Network Working Group
Internet-Draft
Intended status: Experimental
Expires: May 30, 2009

D. Farinacci
D. Meyer
J. Zwiebel
cisco Systems
S. Venaas
Uninett

November 26, 2008

LISP for Multicast Environments
draft-farinacci-lisp-multicast-01.txt

Status of this Memo

By submitting this Internet-Draft, each author represents that any applicable patent or other IPR claims of which he or she is aware have been or will be disclosed, and any of which he or she becomes aware will be disclosed, in accordance with Section 6 of BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt.

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

This Internet-Draft will expire on May 30, 2009.

Copyright Notice

Abstract

This draft describes how inter-domain multicast routing will function in an environment where Locator/ID Separation is deployed using the LISP architecture.

Table of Contents

1. Requirements Notation 3
2. Introduction 4
3. Definition of Terms 6
4. Basic Overview 9
5. Source Addresses versus Group Addresses 12
6. Locator Reachability Implications on LISP-Multicast 13
7. Multicast Protocol Changes 14
8. LISP-Multicast Data-Plane Architecture 16
8.1. ITR Forwarding Procedure 16
8.2. ETR Forwarding Procedure 16
8.3. Replication Locations 17
9. LISP-Multicast Interworking 18
9.1. LISP and non-LISP Mixed Sites 18
9.1.1. LISP Source Site to non-LISP Receiver Sites 19
9.1.2. Non-LISP Source Site to non-LISP Receiver Sites . . . 20
9.1.3. Non-LISP Source Site to Any Receiver Site 21
9.1.4. Unicast LISP Source Site to Any Receiver Sites . . . 21
9.1.5. LISP Source Site to Any Receiver Sites 22
9.2. LISP Sites with Mixed Address Families 22
9.3. Making a Multicast Interworking Decision 24
10. Considerations when RP Addresses are Embedded in Group Addresses 25
11. Taking Advantage of Upgrades in the Core 26
12. Security Considerations 27
13. Acknowledgments 28
14. References 29
14.1. Normative References 29
14.2. Informative References 29
Authors' Addresses 31
Intellectual Property and Copyright Statements 32

1. Requirements Notation

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].

2. Introduction

The Locator/ID Separation Architecture [LISP] provides a mechanism to separate out Identification and Location semantics from the current definition of an IP address. By creating two namespaces, an EID namespace used by sites and a Locator (RLOC) namespace used by core routing, the core routing infrastructure can scale by doing topological aggregation of routing information.

Since LISP creates a new namespace, a mapping function must exist to map a site's EID prefixes to its associated locators. For unicast packets, both the source address and destination address must be mapped. For multicast packets, only the source address needs to be mapped. The destination group address doesn't need to be mapped because the semantics of an IPv4 or IPv6 group address are logical in nature and not topology-dependent. Therefore, this specifications focuses on to map a source EID address of a multicast flow during distribution tree setup and packet delivery.

This specification will address the following scenarios:

1. How a multicast source host in a LISP site sends multicast packets to receivers inside of its site as well as to receivers in other sites that are LISP enabled.

2. How inter-domain (or between LISP sites) multicast distribution trees are built and how forwarding of multicast packets leaving a source site toward receivers sites is performed.

3. What protocols are affected and what changes are required to such multicast protocols.

4. How ASM-mode, SSM-mode, and Bidir-mode service models will operate.

5. How multicast packet flow will occur for multiple combinations of LISP and non-LISP capable source and receiver sites, for example:

   A. How multicast packets from a source host in a LISP site are sent to receivers in other sites when they are all non-LISP sites.

   B. How multicast packets from a source host in a LISP site are sent to receivers in both LISP-enabled sites and non-LISP sites.

   C. How multicast packets from a source host in a non-LISP site are sent to receivers in other sites when they are all LISP-

enabled sites.

D. How multicast packets from a source host in a non-LISP site are sent to receivers in both LISP-enabled sites and non-LISP sites.

This specification focuses on what changes are needed to the multicast routing protocols to support LISP-Multicast as well as other protocols used for inter-domain multicast, such as Multi-protocol BGP (MBGP) [RFC4760]. The approach proposed in this specification requires no changes to the multicast infrastructure inside of a site when all sources and receivers reside in that site, even when the site is LISP enabled. That is, internal operation of multicast is unchanged regardless of whether or not the site is LISP enabled or whether or not receivers exist in other sites which are LISP-enabled.

Therefore, we see changes only to PIM-ASM [RFC4601], MSDP [RFC3618], and PIM-SSM [RFC4607]. Bidir-PIM [RFC5015], which typically does not run in an inter-domain environment is not addressed in depth in this version of the specification.

Also, the current version of this specification does not describe multicast-based Traffic Engineering relative to the TE-ITR and TE-ETR descriptions in [LISP].

3. Definition of Terms

The terminology in this section is consistent with the definitions in [LISP] but is extended specifically to deal with the application of the terminology to multicast routing.

LISP-Multicast: a reference to the design in this specification. That is, when any site that is participating in multicast communication has been upgraded to be a LISP site, the operation of control-plane and data-plane protocols is considered part of the LISP-Multicast architecture.

Endpoint ID (EID): a 32-bit (for IPv4) or 128-bit (for IPv6) value used in the source address field of the first (most inner) LISP header of a multicast packet. The host obtains a destination group address the same way it obtains one today, as it would when it is a non-LISP site. The source EID is obtained via existing mechanisms used to set a host's "local" IP address. An EID is allocated to a host from an EID prefix block associated with the site the host is located in. An EID can be used by a host to refer to another host, as when it joins an SSM (S-EID,G) route using IGMP version 3 [RFC4604]. LISP uses Provider Independent (PI) blocks for EIDs; such 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.

