v
    +-----------+          +-----------+
    |    BGP    |          |    BGP    |
    |   ASBR1   |   . . .  |   ASBR2   |
    +-----------+          +-----------+
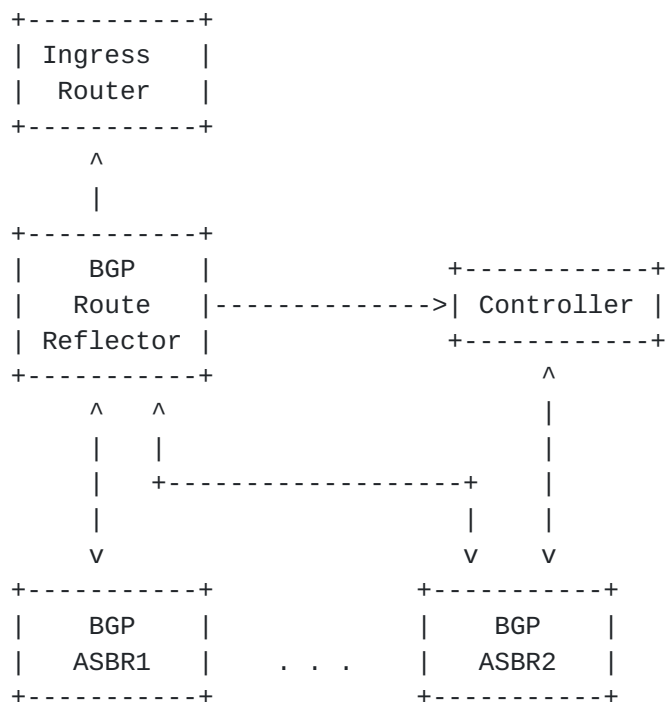```

                    Figure 3: Selective iBGP NH rewrite

5.2.  Single-hop eBGP

   In Figure 2 the ASBR{1,5} and ASBR{2,5} links are examples for
   single-hop eBGP advertisements.

   o  ASBR5 advertises a BGP service (SAFI-1) route {172.16/12} to ASBR1
      with a BGP next-hop of 10.0.0.9.  When ASBR1 re-advertises this
      BGP service route towards its iBGP mesh (R{1,2}) it does not
      overwrite the BGP next-hop, but rather leave it unchanged.

   o  ASBR1 advertises a BGP-LU route {10.0.0.9/32, label 100} with a
      BGP next hop of 192.168.1.11.  ASBR1 programs a MPLS forwarding
      state of 'POP and forward' to 10.0.0.9 for the advertised label
      100.

   o  ASBR5 advertises a BGP service (SAFI-1) route {172.16/12} to ASBR2
      with a BGP next-hop of 10.0.0.11.  When ASBR2 re-advertises this
      BGP service route towards its iBGP mesh (R{1,2}) it does not
      overwrite the BGP next-hop, but rather leave it unchanged.

   o  ASBR2 advertises BGP-LU route {10.0.0.11/32, label 101} with a BGP
      next hop of 192.168.1.12.  ASBR2 programs a MPLS forwarding state
      of 'POP and forward' to 10.0.0.11 for the advertised label 101.

   o  Should the operator already be redistributing egress links into
      the network for purposes of BGP next-hop resolution, the BGP-LU
      route {10.0.0.9/32, label 100} will now take precedence due to LPM
      over the previous redistributed prefix {10.0.0.8/31}. If the BGP
      next-hop prefix {10.0.0.9/32} were to be redistributed as-is, then
      standard protocol best-path and preference selection mechanisms
      will be exhausted in order to select the best-path.

   o  In general, ASBR1 may receive advertisements for the route to
      172.16/12 from ASBR3 and ASBR4, as well as from ASBR5.  One of
      these other advertisements may be chosen as the best path by the
      BGP decision process.  In order to allow ASBR1 to re-advertise the
      route to 172.16/12 received from ASBR5 with next-hop 10.0.0.9,
      independent of the other advertisements received, ASBR1 and R{1,2}
      need to support the BGP add-paths extension.
      [I-D.ietf-idr-add-paths].

5.3.  Multi-hop eBGP

   Todays operational practice for load-sharing across parallel links is
   to configure a single multi-hop eBGP session between a pair of
   routers.  The IP addresses for the Multi-hop eBGP session are
   typically sourced from the loopback IP interfaces.  Note that those
   IP addresses do not share an IP subnet.  Most often those loopback IP

addresses are most specific host routes.  Since the BGP next-hops of
the received BGP service routes are typically rewritten to the remote
routers loopback IP address they cannot get immediatly resolved by
the receiving router.  To overcome this, the operator configures a
static route with next-hops pointing to each of the remote-IP
addresses of the underlying links.

