Network Working Group
Internet-Draft
Intended status: Informational
Expires: July 7, 2013

F. Coras
A. Cabellos-Aparicio
J. Domingo-Pascual
Technical University of Catalonia
F. Maino
D. Farinacci
cisco Systems

January 3, 2013

LISP Replication Engineering
draft-coras-lisp-re-01

Abstract

This document describes a method to build and optimize inter-domain LISP router distribution trees for locator-based unicast and multicast replication of EID-based multicast packets.

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 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 July 7, 2013.

Copyright Notice

Table of Contents

1. Introduction

The Locator/Identifier Separation Protocol (LISP) [I-D.ietf-lisp] provides the mechanisms for the separation of Location and Identity semantics presently overloaded by IP addresses. The split results in the creation of two namespaces that unambigously identify edge-site network objects, Endpoint IDs (EIDs), and core routing objects, Routing LOCators (RLOCs). Apart from aiding the scalablity of the core routing infrastructure, the decoupling also enables the (re)implementation of new or existing inter-domain routing mechanisms.

One such mechanism is inter-domain IP source-specific multicast (SSM) [RFC4607]. In this sense, [I-D.ietf-lisp-multicast] defines the procedures carried out for delivering multicast packets from a source host in a LISP site to receivers residing in the same domain or in other LISP or non-LISP sites when an underlying multicast infrastructure exists. The signaling protocol it specifies for conveying (S-EID,G) state and building the distribution tree connecting the xTRs of the source and receiver domains is PIM [RFC4601]. An alternative method that uses Map-Requests instead of PIM for propagating (S-EID,G) state from multicast receiver site ETRs to source site ITR is established in [I-D.farinacci-lisp-mr-signaling].

Although desirable to use multicast routing in the core network when available, a mismatch between the multicast capabilities of receiver ETRs and source ITR might impede their multicast interconnection. In such a case, unicast RLOC encapsulation will be necessary to deliver multicast packets directly to the ETRs. This however leads to high ITR head-end replication for large sets of receiving ETRs. Therefore, to reduce the replication load of the ITR and scale the service with the number of multicast receivers, the ITR may choose to offload replication to a set of RTRs.

The current document describes how multicast RTRs can be used to

build an inter-domain distribution tree rooted at the ITR that can perform unicast and/or multicast encapsulated replication of multicast packets. This concept, of distributing the replication load from ITR to other RTRs downstream on the core overlay distribution tree, is known as Replication Engineering or LISP-RE. Since unicast replication in such overlays can be suboptimal when compared to the underlay network, methods to optimize packet delivery over the distribution tree are also presented.

This specification does not describe how (S-EID,G) state is built in source and receiver domains, nor does it describe how such state is propagated from receiver ETRs to source ITR. This specification

defines how (S-EID,G) map-cache state is built in the ITR, RTRs and ETRs participating in the overlay distribution tree when they are not connectable by multicast.

2. Definition of Terms

The terminology in this document is consistent with the definitions in [I-D.ietf-lisp] and [I-D.ietf-lisp-multicast] however, it is extended to account for LISP-RE concepts:

Delivery Group (DG): The outer destination address of a multicast encapsulated multicast packet with an EID source.

Replication Target: A unicast locator or Delivery Group towards which a multicast packet may be encapsulated and replicated.

Replication List: A locator-set associated to a multicast map-cache entry (S-EID,G) as defined in [I-D.farinacci-lisp-te]. Locators in the set may be either unicast RLOCs or Delivery Groups. When used by a LISP router, a multicast packet matching the map-cache entry must be replicated to all members of the set. The unicast RLOCs are used to encapsulate multicast packets for unicast delivery on the underlying network. Delivery Groups are used to encapsulate multicast packets for multicast delivery on the underlay.

Unicast Replication: Is the notion of replicating a multicast packet with an EID source address at an ITR or RTR by encapsulating it

into a unicast packet. That is, the oif-list of a multicast map-cache entry can not only have interfaces present for link-layer replication and multicast encapsulation but also for unicast encapsulation.