Routing Locator (RLOC): the IPv4 or IPv6 address of an ingress tunnel router (ITR), the router in the multicast source host's site that encapsulates multicast packets. It is the output of a 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 Provider Assigned (PA) addresses. Multiple RLOCs can be assigned to the same ITR device or to multiple ITR devices at a site.

Ingress Tunnel Router (ITR): a router which accepts an IP multicast packet with a single IP header (more precisely, an IP packet that does not contain a LISP header). The router treats this "inner" IP destination multicast address opaquely so it doesn't need to perform a map lookup on the group address because it is topologically insignificant. The router then prepends an "outer" IP header with one of its globally-routable RLOCs as the source address field. This RLOC is known to other multicast receiver

sites which have used the mapping database to join a multicast tree for which the ITR is the root. In general, an ITR receives IP packets from site end systems on one side and sends LISP-encapsulated multicast IP packets out all external interfaces which have been joined.

An ITR would receive a multicast packet from a source inside of its site when 1) it is on the path from the multicast source to internally joined receivers, or 2) when it is on the path from the multicast source to externally joined receivers.

Egress Tunnel Router (ETR): a router that is on the path from a multicast source host in another site to a multicast receiver in its own site. An ETR accepts a PIM Join/Prune message from a site internal PIM router destined for the source's EID in the multicast source site. The ETR maps the source EID in the Join/Prune message to an RLOC address based on the EID-to-RLOC mapping. This sets up the ETR to accept multicast encapsulated packets from the ITR in the source multicast site. A multicast ETR decapsulates multicast encapsulated packets and replicates them on interfaces leading to internal receivers.

xTR: is a reference to an ITR or ETR when direction of data flow is not part of the context description. xTR refers to the router that is the tunnel endpoint. Used synonymously with the term "Tunnel Router". For example, "An xTR can be located at the Customer Edge (CE) router", meaning both ITR and ETR functionality can be at the CE router.

LISP Header: a term used in this document to refer to the outer IPv4 or IPv6 header, a UDP header, and a LISP header. An ITR prepends headers and an ETR strips headers. A LISP encapsulated multicast packet will have an "inner" header with the source EID in the source field; an "outer" header with the source RLOC in the source field: and the same globally unique group address in the destination field of both the inner and outer header.

(S,G) State: the formal definition is in the PIM Sparse Mode [RFC4601] specification. For this specification, the term is used generally to refer to multicast state. Based on its topological location, the (S,G) state resides in routers can be either (S-EID,G) state (at a location where the (S,G) state resides) or (S-RLOC,G) state (in the Internet core).

(S-EID,G) State: refers to multicast state in multicast source and receiver sites where S-EID is the IP address of the multicast source host (its EID). An S-EID can appear in an IGMPv3 report, an MSDP SA message or a PIM Join/Prune message that travels inside

of a site.

(S-RLOC,G) State: refers to multicast state in the core where S is a source locator (the IP address of a multicast ITR) of a site with a multicast source. The (S-RLOC,G) is mapped from (S-EID,G) entry by doing a mapping database lookup for the EID prefix that S-EID maps to. An S-RLOC can appear in a PIM Join/Prune message when it travels from an ETR to an ITR over the Internet core.

uLISP Site: a unicast only LISP site according to [LISP] which has not deployed the procedures of this specification and therefore, for multicast purposes, follows the procedures from Section 9.

mPTR: this is a multicast PTR that is responsible for advertising a very coarse EID prefix which non-LISP and uLISP sites can target their (S-EID,G) PIM Join/Prune message to. mPTRs are used so LISP source multicast sites can send multicast packets using source addresses from the EID namespace. mPTRs act as Proxy ETRs for supporting multicast routing in a LISP infrastructure.

Mixed Locator-Sets: this is a locator-set for a LISP database mapping entry where the RLOC addresses in the locator-set are in both IPv4 and IPv6 format.

Unicast PIM Join/Prune Message: this is a standard PIM Join/Prune message (encapsulated in an IP header with protocol number 103) which is sent by ETRs at multicast receiver sites to an ITR at a multicast source site. This message is sent periodically as long as there are interfaces in the oif-list for the (S-EID,G) entry the ETR is joining for.

4. Basic Overview

LISP, when used for unicast routing, increases the site's ability to control ingress traffic flows. Egress traffic flows are controlled by the IGP in the source site. For multicast, the IGP coupled with PIM can decide which path multicast packets ingress. By using the traffic engineering features of LISP, a multicast source site can control the egress of its multicast traffic. By controlling the priorities of locators from a mapping database entry, a source multicast site can control which way multicast receiver sites join to the source site.

At this point in time, we don't see a requirement for different locator-sets, priority, and weight policies for multicast than we have for unicast.

The fundamental multicast forwarding model is to encapsulate a multicast packet into another multicast packet. An ITR will encapsulate multicast packets received from sources that it serves in another LISP multicast header. The destination group address from the inner header is copied to the destination address of the outer header. The inner source address is the EID of the multicast source host and the outer source address is the RLOC of the encapsulating ITR.

