Network Working Group
Internet-Draft
Intended status: Experimental
Expires: January 7, 2010

O. Bonaventure
P. Francois
D. Saucez
UCLouvain
July 6, 2009

Preserving the reachability of LISP ETRs in case of failures
draft-bonaventure-lisp-preserve-00.txt

Status of this Memo

This Internet-Draft will expire on January 7, 2010.

Copyright Notice

Abstract

Maintaining reachability of an EID prefix despite the failures of ETRs is a key concern in the LISP architecture. In this document, we first analyse this problem in comparison with traditional routing protocols. Then, we explain how Internet Service Providers could offer a service that preserves the reachability of the LISP ETRs of their customers in case of failures.

Table of Contents

1. Introduction

Measurements performed in ISP networks indicate that link and node failures are frequent events [FAILURES][BGPFRR]. Fortunately, most of these failures have a short duration. However, the more and more stringent Service Level Agreements (SLAs) requested by users of IP networks have forced researchers and router vendors to develop various kinds of fast route techniques that allow a network to quickly recover after a node or link failure [RFC4090] [I-D.ietf-rtgwg-ipfrr-framework] [RECOVERY].

The Locator/Identifier Separation Protocol (LISP) [I-D.ietf-lisp] is being developed within the LISP working group of the IETF. LISP relies on two principles. First, Endpoint Identifiers (EIDs) are allocated to hosts while Routing Locators (RLOCs) are allocated to LISP Ingress/Egress Tunnel Routers (xTRs). The EIDs are not directly routable on the global Internet, only the RLOCs are routable. Second, LISP relies on map and encaps. Hosts are located on sites and are served by xTRs. When host A.1 in site A needs to send a packet to host B.2 in site B, its packet is intercepted by the Ingress Tunnel Router (ITR) that serves its site. This ITR will query a mapping system to find the RLOC of the Egress Tunnel Router (ETR) that serves EID B.2. Once the RLOC of the ETR serving B's site is known, the ITR will encapsulate the packet using the encapsulation defined in [I-D.ietf-lisp] so that it can reach B's ETR. B's ETR will decapsulate the packet and forward it to host B.

Recovery in case of failures is also one of the problems being discussed within the LISP working group. More precisely, the working group is working on techniques to verify the reachability of the destination ETRs for a given EID prefix. The current draft, [I-D.ietf-lisp], uses several locator reachability bits in the header of all data encapsulated packets to allow an ITR to indicate to a remote ETR the xTRs on the ITR's site that are known to be reachable and unreachable. For another discussion of the reachability problem, see [I-D.meyer-loc-id-implications]

This reachability problem can be better understood by comparing it with the operation of traditional routing protocols in the network shown in Figure 1. In this picture, the stars indicate domain boundaries.

