Network Working Group
Internet-Draft
Intended status: Experimental
Expires: May 12, 2008

S. Brim
D. Farinacci
D. Meyer
Cisco Systems, Inc.
J. Curran
ServerVault

November 9, 2007

EID Mappings Multicast Across Cooperating Systems for LISP
draft-curran-lisp-emacs-00

Status of this Memo

Copyright Notice

Abstract

One of the potential problems with the "map-and-encapsulate" approaches to routing architecture is that there is a significant chance of packets being dropped while a mapping is being retrieved. Some approaches pre-load ingress tunnel routers with at least part of the mapping database. Some approaches try to solve this by providing

intermediate "default" routers which have a great deal more knowledge than a typical ingress tunnel router. This document proposes a scheme which does not drop packets yet does not require a great deal of knowledge in any router. However, there are still some issues that need to be worked out.

Requirements Language

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 RFC 2119 [RFC2119].

Table of Contents

1. Introduction

One of the potential problems with the "map-and-encapsulate" approaches to routing architecture is that there is a significant chance of packets being dropped while a mapping is being retrieved. Some approaches pre-load ingress tunnel routers (ITRs) with at least part of the mapping database. Some approaches try to solve this by providing intermediate "default" routers which have a great deal more knowledge than a typical ingress tunnel router. This document proposes a scheme which does not drop packets yet does not require a great deal of knowledge in any router. However, there are still some issues that need to be worked out.

2. Problem Statement

LISP [I-D.farinacci-lisp] assumes a mechanism for obtaining mappings from EID to RLOC exists, but does not require or assume any specific mapping mechanism. Among those proposed for use with LISP are LISP-ALT [I-D.fuller-lisp-alt],NERD [I-D.lear-lisp-nerd], CONS [I-D.meyer-lisp-cons], and APT [I-D.jen-apt]. Others have also been considered.

These mechanisms attempt in various ways to balance database size and churn with the delay of looking up a mapping for the first packet between two sites. If complete mapping information is pushed all the way to the ITRs, then there is no delay in looking up the first mapping, but each ITR must hold a large amount of information and be able to keep it up to date. If mapping information is not pushed at all (as in CONS), then an ITR need only hold the information it decides to cache, but there may be significant delays in retrieving a mapping for the first packets sent between two sites, and those packets may be dropped. Hybrid schemes, where mapping information is pushed partway to the ITRs, have been proposed, but the tradeoff between database size/churn and lookup delay is still not solved satisfactorily.

"Default forwarders" have been proposed in CONS, APT, and CRIO [CRIO]. These are intermediate forwarding points. The intent is that if an ITR does not have a mapping for a packet, it will forward the packet to the default forwarder. The assumption is that the default forwarder serves an aggregate of endpoints, and will thus have better knowledge of how to reach the destination. This eliminates mapping lookup delay and the possibility of dropped packets, at the cost of possibly having the first packets sent between two sites take a longer path. There are two kinds of default forwarders, those that represent multiple sources and those that represent multiple destinations.

- Source-side default forwarders (SSDFs) serve a group of sources, for example a site or all customers of an IP service provider. The belief is that a source-side default forwarder can have more mapping entries than the usual ITR, either because more can be pushed to it or because it will have more entries cached, since it is serving more queries. If it does not have a mapping for the destination it will use one of the mapping mechanisms on behalf of the source. Source-side default forwarders do not actually change the problem, they simply move the problem from the ITR to an intermediary. The same tradeoff, of having high rate/state versus dropping packets versus delay, is still there. The advantage is that they concentrate the problem so that costs can be concentrated as well, but in most cases an SSDF would have performance requirements at the level of a high end router. Valid packets can still be dropped if the SSDF does not itself have a mapping. Another disadvantage is that since they offer a general "default" route, bogus packets will get forwarded to and possibly through them instead of being dropped (for no route).

- Destination-side default forwarders (DSDFs) serve a group of destinations. For example, if a source sends a packet to 192.168.100.1 and its ITR does not have a mapping entry for that packet, the ITR might forward the packet to a default forwarder responsible for all of 192.168.0.0/16. A destination-side forwarder has a mapping database which is complete, but only for a subset of the Internet, so it does not have the high performance requirements of a mainstream source-side default forwarder. Most bogus packets are not forwarded because ITRs will only have routes to DSDFs for valid EID prefixes. A potential downside to DSDFs is that since they represent an aggregate of destinations, the path to the destination through the DSDF may see some stretch.