The LISP-Multicast architecture will follow this high-level protocol and operational sequence:

1. Receiver hosts in multicast sites will join multicast content the way they do today, they use IGMP. When they use IGMPv3 where they specify source addresses, they use source EIDs, that is they join (S-EID,G). If the S-EID is a local multicast source host. If the multicast source is external to this receiver site, the PIM Join/Prune message flows toward the ETRs, finding the shortest exit (that is the closest exit for the Join/Prune message but it is the closest entrance for the multicast packet to the receiver).

2. The ETR does a mapping database lookup for S-EID. If the mapping is cached from a previous lookup (from either a previous Join/ Prune for the source multicast site or a unicast packet that went to the site), it will use the RLOC information from the mapping. The ETR will use the same priority and weighting mechanism as for unicast. So the source site can decide which way multicast packets egress.

3. The ETR will build two PIM Join/Prune messages, one that contains a (S-EID,G) entry that is unicast to the ITR that matches the RLOC the ETR selects, and the other which contains a (S-RLOC,G) entry so the core network can create multicast state from this ETR to the ITR.

4. When the ITR gets the unicast Join/Prune message (see Section 3 for formal definition), it will process (S-EID,G) entries in the message and propagate them inside of the site where it has explicit routing information for EIDs via the IGP. When the ITR receives the (S-RLOC,G) PIM Join/Prune message it will process it like any other join it would get in today's Internet. The S-RLOC address is the IP address of this ITR.

5. At this point there is (S-EID,G) state from the joining host in the receiver multicast site to the ETR of the receiver multicast site. There is (S-RLOC,G) state across the core network from the ETR of the multicast receiver site to the ITR in the multicast source site and (S-EID,G) state in the source multicast site. Note, the (S-EID,G) state is the same S-EID in each multicast site. As other ETRs join the same multicast tree, they can join through the same ITR (in which case the packet replication is done in the core) or a different ITR (in which case the packet replication is done at the source site).

6. When a packet is originated by the multicast host in the source site, it will flow to one or more ITRs which will prepend a LISP header by copying the group address to the outer destination address field and insert its own locator address in the outer source address field. The ITR will look at its (S-RLOC,G) state, where S-RLOC is its own locator address, and replicate the packet on each interface a (S-RLOC,G) joined was received on. The core has (S-RLOC,G) so where fanout occurs to multiple sites, a core router will do packet replication.

7. When either the source site or the core replicates the packet, the ETR will receive a LISP packet with a destination group address. It will also decapsulate packets because it has receivers for the group. Otherwise, it would have not received the packets because it would not have joined. The ETR decapsulates and does a (S-EID,G) lookup in its multicast FIB to forward packets out one or more interfaces to forward the packet to internal receivers.

This architecture is consistent and scalable with the architecture presented in [LISP] where multicast state in the core operates on locators and multicast state at the sites operates on EIDs.

Alternatively, [LISP] does present a mechanism where (S-EID,G) state can reside in the core through the use of RPF-vectors [RPFV] in PIM Join/Prune messages. However, this will require EID state in core as well as the use of RPF-vector formatted Join/Prune messages which are not the default implementation choice. So we choose a design that can allow the separation of namespaces as unicast LISP provides. It will be at the expense of creating new (S-RLOC,G) state when ITRs go unreachable. See Section 5 for details.

However, we have some observations on the algorithm above. We can scale the control plane but at the expense of sending data to sites which may have not joined the distribution tree where the encapsulated data is being delivered. For example, one site joins (S-EID1,G) and another site joins (S-EID2,G). Both EIDs are in the same multicast source site. Both multicast receiver sites join to the same ITR with state (S-RLOC,G) where S-RLOC is the RLOC for the ITR. The ITR joins both (S-EID1,G) and (S-EID2,G) inside of the site. The ITR receives (S-RLOC,G) joins and populates the oif-list state for it. Since both (S-EID1,G) and (S-EID2, G) map to the one (S-RLOC,G) packets will be delivered by the core to both multicast receiver sites even though each have joined a single source-based distribution tree. This behavior is a consequence of the many-to-one mapping between S-EIDs and a S-RLOC.

There is a possible solution to this problem which reduces the number of many-to-one occurrences of (S-EID,G) entries aggregating into a single (S-RLOC,G) entry. If a physical ITR can be assigned multiple RLOC addresses and these addresses are advertised in mapping database entries, then ETRs at receiver sites have more RLOC address options and therefore can join different (RLOC,G) entries for each (S-EID,G) entry joined at the receiver site. It would not scale to have a one-to-one relationship between the number of S-EID sources at a source site and the number of RLOCs assigned to all ITRs at the site, but we can reduce the "n" to a smaller number in the "n-to-1" relationship. And in turn, reduce the opportunity for data packets to be delivered to sites for groups not joined.

5. Source Addresses versus Group Addresses

Multicast group addresses don't have to be associated with either the EID or RLOC namespace. They actually are a namespace of their own that can be treated as logical with relatively opaque allocation. So, by their nature, they don't detract from an incremental deployment of LISP-Multicast.

As for source addresses, as in the unicast LISP scenario, there is a decoupling of identification from location. In a LISP site, packets are originated from hosts using their allocated EIDs, those addresses are used to identify the host as well as where in the site's topology the host resides but not how and where it is attached to the Internet.

Therefore, when multicast distribution tree state is created anywhere in the network on the path from the any multicast receiver to a multicast source, EID state is maintained at the source and receiver multicast sites, and RLOC state is maintained in the core. That is, a multicast distribution tree will be represented as a 3-tuple of {(S-EID,G) (S-RLOC,G) (S-EID,G)} where the first element of the 3-tuple is the state stored in routers from the source to one or more ITRs in the source multicast site, the second element of the 3-tuple is the state stored in routers downstream of the ITR, in the core, to all LISP receiver multicast sites, and the third element in the 3-tuple is the state stored in the routers downstream of each ETR, in each receiver multicast site, reaching each receiver. Note that (S-EID,G) is the same in both the source and receiver multicast sites.