Overlay Distribution Tree: A degree-constrained spanning tree that represents the path followed by unicast and/or multicast encapsulated multicast packets from the root (ITR) to the leaves (ETRs) through intermediary nodes (RTRs). The ITR and RTRs unicast and/or multicast replicate packets to their tree children. Such tree is built and maintained by the overlay distribution tree controller either by using LISP signaling defined in [I-D.ietf-lisp-multicast] and [I-D.farinacci-lisp-mr-signaling] or by programming the mapping database directly by using ELPs to describe network-wide fanout.

Distribution Tree Controller (DTC): A central entity that manages the overlay distribution tree, such entity can be either the ITR or an external orchestration system.

LISP Replication Node: A router (either the ITR or an RTR) participating and replicating packets downstream in the distribution tree.

Multicast Ingress Tunnel Router (ITR): An ITR as specificed in [I-D.ietf-lisp] that participates as the root in the distribution tree.

Multicast Egress Tunnel Router (ETR): An ETR as specified in [I-D.ietf-lisp] that participates as a leaf in the distribution tree. ETR are the only members of the tree that do not unicast replicate.

Multicast Re-encapsulating Tunnel Router (RTR): An RTR as specified in [I-D.farinacci-lisp-te] that participates as an intermediary node in the distribution tree.

Explicit Locator Path (ELP): an explicit and strictly ordered list

of replication targets a packet must travel to. An ELP may be used to source route a LISP-encapsulated packet on an explicit path of RTRs, however the path between two RTRs is determined by the underlying routing protocol. ELP format is described in [I-D.ietf-lisp-lcaf] and their use in [I-D.farinacci-lisp-te].

3. Overview

This specification describes a method to diminish the replication load of the ITR by using RTRs to build an inter-domain distribution tree. The tree is centrally managed either by the ITR itself or by an external orchestration system. An advantage of this orchestration system is that it offloads signaling from the ITR. The entity that manages the tree is generally referred to as the distribution tree controller (DTC).

In order to offload unicast replication of multicast packets the DTC uses a ITR and a set of RTRs. RTRs willing to participate in the distribution tree associated to the (S-EID,G) multicast channel must join the distribution tree by sending a Map-Request/Join-Request [I-D.farinacci-lisp-mr-signaling] to the DTC. Using this procedure the DTC learns the RLOCs of the available RTRs. Additionally, the DTC must learn the replication capacity of each RTR using out-of-band signaling or by manual configuration.

Given that the ITR and RTRs have a limited replication capacity the distribution tree is a degree-constrained spanning-tree. This means that the root is the ITR, the intermediary members are RTRs while leaves are always ETRs. Multicast packets are addressed to (S-EID,G) and are unicast and/or multicast encapsulated when being transported downstream the tree.

In order to build and maintain the overlay distribution tree the DTC must configure state in the replication nodes (ITR and RTRs). This is done by means of the signaling specified in [I-D.ietf-lisp-multicast] and [I-D.farinacci-lisp-mr-signaling]. Particularly, the DTC receives Map-Requests from RTRs (also from the ITR if the DTC is an external orchestration system) addressed to (S-EID,G). Upon inspection of the source RLOC of the Map-Request the controller determines the originating ITR/RTR and generates an ad-hoc

Map-Reply containing the specific replication list for that particular node according to the topology of the tree. For a LISP replication node, the replication list specifies the set of RTRs/ETRs to which it has to replicate packets, i.e., its overlay distribution tree children. Alternatively, an external orchestration system may directly program the mapping database with ELPs that describe the topology of the overlay distribution tree. Ways of achieving this will be discussed in future versions of the document.

The DTC determines the specific topology of the overlay distribution tree using a centralized algorithm. The only requirements for this algorithm are that it builds a tree that guarantees that ETRs receive the encapsulated multicast packets, that the replication capacity of the ITR and RTRs is not exceeded and that forwarding loops are avoided.

In some cases the network administrator may want an optimized overlay distribution tree, although this specification does not standardize any particular algorithm it suggests one in Section 5.2. In order to build an optimized tree this algorithm makes use of the distance (e.g., latency) between the tree members and the amount of multicast receivers connected to each ETR. Such metrics are not provided by LISP and therefore must be obtained using out-of-band signaling.

