

Network Working Group
Internet-Draft
Intended status: Experimental
Expires: June 23, 2014

D. Saucez
INRIA
O. Bonaventure
UCLouvain
L. Iannone
Telecom ParisTech
C. Filsfils
Cisco Systems
December 20, 2013

LISP ITR Graceful Restart
draft-saucez-lisp-itr-graceful-03.txt

Abstract

The Locator/ID Separation Protocol (LISP) is a map-and-encap mechanism to enable communications between hosts identified with their Endpoint IDentifier (EID) over the Internet where EIDs are not routable. To do so, packets toward EIDs are encapsulated in packets with routing locators (RLOCs) to form dynamic tunnels. An Ingress Tunnel Router (ITR) that encapsulates EID packets determines tunnel endpoints via mappings that associate EIDs to RLOCs. Before encapsulating a packet, the ITR queries the mapping system to obtain the mapping associated to the EID of the packet it must encapsulate. Such mapping is cached by the ITR in its local EID-to-RLOC cache for any subsequent encapsulation for the same EID. LISP is scalable because EID-to-RLOC mappings are cached on ITRs. Initially, the cache is empty and is populated progressively according to the traffic traversing the ITR. However, after an ITR is restarted, e.g., for maintenance reason, its cache is empty which means that all packets that are re-routed to the freshly restarted ITR will cause cache misses and a potentially high loss rate. In this draft, we present mechanisms to reduce the negative impact on traffic caused by the restart of an ITR in a LISP network.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of [BCP 78](#) and [BCP 79](#).

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any

Internet-Draft

LISP Graceful Restart

December 2013

time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on June 23, 2014.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to [BCP 78](#) and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1.	Introduction	2
2.	Definition of terms	3
2.1.	LISP Definition of Terms	4
3.	Problem Statement	6
4.	ITR Graceful Restart	7
5.	Security Considerations	8
6.	Conclusion	9
7.	Acknowledgments	9
8.	References	9
8.1.	Normative References	9
8.2.	Informative References	9
	Authors' Addresses	10

[1.](#) Introduction

The Locator/ID Separation Protocol (LISP) [[RFC6830](#)] relies on two principles. First, Endpoint Identifiers (EIDs) are allocated to hosts while Routing Locators (RLOCs) are allocated to LISP Ingress Tunnel Routers (ITR) and Egress Tunnel Routers (ETR). EIDs are not directly routable on the global Internet, only RLOCs are. Second, LISP relies on mapping and encapsulation. Hosts are located on sites

and are served by ITRs and ETRs. When host A.1 in site A needs to send a packet to host B.2 in site B, its packet is intercepted by an ITR that serves its site. The ITR queries a mapping system to find the RLOC of the ETR that serves EID B.2. Once the RLOC of the ETR serving B's site is known, the ITR encapsulates the packet using the

encapsulation defined in [[RFC6830](#)] so that it can reach B's ETR. B's ETR decapsulates the packet and forwards it to host B.

Packets from a LISP site are routed to their closest ITR by the mean of the routing system (e.g., IGP). In case of an ITR that just booted (either because it has just been added to the network or because it has been restarted due to maintenance) a large portion of the traffic can potentially be routed to the freshly started ITR. However, in this case, its EID-to-RLOC cache is empty. While with traditional routing, such a massive redirection has minor impact on the traffic (except for path stretch and latency), in the context of LISP, this can cause a high volume of cache misses (i.e., no EID-to-RLOC cache entry matching the destination RLOC) resulting in a high volume of dropped packets, hence, potentially leading to severe traffic disruption. Furthermore, such a high number of cache misses triggers a burst of Map-Requests that may overload the mapping system (or Map Resolvers if [[RFC6833](#)] is used).

This memo opens the question about how to perform graceful (re)start of ITRs in LISP networks. It aims at documenting the problem of ITR (re)start with the associated risk of "miss storm" and discusses EID-to-RLOC cache synchronization solutions to provide ITR graceful restart without overwhelming the mapping system and without high packet losses.

[2.](#) Definition of terms

This section introduces the definition of the main elements and terms used throughout the whole document. More specifically, hereafter the terms introduced by this document are defined, while in [Section 2.1](#) the definitions related to the LISP's architecture are provided in order to ease the read of the present document.

EID-to-RLOC cache miss storm: A sudden raise of the cache miss rate at an ITR to a level significantly higher than the rate observed at steady state on the ITR.

Map-Request storm: The side effect of a EID-to-RLOC cache miss storm, is the generation of a high number of Map-Requests, which is called a Map-Request storm.