In Figure 2 both ASBR{1,3} links are examples of a multi-hop eBGP
advertisement.  In order to advertise a distinct label for a common
FEC throughout the iBGP mesh, ASBR1 and all the receiving iBGP
routers need to support the BGP Add-paths extension.
[I-D.ietf-idr-add-paths].

o  ASBR3 advertises a BGP service (SAFI-1) route {172.16/12} over
   multi-hop eBGP to ASBR1 with a BGP next-hop of 192.168.1.13.  When
   ASBR1 re-advertises this BGP service route towards its iBGP mesh
   (R{1,2}) it does not overwrite the BGP next-hop, but rather leave
   it unchanged.  Note that the iBGP routers SHOULD support the BGP
   Add-paths extensions [I-D.ietf-idr-add-paths].  such that ASBR can
   re-advertise all paths to the SAFI-1 route {172.16/12}.

o  For link #1, ASBR1 advertises into its iBGP mesh a BGP-LU route
   {192.168.1.13/32, label 102} with a BGP next hop of 192.168.1.11.
   To differentiate this from the link #2 route-advertisement (which
   contains the same FEC) it is setting the path-ID to 1.  ASBR1
   programs a MPLS forwarding state of 'POP and forward' to 10.0.0.3
   for the advertised label 102.

o  For link #2, ASBR1 advertises into its iBGP mesh a BGP-LU route
   {192.168.1.13/32, label 103} with a BGP next hop of 192.168.1.11.
   To differentiate this from the link #1 route-advertisement (which
   contains the same FEC) it is setting the path-ID to 2.  ASBR1
   programs a MPLS forwarding state of 'POP and forward' to 10.0.0.5
   for the advertised label 103.

o  Should the operator already be redistributing static routes into
   the network, the BGP next-hop {192.168.1.13} may already be
   resolvable.  It is then that standard protocol best-path and
   preference selection mechanisms will be exhausted in order to
   select the best-path.

## 5.4.  Grouping of Peers

In addition to offering a distinct BGP-LU label for each egress link,
an ASBR MAY want to advertise a BGP-LU label which represents a load-
balancing forwarding action across a set of peers.  The difference is
here that the ingress node gives up individual link control, but
rather delegates the load-balancing decision to a particular egress

router which has the freedom to send the traffic down to any link in
the Peer Set as identified by the BGP-LU label.

Assume that ASBR1 wants to advertise a label identifying the Peer Set
{ASBR3, ASBR4, ASBR5}.

o  For the two ASBR{1,3} links in Figure 2, belonging to Peer Set 1,
   ASBR1 advertises a single BGP-LU route {192.168.1.13/32, label
   104} with a BGP next hop of 192.168.1.11.  To differentiate this
   from the ASBR{1,3} single link route-advertisements (which
   contains the same FEC) it is setting the path-ID to 3 and
   attaching a Peer-Set Community 'Peer Set 1'.

o  For the ASBR{1,4} link in Figure 2, ASBR1 advertises a BGP-LU
   route {10.0.0.7/32, label 104} with a BGP next hop of
   192.168.1.11.  To differentiate this from the ASBR{1,4} single
   link route-advertisements (which contains the same FEC) it is
   setting the path-ID to 2 and attaching a Peer-Set Community 'Peer
   Set 1'.

o  For the ASBR{1,5} link in Figure 2, ASBR1 advertises a BGP-LU
   route {10.0.0.9/32, label 104} with a BGP next hop of
   192.168.1.11.  To differentiate this from the ASBR{1,5} single
   link route-advertisements (which contains the same FEC) it is
   setting the path-ID to 2 and attaching a Peer-Set Community 'Peer
   Set 1'.

Finally ASBR1 programs a MPLS forwarding state of 'POP and load-
balance' to {10.0.0.3, 10.0.0.5, 10.0.0.7, 10.0.0.9} for the
advertised label 104.

## 6.  Egress Link Protection

It is desirable to provide a local-repair based protection scheme, in
case a redundant path is available to reach a peer AS.  Protection
may be applied at multiple levels in the routing stack.  Since the
ASBR has insight into both BGP-LU and BGP service advertisements,
protection can be provided at the BGP-LU, at the BGP service or both
levels.

## 6.1.  FRR backup routes

Assume the network operator wants to provide a local-repair next-hop
for the 172.16/12 BGP service route at ASBR1.  The active route
resolves over the parallel links towards ASBR3.  In case the link #1
between ASBR{1,3} fails there are now several candidate backup paths
providing protection against link or node failure.

### 6.1.1.  Local links

Assuming that the remaining link #2 between ASBR{1,3} has enough
capacity, and link-protection is sufficient, this link MAY serve as
temporary backup.

However if node-protection or additional capacity is desired, then
the local link between ASBR{1,4} or ASBR{1,5} MAY be used as
temporary backup.