The concatenation/mapping from the first element to the second element of the 3-tuples is done by the ITR and from the second element to the third element is done at the ETRs.

6. Locator Reachability Implications on LISP-Multicast

Multicast state as it is stored in the core is always (S,G) state as it exists today or (S-RLOC,G) state as it will exist when LISP sites are deployed. The core routers cannot distinguish one from the other. They don't need to because it is state that RPFs against the core routing tables in the RLOC namespace. The difference is where the root of the distribution tree for a particular source is. In the traditional multicast core, the source S is the source host's IP address. For LISP-Multicast the source S is a single ITR of the multicast source site.

An ITR is selected based on the LISP EID-to-RLOC mapping used when an ETR propagates a PIM Join/Prune message out of a receiver multicast site. The selection is based on the same algorithm an ITR would use to select an ETR when sending a unicast packet to the site. In the unicast case, the ITR can change on a per-packet basis depending on the reachability of the ETR. So an ITR can change relatively easily using local reachability state. However, in the multicast case, when an ITR goes unreachable, new distribution tree state must be built because the encapsulating root has changed. This is more significant than an RPF-change event, where any router would typically locally change its RPF-interface for its existing tree state. But when an encapsulating LISP-Multicast ITR goes unreachable, new distribution state must be rebuilt and reflect the new encapsulator. Therefore, when an ITR goes unreachable, all ETRs that are currently joined to that ITR will have to trigger a new Join/Prune message for (S-RLOC,G) to the new ITR as well as send a unicast Join/Prune message telling the new ITR which (S-EID,G) is being joined.

This issue can be mitigated by using anycast addressing for the ITRs so the problem does reduce to an RPF change in the core, but still requires a unicast Join/Prune message to tell the new ITR about (S-EID,G). The problem with this approach is that the ETR really doesn't know when the ITR has changed so the new anycast ITR will get the (S-EID,G) state only when the ETR sends it the next time during its periodic sending procedures.

7. Multicast Protocol Changes

A number of protocols are used today for inter-domain multicast routing:

IGMPv1-v3, MLDv1-v2: These protocols do not require any changes for LISP-Multicast for two reasons. One being that they are link-local and not used over site boundaries and second they advertise group addresses that don't need translation. Where source addresses are supplied in IGMPv3 and MLDv2 messages, they are semantically regarded as EIDs and don't need to be converted to RLOCs until the multicast tree-building protocol, such as PIM, is received by the ETR at the site boundary. Addresses used for IGMP and MLD come out of the source site's allocated addresses which are therefore from the EID namespace.

MBGP: Even though MBGP is not a multicast routing protocol, it is used to find multicast sources when the unicast BGP peering topology and the multicast MBGP peering topology are not congruent. When MBGP is used in a LISP-Multicast environment, the prefixes which are advertised are from the RLOC namespace. This allows receiver multicast sites to find a path to the source multicast site's ITRs. MBGP peering addresses will be from the RLOC namespace.

MSDP: MSDP is used to announce active multicast sources to other routing domains (or LISP sites). The announcements come from the PIM Rendezvous Points (RPs) from sites where there are active multicast sources sending to various groups. In the context of LISP-Multicast, the source addresses advertised in MSDP will semantically be from the EID namespace since they describe the identity of a source multicast host. It will be true that the state stored in MSDP caches from core routers will be from the EID namespace. An RP address inside of site will be from the EID namespace so it can be advertised and reached by internal unicast routing mechanism. However, for MSDP peer-RPF checking to work properly across sites, the RP addresses must be converted or mapped into a routable address that is advertised and maintained in the BGP routing tables in the core. MSDP peering addresses can come out of either the EID or a routable address namespace. And the choice can be made unilaterally because the ITR at the site will determine which namespace the destination peer address is out of by looking in the mapping database service.

PIM-SSM: In the simplest form of distribution tree building, when PIM operates in SSM mode, a source distribution tree is built and maintained across site boundaries. In this case, there is a small modification to the operation of the PIM protocol (but not to any

message format) to support taking a Join/Prune message originated inside of a LISP site with embedded addresses from the EID namespace and converting them to addresses from the RLOC namespace when the Join/Prune message crosses a site boundary. This is similar to the requirements documented in [MNAT].

PIM-Bidir: Bidirectional PIM is typically run inside of a routing domain, but if deployed in an inter-domain environment, one would have to decide if the RP address of the shared-tree would be from the EID namespace or the RLOC namespace. If the RP resides in a site-based router, then the RP address is from the EID namespace. If the RP resides in the core where RLOC addresses are routed, then the RP address is from the RLOC namespace. This could be easily distinguishable if the EID address were well-known address allocation block from the RLOC namespace. Also, when using Embedded-RP for RP determination [RFC3956], the format of the group address could indicate the namespace the RP address is from. However, refer to Section 10 for considerations core routers need to make when using Embedded-RP IPv6 group addresses. With respect to DF-election in Bidir PIM, no changes are required since all messaging and addressing is link-local.

