Routing Area Working Group
Internet-Draft
Intended status: Standards Track
Expires: November 6, 2013

H. Gredler, Ed.
Juniper Networks, Inc.
S. Amante
Level 3 Communications, Inc.
T. Scholl
Amazon
L. Jalil
Verizon

May 5, 2013

Advertising MPLS labels in IGPs
draft-gredler-rtgwg-igp-label-advertisement-04

Abstract

Historically MPLS label distribution was driven by session oriented protocols. In order to obtain a particular routers label binding for a given destination FEC one needs to have first an established session with that node.

This document describes a mechanism to distribute FEC/label mappings through flooding protocols. Flooding protocols publish their objects for an unknown set of receivers, therefore one can efficiently scale label distribution for use cases where the receiver of label information is not directly connected.

Application of this technique are found in the field of backup (Bypass, R-LFA) routing, Label switched path stitching, egress protection, explicit routing and egress ASBR link selection.

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

MPLS label allocations are predominantly distributed by using the LDP [RFC5036], RSVP [RFC5151] or labeled BGP [RFC3107] protocol. All of those protocols have in common that they are session oriented, which means that in order to learn the Label Information database of a particular router one needs to have a direct control-plane session using the given protocol.

There are a couple of practical use cases where the consumer of a MPLS label allocation may not be adjacent to the router having allocated the label. Bringing up an explicit session using existing label distribution protocols between the non-adjacent label allocator and the label consumer is the existing remedy for this dilemma.

For LDP protection routing LDP next next hop labels [NNHOP] have been proposed to provide the 2 hop neighborhood labels. While the 2 hop neighborhood provides good backup coverage for the typical network operator topology it is inadequate for some sparse for example ring like topologies.

Depending on the application, retrieval and setup of forwarding state of such >1 hop label allocations may only be transient. As such configuring and un-configuring the explicit session is an operational burden and therefore should be avoided.

The use cases described in this document are equally applicable to IPv4 and IPv6 carried over MPLS. Furthermoe the proposed use of distributing MPLS Labels using IGP prototocols adheres the architectural principles laid out in [RFC3031].

2. Motivation and Applicability

It may not be immediate obvious, however introduction of Remote LFA [I-D.ietf-rtgwg-remote-lfa] technology has implied important changes for an IGP implementation. Previously the IGP had a one-way communication path with the LDP module. The IGP supplies tracking routes and LDP selects the best neighbor based upon FEC to tracking routes exact matching results. Remote LFA changes that relationship such that there is a bi-directional communication path between the IGP and LDP. Now the IGP needs to learn about if a label switched path to a given destination prefix has been established and what the ingress label for getting there is. The IGP needs to push that label for the tracking routes of destinations beyond a remote LFA neighbor.

Since the IGP is now aware of label switched paths and it does create forwarding state based on label information it makes sense to

distribute label switched paths by the IGP as well.

3. Use cases for IGP label distribution

This section lists example use cases which illustrate IGP distribution of MPLS label switched paths.

3.1. Increase LFA backup coverage using 'Directed Forwarding'

Deployment of Loop free alternate backup technology [RFC5286] results in backup graphs whose coverage is highly dependent on the underlying Layer-3 topology. Typical network deployments provide backup coverage less than 100 percent (see RFC 6571 Section 4.3 for Results [RFC6571]) for IGP destination prefixes.

By closer examining the coverage gaps from the referenced production network topologies, it becomes obvious that most topologies lacking backup coverage are close to ring shaped topologies (Figure 1).

Remote LFA [I-D.ietf-rtgwg-remote-lfa] has introduced the notion of a "remote" LFA neighbor. This helper router which is both in P and Q space could forward the traffic to the final destination. Router 'H' is in P space, however due to the actual metric allocation router 'H' is not in Q space.

```
              +-----+
              |  D  |
              +-----+
             /       \
            / M1      \ M4 >= (M1 + M2 + M3)
           /           \
    +-----+             +-----+
    | PLR |             |  H  |
    +-----+             +-----+
           \           /
            \ M2      / M3
             \       /
              +-----+
              |  E  |
              +-----+
```

Figure 1: Coverage gap analysis

The protection router (PLR) evaluates for a primary path to destination 'D' if {E -> H -> D} is a viable backup path. Because the metric M4 {H -> D} is higher than the sum of the original primary path and the path from router 'H' to the PLR, this particular path

would result in a loop and therefore is rejected.