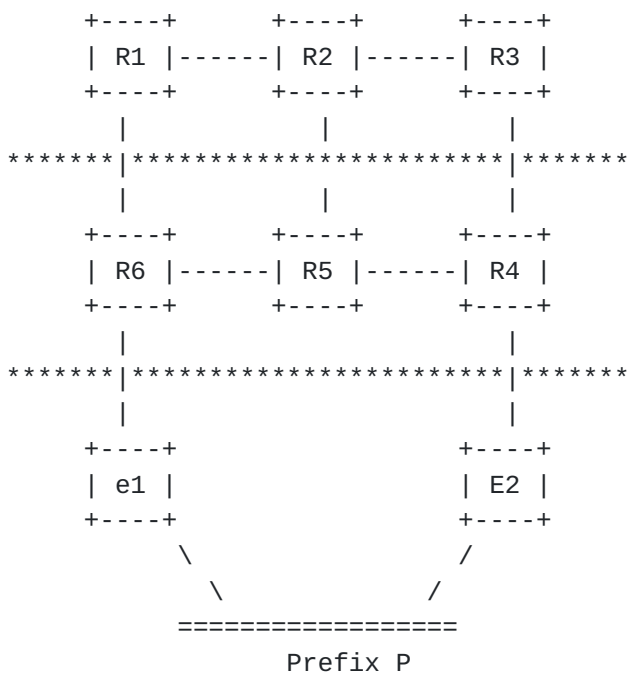

Figure 1: A simple network

Figure 1 shows a simple network with 8 routers and one LAN containing a single prefix P. With traditional routing protocols, the prefix P will be advertised by both E1 and E2 via BGP. If E1 and E2 are up, P will be reachable via both routers. If E1 (resp. E2) fails, then all the packets destined to P will be sent via E2 (resp. E1). In such a network, the reachability of P is maintained despite the failures of E1 or E2 because :

- routers E1 and E2 send messages about the reachability of P in the entire network
- all routers of the network have an entry for prefix P inside their Forwarding Information Base (FIB)

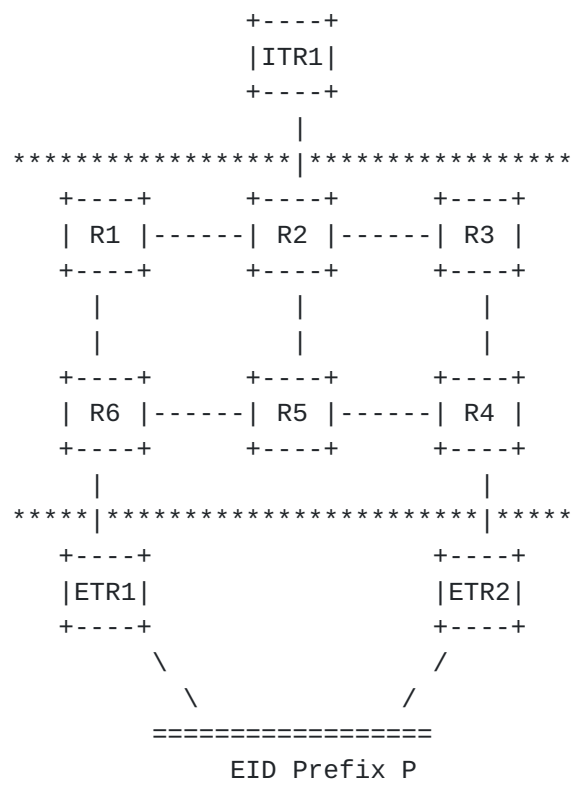

Figure 2: A simple network with LISP routers

Now, let us assume that E1 and E2 are LISP ETRs and that P is an EID prefix. We also add an ITR connected to R2 as shown in Figure 2. Since both the network of Figure 1 and of Figure 2 have the same topology, they should be able to maintain reachability even in case of failures. Unfortunately, there are several important differences :

1. the routers are managed by three different autonomous entities and different IGPs are used : one for R1-R6, another one for ETR1 and ETR2 and a third for the network that contains ITR1. Three different routing protocols are used and only aggregated RLOCs are advertised accross the boundaries represented by stars in the figure.

2. The packets sent towards EID prefix P are encapsulated in packets destined to ETR1 or ETR2. There is no entry for prefix P in the FIB or routers R1-R6. ITR1 has one entry for P inside its LISP mapping cache. Only ETR1 and ETR2 can reach directly EID prefix P.

We assume that the middle network uses an IGP to advertise the reachability of all the routers (R1-R6) and of the directly attached

customers (i.e. ITR1, ETR1 and ETR2). This is a very common design. For the routers R1-R6, ETR1, ETR2 ad ITR1 are different RLOCs and none of these routers is aware of the fact that LISP data encapsulated packets sent to ETR1 can also be sent to ETR2.

The network of Figure 2 is sufficiently redundant to preserve the reachability of EID prefix P in case of the failure of ETR1, ETR2, R6 or R4. Let us analyse how LISP would react to these four failures :

- Failure of ETR1. In this case, ETR2 can notice the failure by either having an iBGP or BFD session with ETR1 or participating in the same IGP. Once ETR2 has detected the failure of ETR1, it changes its locator reachability bits so that ITR1 is also informed and can redirect the packets destined to EID prefix P via ETR2. The time required to inform ITR1 will depend on both the local failure detection time and the current packet transmission rate between ETR2 and ITR1. This only works, of course, if traffic is bidirectionnal.