Synchronization Set: The set of ITRs that are potentially on the path of the same traffic should have their EID-to-RLOC cache synchronized in order to avoid EID-to-RLOC cache miss storms.

ITR Restart: Generic term indicating an ITR that has just completed the bootstrap phase and resuming normal operation. It can be either an ITR that has been added to the network (hence,

actually at its first boot as part of the specific network) or an ITR actually re-booting due to various reasons such as maintenance or outage.

[2.1.](#) LISP Definition of Terms

LISP operates on two name spaces and introduces several new network elements. This section provides high-level definitions of the LISP name spaces and network elements and as such, it MUST NOT be considered as an authoritative source. The reference to the authoritative document for each term is included in every term description.

Ingress Tunnel Router (ITR) [[RFC6830](#)]: An ITR is a router that resides in a LISP site. Packets sent by sources inside of the LISP site to destinations outside of the site are candidates for encapsulation by the ITR. The ITR treats the IP destination address as an EID and performs an EID-to-RLOC mapping lookup. The router then prepends an "outer" IP header with one of its globally routable RLOCs in the source address field and the result of the mapping lookup in the destination address field. Note that this destination RLOC MAY be an intermediate, proxy device that has better knowledge of the EID-to-RLOC mapping closer to the destination EID. In general, an ITR receives IP packets from site end-systems on one side and sends LISP-encapsulated IP packets toward the Internet on the other side. Specifically, when a service provider prepends a LISP header for Traffic Engineering purposes, the router that does this is also regarded as an ITR. The outer RLOC the ISP

ITR uses can be based on the outer destination address (the originating ITR's supplied RLOC) or the inner destination address (the originating hosts supplied EID).

Egress Tunnel Router (ETR) [[RFC6830](#)]: An ETR is a router that accepts an IP packet where the destination address in the "outer" IP header is one of its own RLOCs. The router strips the "outer" header and forwards the packet based on the next IP header found. In general, an ETR receives LISP-encapsulated IP packets from the Internet on one side and sends decapsulated IP packets to site end-systems on the other side. ETR functionality does not have to be limited to a router device. A server host can be the endpoint of a LISP tunnel as well.

Routing LOcator (RLOC) [[RFC6830](#)]: A RLOC is an IPv4 [[RFC0791](#)] or IPv6 [[RFC2460](#)] address of an egress tunnel router (ETR). A RLOC is the output of an EID-to-RLOC mapping lookup. An EID maps to one or more RLOCs. Typically, RLOCs are numbered from topologically aggregatable blocks that are assigned to a site

at each point to which it attaches to the global Internet; where the topology is defined by the connectivity of provider networks, RLOCs can be thought of as PA addresses. Multiple RLOCs can be assigned to the same ETR device or to multiple ETR devices at a site.

Endpoint ID (EID) [[RFC6830](#)]: An EID is a 32-bit (for IPv4) or 128-bit (for IPv6) value used in the source and destination address fields of the first (most inner) LISP header of a packet. The host obtains a destination EID the same way it obtains an destination address today, for example through a Domain Name System (DNS) [[RFC1034](#)] lookup or Session Invitation Protocol (SIP) [[RFC3261](#)] exchange. The source EID is obtained via existing mechanisms used to set a host's "local" IP address. An EID used on the public Internet must have the same properties as any other IP address used in that manner; this means, among other things, that it must be globally unique. An EID is allocated to a host from an EID-prefix block associated with the site where the host is located. An EID can be used by hosts to refer to other hosts. EIDs MUST NOT be used as LISP RLOCs. Note that EID blocks MAY be assigned in a hierarchical manner, independent of the network topology, to facilitate

scaling of the mapping database. In addition, an EID block assigned to a site may have site-local structure (subnetting) for routing within the site; this structure is not visible to the global routing system. In theory, the bit string that represents an EID for one device can represent an RLOC for a different device. As the architecture is realized, if a given bit string is both an RLOC and an EID, it must refer to the same entity in both cases. When used in discussions with other Locator/ID separation proposals, a LISP EID will be called a "LEID". Throughout this document, any reference to "EID" refers to an LEID.

EID-to-RLOC cache [[RFC6830](#)]: The EID-to-RLOC cache is a short-lived, on-demand table in an ITR that stores, tracks, and is responsible for timing-out and otherwise validating EID-to-RLOC mappings. This cache is distinct from the full "database" of EID-to-RLOC mappings, it is dynamic, local to the ITR(s), and relatively small while the database is distributed, relatively static, and much more global in scope.