4. Overlay Distribution Tree

This section describes how the DTC can build an overlay distribution tree using the signaling and mechanisms defined in [I-D.ietf-lisp-multicast] and [I-D.farinacci-lisp-mr-signaling].

4.1. LISP Replication Node Database

The DTC maintains per (S-EID,G) multicast channel a LISP Replication Node Database (LRND) that stores information about the distribution tree state. This information includes among others the RLOCs of the ITR, RTRs and ETRs that constitute the distribution tree and define the overlay replication topology (i.e., the parent-child relations). Said data may be obtained by the DTC from the standard signaling

messages exchanged with the RTRs and ETRs. Additionally, by means of out-of-band signalling the DTC should obtain information about the replication capacity of RTRs.

If the operator chooses to build an optimized tree, more information must be available at the LRND, this is further discussed in Section 5.2.

4.2. Building the Distribution Tree

This section describes the procedures followed by ETRs and RTRs when attaching to the distribution tree. All procedures assume that the DTC has a LRND consistent with the state of the overlay distribution tree and is aware of the replication capacity of participating RTRs.

The decision of an RTR to join the overlay distribution tree depends on out-of-band signalling (e.g., orchestration system, manual configuration). But, its attachment to the distribution tree is done by means of one of the following two procedures:

1. The RTR explicitly signals the ITR by sending a Join-Request for (S-EID,G) and is replied to with a replication list.

2. If an orchestration system programs the mapping database with ELPs describing the overlay distribution tree, an RTR Map-Requests for (S-EID,G) and receives as reply an ELP that defines its distribution tree fanout. Ways of encoding the tree topology into ELPs will be discussed in future versions of this document.

For RTRs using option 1 the DTC, an ITR in this case, will perform the same processing as for joining ETRs. The following sequence of steps is used to attach an ETR to the overlay distribution tree:

1. The DTC receives a Map-Request/Join-Request for (S-EID,G) from an ETR.

2. If multicast replication is requested, the DTC proceeds as defined in [I-D.farinacci-lisp-mr-signaling] and no further steps are taken.

3. If unicast replication is requested, the DTC must choose a

position for the ETR in the distribution tree topology. Specifically, it initiates a search within the LRND for a node (either the ITR or a RTR) with enough spare replication capacity that will replicate multicast traffic towards the ETR. This tree member will become the parent of the ETR and once it is selected the LRND is updated accordingly. The search algorithm depends on operational requirements and this specification does not standardize one, however Section 5.2 provides an example algorithm. Note also that certain algorithms may require the complete or partial re-shape the tree based on certain performance metrics.

4. The DTC must create/update the (S-EID,G) associated replication state for the selected parent using the mechanisms defined in [I-D.ietf-lisp] and [I-D.farinacci-lisp-mr-signaling] (e.g., Solicit-Map-Request). This results in the parent sending a Map-Request for (S-EID,G), in turn, the DTC Map-Replies with an ad-hoc replication list of locator-sets according to topology stored at the LRND. If the algorithm results in a complete or partial re-shape of the tree then state at multiple tree members must be created/updated. In order to avoid packet loss this must be done synchronously. It will be discussed in future versions of this document how to achieve this.

5. Once the distribution tree is configured to replicate multicast traffic to the ETR the DTC Map-Replies (as specified in [I-D.farinacci-lisp-mr-signaling]) with the destination EID-prefix set to (parent-RLOC, ETR-RLOC).

When a LISP replication node signals its departure from the tree, the information stored at the LRND is updated accordingly. For ETRs, the state of the parent member must be updated as described in step 4. For RTRs both the state of the parent and its children must be updated however, such updates may result in packet loss. Moreover, certain optimization algorithms may result in a re-shape of the tree and as a consequence the state of multiple tree members must be created/updated according to the new topology. How to manage these updates with no packet loss will be discussed in future versions of this document.

