          LDP Extension for Inter-Area Label Switched Paths (LSPs)

Status of This Memo

   This document specifies an Internet standards track protocol for the
   Internet community, and requests discussion and suggestions for
   improvements. Please refer to the current edition of the "Internet
   Official Protocol Standards" (STD 1) for the standardization state
   and status of this protocol. Distribution of this memo is unlimited.

Abstract

   To facilitate the establishment of Label Switched Paths (LSPs) that
   would span multiple IGP areas in a given Autonomous System (AS), this
   document describes a new optional Longest-Match Label Mapping
   Procedure for the Label Distribution Protocol (LDP).

   This procedure allows the use of a label if the Forwarding
   Equivalence Class (FEC) Element matches an entry in the Routing
   Information Base (RIB). Matching is defined by an IP longest-match
   search and does not mandate an exact match.

Table of Contents

1. Introduction

    Link state Interior Gateway Protocols (IGPs) such as OSPF [OSPFv2]
    and IS-IS [IS-IS] allow the partition of an autonomous system into
    areas or levels so as to increase routing scalability within a
    routing domain.

    However, [LDP] recommends that the IP address of the FEC Element
    should *exactly* match an entry in the IP Routing Information Base
    (RIB). According to [LDP], section 3.5.7.1 ("Label Mapping Messages
    Procedures"):

        An LSR [Label Switching Router] receiving a Label Mapping message
        from a downstream LSR for a Prefix SHOULD NOT use the label for
        forwarding unless its routing table contains an entry that exactly
        matches the FEC Element.

    Therefore, MPLS LSPs between Label Edge Routers (LERs) in different
    areas/levels are not set up unless the specific (e.g., /32 for IPv4)
    loopback addresses of all the LERs are redistributed across all
    areas.

    The problem statement is discussed in section 4. Then, in section 5
    we extend the Label Mapping Procedure defined in [LDP] so as to
    support the setup of contiguous inter-area LSPs while maintaining IP
    prefix aggregation on the ABRs. This consists of allowing for
    longest-match-based Label Mapping.

2. Conventions Used in This Document

    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].

3. Terminology

    IGP Area: OSPF Area or IS-IS level

    ABR: OSPF Area Border Router or IS-IS L1/L2 router

    LSP: Label Switched Path

    Intra-area LSP: LSP that does not traverse any IGP area boundary.

    Inter-area LSP: LSP that traverses at least one IGP area boundary.

4. Problem Statement

   Provider-based MPLS (Multiprotocol Label Switching) networks are
   expanding with the success of Layer 3 Virtual Private Networks
   [L3-VPN] and the new deployments of Layer 2 VPNs ([VPLS-BGP],
   [VPLS-LDP]). Service providers' MPLS backbones are significantly
   growing both in terms of density with the addition of Provider Edge
   (PE) routers to connect new customers and in terms of footprint as
   traditional Layer 2 aggregation networks may be replaced by IP/MPLS
   networks. As a consequence many providers need to introduce IGP
   areas. Inter-area LSPs (that is, LSPs that traverse at least two IGP
   areas) are required to ensure MPLS connectivity between PEs located
   in distinct IGP areas.

   To set up the required MPLS LSPs between PEs in different IGP areas,
   service providers currently have three solutions: 1) LDP with IGP
   route leaking, 2) BGP [MPLS-BGP] over LDP with MPLS hierarchy, and 3)
   inter-area RSVP-TE (Resource Reservation Protocol-Traffic Engineering
   [RSVP-TE]).

   IGP route leaking consists of redistributing all specific PE loopback
   addresses across area boundaries. As a result, LDP finds in the RIB
   an exact match for its FEC and sets up the LSP. As a consequence,
   the potential benefits that a multi-area domain may yield are
   significantly diminished since a lot of addresses have to be
   redistributed by ABRs, and the number of IP entries in the IGP Link
   State Database (LSDB), RIB, and Forwarding Information Base (FIB)
   maintained by every LSR of the domain (whatever the area/level it
   belongs to) cannot be minimized.