EID-to-RLOC Database [[RFC6830](#)]: The EID-to-RLOC database is a global distributed database that contains all known EID-prefix to RLOC mappings. Each potential ETR typically contains a small piece of the database: the EID-to-RLOC mappings for the EID prefixes "behind" the router. These map to one of the router's own, globally visible, IP addresses. The same database mapping

entries MUST be configured on all ETRs for a given site. In a steady state the EID-prefixes for the site and the locator-set for each EID-prefix MUST be the same on all ETRs. Procedures to enforce and/or verify this are outside the scope of this document. Note that there MAY be transient conditions when the EID-prefix for the site and locator-set for each EID-prefix may not be the same on all ETRs. This has no negative implications since a partial set of locators can be used.

[3.](#) Problem Statement

LISP is a map-and-encap mechanism where an ITR dynamically learns the mappings when it receives a packet for a destination EID for which it did not do encapsulation before. When such a packet is received, a cache miss occurs and the ITR sends a Map-Request to the mapping

system to retrieve the mapping that corresponds to the destination of the packet that caused the cache miss. The ITR then caches the mapping for any subsequent packet toward the same destination. LISP [[RFC6830](#)] does not specify how a packet that causes a cache miss must be handled. However, to the best of our knowledge, the current implementations drop packets causing a cache miss. The consequences of such a current practice in case of cache miss is two-fold. On the one hand, misses imply packet losses and hence performance issues. On the other hand, due to the consequent Map-Request, cache misses cause load on the mapping system.

When an ITR restarts, its EID-to-RLOC cache is initially empty, and is populated, growing in size, progressively with the traffic. However, because mappings have a limited lifetime, the EID-to-RLOC cache size converges to a stable value and it is expected to always observe misses. As shown in [[Networking12](#)], at the steady state, networks experience a rather stable, and limited, miss rate. However, when an ITR is restarted, e.g., for a maintenance operation, a cache miss storm can be observed. A EID-to-RLOC cache miss storm is a phenomenon during which the miss rate is significantly higher than the miss rate normally observed in the network. A miss storm has two severe side effects, first, it abruptly increases the load on the mapping system, and second, many packets are dropped, which causes performance issues. When an ITR is restarted, actually two cache miss storms can be observed. The first one happens when the ITR is stopped (or fails); while the second one happens when the ITR is again available for encapsulation. The first EID-to-RLOC cache miss storm is due to the fact that all the traffic is suddenly redirected to the other ITRs in the network, which might not have the mappings for all the EIDs of ongoing communications. The second EID-to-RLOC cache miss storm can be observed when the ITR is restarted, because it might have to encapsulate all the traffic redirected to it. As a matter of fact, when the ITR is freshly restarted, its

cache is empty meaning that every packet will cause misses at that particular time.

Cache misses are normal in a LISP network. However, these misses normally happen only when the first packet of the first flow toward an EID is received by an ITR which have no significant impact on the traffic at steady state in the network. On the contrary, when an ITR restarts, cache misses happen on elapsing, potentially high

throughput, flows for which high loss rate is not acceptable. For this particular reason, techniques must be applied to avoid EID-to-RLLOC cache miss storm upon ITRs restarts.

It can be argued that if a router fails and is out of order for a long time, avoiding the EID-to-RLLOC cache miss storm, which lasts in the order of minutes, is not worth. This is not actually accurate. When a router fails, there are usually already deployed backup solutions in order to re-direct the traffic instantaneously, with almost no losses. Such redirection remains in place until the failure is fixed, without any consequence on the traffic except for using a different path. Similarly, when the router is back online, booting, traffic will flow again through it only when the state of the router is consistent with the rest of the network, making re-directing the traffic through it disruptionless. All of this is not true for ITRs. Even if with existing techniques we are able to re-direct the traffic with no losses, the LISP encapsulation engine will drop packets because of the lack of mappings in the cache, creating traffic disruption and a raise in signaling traffic on the mapping system.

In this memo, we open the discussion on techniques that can be used to avoid EID-to-RLLOC cache miss storms in the case of a planned ITR restart. In other words, we discuss how to achieve ITR graceful restart.

[4.](#) ITR Graceful Restart

The addition of an ITR causes the traffic to be redirected to the freshly started ITR and hence risks to cause miss storm. As the cache of an ITR is empty when it starts, every received packet potentially causes a miss. We can isolate three techniques to protect the network from miss storm when an ITR is added (or restarted) in the network. All the ITRs that are potentially used by the same node in the network are grouped in synchronization sets.

- o Non-volatile mapping storage: when an ITR has to be stopped, its EID-to-RLLOC cache is stored on a non-volatile medium (e.g., a hard drive) such that when it is restarted, it can load the EID-to-RLLOC cache to be equivalent of the cache it had before it restarted.