Now consider that router 'H' would advertise a label for FEC 'D', which has the semantics that H will POP the label and forward to the destination node 'D'. This is done irrespective of the underlying IGP metric 'M4' it is a 'strict forwarding' label. The PLR router can now construct a label stack where the outermost label provides transport to router 'H'. The next label on the MPLS stack is the IGP learned 'strict forwarding label' label. Note that the label 'strict forwarding' semantics are similar to a 1-hop ERO (Explicit route object). The Remote 'LFA' calculation would need to get changed, such that even if a node is not in PQ space, but rather in P space, it may get used as a backup neighbor if it advertises a strict forwarding label to the final destination. A recursive version of the algorithm is applicable as well as long a node in P space has some non looping LSP path to the final destination. The PLR router can now program a backup path irrespective of the undesirable underlying layer-3 topology.

Using existing tunnels for backup routing has been previously described in [I-D.bryant-ipfrr-tunnels]. Section 5.2.3 'Directed forwarding' describes an option to insert a single MPLS label between the tunnel and the payload. Traffic may thereby be directed to a particular neighbor. The mechanism described in this document, is an MPLS specific manifestation of 'Directed forwarding'.

3.2. Egress ASBR Link Selection

In the topology described in Figure 2. router 'S' is facing a dilemma. Router S receives a BGP route from all of its 4 upstream routers. Using existing mechanism the provider owning AS1 can control the loading of its direct links *to* its ASBR1 and ASBR2, however it cannot control the load of the links beyond the ASBRs, except manually tweaking the eBGP import policy and filtering out a certain prefix. It would be more desirable to have visibility of all four BGP paths and be able to control the loading of those four paths using Weighted ECMP. Note that the computation of the 'Weight' percentage and the component doing this computation (Router embedded or SDN) is outside the scope of this document.

If all the ASes would be under one common administrative control then the network operator could deploy a forwarding hierarchy by using [RFC3107] to learn about the remote-AS BGP nexthop addresses and associated labels. An ingress router 'S' would then stack the transport label to its local egress ASBR and the remote ASBR supplied label. In reality it is hard to convince a peering AS to deploy another protocol just in order to easier control the egress load on the WAN links for the ingress AS.

A 'strict forwarding' paradigm would solve this problem: An Egress ASBR (e.g. ASBR 1 and 2) allocates a strict forwarding label toward all of its peering ASes and advertises it into its local IGP. The forwarding state of all those labels is to POP off the label and forward to the respective interface. The ingress router 'S' then builds a MPLS label stack by combining its local transport label to ASBR1 or ASBR2 with the IGP learned label pointing to the remote-AS ASBR.

```
           -------------traffic-flow--------->
           <-----------signaling-flow---------

                            :
                            :      AS3
                            :   +-------+
           AS1             _:___+ ASBR3 |
                          / :   +-------+
                 +-------+  :
                 | ASBR1 |  :      AS4
                 +-------+  :   +-------+
                /         \_:___+ ASBR4 |
               /            :   +-------+
              /             :
       +-----+              :
       |  S  |              :
       +-----+              :      AS5
              \             :   +-------+
               \           _:___+ ASBR5 |
                \         / :   +-------+
                 +-------+  :
                 | ASBR2 |  :      AS6
                 +-------+  :   +-------+
                          \_:___+ ASBR6 |
                            :   +-------+
                            :
```

Figure 2: Egress ASBR Link selection

ASBR {1,2} may want to periodically check the liveliness state to the endpoint of the label (ASBR {3,4,5,6}) which they are advertising. BFD Echo mode [RFC5880] is suitable technology to ensure liveliness state of undirectional links.