- Failure of R6. To detect such failures, since ETR1 does not participate in the ISP's IGP, it needs to use a mechanism to verify that its upstream router is alive. This can be achieved for example by having a BGP session between ETR1 and R6 possibly coupled with a fast failure detection mechanism such as BFD [I-D.ietf-bfd-base]. Once ETR1 has detected the failure of R6, it must inform ETR2. The method used to inform ETR2 is not specified by LISP, but is important from a deployment viewpoint. For example, ETR1 could withdrawing the default route learned from R6 from the site's IGP. ETR2 can then update the loc-reach bits of the LISP encapsulated packets that it sends. ITR1 will stop sending LISP data encapsulated packets to ETR1 as soon as it has received the updated loc-reach bits.

In practice, the time required to detect and recover from such failures can be longer than a round-trip-time. It would be desirable in some environments to have a shorter recovery time. Unfortunately, the classical techniques [RECOVERY] deployed in IP and MPLS networks are not directly applicable to preserve the reachability of the EIDs behind the unreachable ETR.

In this document, we first analyse several solutions based on anycast that can be used by an ISP to preserve the reachability to LISP ETRs in case and failures and discuss their advantages and drawbacks. Then, we propose a rewriting technique that can be deployed by ISPs to ensure that the EIDs of their customers remain reachable despite that some of their LISP ETRs are unreachable.

2. Using anycast to preserve reachability of EID prefixes in case of failure

A first possible approach to preserve the reachability of EID prefixes in case of link or node failures in the service provider network to which the ETR is attached is to use anycast routing. The figure below shows a simplified network using the terminology used by BGP/MPLS VPNs [RFC2547]. The ISP network contains three Provider (P) routers, 3 Provider Edge (PE) routers and two LISP ETRs. The two LISP ETRs are responsible for the same EID prefix P.

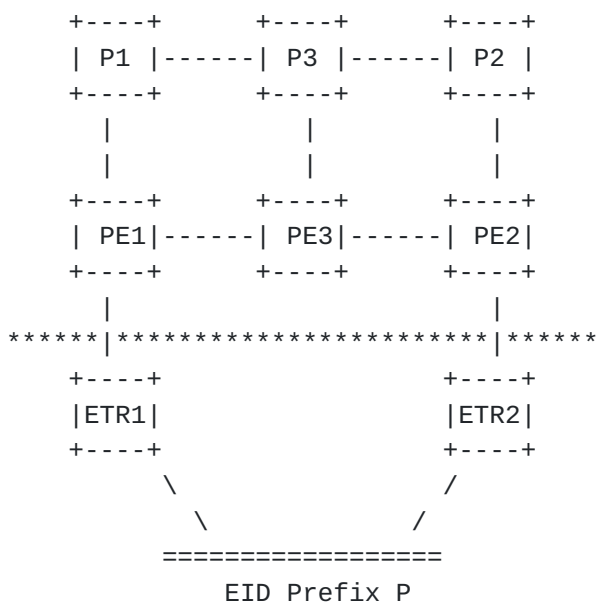

Figure 3: A simple network with two ETRs

A first solution to ensure that ETR2 remains reachable when ETR1 becomes unreachable is to use an anycast address for the RLOC used by both ETR1 and ETR2. For example, with IPv4 a single anycast /32 would be allocated to both ETR1 and ETR2. This solution clearly ensures that all LISP data encapsulated packets will reach an ETR attached to EID prefix P as long as either ETR remains reachable. However, it has several important drawbacks :

- As ETR1 and ETR2 use the same anycast address, the site cannot engineer the incoming traffic toward EID prefix p by tuning its mapping replies.