Destination-side default forwarders look like a good idea if some issues can be dealt with. They can eliminate the possibility of dropped packets. Delay for first packets exchanged between sites has a possibility of being long for some sites, depending on how DSDFs are organized. They still hold part of the mapping database, and need to maintain its accuracy. Also a mechanism is needed for sources, and the routers near sources, to determine which SSDF handles which prefixes. Finally, the mechanism by which the forwarding path is moved toward optimality needs to be secure.

This draft proposes a mechanism using destination-side default forwarders that has low "rate*state" overhead, has easy DSDF location, and controls path stretch. It is called "EID-mappings Multicast Across Cooperating Systems for LISP", or EMACS-LISP.

3. Overview

The mechanism by which DSDFs forward packets to the appropriate ETRs is bidirectional PIM [RFC5015] multicast trees. Briefly:

- In order to keep the number of multicast trees reasonable, each tree handles a subset of the entire EID address space. In IPv4 this might be a /16.

- An ETR responsible for an EID prefix, for example 192.168.100.0/24, joins an appropriate multicast group for an including prefix, for example 192.168.0.0/16. An ETR responsible for an EID prefix larger than an including prefix, or for multiple EID prefixes in different including prefixes, will need to join multiple groups. Multiple ETRs for a site might join the group.

- The bidirectional PIM tree rendezvous point addresses, and the groups they are rendezvous points for, are advertised in multiprocotol eBGP. This instance of eBGP runs in an overlay GRE infrastructure, distinct from the eBGP instance which will be used for normal RLOC routing in the Internet core.

- When an ITR needs to forward a packet and does not have a LISP EID->RLOC mapping, it uses an algorithm to find the correct multicast group, and sends the packet to that group, on the overlay GRE infrastructure. The outer destination RLOC is the multicast address. The outer source RLOC is the ITR's.

- The packets travel to all registered recipients. Most of them examine the packet and realize they are not responsible for the destination, so they throw it away. Among the ETRs which are responsible for the EID prefix, one or more will send a LISP Map-Reply back to the originating ITR, providing a specific mapping, so that the ITR can send all further packets directly.

Thus the first one or two packets sent between two sites will experience more delay than following packets, but no packets are dropped unless they should be.

Details are added, and issues discussed, in the following sections.

4. Detailed Discussion

4.1. Assignment of rendezvous point addresses to groups

Rendezous point addresses (RPAs) are not necessarily related to anything physical. Their determination does not need to be covered

in this document, as long as they can be advertised in the overlay eBGP instance.

4.2. Determination of the Right Group to Join

An ETR must join one or more multicast groups in order to receive packets to the EID prefixes it is responsible for. There must be agreement among the ETRs for a prefix and the ITRs that want to send to that prefix on how the correct multicast group is determined. To avoid mapping retrieval delay, the multicast group must be determinable without querying a server. The following mechanism for mapping from destination EID to multicast group address MUST be supported for 32-bit EIDs:

- There is a /16 in IPv4 multicast address space allocated for use by this protocol, for example 238.1.0.0/16. There are 64k groups, one for each of 64k "including" prefixes.

- The RP addresses of all valid groups are advertised in eBGP, along with group addresses. The next_hop for a group address is the RP's address.

- An ETR responsible for an EID prefix will mask out the higher order 16 bits of that EID prefix, and OR those bits into the lower order 16 bits of 238.1.0.0 to get the group to join. For example, for the EID prefix 192.168.100.0/24, the ETR will join multicast group 238.1.192.168.

- If an ETR handles traffic to an EID prefix shorter than a /16, it will join all groups necessary to cover it.

- The ETR joins those groups, on the GRE overlay.

A similar mechanism can be defined for IPv6. Only valid groups, known to contain EID prefixes participating in LISP, will be advertised in the eBGP instance. Therefore it is all right to define the IPv6 mechanism in a way that allows for a large number of groups.

4.3. Sending a Packet

ITRs participate in the eBGP instance running on the GRE overlay, so that they receive information on available groups and rendezvous point addresses. Packets for which the ITR does not have a direct lisp EID->RLOC mapping cached, does not have a route to an appropriate multicast group, and does not have a direct route for, and are dropped. To send a packet on the multicast overlay, the ITR encapsulates the packet with a LISP header. The destination address is the group determined by the same algorithm as above (Section 4.2).

The source address is an RLOC for the ITR, or at least an RLOC for the source site.

4.4. Path Stretch

The first packets sent between two sites will be multicast. Depending on how the multicast tree is assembled this may not be a direct path. How sub-optimal these paths would be is for further study.