### 6.1.2.  Remote BGP-LU labels

ASBR1 is both originator and receiver of BGP routing information.
For this protection method it is required that the ASBRs support the
[I-D.ietf-idr-best-external] behavior.  ASBR1 receives both the BGP-
LU and BGP service routes from ASBR2 and therefore can use the ASBR2
advertised label as a backup path given that ASBR1 has a tunnel
towards ASBR2.

### 6.1.3.  Local IP forwarding tables

For protecting plain unicast (Internet) routing information a very
simple backup scheme could be to recurse to the relevant IP
forwarding table and do an IP lookup to further determine a new
egress link.

### 7.  Dynamic link utilization

For a software component which controls the egress link selection it
may be desirable to know about a particular egress link current
utilization, such that it can adjust the traffic that gets sent to a
particular interface.

In [I-D.ietf-idr-link-bandwidth] a community for reporting link-
bandwidth is specified.  Rather than reporting the static bandwidth
of the link, the ASBRs shall report the available bandwidth as seen
by the data-plane via the link-bandwidth community in their BGP-LU
update message.

It is crucial that ingress routers learn quickly about congestion of
an egress link and hence it is desired to get timely updates of the
advertised per-link BGP-LU routes carrying the available bandwidth
information when the available bandwidth crosses a certain
(preconfigured) threshold.

Controllers may also utilize the link-bandwidth community among other
common mechanisms to retrieve data-plane statistics (e.g.  SNMP,
NETCONF)

## 8.  Acknowledgements

Many thanks to Yakov Rekhter, Chris Bowers and Jeffrey (Zhaohui) Zhang for their detailed review and insightful comments.

## 9.  IANA Considerations

This documents does not request any action from IANA.

## 10.  Security Considerations

This document does not introduce any change in terms of BGP security.

## 11.  References

### 11.1.  Normative References

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

[RFC3107]   Rekhter, Y. and E. Rosen, "Carrying Label Information in
            BGP-4", RFC 3107, May 2001.

### 11.2.  Informative References

[I-D.filsfils-spring-segment-routing-central-epe]
            Filsfils, C., Previdi, S., Patel, K., Aries, E.,
            shaw@fb.com, s., Ginsburg, D., and D. Afanasiev, "Segment
            Routing Centralized Egress Peer Engineering", draft-
            filsfils-spring-segment-routing-central-epe-03 (work in
            progress), January 2015.

[I-D.ietf-grow-bmp]
            Scudder, J., Fernando, R., and S. Stuart, "BGP Monitoring
            Protocol", draft-ietf-grow-bmp-07 (work in progress),
            October 2012.

[I-D.ietf-idr-add-paths]
            Walton, D., Retana, A., Chen, E., and J. Scudder,
            "Advertisement of Multiple Paths in BGP", draft-ietf-idr-
            add-paths-10 (work in progress), October 2014.

[I-D.ietf-idr-best-external]
            Marques, P., Fernando, R., Chen, E., Mohapatra, P., and H.
            Gredler, "Advertisement of the best external route in
            BGP", draft-ietf-idr-best-external-05 (work in progress),
            January 2012.

   [I-D.ietf-idr-link-bandwidth]
              Mohapatra, P. and R. Fernando, "BGP Link Bandwidth
              Extended Community", draft-ietf-idr-link-bandwidth-06
              (work in progress), January 2013.

   [I-D.ietf-mpls-seamless-mpls]
              Leymann, N., Decraene, B., Filsfils, C., Konstantynowicz,
              M., and D. Steinberg, "Seamless MPLS Architecture", draft-
              ietf-mpls-seamless-mpls-07 (work in progress), June 2014.

Authors' Addresses

   Hannes Gredler (editor)
   Juniper Networks, Inc.
   1194 N. Mathilda Ave.
   Sunnyvale, CA  94089
   US


   Email: hannes@juniper.net


   Kaliraj Vairavakkalai
   Juniper Networks, Inc.
   1194 N. Mathilda Ave.
   Sunnyvale, CA  94089
   US


   Email: kaliraj@juniper.net


   Chandra Ramachandran
   Juniper Networks, Inc.
   Electra, Exora Business Park Marathahalli - Sarjapur Outer Ring Road
   Bangalore, KA  560103
   India


   Email: csekar@juniper.net


   Balaji Rajagopalan
   Juniper Networks, Inc.
   Electra, Exora Business Park Marathahalli - Sarjapur Outer Ring Road
   Bangalore, KA  560103
   India


   Email: balajir@juniper.net

Luyuan Fang
Microsoft
5600 148th Ave NE
Redmond, WA  98052
US


Email: lufang@microsoft.com


Ebben Aries
Facebook
1601 Willow Road
Menlo Park, CA  94025
US


Email: exa@fb.com