3.3. Tail end protection of BGP service routes

[I-D.minto-2547-egress-node-fast-protection] describes how PE routers advertising their labeled routes could get protected from node-

failures. This is a local repair technology being dependent upon successful construction of a LFA path from any PLR to the 'protector PE' in a network.

```
                              >>>>>>>>CTX-label>>>>>>
                             //                      \\
                            //                        \\
+------+   +------+   +------+   +------+   +------+   +------+
| CE1  +---+PEingr+---+PEprot+---+  P   +---+ PLR  +-X-+PEegr +
+------+   +------+   +------+   +------+   +------+   +------+
              \\              \                        /  //
                >>>>>>>>>>>>>>>>>>>>>>primary-LSP>>>>>>>>
                                \                   /
                                 \    +------+     /
                                  \___+  CE2 +____/
                                      +------+
```

Figure 3: Backup Context advertisement

Assume a primary LDP LSP from the 'ingress PE' router to the 'egress PE' router. Now consider the FRR calculation from the 'PLR' router if its direct link to the 'egress PE' router fails (X) or the entire 'egress PE' goes down. The 'PLR' cannot find a LFA path to local-repair the traffic to the 'protector PE'. This is because the 'protector PE' router has not yet converged, and hence would want to forward the traffic to the original PE egress router, such that a temporal forwarding loop would be established.

Using IGP advertisement of MPLS Labels the 'protector PE' router can advertise a Label which identifies backup traffic such that arriving traffic, can be forwarded using a context specific forwarding table, rather than the main LSP transit table. The advertised context label is a unidirectional pointer to the 'egress PE' router. The LFA calculation of the PLR gets augmented such that it considers advertised labels pointing to the original tail-end of the LSP. The network learns thereby an egress LSP point which is is as good as the original egress LSP point.

3.4. Explicit Path Routing through Label Stacking

IGP advertised strict forwarding labels can be utilized for constructing simple EROs via virtue of the MPLS label stack. In a classical traffic engineering problem (Figure 4) is illustrated. The best IGP path between {S,D} is {S, R3, R4, D}. Unfortunately this path is congested. It turns out that the links {S, R1}, {R1, R4} and {R2, R4} do have some spare capacity. In the past a C-SPF

calculation would have passed the ERO {S, R1, R4, R2, D} down to RSVP for signaling. One of the features that RSVP provides, is that it keeps track of all the reservations over a particular link, enabling bandwidth reservations of all ingress/egress pairs in a network. What is a feature for bandwidth reservations, may become a scaling harm, as the RSVP signaled paths may not be shared with other nodes in the network. This is a use case for constructing explicit routed paths, without the need to neither track per LSP control-plane state for each link, nor to program per LSP forwarding state.

```
                 +----+         +----+
                 | R1 +---------+ R2 |
                 +----+    2    +----+
                /      \           |  \
               / 2      \          |   \ 2
              /          \         |    \
       +-----+            \        |     +-----+
       |  S  |             \ 5     | 5   |  D  |
       +-----+              \      |     +-----+
              \              \     |    /
               \ 1            \    |   / 1
                \              \   |  /
                 +----+         +----+
                 | R3 +---------+ R4 |
                 +----+    1    +----+
```

Figure 4: Explicit Routing using Label stacking

Consider now every router along the path does advertise a strict forwarding label for its direct neighbor. Router S could now construct a couple of paths for avoiding the hot links without explicitly signaling them.

o {S, R1, R2, D}

o {S, R1, R4, D}

o {S, R1, R4, R2, D}

Note that not every hop in the ERO needs to be unique label in the label stack. This is undesired as existing forwarding hardware technology has got upper limits how much labels can get pushed on the label stack. In fact an existing tunnel (for example LDP tunnel {S, R1, R2} can be reused for certain path segments.

3.5. Link and Node Protection LSPs

In a network that is utilizing IGP advertised labels, it is still critical to perform fast restoration, with packet forwarding restoration times that are comparable or better than those of RSVP Fast Re-Route (FRR) [RFC4090]. In a classic link failure scenario (Figure 5) is illustrated. The best IGP path between {S,D} is {S, R3, R4, D}. When the directly adjacent link between R3 to R4 experiences a failure, (e.g.: fiber cut), the length of time to restore packet forwarding, from S to D, is dependent on several factors: propagation delay during forwarding of new Link State PDU's; artificial delays that may be introduced during the flooding of Link State PDU's (i.e.: LSP/LSA generation intervals, LSA/LSP transmit and retransmit pacing); artificial delays that may be introduced during CSPF computations (i.e.: SPF throttling) and, finally, time necessary to program new label forwarding entries in hardware. The overall length of IGP convergence time, in particular due to artificial delays introduced by various IGP timers that could have been manipulated by operators, will be substantially worse than those observed in networks who have deployed RSVP Fast Re-Route for Link and/or Node Protection.