   Service providers may also set up these inter-area LSPs by using MPLS
   hierarchy with BGP [MPLS-BGP] as a label distribution protocol
   between areas. The BGP next hop would typically be the ABRs, and the
   BGP-created LSPs would be nested within intra-area LSPs set up by LDP
   between PEs and ABRs and between ABRs.

   This solution is not adequate for service providers which don't want
   to run BGP on their provider routers as it requires BGP on all ABRs.
   In addition, MPLS hierarchy does not allow locally protecting the LSP
   against ABR failures (IP/LDP Fast Reroute), and hence ensuring sub-
   50ms recovery upon ABR failure. The resulting convergence time may
   not be acceptable for stringent Service Level Agreements (SLAs)
   required for voice or mission-critical applications. Finally, this
   solution requires a significant migration effort for service
   providers that started with LDP and IGP route leaking to quickly set
   up the first inter-area LSPs.

Service providers may also set up these inter-area LSPs by using
inter-area RSVP-TE [RSVP-TE]. This is a relevant solution when RSVP-
TE is already used for setting up intra-area LSPs, and inter-area
traffic engineering features are required. In return, this is not a
desired solution when LDP is already used for setting up intra-area
LSPs, and inter-area traffic engineering features are not required.

To avoid the above drawbacks, there is a need for an LDP-based
solution that allows setting up contiguous inter-area LSPs while
avoiding leaking of specific PE loopback addresses across area
boundaries, thereby keeping all the benefits of IGP hierarchy.

In that context, this document defines a new LDP Label Mapping
Procedure so as to support the setup of contiguous inter-area LSPs
while maintaining IP prefix aggregation on the ABRs. This procedure
is similar to the one defined in [LDP] but performs an IP longest
match when searching the FEC element in the RIB.

5. Longest-Match Label Mapping Message Procedure

This document defines a new Label Mapping Procedure for [LDP]. It is
applicable to IPv4 and IPv6 prefix FEC elements (address families 1
and 2 as per the "Address Family Numbers" registry on the IANA site).
It SHOULD be possible to activate/deactivate this procedure by
configuration, and it SHOULD be deactivated by default. It MAY be
possible to activate it on a per-prefix basis.

With this new Longest-Match Label Mapping Procedure, an LSR receiving
a Label Mapping message from a neighbor LSR for a Prefix Address FEC
Element FEC1 SHOULD use the label for MPLS forwarding if its routing
table contains an entry that matches the FEC Element FEC1 and the
advertising LSR is a next hop to reach FEC1. If so, it SHOULD
advertise the received FEC Element FEC1 and a label to its LDP peers.

By "matching FEC Element", one should understand an IP longest match.
That is, either the LDP FEC element exactly matches an entry in the
IP RIB or the FEC element is a subset of an IP RIB entry. There is
no match for other cases (i.e., if the FEC element is a superset of a
RIB entry, it is not considered a match).

Note that LDP re-advertises to its peers the specific FEC element
FEC1, and not the aggregated prefix found in the IP RIB during the
longest-match search.

Note that with this Longest-Match Label Mapping Procedure, each LSP
established by LDP still strictly follows the shortest path(s)
defined by the IGP.

FECs selected by this Longest-Match Label Mapping Procedure are
distributed in an ordered way. In case of LER failure, the removal
of reachability to the FEC occurs using LDP ordered label
distribution mode procedures. As defined in [LDP] in section A.1.5,
the FEC will be removed in an ordered way through the propagation of
Label Withdraw messages. The use of this (un)reachability
information by application layers using this MPLS LSP (e.g.,
[MP-BGP]) is outside the scope of this document.

As per [LDP], LDP already has some interactions with the RIB. In
particular, it needs to be aware of the following events:

   - prefix up when a new IP prefix appears in the RIB,

   - prefix down when an existing IP prefix disappears,

   - next-hop change when an existing IP prefix has a new next hop
     following a routing change.