PIM-ASM: The way ASM mode PIM, the most popular form of PIM, is deployed in the Internet today is by having shared-trees within a site and using source-trees across sites. By the use of MSDP and PIM-SSM techniques described above, we can get multicast connectivity across LISP sites. Having said that, that means there are no special actions required for processing (*,G) or (S,G,R) Join/Prune messages since they all operate against the shared-tree which is site resident. This is also true for the RP-mapping mechanisms Auto-RP and BSR.

Based on the protocol description above, the conclusion is that there are no protocol message format changes, just a translation function performed at the control-plane. This will make for an easier and faster transition for LISP since fewer components in the network have to change.

It should also be stated just like it is in [LISP] that no host changes, whatsoever, are required to have a multicast source host send multicast packets and for a multicast receiver host to receive multicast packets.

8. LISP-Multicast Data-Plane Architecture

The LISP-Multicast data-plane operation conforms to the operation and packet formats specified in [LISP]. However, encapsulating a multicast packet from an ITR is a much simpler process. The process is simply to copy the inner group address to the outer destination address. And to have the ITR use its own IP address (its RLOC), and as the source address. The process is simpler for multicast because there is no EID-to-RLOC mapping lookup performed during packet forwarding.

In the decapsulation case, the ETR simply removes the outer header and performs a multicast routing table lookup on the inner header (S-EID,G) addresses. Then the oif-list for the (S-EID,G) entry is used to replicate the packet on site-facing interfaces leading to multicast receiver hosts.

There is no Data-Probe logic for ETRs as there can be in the unicast forwarding case.

8.1. ITR Forwarding Procedure

The following procedure is used by an ITR, when it receives a multicast packet from a source inside of its site:

1. A multicast data packet sent by a host in a LISP site will have the source address equal to the host's EID and the destination address equal to the group address of the multicast group. It is assumed the group information is obtained by current methods. The same is true for a multicast receiver to obtain the source and group address of a multicast flow.

2. When the ITR receives a multicast packet, it will have both S-EID state and S-RLOC state stored. Since the packet was received on a site-facing interface, the RPF lookup is based on the S-EID state. If the RPF check succeeds, then the oif-list contains interfaces that are site-facing and external-facing. For the site-facing interfaces, no LISP header is prepended. For the external-facing interfaces a LISP header is prepended. When the ITR prepends a LISP header, it uses its own RLOC address as the source address and copies the group address supplied by the IP header the host built as the outer destination address.

8.2. ETR Forwarding Procedure

The following procedure is used by an ETR, when it receives a multicast packet from a source outside of its site:

1. When a multicast data packet is received by an ETR on an external-facing interface, it will do an RPF lookup on the S-RLOC state it has stored. If the RPF check succeeds, the interfaces from the oif-list are used for replication to interfaces that are site-facing as well as interfaces that are external-facing (this ETR can also be a transit multicast router for receivers outside of its site). When the packet is to be replicated for an external-facing interface, the LISP encapsulation header are not stripped. When the packet is replicated for a site-facing interface, the encapsulation header is stripped.

2. The packet without a LISP header is now forwarded down the (S-EID,G) distribution tree in the receiver multicast site.

8.3. Replication Locations

Multicast packet replication can happen in the following topological locations:

- In an IGP multicast router inside a site which operates on S-EIDs.
- In a transit multicast router inside of the core which operates on S-RLOCs.
- At one or more ETR routers depending on the path a Join/Prune message exits a receiver multicast site.
- At one or more ITR routers in a source multicast site depending on what priorities are returned in a Map-Reply to receiver multicast sites.

In the last case the source multicast site can do replication rather than having a single exit from the site. But this only can occur when the priorities in the Map-Reply are modified for different receiver multicast site so that the PIM Join/Prune messages arrive at different ITRs.

This policy technique, also used in [ALT] for unicast, is useful for multicast to mitigate the problems of changing distribution tree state as discussed in Section 6.

9. LISP-Multicast Interworking

This section will describe the multicast corollary to [INTWORK] which describes the interworking of multicast routing among LISP and non-LISP sites.

9.1. LISP and non-LISP Mixed Sites

Since multicast communication can involve more than two entities to communicate together, the combinations of interworking scenarios are more involved. However, the state maintained for distribution trees at the sites is the same regardless of whether or not the site is LISP enabled or not. So most of the implications are in the core with respect to storing routable EID prefixes from either PA or PI blocks.

Before we enumerate the multicast interworking scenarios, we must define 3 deployment states of a site:

- A non-LISP site which will run PIM-SSM or PIM-ASM with MSDP as it does today. The addresses for the site are globally routable.

- A site that deploys LISP for unicast routing. The addresses for the site are not globally routable. Let's define the name for this type of site as a uLISP site.

- A site that deploys LISP for both unicast and multicast routing. The addresses for the site are not globally routable. Let's define the name for this type of site as a LISP-Multicast site.

We will not consider a LISP site enabled for multicast purposes only but do consider a uLISP site as documented in [INTWORK]. In this section we don't discuss how a LISP site sends multicast packets when all receiver sites are LISP-Multicast enabled; that has been discussed in previous sections.

The following scenarios exist to make LISP-Multicast sites interwork with non-LISP-Multicast sites:

1. A LISP site must be able to send multicast packets to receiver sites which are a mix of non-LISP sites and uLISP sites.

2. A non-LISP site must be able to send multicast packets to receiver sites which are a mix of non-LISP sites and uLISP sites.

3. A non-LISP site must be able to send multicast packets to receiver sites which are a mix of LISP sites, uLISP sites, and non-LISP sites.