- Anycast cannot be used if ETR1 and ETR2 are attached to two different ISPs. Unfortunately, it can be expected that owners of sites will often attach their ETRs to different ISP networks to

have technical and economical redundancy. Anycast could probably be used if ETR1 and ETR2 were located in the same IGP area (often equivalent to the same POP in large ISP networks).

To allow a site to continue to engineer its incoming traffic, an alternative could be to use two anycast addresses as RLOCs for the site's ETRs. PE1 (resp. PE2) would advertise in the ISP's IGP two addresses for ETR1 (resp. ETR2) : ETR1's RLOC (resp. ETR2's RLOC) with a low IGP distance and ETR2's RLOC (resp. ETR1's RLOC) with a very high IGP distance. With those advertisements, ETR1 and ETR2 are both used when they are up. If ETR1 becomes unreachable, the provider's IGP will converge and all packets sent to its RLOC will be automatically rerouted to ETR2 which also supports the same RLOC. Unfortunately, this solution has the following drawbacks :

- It increases the size of the IGP, especially when ETR1 and ETR2 are not in the same POP/area.

- It cannot be used when ETR1 and ETR2 are attached to two different ISPs.

For these reasons, anycast cannot be considered as a technique that totally fulfills the role of preserving the reachability of multihomed EID prefixes.

3. Rewriting to preserve the reachability of EID prefixes

To preserve the reachability of EID prefixes in case of failures of either the link or the router that connects an ETR to its provider, we need to ensure that the packets destined to the RLOC of an ETR that became unreachable can be rerouted efficiently by routers in the provider's network. We consider three reference environments where our solution must be applicable :

- A network where the two ETRs are attached to the same POP of one ISP
- A network where the two ETRs are attached to different POPs of the same ISP
- A network where the two ETRs are attached to different ISPs

The more general case is the third one. In the remainder of this section, we will mainly discuss the topology shown in Figure 4.

A solution to preserve the reachability of these ETRs in case of link/router failures must be applicable to these three deployment scenarios. We consider two different types of failures :

- Failure of the link between an ETR and its PE router, such as PE1-E1 in Figure 4. From the viewpoint of the ISP network, the failure of a link between a PE and an ETR is equivalent to the failure of the ETR itself.
- Failure of the PE router to which an ETR is attached, such as PE1 in Figure 4. In this case, all the ETRs attached to the PE router become unreachable.

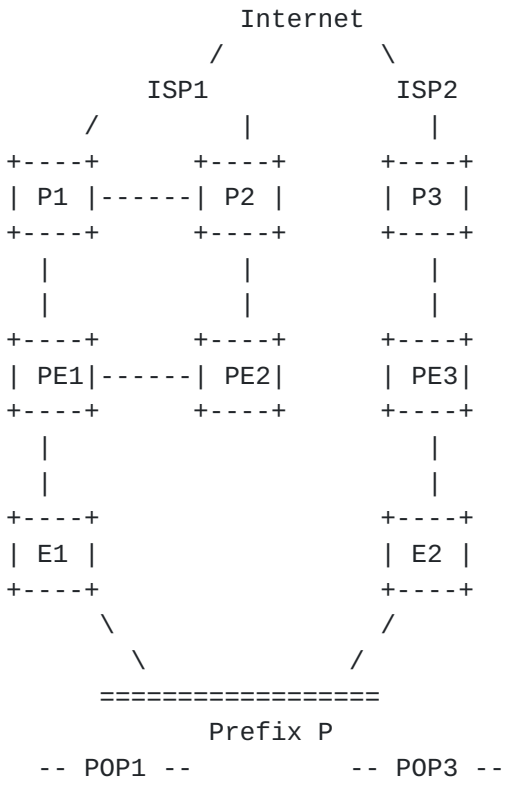

Figure 4: A network with two LISP ETRs attached to different ISPs

3.1. Rewriting interface

Our technique to preserve the reachability of EID prefixes despite link and node failures relies on a new type of virtual interface that we call a rewriting interface. Besides real physical interfaces, routers often have virtual interfaces such as tunnel interfaces. When the nexthop of a packet is a tunnel interface, this packet is encapsulated and the encapsulated packet is sent towards the tunnel destination.