5. Distribution Tree Optimization

Operators wishing to improve the performance of the distribution tree need to implement at least one topology discovery mechanism and choose a set of optimization algorithms. Due to the centralized group management, on-line switching between optimization algorithms

may be possible in accordance to the required performance. However, their use is dependent on the presence of overlay topological information. The following logical modules need to be implemented in order to support overlay optimizations with LISP-RE:

Topology Discovery Coordinator: It is in charge of organizing the topology measurements and building a database that stores the topological distances (i.e., a metric must be chosen) between overlay members.

Distribution Tree Computation Unit: It is a component that with the help of an algorithm or heuristic, given as input the topology of the overlay and a constraint, or constraint set, can compute an optimal, or close to optimal, degree-constrained minimum spanning tree that may be used for multicast content distribution.

Whether to implement the above modules in the ITR or in other network elements is the decision of the network administrator.

5.1. Topology Discovery

The present document does not specify any topology discovery mechanisms. Both active and passive topology measurements could be used. A choice between the two, of the policy and admission control used or of the network element in charge of coordinating these measurements could be made in the future based on practical experience. Alternatively, precomputed network maps like the ones offered by [IPLANE] and/or out-of-band signalling may be used.

5.2. Optimization Algorithm

The current document does not recommend an optimization algorithm. However, it provides as an example a low computation cost heuristic, which, in the scenarios simulated in [LCAST-TR], can produce latencies between the ITR and the multicast receivers close to unicast ones. Its choice is to be influenced by operational requirements and the hardware constraints of the equipment in charge of running it. Future experiments might result in a recommendation.

In what follows, we use the term "distance" when referring to a relative length or amplitude of a metric, observed on a path connecting two points, but when the exact nature of the metric is of no interest.

Considering as goal the delivery of content for delay sensitive applications, the function the algorithm minimizes is the maximum

distance (e.g. latency or number of AS hops) from a multicast receiver to the ITR source. Notice that the reference is the

multicast receiver host and not an ETR. Thus, what matters in deciding a member's position in the distribution tree is not solely its distance to the ITR but also the number of multicast receivers it serves. Then, a router close to the source but serving few receivers might find itself lower in the distribution tree than another with a slightly higher distance to the source but with a larger receiver set. The algorithm optimizes the quality of experience for multicast receivers and not for tunnel routers.

The problem described above, that searches for a minimum average distance, degree-bounded spanning tree (MADDBST), can be formally stated as:

Definition: Given an undirected complete graph G=(V,E), a designated vertex r belonging to V, for all vertices v in V, a degree bound d(v) <= dmax, dmax a positive integer, a vertex weight function c(v) with positive integer values, and an edge weight function w(e) with positive values, for all edges e in E. Let W(r,v,T) represent the cost of the path linking r and v in the spanning tree T. Find the spanning tree T of G, routed at r, satisfying that d(v) <= dmax and the distance to the source per multicast receiver is minimized.

The heuristic used to solve this problem works by incrementally growing a tree, starting at the root node r, until it becomes a spanning tree. For each node v, not yet a tree member, it selects a potential parent node u in the tree T, such that the distance per receiver to r, is minimized. At each step, the node with the smallest metric value is added to the tree and the parent selection is redone. The pseudocode of the heuristic is provided in Appendix A.

[SHI] and [BAN] have previously defined and solved similar optimization problems. Shi et al. [SHI] also prove that a particular instance of the problem, where all vertices have weight 1, is NP-complete for degree constraints 2 <= dmax <= |V|-1.

The algorithm can optimize an unicast overlay however, it should not be used to optimize multicast underlay delivery. As a result, if

multicast is used as underlay between part of the overlay members, once one of the members of such Delivery Group is added to the distribution tree, the others should be marked as attached also. These nodes should receive multicast encapsulated multicast packets from the chosen node over the underlying multicast distribution tree.

Finally, since the RTRs do not replicate packets for multicast receiver hosts, prior to applying the MADDBST heuristic, a Minimum Spanning Tree (MST) algorithm should be used to compute the RTR

distribution tree. In this case, the MADDBST heuristic should start attaching ETRs having as input the tree resulting from MST.