4.5. Requirement for Multicast Deployment

A potential concern with this approach is that it will require multicast support in all of the routers in the Internet core. However, the only nodes interested in the multicast routes are xTRs. They will participate in an overlay tunneled infrastructure, for example over GRE. eBGP, bidirectional PIM, and the multicast packets themselves would all travel over this tunneled infrastructure. The only nodes that need know or care about multicast are the ones that want to use it. The overhead of constructing and maintaining the GRE overlay is for further study.

4.6. Protection against Snoopers

Without any join filters, it is possible for anyone to join any group. A node could join all groups in order to find out which sites are talking to which other sites. This is sometimes not acceptable. Therefore group join filters are required. At the points where a particular ETR will join, "join" messages to the groups for EID prefixes which that ETR handles MUST be allowed, but more SHOULD not be. This configuration is limited, simple, related to configuration that will already be done, and scalable.

4.7. ETR Initial Packet Forwarding

When a packet is multicast it is distributed to all potentially interested ETRs. ETRs that are not responsible for an EID prefix containing the packet's destination address will discard the packet. ETRs which do handle traffic for the destination EID SHOULD decide among themselves who will forward the packet, since duplicate packets can sometimes be a problem. They do so by position in the LISP RLOC-set (see LISP [I-D.farinacci-lisp]). The ETR which is active and has the lowest ordinal position in the RLOC-set will forward the packet.

Note: there can be a serious problem if a small site has an EID prefix in the same /16 including prefix as a large site. In that case, the small site's ETRs would get all of the initial traffic to the large site and have to throw it all away. This could overwhelm a

small site's ETR. This problem, and possibly how to focus some multicast groups, is for further study.

4.8. Responding with a Map-Reply

At least one receiving ETR SHOULD unicast a Map-Reply directly back to the originating ITR, so that future packets will be sent directly to the ETR, not via the multicast infrastructure. As with packet delivery (Section 4.7), the active ETR with the lowest ordinal position in the LISP RLOC-set will be the one to respond.

4.9. Authentication of the Map-Reply

Because an initial packet can go to multiple sites, an ITR SHOULD authenticate any received Map-Reply messages. Otherwise it may misroute future packets. Nonces are not an adequate measure. Some kind of signature is required on the content of the Map-Reply. The trade-offs here are for further study.

4.10. Transition Scenarios

To be filled in. Possibilities include using LISP-ALT as an intermediate approach, using alternative DNS resources, and sending initial packets via multiple paths.

5. IANA Considerations

This section will be filled in later.

A new SAFI will be needed to carry both rendezvous point addresses and group addresses.

A set of at least 64k multicast groups is needed, and it would be better if a /8 were allocated, to be sure. Allocation of 238.0.0.0/8 is requested.

A similar request for IPv6 address space will be made after further study.

Note to RFC Editor: this section may be removed on publication as an RFC.

6. Security Considerations

To be filled in. Known issues are

- Revealing first packets to destinations which are not the source's intended destination.
- Inviting map-reply responses off the path between source and destination.
- Denial of service attacks on the multicast infrastructure.
- Potentially overloading ETRs with unwanted traffic.

7. Contributors

The authors are grateful for the help of those who offered comments, notably Vince Fuller, Eliot Lear, Darrel Lewis, and David Oran.

8. References

8.1. Normative References

[I-D.farinacci-lisp] Farinacci, D., "Locator/ID Separation Protocol (LISP)", draft-farinacci-lisp-04 (work in progress), August 2007.

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

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

8.2. Informative References

[CRIO] Zhang, X., Francis, P., Wang, J., and K. Yoshida, "Scaling IP Routing with the Core Router-Integrated Overlay", November 2006.

[I-D.fuller-lisp-alt] Fuller, V., "LISP-ALT", Internet-Draft not yet published.

[I-D.jen-apt] Jen, D., "APT: A Practical Transit Mapping Service", draft-jen-apt-00 (work in progress), July 2007.

[I-D.lear-lisp-nerd] Lear, E., "NERD: A Not-so-novel EID to RLOC Database", draft-lear-lisp-nerd-02 (work in progress),

September 2007.

[I-D.meyer-lisp-cons]
Brim, S., "LISP-CONS: A Content distribution Overlay Network Service for LISP", draft-meyer-lisp-cons-02 (work in progress), September 2007.

Authors' Addresses

Scott Brim
Cisco Systems, Inc.

Email: swb@employees.org

Dino Farinacci
Cisco Systems, Inc.

Email: dino@cisco.com

David Meyer
Cisco Systems, Inc.

Email: dmm@1-4-5.net

John Curran
ServerVault

Email: jcurran@istaff.org

Full Copyright Statement

Intellectual Property

Acknowledgment

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