A rewriting virtual interface is configured with :

o a primary address

o a (set of) alternate addresses

A rewriting interface can only be used by packets whose destination address is equal to the primary address of the rewriting interface. When such a packet is to be forwarded by the rewriting interface, its destination address is replaced by one of alternate addresses known for this interface. Of course, the IP and UDP checksums of the rewritten packets are updated. When selecting an alternate address,

the router should prefer an alternate address that it knows (e.g. based on its own routing table or thanks to other information) to be reachable. The rewritten packet is then forwarded towards its new destination.

Instead of using a rewriting interface, another solution could have been to encapsulate the packet destined to the failed address towards the alternate. However, using a second level of encapsulation would like cause MTU problems. For this reason, we chose to rewrite part of the LISP header. From an implementation viewpoint, rewriting part of a LISP header is similar to the operation performed by a Network Address Translator. Given the current interest in carrier-grade NAT, it can be expected that efficient hardware-based NAT implementations will appear.

The operation of the rewriting interface is discussed in more details in section Section 4.3.

3.2. Link and ETR failures

In this section, we describe informally the principle of our solution. The details are discussed later. To maintain reachability of EID prefix when the link between one of its ETR and the associated PE fails, we propose to install a rewriting interface on the upstream PE. Consider for example Figure 4 and that E1 is the ETR whose reachability needs to be preserved. This can be achieved as follows :

- PE1 is configured with a rewriting interface having E1's RLOC as primary address and E2's RLOC as alternate address. A static route for this rewriting interface is configured on PE1, but this route has a high administrative distance so that the route is not installed in the FIB when E1 is up.

- When the link between PE1 and E1 fails, PE1's rewriting interface is still up. Thus, PE1 continues to announce E1's RLOC as being reachable in the IGP. This ensures that packets destined to E1 still reach PE1. However, the rewriting interface replaces the physical interface as the nexthop for E1 in PE1's FIB.

- When a LISP data encapsulated packet destined to E1 arrives while E1 is unreachable, PE1 forwards this packet over its rewriting interface. This interface rewrites the destination RLOC of this LISP data encapsulated packet with E2's RLOC as destination address and the packet is forwarded to E2.

- When E1 becomes again reachable, the physical interface towards E1 replaces the rewriting interface as the nexthop for E1 in PE1's

FIB and the rewriting stops. Rewriting could also stop by removing the rewriting interface e.g. after the expiration of a timer.

It should be noted that this solution is purely local on the PE router attached to the ETR responsible for the EID prefix whose reachability must be preserved in case of failures. No additional prefix needs to be advertised in the IGP. Thus, there are no scalability issues with this solution.

3.3. PE failures

To maintain reachability of an EID prefix when the PE attached to one ETR fails, we cannot use the solution described above as the PE is not reachable anymore. To solve this problem, we introduce a rewriting PE. A rewriting PE is a PE router that is configured with a rewriting interface whose primary address is the address of an ETR attached to another PE router. The rewriting PE will usually be located in the same POP as the PE that must be protected. For example, let us consider the failure of PE1 in Figure 4 and assume that PE2 is the rewriting PE :

- PE2 is configured with one rewriting interface having :
  - E1's RLOC as primary address
  - E2's RLOC as alternate address
- E1's RLOC is advertised as an anycast address by both PE1 and PE2 that acts as a rewriting router. PE2's advertisement has a high IGP distance such that PE1's advertisement is always preferred inside the ISP network. Furthermore, the rewriting interface has a high administrative distance and thus PE2 does not install a FIB entry towards this rewriting interface.
- When PE1 becomes unreachable, the IGP converges and PE2 becomes the only router that advertises E1's RLOC. It thus receives all packets destined to E1's RLOC. These packets are rewritten by the rewriting interface and forwarded to E2's RLOC.
- When PE1 comes back, it readvertises the reachability of E1's RLOC. PE2 prefers PE1's advertisement and stops receiveing packets destined to E1's RLOC.

