Inter-Domain Routing H. Gredler, Ed.
Internet-Draft K. Vairavakkalai
Intended status: Informational C. Ramachandran
Expires: April 30, 2015 B. Rajagopalan
Juniper Networks, Inc.
L. Fang
Microsoft
October 27, 2014

# Egress Peer Engineering using BGP-LU
draft-gredler-idr-bgplu-epe-01

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 [RFC3209] and CR-LDP [RFC3212] are utilized. This documents outlines how MPLS routers may use the BGP labeled unicast protocol (BGP-LU) [RFC3107] for doing traffic-engineering on inter-AS links.

## 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].

Copyright Notice



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 visible 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 operator 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
                                    :                  :
                                    :      +-----+     :
                              /IP1--:-IP2--|ASBR3|     :
   +-----+             +-----+-IP3--:-IP4--+-----+-----------+-----+
   | R1  +-------------+ASBR1|      :                        |ASBR6|
   +--+--+             +--+--+-IP5--:-IP6--+-----+-----------+-----+
      |                   |   \     :      |ASBR4|     :    /
      |                   |    \    :      +-----+     :   /
      |                   |     IP7-                    ---
      |                   |         \ ................ /
      |                   |          IP8-           ---
      |                   |         :    \         /   :
      |                   |         :     \       /    :
   +--+--+             +--+--+      :      +--+--+     :
   | R2  +-------------+ASBR2|-IP9--:-IP10-|ASBR5|     :
   +-----+             +-----+      :      +-----+     :
                                    :                  :
                                    :        AS3       :
                                    :                  :
```

Figure 2: Sample Topology

### 3.1. Loopback IP addresses and Router-IDs

- o R1: 192.168.1.1

- o R2: 192.168.1.2

- o ASBR1: 192.168.1.11
- o ASBR2: 192.168.1.12
- o ASBR3: 192.168.1.13
- o ASBR4: 192.168.1.14
- o ASBR5: 192.168.1.15
- o ASBR5: 192.168.1.15

### 3.2. Link IP addresses

- o ASBR1 to ASBR3 link #1: 10.0.0.1, 10.0.0.2
- o ASBR1 to ASBR3 link #2: 10.0.0.3, 10.0.0.4





- o ASBR1 to ASBR4 link: 10.0.0.5, 10.0.0.6
- o ASBR1 to ASBR5 link: 10.0.0.7, 10.0.0.8
- o ASBR2 to ASBR5 link: 10.0.0.9, 10.0.0.10

## 4. Egress Link Advertisement

An ASBR assigns for each of its egress links 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. 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.

### 4.1. 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 10.0.0.8. 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 {10.0.0.8/32, label 100} with a BGP next hop of 192.168.1.11. ASBR1 programs a MPLS forwarding state of 'POP and forward' to 10.0.0.8 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 10.0.0.10. 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 {10.0.0.10/32, label 101} with a BGP next hop of 192.168.1.12. ASBR2 programs a MPLS forwarding state of 'POP and forward' to 10.0.0.10 for the advertised label 101.

- o Note that in order for ASBR1 to advertise towards its | iBGP mesh multiple next hops (10.0.0.8, 10.0.0.10) for the route to





172.16/12, ASBR1 and R{1,2} have to support the BGP Add-paths extension. [I-D.ietf-idr-add-paths].

## 4.2. 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. Since the BGP next-hops of the received BGP service routes are typically not on a common IP subnet, 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. 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 10.0.0.2 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 10.0.0.4 for the advertised label 103.

### 4.3. Grouping of Peers

In addition to offer 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 all links going to a particular Peer.





For link #1 and link #2 in Figure 2, ASBR1 advertises a 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 link #1 and link #2 route-advertisements (which contains the same FEC) it is setting the path-ID to 3. ASBR1 programs a MPLS forwarding state of 'POP and load-balance' to 10.0.0.2 and 10.0.0.4 for the advertised label 104.

## 5. 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 in both BGP-LU and BGP service advertisements, protection can be provided at the BGP-LU, at the BGP service or both levels.

### 5.1. 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 resolve 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.

#### 5.1.1. 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.

#### 5.1.2. 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.

#### 5.1.3. 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.

## 6. 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.

## 7. Acknowledgements

Many thanks to Yakov Rekhter for his detailed review and insightful comments

## 8. IANA Considerations

This documents does not request any action from IANA.

## 9. Security Considerations

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

## 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.

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

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.

[RFC3212] Jamoussi, B., Andersson, L., Callon, R., Dantu, R., Wu, L., Doolan, P., Worster, T., Feldman, N., Fredette, A., Girish, M., Gray, E., Heinanen, J., Kilty, T., and A. Malis, "Constraint-Based LSP Setup using LDP", RFC 3212, January 2002.

## 10.2. Informative References

[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