- o ITR deflection: when a miss occurs at an ITR while it is starting

up, the ITR deflects the packet that caused a miss to an ITR in its synchronization set and, in parallel, sends a Map-Request for the EID that caused the miss. Note that the Map-Request can even be sent to another ITR of the site or a Map Resolver working in proxy mode. In this manner mapping retrieval latency can be shortened.

- o ITR cache synchronization: upon startup, the ITR synchronizes its cache with the other ITRs in its synchronization set. The ITR is marked as available only after the cache is synchronized.

The non-volatile storage offers the advantage to be transparent for the network and is adapted to short unavailability periods (e.g., the ITR reboots after an upgrade). However, this technique is not adapted for long unavailability periods where most of the entries might be outdated and new prefixes unknown, or when an ITR is added for the first time in the network. This technique is thus recommended only for network with a low mapping caching dynamics.

Traffic deflection to other ITRs (or a PxTR) upon misses causes several issues. On the one hand, the ITR that is restarting must determine the ITR to which the packet must be deflected. On the other hand, packets must be marked as deflected in order to avoid loops. In addition, the ITR must determine its graceful restart period such that it stops deflecting traffic once at steady state. The deflection from one ITR to another can be done directly in LISP where the ITR that started LISP encapsulates and forwards the packet to another ITR. This last ITR must then also run the ETR functionality to decapsulate the packet.

ITR EID-to_RLOC cache synchronization is the most adapted to graceful restart. When the ITR starts, it sends requests to an ITR in its synchronization set (or its MR) to obtain the full cache. When the synchronization is finished, the ITR advertises itself as an ITR in the network such that the ITR does receive traffic to encapsulate only once its cache is synchronized.

5. Security Considerations

Security considerations have to be written accordingly to the technique finally chosen for ITR graceful restart. However, as a general security recommendation, we can say that mappings must be authenticated in order to avoid relay attacks or denial of service. However, ITR graceful restart should not introduce any new threat in the core LISP mechanism.

[6.](#) Conclusion

In this memo, we highlighted the implication of the addition or the restart of an ITR in a LISP network. When an ITR is added into a LISP network, its EID-to-RLOC cache is initially empty. Therefore, when on-going flows are routed to the freshly started ITR, their packets cause potential miss storms which result in packet drops and mapping system overload. To tackle this issue, we propose and discuss three different techniques to reduce the impact of a planed ITR restart.

[7.](#) Acknowledgments

The authors would like to acknowledge Dino Farinacci, Vince Fuller, Darrel Lewis, Fabio Maino, and Simon van der Linden.

[8.](#) References

[8.1.](#) Normative References

- [RFC6830] Farinacci, D., Fuller, V., Meyer, D., and D. Lewis, "The Locator/ID Separation Protocol (LISP)", [RFC 6830](#), January 2013.
- [RFC6833] Fuller, V. and D. Farinacci, "Locator/ID Separation Protocol (LISP) Map-Server Interface", [RFC 6833](#), January 2013.

[8.2.](#) Informative References

- [Networking12]
Saucez, D., Kim, J., Iannone, L., Bonaventure, O., and C. Filsfils, "A local Approach to Fast Failure Recovery of LISP Ingress Tunnel Routers", The 11th International Conference on Networking (Networking'12) , May 2012, <[\[Networking12\]](#)>.
- [RFC0791] Postel, J., "Internet Protocol", STD 5, [RFC 791](#), September 1981.
- [RFC1034] Mockapetris, P., "Domain names - concepts and facilities", STD 13, [RFC 1034](#), November 1987.
- [RFC2460] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", [RFC 2460](#), December 1998.

Internet-Draft

LISP Graceful Restart

December 2013

[RFC3261] Rosenberg, J., Schulzrinne, H., Camarillo, G., Johnston, A., Peterson, J., Sparks, R., Handley, M., and E. Schooler, "SIP: Session Initiation Protocol", [RFC 3261](#), June 2002.

Authors' Addresses

Damien Saucez
INRIA
2004 route des Lucioles BP 93
Sophia Antipolis Cedex 06902
France

Email: damien.saucez@inria.fr

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>

Luigi Iannone
Telecom ParisTech
23, Avenue d'Italie
75013 Paris
France

Email: luigi.iannone@telecom-paristech.fr

Clarence Filsfils
Cisco Systems
Brussels 1000
Belgium

Email: cf@cisco.com

Saucez, et al.

Expires June 23, 2014

[Page 10]