In those networks that use RSVP FRR, there are pre-established Bypass LSP's to immediately restore packet forwarding on an alternate path, until a later time when a head-end LSR is able to signal a new LSP that is routed around the failure. In the below example, an RSVP FRR Bypass LSP may be pre-established along {R3, R1, R2, R4} to provide Link Protection of the R3 to R4 link. When that link fails, R3 will immediately start forwarding traffic along the {R3, R1, R2, R4} Bypass LSP while simultaneously signaling in the Control Plane to the Head-End LSR, S, that the R3 to R4 link has failed. This allows time for S to run CSPF to calculate a new, optimal forwarding path around the link failure; signal a new LSP through intermediate LSRs; and, finally, S may perform "make-before-break" to start forwarding traffic on the new LSP.

```
                 +----+         +----+
                 | R1 +---------+ R2 |
                 +----+    2    +----+
                /   |              |  \
               / 2  |              |   \ 2
              /     |              |    \
      +-----+       |              |     +-----+
      |  S  |       | 4            | 5   |  D  |
      +-----+       |              |     +-----+
              \     |              |    /
               \ 1  |              |   / 1
                \   |              |  /
                 +----+         +----+
                 | R3 +---------+ R4 |
                 +----+    1    +----+
```

Figure 5: Protection LSPs using Label stacking

A method is required achieve fast restoration, immediately after a node or link failure, in a network utilizing IGP labels. One method to achieve this goal is discussed in Section 3.1, Increase LFA backup coverage using 'Directed Forwarding'. In this case, R2 would need to advertise a FEC for D with a Directed Forwarding label, 200, for its output interface to D. In addition, R1 would also need to advertise a Directed Forwarding Label, 100, to R2 out its directly attached interface: R1 to R2. At this point, R3 would compute a secondary path to D perhaps using the path {R3, R1, R2, D}. R3 could then pre-program its LFIB with a primary LSP from R3 to R4 and simultaneously pre-program its LFIB with a secondary, fast-restoration path {R3, R1, R2, D}. In this scenario, when a link failure of R3 to R4 occurs, R3 would perform the appropriate label operations to restore packet forwarding along the alternate path: {R3, R1, R2, D}. Specifically, R3 would swap 1 label, (the received label from S would be swapped with 200), push one label, (the DF label advertised by R1 to forward the packet out its directly connected link to R2). This would result in the following label stack as the packet is being transmitted by R3 out to R1: {100, 200}.

3.6. Stitching MPLS Label Switched Path Segments

One of the shortcomings of existing traffic-engineering solutions is that existing label switched paths cannot get advertised and shared by many ingress routers in the network. In the example network (Figure 6) a LSP with an ERO of {R4, R2, R6} has been established in order to utilize two unused north / south links. The only way to attract traffic to that LSP is to advertise the LSP as a forwarding adjacency. This causes loss of the original path information which might be interesting for a potential router which might wants to use

this LSP for backup purposes. A computing router would need to have all underlying fate-sharing and bandwidth utilization information.

```
                 +----+         +----+         +----+
                 | R1 +---------+ R2 +---------+ R5 |
                 +----+    2    +----+    2    +----+
                /      \           |  \              \
               / 2      \          |   \              \ 2
              /          \         |    \              \
        +----+            \        |     \              +----+
        | S  |             \ 5     | 5    \ 5           | D  |
        -----+              \      |       \            +----+
              \              \     |        \          /
               \ 1            \    |         \        / 1
                \              \   |          \      /
                 +----+         +----+         +----+
                 | R3 +---------+ R4 |---------+ R6 |
                 +----+    1    +----+    1    +----+
```

Figure 6: Advertising path segments

The IGP on R4 can now advertise the LSP segment by advertising its ingress label and optionally pass the original ERO, such that any upstream router can do their fate-sharing computations. Potential ingress routers now can use this LSP as a segment of the overall LSP. Furthermore ingress routers can combine label advertisements from different routers along the path. For example router S could stacks its LDP path to R2 {S, R1, R2} plus the IGP learned RSVP LSP {R4, R5, R6} plus a strict forwarding label {R6, D}.