4. A uLISP site must be able to send multicast packets to receiver sites which are a mix of LISP sites, uLISP sites, and non-LISP sites.

5. A LISP site must be able to send multicast packets to receiver sites which are a mix of LISP sites, uLISP sites, and non-LISP sites.

9.1.1. LISP Source Site to non-LISP Receiver Sites

In the first scenario, a site is LISP capable for both unicast and multicast traffic and as such operates on EIDs. Therefore there is a possibility that the EID prefix block is not routable in the core. For LISP receiver multicast sites this isn't a problem but for non-LISP or uLISP receiver multicast sites, when a PIM Join/Prune message is received by the edge router, it has no route to propagate the Join/Prune message out of the site. This is no different than the unicast case that LISP-NAT in [INTWORK] solves.

LISP-NAT allows a unicast packet that exits a LISP site to get its source address mapped to a globally routable address before the ITR realizes that it should not encapsulate the packet destined to a non-LISP site. For a multicast packet to leave a LISP site, distribution tree state needs to be built so the ITR can know where to send the packet. So the receiver multicast sites need to know about the multicast source host by its routable address and not its EID address. When this is the case, the routable address is the (S-RLOC,G) state that is stored and maintained in the core routers. It is important to note that the routable address for the host cannot be the same as an RLOC for the site. Because we want the ITRs to process a received PIM Join/Prune message from an external-facing interface to be propagated inside of the site so the site-part of the distribution tree is built.

Using a globally routable source address allows non-LISP and uLISP multicast receiver to join, create, and maintain a multicast distribution tree. However, the LISP multicast receiver site will want to perform an EID-to-RLOC mapping table lookup when a PIM Join/Prune message is received on a site-facing interface. It does this because it wants to find a (S-RLOC,G) entry to Join in the core. So we have a conflict of behavior between the two types of sites.

The solution to this problem is the same as when an ITR wants to send a unicast packet to a destination site but needs determine if the site is LISP capable or not. When it is not LISP capable, the ITR does not encapsulate the packet. So for the multicast case, when ETR receives a PIM Join/Prune message for (S-EID,G) state, it will do a mapping table lookup on S-EID. In this case, S-EID is not in the

mapping database because the source multicast site is using a routable address and not an EID prefix address. So the ETR knows to simply propagate the PIM Join/Prune message to a external-facing interface without converting the (S-EID,G) because it is an (S,G) where S is routable and reachable via core routing tables.

Now that the multicast distribution tree is built and maintained from any non-LISP or uLISP receiver multicast site, the way packet forwarding model is performed can be explained.

Since the ITR in the source multicast site has never received a unicast PIM Join/Prune message from any ETR in a receiver multicast site, it knows there are no LISP-Multicast receiver sites. Therefore, there is no need for the ITR to encapsulate data. Since it will know a priori (via configuration) that its site's EIDs are not routable, it assumes that the multicast packets from the source host are sent by a routable address. That is, it is the responsibility of the multicast source host's system administrator to ensure that the source host sends multicast traffic using a routable source address. When this happens, the ITR acts simply as a router and forwards the multicast packet like an ordinary multicast router.

There is an alternative to using a LISP-NAT scheme just like there is for unicast [INTWORK] forwarding by using Proxy Tunnel Routers (PTRs). This can work the same way for multicast routing as well, but the difference is that non-LISP and uLISP sites will send PIM Join/Prune messages for (S-EID,G) which make their way in the core to PTRs. Let's call this use of a PTR as a "Multicast PTR" (or mPTR). Since the PTRs advertise very coarse EID prefixes, they draw the PIM Join/Prune control traffic making them the target of the distribution tree. To get multicast packets from the LISP source multicast sites, the tree needs to be built on the path from the mPTR to the LISP source multicast site. To make this happen the mPTR acts as a "Proxy ETR" (where in unicast it acts as a "Proxy ITR").

The existence of mPTRs in the core allows LISP source multicast site ITRs to encapsulate multicast packets so the state built between the ITRs and the mPTRs is (S-RLOC,G) state. Then the mPTRs can decapsulate packets and forward natively to the non-LISP and uLISP receiver multicast sites.

9.1.2. Non-LISP Source Site to non-LISP Receiver Sites

Clearly non-LISP multicast sites can send multicast packets to non-LISP receiver multicast sites. That is what they do today. However, discussion is required to show how non-LISP multicast sites send multicast packets to uLISP receiver multicast sites.

Since uLISP receiver multicast sites are not targets of any (S,G) state, they simply send (S,G) PIM Join/Prune messages toward the non-LISP source multicast site. Since the source multicast site, in this case has not been upgraded to LISP, all multicast source host addresses are routable. So this case is simplified to where a uLISP receiver multicast site looks to the source multicast site as a non-LISP receiver multicast site.

9.1.3. Non-LISP Source Site to Any Receiver Site

When a non-LISP source multicast site has receivers in either a non-LISP/uLISP site or a LISP site, one needs to decide how the LISP receiver multicast site will attach to the distribution tree. We know from Section 9.1.2 that non-LISP and uLISP receiver multicast sites can join the distribution tree, but a LISP receiver multicast site ETR will need to know if the source address of the multicast source host is routable or not. We showed in Section 9.1.1 that an ETR, before it sends a PIM Join/Prune message on an external-facing interface, does a EID-to-RLOC mapping lookup to determine if it should convert the (S,G) state from a PIM Join/Prune message received on a site-facing interface to a (S-RLOC,G). If the lookup fails, the ETR can conclude the source multicast site is a non-LISP site so it simply forwards the Join/Prune message (it also doesn't need to send a unicast Join/Prune message because there is no ITR in a non-LISP site and there is namespace continuity between the ETR and source).