With this Longest-Match Label Mapping Message Procedure, multiple
FECs may be concerned by a single RIB prefix change. The LSR MUST
check all the FECs that are a subset of this RIB prefix. So, some
LDP reactions following a RIB event are changed:

   - When a new prefix appears in the RIB, the LSR MUST check if this
     prefix is a better match for some existing FECs. For example,
     the FEC elements 192.0.2.1/32 and 192.0.2.2/32 used the IP RIB
     entry 192.0.2.0/24, and a new more specific IP RIB entry
     192.0.2.0/26 appears. This may result in changing the LSR used
     as next hop and hence the Next Hop Label Forwarding Entry
     (NHLFE) for this FEC.

   - When a prefix disappears in the RIB, the LSR MUST check all FEC
     elements that are using this RIB prefix as best match. For each
     FEC, if another RIB prefix is found as best match, LDP MUST use
     it. This may result in changing the LSR used as next hop and
     hence the NHLFE for this FEC. Otherwise, the LSR MUST remove
     the FEC binding and send a Label Withdraw message.

   - When the next hop of a RIB prefix changes, the LSR MUST change
     the NHLFE of all the FEC elements using this prefix.

Future work may define new management objects to the MPLS LDP MIB
modules [LDP-MIB] to activate/deactivate this Longest-Match Label
Mapping Message Procedure, possibly on a per-prefix basis.

6. Application Examples

6.1. Inter-Area LSPs

   Consider the following example of an autonomous system with one
   backbone area and two edge areas:

```
                         Area "B"

                  Level 2 / Backbone area


               +--------------------------+
    Area "A" |                            |  Area "C"
             |                            |
    Level 1  |                            |  Level 1 / area
             |            P1              |
  +---------+|                   +-------------+
  |         |            P2      |          PE1 | 192.0.2.1/32
  |         |                    |              |
  |PE4    ABR2                 ABR1          PE2 | 192.0.2.2/32
  |         |            P3      |              |
  |         |                    |          PE3 | 192.0.2.3/32
  +---------+|                   +-------------+
             |                    |
             +--------------------------+
```
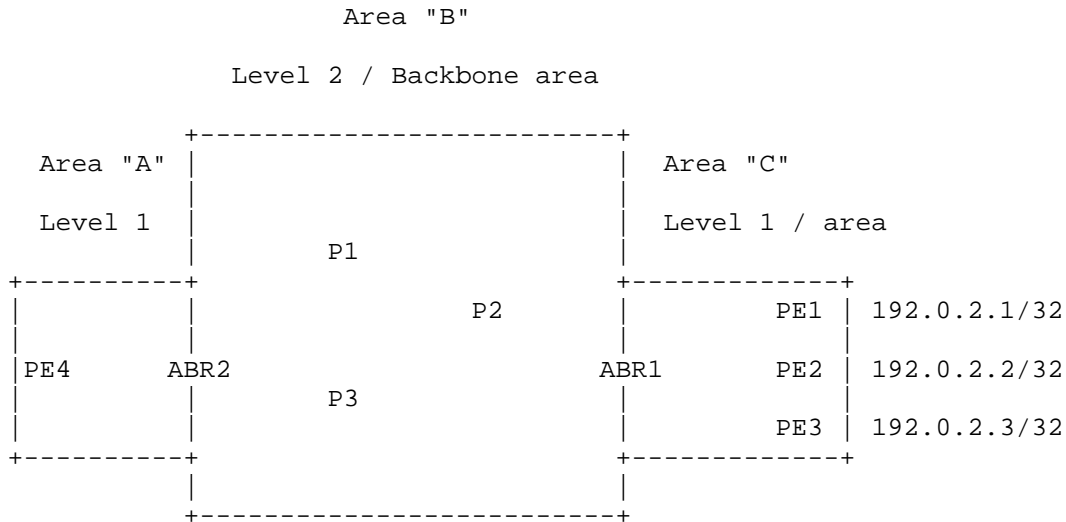
              Figure 1: An IGP domain with two areas
                   attached to the Backbone Area.

   Note that this applies equally to IS-IS and OSPF. An ABR refers here
   either to an OSPF ABR or to an IS-IS L1/L2 node.

   All routers are MPLS enabled, and MPLS connectivity (i.e., an LSP) is
   required between all PE routers.