4. Protocol issues

In this section, we discuss in more details the protocols and mechanisms that are required to implement the solution described informally in the previous section. We first discuss how a PE can verify the reachability of ETRs. Then we discuss how a rewriting router can learn the rewriting address that it should use when an ETR becomes unreachable. Finally we explain how the RLOC of the unreachable ETR needs to be rewritten and propose a small change to the LISP header for this.

4.1. Verifying the reachability of ETRs

The first router that needs to detect the unreachability of a LISP ETR is the PE router directly connected to it. Several mechanisms can be used to detect this unreachability : physical layer information (if available), BFD or a single hop eBGP session could be established between the PE and the ETR. No prefix will be advertised by the ETR on this eBGP session, but the PE may advertise a default route or its full BGP (RLOC) routing table.

However, the rewriting PE router could also need to verify the reachability of the ETR that owns the RLOC that it will rewrite if the primary ETR becomes unreachable due to the failure of its attached PE. This is especially important when the the rewriting PE knows several alternate ETR routers. If it only knows a single alternate ETR and the primary fails, the only solution is to rewrite the packets towards the only alternate ETR. This alternate ETR can be located in the same POP, in another POP or in another ISP. Thus, the rewriting PE cannot always rely on its routing table to verify the reachability of such a distant ETR.

To allow a PE to know which of the alternate addresses for a given primary address are alive, we propose to use multihop eBGP sessions to distribute the reachability information of each ETR. Reachability information could be distributed as follows :

- Each LISP site, containing at least one EID prefix and several ETRs is allocated a unique route target.

- Each ETR has a single-hop BGP session with its attached PE router. On this eBGP session, the ETR advertises only its own RLOC with the allocated route target.

- The PE routers and the routers with rewriting interfaces are part of an iBGP mesh (e.g. based on route reflectors) where the routes received by the ETRs are distributed with their route target.

- The route reflectors of different ASes that host LISP ETRs can exchange the routes received from their ETRs by using multihop eBGP sessions.

- A rewriting router only needs to receive reachability information for alternate addresses that it supports. This can be achieved by requesting in the iBGP mesh all the routes with a list of route targets.

The next version of this document will analyse this problem in more details

4.2. Advertising the backup ETR

In the previous section, we have assumed that the PE and the rewriting router were configured with several information. Such a manual configuration is possible, but in practice it would be useful to allow some of these routers to automatically learn some of this information. For example, it would be useful for a PE router to learn automatically the backup RLOCs to be used in case of failure of one of its directly attached ETRs. This can be achieved by either :

- developing a new protocol to advertise these backup RLOCs to be rewritten

- using BGP and defining a new address family that allows BGP to carry this kind of information

- extending the Map-Request/Map-Reply and allow the PE to query the ETR for its alternate ETR

The next version of this document will analyse in more detailed the advantages and drawbacks of each of these two approaches.

4.3. Destination RLOC rewriting

Our solution rewrites the destination RLOC of LISP packets once the destination of this packet has been found unreachable. This rewriting raises several questions as discussed in the following sections.

4.3.1. Which packets should be rewritten ?

A LISP ETR will receive different types of packets and we need to define which packets should be rewritten by the rewriting router. LISP encapsulated data packets should be rewritten. However, we need to ensure that when multiple failures occur LISP encapsulated data packets do not loop between rewriting routers. This can be achieved

by reserving one bit in the LISP header, called the Deflection (D) bit. When an ITR sends a data encapsulated packet, it sets the D bit to false. When a rewriting router receives a LISP data encapsulated with the D bit set to false, it can rewrite the destination address of the packet. If the D bit is set to true, the packet must be dropped. LISP control packets, i.e. Map-Request and Map-Reply packets, do not need to be rewritten as they are targeted at the ETR itself and not at hosts behind the ETR. Non-LISP packets destined to the ETR do not need to be rewritten either.

Upon reception of packets with the D bit set, the ETR knows that the packets have been deflected by upstream routers, likely due to an upstream failure. This ETR will soon detect the failure by other means (e.g. the primary ETR stops advertising its default route in the site's IGP).