9.1.4. Unicast LISP Source Site to Any Receiver Sites

In the last section, it was explained how an ETR in a multicast receiver site can determine if a source multicast site is LISP-enabled by looking into the mapping database. When the source multicast site is a uLISP site, it is LISP enabled but the ITR, by definition is not capable of doing multicast encapsulation. So for the purposes of multicast routing, the uLISP source multicast site is treated as non-LISP source multicast site.

Non-LISP receiver multicast sites can join distribution trees to a uLISP source multicast site since the source site behaves, from a forwarding perspective, as a non-LISP source site. This is also the case for a uLISP receiver multicast site since the ETR does not have multicast functionality built-in or enabled.

Special considerations are required for LISP receiver multicast sites since they think the source multicast site is LISP capable, the ETR cannot know if ITR is LISP-Multicast capable. To solve this problem, each mapping database entry will have a multicast 2-tuple (Mpriority, Mweight) per RLOC. When the Mpriority is set to 255, the site is considered not multicast capable. So an ETR in a LISP receiver

multicast site can distinguish whether a LISP source multicast site is LISP-Multicast site from a uLISP site.

9.1.5. LISP Source Site to Any Receiver Sites

When a LISP source multicast site has receivers in LISP, non-LISP, and uLISP receiver multicast sites, it has a conflict about how it sends multicast packets. The ITR can either encapsulate or natively forward multicast packets. Since the receiver multicast sites are heterogeneous in their behavior, one packet forwarding mechanism cannot satisfy both. However, if a LISP receiver multicast site acts like a uLISP site then it could receive packets like a non-LISP receiver multicast site making all receiver multicast sites have homogeneous behavior. However, this poses the following issues:

- LISP-NAT techniques with routable addresses would be required in all cases.
- Or alternatively, mPTR deployment would be required forcing coarse EID prefix advertisement in the core.
- But what is most disturbing is that when all sites that participate are LISP-Multicast sites but then a non-LISP or uLISP site joins the distribution tree, then the existing joined LISP receiver multicast sites would have to change their behavior. This would create too much dynamic tree-building churn to be a viable alternative.

So the solution space options are:

1. Make the LISP ITR in the source multicast site send two packets, one that is encapsulated with (S-RLOC,G) to reach LISP receiver multicast sites and another that is not encapsulated with (S-EID,G) to reach non-LISP and uLISP receiver multicast sites.
2. Make the LISP ITR always encapsulate packets with (S-RLOC,G) to reach LISP-Multicast sites and to reach mPTRs that can decapsulate and forward (S-EID,G) packets to non-LISP and uLISP receiver multicast sites.

9.2. LISP Sites with Mixed Address Families

A LISP database mapping entry that describes the locator-set, Mpriority and Mweight per locator address (RLOC), for an EID prefix associated with a site could have RLOC addresses in either IPv4 or IPv6 format. When a mapping entry has a mix of RLOC formatted addresses, it is an implicit advertisement by the site that it is a dual-stack site. That is, the site can receive IPv4 or IPv6 unicast

packets.

To distinguish if the site can receive dual-stack unicast packets as well as dual-stack multicast packets, the Mpriority value setting will be relative to an IPv4 or IPv6 RLOC See [LISP] for packet format details.

If you consider the combinations of LISP, non-LISP, and uLISP sites sharing the same distribution tree and considering the capabilities of supporting IPv4, IPv6, or dual-stack, the number of total combinations grows beyond comprehension.

Using some combinatorial math, we have the following profiles of a site and the combinations that can occur:

1. LISP-Multicast IPv4 Site

2. LISP-Multicast IPv6 Site

3. LISP-Multicast Dual-Stack Site

4. uLISP IPv4 Site

5. uLISP IPv6 Site

6. uLISP Dual-Stack Site

7. non-LISP IPv4 Site

8. non-LISP IPv6 Site

9. non-LISP Dual-Stack Site

Lets define (m n) = m!/(n!*(m-n)!), pronounced "m choose n" to illustrate some combinatorial math below.

When 1 site talks to another site, the combinatorial is (9 2), when 1 site talks to another 2 sites, the combinatorial is (9 3). If sum this up to (9 9), we have:

(9 2) + (9 3) + (9 4) + (9 5) + (9 6) + (9 7) + (9 8) + (9 9) =

36 + 84 + 126 + 126 + 84 + 36 + 9 + 1

Which results in the total number of cases to be considered at 502.

This combinatorial gets even worse when you consider a site using one

address family inside of the site and the xTRs use the other address family (as in using IPv4 EIDs with IPv6 RLOCs or IPv6 EIDs with IPv4 RLOCs).

To rationalize this combinatorial nightmare, there are some guidelines which need to be put in place:

- Each distribution tree shared among sites will either be an IPv4 distribution tree or an IPv6 distribution tree. Therefore, we can avoid head-end replication by building and sending packets on each address family based distribution tree. Even though there might be an urge to do multicast packet translation from one address family format to the other, it is a non-viable over-complicated urge.

- All LISP sites on a multicast distribution tree must share a common address family which is determined by the source site's locator-set in its LISP database mapping entry. All receiver multicast sites will use the best RLOC priority controlled by the source multicast site. This is true when the source site is either LISP-Multicast or uLISP capable. This means that priority-based policy modification is prohibited.