   In the "egress" area "C", the records available are:

```
IGP RIB                      LDP FEC elements:
  192.0.2.1/32                   192.0.2.1/32
  192.0.2.2/32                   192.0.2.2/32
  192.0.2.3/32                   192.0.2.3/32
```

   The area border router ABR1 advertises in the backbone area:
      - the aggregated IP prefix 192.0.2.0/26 in the IGP
      - all the specific IP FEC elements (/32) in LDP

   In the "backbone" area "B", the records available are:

   IGP RIB   LDP FEC elements:
     192.0.2.0/26   192.0.2.1/32
                                     192.0.2.2/32
                                     192.0.2.3/32

   The area border router ABR2 advertises in the area "A":
      - an aggregated IP prefix 192.0.2.0/24 in the IGP
      - all the individual IP FEC elements (/32) in LDP

   In the "ingress" area "A", the records available are:

   IGP RIB   LDP FEC elements:
     192.0.2.0/24   192.0.2.1/32
                                     192.0.2.2/32
                                     192.0.2.3/32

   In this situation, one LSP is established between the ingress PE4 and
   every egress PE of area C while maintaining IP prefix aggregation on
   the ABRs.

6.2. Use of Static Routes

   Consider the following example where a LER is dual-connected to two
   LSRs:

                         +--------LSR1----
                         | |
                        LER |
                         | |
                         +-------LSR2----
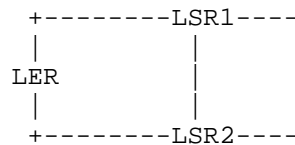
              Figure 2: LER dual-connected to two LSRs.

   In some situations, especially on the edge of the network, it is
   valid to use static IP routes between the LER and the two LSRs. If
   necessary, the Bidirectional Forwarding Detection protocol [BFD] can
   be used to quickly detect loss of connectivity.

   The LDP specification defined in [LDP] would require on the ingress
   LER the configuration and the maintenance of one IP route per egress
   LER and per outgoing interface.

   The Longest-Match Label Mapping Procedure described in this document
   only requires one IP route per outgoing interface.

7. Caveats for Deployment

7.1. Deployment Considerations

   LSRs compliant with this document are backward compatible with LSRs
   that comply with [LDP].

   For the successful establishment of end-to-end MPLS LSPs whose FECs
   are aggregated in the RIB, this specification must be implemented on
   all LSRs in all areas where IP aggregation is used. If an LSR on the
   path does not support this procedure, then the LSP initiated on the
   egress LSR stops at this non-compliant LSR. There are no other
   adverse effects.

   This extension can be deployed incrementally:

      - It can be deployed on a per-area or per-routing-domain basis and
        does not necessarily require an AS-wide deployment. For
        example, if all specific IP prefixes are leaked in the IGP
        backbone area and only stub areas use IP aggregation, LSRs in
        the backbone area don't need to be compliant with this document.

      - Within each routing area, LSRs can be upgraded independently, at
        any time, in any order, and without service disruption. During
        deployment, if those LSPs are already used, one should only make
        sure that ABRs keep advertising the specific IP prefixes in the
        IGP until all LSRs of this area are successfully upgraded.
        Then, the ABRs can advertise the aggregated prefix only and stop
        advertising the specific ones.

   A service provider currently leaking specific LER loopback addresses
   in the IGP and considering performing IP aggregation on ABR should be
   aware that this may result in suboptimal routing as discussed in
   [RFC2966].

7.2. Routing Convergence Time Considerations

   IP and MPLS traffic restoration time is based on two factors: the
   Shortest Path First (SPF) calculation in the control plane and
   Forwarding Information Base (FIB) / Label FIB (LFIB) update time in
   the forwarding plane. The SPF calculation scales $O(N*Log(N))$ where N
   is the number of Nodes. The FIB/LFIB update scales $O(P)$ where P is
   the number of modified prefixes. Currently, with most routers
   implementations, the FIB/LFIB update is the dominant component
   [IGP-CONV], and therefore the bottleneck that should be addressed in
   priority. The solution documented in this document reduces the link
   state database size in the control plane and the number of FIB
   entries in the forwarding plane. As such, it solves the scaling of

pure IP routers sharing the IGP with MPLS routers. However, it does
not decrease the number of LFIB entries so is not sufficient to solve
the scaling of MPLS routers. For this, an additional mechanism is
required (e.g., introducing some MPLS hierarchy in LDP). This is out
of scope for this document.