6. Security Considerations

Security concerns for LISP-RE the same as for [I-D.ietf-lisp-multicast] and [I-D.farinacci-lisp-mr-signaling].

7. IANA Considerations

This memo includes no request to IANA.

8. Acknowledgements

TODO

9. References

9.1. Normative References

[I-D.farinacci-lisp-mr-signaling]
Farinacci, D. and M. Napierala, "LISP Control-Plane Multicast Signaling", draft-farinacci-lisp-mr-signaling-00 (work in progress), July 2012.

[I-D.farinacci-lisp-te]
Farinacci, D., Lahiri, P., and M. Kowal, "LISP Traffic Engineering Use-Cases", draft-farinacci-lisp-te-01 (work

in progress), July 2012.

[I-D.ietf-lisp]
Farinacci, D., Fuller, V., Meyer, D., and D. Lewis, "Locator/ID Separation Protocol (LISP)", draft-ietf-lisp-24 (work in progress), November 2012.

[I-D.ietf-lisp-lcaf]
Farinacci, D., Meyer, D., and J. Snijders, "LISP Canonical Address Format (LCAF)", draft-ietf-lisp-lcaf-00 (work in progress), August 2012.

[I-D.ietf-lisp-multicast]
Farinacci, D., Meyer, D., Zwiebel, J., and S. Venaas, "LISP for Multicast Environments", draft-ietf-lisp-multicast-14 (work in progress),

February 2012.

[RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, August 2006.

[RFC4607] Holbrook, H. and B. Cain, "Source-Specific Multicast for IP", RFC 4607, August 2006.

9.2. Informative References

[BAN] Banerjee, S., Kommareddy, C., Kar, K., Bhattacharjee, B., and S. Khuller, "Construction of an efficient overlay multicast infrastructure for real-time applications", INFOCOM , 2002.

[IPLANE] Madhyastha, H., Katz-Bassett, E., Anderson, T., Krishnamurthy, A., and A. Venkataramani, "iPlane: An Information Plane for Distributed Services", USENIX OSDI , 2009.

[LCAST-TR]
Coras, F., Cabellos, A., Domingo, J., Maino, F., and D. Farinacci, "Inter-Domain Multicast: LISP Edge Based Trees", Technical

Report http://personals.ac.upc.edu/fcoras/lcast-tr.pdf, 2012.

[SHI] Shi, S., Turner, J., and M. Waldvogel, "Dimensioning server access bandwidth and multicast routing in overlay networks", NOSSDAV , 2001.

Appendix A. MADDBST heuristic

```
INPUT: G = (V,E); r; dmax; w(u,v); c(v); u, v in V
OUTPUT: T

  FOREACH v in V DO
    delta(v) = w(r,v)/c(v);
    p(v) = r;
  END FOREACH

  T takes (U = {r}, D={});
  WHILE U != V DO
    LET u in U-V be the vertex with the smallest delta(u);
    U = U U {u}; L = L U {(p(u),u)};
    FOREACH v in V-U DO
      delta(v) = infinity;
      FOREACH u in U DO
        IF  d(u) < dmax and
```

```
                    W{r,u,T} + w(u,v)/c(v) < delta(v) THEN
                  delta(v) = W{r,u,T} + w(u,v)/c(v);
                  p(v) = u;
                END IF
              END FOR
            END FOR
          END WHILE
```

Figure 1

Authors' Addresses

Florin Coras
Technical University of Catalonia
C/Jordi Girona, s/n
BARCELONA 08034
Spain

Email: fcoras@ac.upc.edu

Albert Cabellos-Aparicio
Technical University of Catalonia
C/Jordi Girona, s/n
BARCELONA 08034
Spain

Email: acabello@ac.upc.edu

Jordi Domingo-Pascual
Technical University of Catalonia
C/Jordi Girona, s/n
BARCELONA 08034
Spain

Email: jordi.domingo@ac.upc.edu

Fabio Maino
cisco Systems

Tasman Drive
San Jose, CA 95134
USA

Email: fmaino@cisco.com

Dino Farinacci
cisco Systems
Tasman Drive
San Jose, CA 95134
USA

Email: farinacci@gmail.com