- When the source site is not LISP capable, it is up to how receivers find the source and group information for a multicast flow. That mechanism decides the address family for the flow.

9.3. Making a Multicast Interworking Decision

This Multicast Interworking section has shown all combinations of multicast connectivity that could occur. As you might have already concluded, this can be quite complicated and if the design is too ambitious, the dynamics of the protocol could cause a lot of instability.

The trade-off decisions are hard to make and we want the same single solution to work for both IPv4 and IPv6 multicast. It is imperative to have an incrementally deployable solution for all of IPv4 unicast and multicast and IPv6 unicast and multicast while minimizing (or eliminating) both unicast and multicast EID namespace state.

Therefore the design decision to go with PTRs for unicast routing and mPTRs for multicast routing seems to be the sweet spot in the solution space so we can optimize state requirements and avoid head-end data replication at ITRs.

10. Considerations when RP Addresses are Embedded in Group Addresses

When ASM and PIM-Bidir is used in an IPv6 inter-domain environment, a technique exists to embed the unicast address of an RP in a IPv6 group address [RFC3956]. When routers in end sites process a PIM Join/Prune message which contain an embedded-RP group address, they extract the RP address from the group address and treat it from the EID namespace. However, core routers do not have state for the EID namespace, need to extract an RP address from the RLOC namespace.

Therefore, it is the responsibility of ETRs in multicast receiver sites to map the group address into a group address where the embedded-RP address is from the RLOC namespace. The mapped RP-address is obtained from a EID-to-RLOC mapping database lookup. The ETR will also send a unicast (*,G) Join/Prune message to the ITR so the branch of the distribution tree from the source site resident RP to the ITR is created.

This technique is no different than the techniques described in this specification for translating (S,G) state and propagating Join/Prune messages into the core. The only difference is that the (*,G) state in Join/Prune messages are mapped because they contain unicast addresses encoded in an Embedded-RP group address.

11. Taking Advantage of Upgrades in the Core

If the core routers are upgraded to support [RPFV] and [JOIN-ATTR], then we can pass EID specific data through the core without, possibly, having to store the state in the core.

By doing this we can eliminate the ETR from unicasting PIM Join/Prune messages to the source site's ITR.

12. Security Considerations

Refer to the [LISP] specification.

13. Acknowledgments

The authors would like to gratefully acknowledge the people who have contributed discussion, ideas, and commentary to the making of this proposal and specification. People who provided expert review were Scott Brim, Greg Shepherd, and Dave Oran. Other commentary from discussions at Summer 2008 Dublin IETF were Toerless Eckert and Ijsbrand Wijnands.

We would also like to thank the MBONED working group for constructive and civil verbal feedback when this draft was presented at the Fall 2008 IETF in Minneapolis. In particular, good commentary came from Tom Pusateri, Steve Casner, Marshall Eubanks, Dimitri Papadimitriou, Ron Bonica, and Lenny Guardino.

14. References

14.1. Normative References

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

[RFC3618] Fenner, B. and D. Meyer, "Multicast Source Discovery Protocol (MSDP)", RFC 3618, October 2003.

[RFC3956] Savola, P. and B. Haberman, "Embedding the Rendezvous Point (RP) Address in an IPv6 Multicast Address", RFC 3956, November 2004.

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

[RFC4604] Holbrook, H., Cain, B., and B. Haberman, "Using Internet Group Management Protocol Version 3 (IGMPv3) and Multicast Listener Discovery Protocol Version 2 (MLDv2) for Source-Specific Multicast", RFC 4604, August 2006.

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

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.

[RFC5015] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", RFC 5015, October 2007.

14.2. Informative References

[ALT] Farinacci, D., Fuller, V., and D. Meyer, "LISP Alternative Topology (LISP-ALT)", draft-fuller-lisp-alt-02.txt (work in progress), April 2008.

[INTWORK] Lewis, D., Meyer, D., and D. Farinacci, "Interworking LISP with IPv4 and IPv6", draft-lewis-lisp-interworking-00.txt (work in progress), December 2007.

[JOIN-ATTR] Wijnands, IJ. and A. Boers, "Format for using TLVs in PIM messages", draft-ietf-pim-join-attributes-03.txt (work in progress), May 2007.

[LISP] Farinacci, D., Fuller, V., Oran, D., and D. Meyer, "Locator/ID Separation Protocol (LISP)", draft-farinacci-lisp-10.txt (work in progress), November 2008.

[MNAT] Wing, D. and T. Eckert, "Multicast Requirements for a Network Address (and port) Translator (NAT)", draft-ietf-behave-multicast-07.txt (work in progress), June 2007.

[RPFV] Wijnands, IJ., Boers, A., and E. Rosen, "The RPF Vector TLV", draft-ietf-pim-rpf-vector-06.txt (work in progress), February 2008.

Authors' Addresses

Dino Farinacci
cisco Systems
Tasman Drive
San Jose, CA
USA

Email: dino@cisco.com

Dave Meyer
cisco Systems
Tasman Drive
San Jose, CA
USA

Email: dmm@cisco.com

John Zwiebel
cisco Systems
Tasman Drive
San Jose, CA
USA

Email: jzwiebel@cisco.com

Stig Venaas
Uninett
Abels gate 5, 4th Floor
N-7465, Trondheim
Norway

Email: Stig.Venaas@uninett.no

Full Copyright Statement

Intellectual Property

Acknowledgment

Funding for the RFC Editor function is provided by the IETF Administrative Support Activity (IASA).