3.7. T-LDP replacement for infrastructure labels

Consider Figure 7. There is a LSP {S, R1, R2, D} which seeks link-protection against failure of the {R1, R2} link using R-LFA.

```
                 +----+         +----+
                 | R1 +----X----+ R2 |
                 +----+    1    +----+
                /   >>>>LSP-A->D>>>>  \
               / 1 //              \\ \ 1
              /   //                \\ \
       +-----+   //                  \> +-----+
       |  S  |                          |  D  |
       +-----+                          +-----+
              \                          /
               \ 2                      / 4
                \                      /
                 +----+         +----+
                 | R3 +---------+ R4 |
                 +----+    3    +----+
```

Figure 7: Avoidance of T-LDP for obtaining infrastructure labels

The Remote LFA Calculations results in the following Node sets.

o Extended P set: {R4}

o Q set: {R2, D, R4}

o PQ set: {R4}

The PLR router (R1) needs to obtain the label-bindings from R4 towards the final destination D in order to push the two LSPs {R1, S, R3, R4} and {R4, D}. State of the art is to establish a targeted LDP session between PLR (R1) and the R-LFA Neighbor (R4). It would be desirable to avoid dynamic bringup of T-LDP sessions. Rather the IGP should supply the corresponding Label Bindings. Furthermore it would be desirable to apply some form of message compression, such that (unlike T-LDP) not per-FEC label bindings need to be exchanged. Applying Label Block style encoding [RFC4761] would be a suitable technology to compress the messaging overhead.

4. Acknowledgements

Many thanks to Yakov Rekhter, Ina Minei, Stephane Likowski and Bruno Decraene for their useful comments.

5. IANA Considerations

This memo includes no request to IANA.

6. Security Considerations

This document does not introduce any change in terms of IGP security. It simply proposes to flood existing information gathered from other protocols via the IGP.

7. References

7.1. Normative References

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

[RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.

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

[RFC4761] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.

[RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.

[RFC5151] Farrel, A., Ayyangar, A., and JP. Vasseur, "Inter-Domain MPLS and GMPLS Traffic Engineering -- Resource Reservation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 5151, February 2008.

[RFC5286] Atlas, A. and A. Zinin, "Basic Specification for IP Fast Reroute: Loop-Free Alternates", RFC 5286, September 2008.

[RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, June 2010.

[RFC6571] Filsfils, C., Francois, P., Shand, M., Decraene, B., Uttaro, J., Leymann, N., and M. Horneffer, "Loop-Free Alternate (LFA) Applicability in Service Provider (SP) Networks", RFC 6571, June 2012.

7.2. Informative References

[I-D.bryant-ipfrr-tunnels]
Bryant, S., Filsfils, C., Previdi, S., and M. Shand, "IP Fast Reroute using tunnels", draft-bryant-ipfrr-tunnels-03

(work in progress), November 2007.

[I-D.ietf-rtgwg-remote-lfa]
Bryant, S., Filsfils, C., Previdi, S., Shand, M., and S. Ning, "Remote LFA FRR", draft-ietf-rtgwg-remote-lfa-01 (work in progress), December 2012.

[I-D.minto-2547-egress-node-fast-protection]
Jeganathan, J. and H. Gredler, "2547 egress PE Fast Failure Protection", draft-minto-2547-egress-node-fast-protection-01 (work in progress), October 2012.

[NNHOP] Chen, E., Shen, N., and A. Tian, "Discovering LDP Next-Nexthop Labels", November 2005, <http://tools.ietf.org/html/draft-shen-mpls-ldp-nnhop-label-02>.

[RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.

Authors' Addresses

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

Email: hannes@juniper.net

Shane Amante
Level 3 Communications, Inc.
1025 Eldorado Blvd
Broomfield, CO 80021
US

Email: shane@level3.net

Tom Scholl
Amazon
Seattle, WA
US

Email: tscholl@amazon.com

Luay Jalil
Verizon
1201 E Arapaho Rd.
Richardson, TX 75081
US

Email: luay.jalil@verizon.com