Compared to [LDP], for all failures except LER failure (i.e., links,
provider routers, and ABRs), the failure notification and the
convergence is unchanged. For LER failure, given that the IGP
aggregates IP routes on ABRs and no longer advertises specific
prefixes, the control plane and more specifically the routing
convergence behavior of protocols (e.g., [MP-BGP]) or applications
(e.g., [L3-VPN]) may be changed in case of failure of the egress LER
node. For protocols and applications which need to track egress LER
availability, several solutions can be used, for example:

- Rely on the LDP ordered label distribution control mode -- as
  defined in [LDP] -- to know the availability of the LSP toward the
  egress LER. The egress to ingress propagation time of that
  unreachability information is expected to be comparable to the IGP
  (but this may be implementation dependent).

- Advertise LER reachability in the IGP for the purpose of the
  control plane in a way that does not create IP FIB entries in the
  forwarding plane.

8. Security Considerations

   The Longest-Match Label Mapping procedure described in this
   document does not introduce any change as far as the Security
   Considerations section of [LDP] is concerned.

9. References

9.1. Normative References

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

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

9.2. Informative References

   [L3-VPN]   Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private
                  Networks (VPNs)", RFC 4364, February 2006.

[MP-BGP]   Bates, T., Chandra, R., Katz, D., and Y. Rekhter,
                  "Multiprotocol Extensions for BGP-4", RFC 4760, January
                  2007.

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

[IS-IS]   Callon, R., "Use of OSI IS-IS for routing in TCP/IP and
                  dual environments", RFC 1195, December 1990.

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

[VPLS-LDP]   Lasserre, M., Ed., and V. Kompella, Ed., "Virtual
                  Private LAN Service (VPLS) Using Label Distribution
                  Protocol (LDP) Signaling", RFC 4762, January 2007.

[RFC2966]   Li, T., Przygienda, T., and H. Smit, "Domain-wide
                  Prefix Distribution with Two-Level IS-IS", RFC 2966,
                  October 2000.

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

[LDP-MIB]   Cucchiara, J., Sjostrand, H., and J. Luciani,
                  "Definitions of Managed Objects for the Multiprotocol
                  Label Switching (MPLS), Label Distribution Protocol
                  (LDP)", RFC 3815, June 2004.

[BFD]   Katz, D. and D. Ward, "Bidirectional Forwarding
                  Detection", Work in Progress, March 2008.

[IGP-CONV]   Francois, P., Filsfils, C., and Evans, J., "Achieving
                  sub-second IGP convergence in large IP networks". ACM
                  SIGCOMM Computer Communications Review, July 2005.

[OSPFv2]   Moy, J., "OSPF Version 2", STD 54, RFC 2328, April
                  1998.

10. Acknowledgments

   The authors would like to thank Yakov Rekhter, Stefano Previdi, Vach
   Kompella, Bob Thomas, Clarence Filsfils, Kireeti Kompella, Luca
   Martini, Sina Mirtorabi, Dave McDysan, Benoit Fondeviole, Gilles
   Bourdon, and Christian Jacquenet for the useful discussions on this
   subject, their reviews, and comments.

Authors' Addresses

   Bruno Decraene
   France Telecom
   38 rue du General Leclerc
   92794 Issy Moulineaux cedex 9
   France

   EMail: bruno.decraene@orange-ftgroup.com

   Jean-Louis Le Roux
   France Telecom
   2, avenue Pierre-Marzin
   22307 Lannion Cedex
   France

   EMail: jeanlouis.leroux@orange-ftgroup.com

   Ina Minei
   Juniper Networks
   1194 N. Mathilda Ave.
   Sunnyvale, CA 94089

   EMail: ina@juniper.net