4.3.2. After a failure, for how long should packets be rewritten ?

In theory, the ITR which is sending packets to the ETR could have learned the mapping up to TTL minutes ago if TTL is the mapping lifetime. Thus, the rewriting entry should remain in the rewriting router for a duration at least equal to the lifetime of the mapping entries if we do not want to loose encapsulated packets. With a default mapping lifetime of 24hours, this duration can be large. In practice however, most of the failures have a short duration and the ETR will become reachable again well before the expiration of the lifetime of its mapping entries.

5. Security Considerations

To be written once the details of the protocols have been specified.

6. Conclusion

In this document, we have first compared the LISP reachability problem with the traditional reachability problem with routing protocols. We have then shown the drawbacks of using anycast to preserve the reachability of LISP ETRs in case of failures. Then, we have proposed to allow PE routers to rewrite the destination address of LISP encapsulated packets to preserve the reachability of the EID prefix in case of failure of one of the responsible ETRs. Further work is required to define the protocols and mechanisms that are necessary to allow ISPs to preserve the reachability of the ETRs of their customers.

7. Acknowledgements

We would like to thank Dave Meyer for his comments on the first version of this draft. This work was partially supported by a Cisco URP grant.

8. References

8.1. Normative References

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

8.2. Informative References

[BGPFRR] Bonaventure , O., Filsfils, C., and P. Francois, "Achieving Sub-50 Milliseconds Recovery Upon BGP Peering Link Failures", Conext 2005 .

[FAILURES] Markopoulou, A., Iannacone, G., Chattacharyya, S., Chuah, C., and C. Diot, "Characterization of Failures in an IP Backbone", INFOCOM 2004.

[I-D.ietf-bfd-base] Katz, D. and D. Ward, "Bidirectional Forwarding Detection", draft-ietf-bfd-base-09 (work in progress), February 2009.

[I-D.ietf-bfd-multihop] Katz, D. and D. Ward, "BFD for Multihop Paths", draft-ietf-bfd-multihop-07 (work in progress), February 2009.

[I-D.ietf-lisp] Farinacci, D., Fuller, V., Meyer, D., and D. Lewis, "Locator/ID Separation Protocol (LISP)", draft-ietf-lisp-01 (work in progress), May 2009.

[I-D.ietf-rtgwg-ipfrr-framework] Shand, M. and S. Bryant, "IP Fast Reroute Framework", draft-ietf-rtgwg-ipfrr-framework-10 (work in progress), February 2009.

[I-D.meyer-loc-id-implications] Meyer, D. and D. Lewis, "Architectural Implications of Locator/ID Separation", draft-meyer-loc-id-implications-01 (work in progress), January 2009.

[RECOVERY] Vasseur, J., Demeester, P., and M. Pickavet, "Network Recovery: Protection and Restoration of Optical, SONET-SDH, IP, and MPLS", Elsevier Science & Technology

Books 2004.

[RFC2547] Rosen, E. and Y. Rekhter, "BGP/MPLS VPNs", RFC 2547, March 1999.

[RFC4984] Meyer, D., Zhang, L., and K. Fall, "Report from the IAB Workshop on Routing and Addressing", RFC 4984, September 2007.

Authors' Addresses

Olivier Bonaventure
UCLouvain
Universite catholique de Louvain, Place Sainte Barbe 2
Louvain-la-Neuve, 1348
Belgium

Email: olivier.bonaventure@uclouvain.be
URI: http://inl.info.ucl.ac.be

Pierre Francois
UCLouvain
Universite catholique de Louvain, Place Sainte Barbe 2
Louvain-la-Neuve, 1348
Belgium

Email: pierre.francois@uclouvain.be
URI: http://inl.info.ucl.ac.be

Damien Saucez
UCLouvain
Universite catholique de Louvain, Place Sainte Barbe 2
Louvain-la-Neuve, 1348
Belgium

Email: damien.saucez@uclouvain.be
URI: http://inl.info.ucl.ac.be