TRILL working group
Internet Draft
Intended status: Standard Track
Expires: Sept 2012

L. Dunbar
D. Eastlake
Huawei
Radia Perlman
Intel
I. Gashinsky
Yahoo

March 5, 2012

Mechanisms for Directory Assisting RBridge
draft-dunbar-trill-scheme-for-directory-assist-00.txt

Status of this Memo

Copyright Notice

Abstract

This draft describes the mechanisms of using directory server(s) to assist RBridge edge in data center environment.

Conventions used in this document

The term ''Subnet'' and ''VLAN'' are used interchangeably in this document because it is common to map one subnet to one VLAN. The term ''TRILL'' and ''RBridge'' are used interchangeably in this document.

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 0.

Table of Contents

1. Introduction

[Directory-Assisted-RBridge] describes the framework of using directory to assist TRILL edge nodes to improve TRILL network

scalability in data center environment. This draft describes mechanisms of using directory to assist RBridge edge nodes in data center environment.

2. Terminology

AF Appointed Forwarder RBridge port

Bridge: IEEE 802.1Q compliant device. In this draft, Bridge is used interchangeably with Layer 2 switch.

DA: Destination Address

DC: Data Center

EoR: End of Row switches in data center. Also known as Aggregation switches in some data centers

FDB: Filtering Database for Bridge or Layer 2 switch

Host: Application running on a physical server or a virtual machine. A host usually has at least one IP address and at least one MAC address.

SA: Source Address

STP: Spanning Tree Protocol

RSTP: Rapid Spanning Tree Protocol

ToR: Top of Rack Switch in data center. It is also known as access switches in some data centers.

VM: Virtual Machines

3. Push Model of Directory Assisted RBridge Edge in DC Environment

Under this model, Directory Server(s) push down the MAC&VLAN <-> RBridgeEdge mapping for all the hosts which might communicate with hosts attached to an RBridge edge node. With this environment, it is recommended that RBridge edge simply drop a data packet (instead of flooding to RBridge domain) if the packet's destination address can't be found in the MAC&VLAN<->RBridgeEdge mapping table.

The mapping entry to be pushed down could leverage the gratuitous ARP reply with extended fields showing the edge RBridge's name, as shown in Table 2.

3.1. Minimize the mapping entries maintained by RBridge Edge

One major drawback of the ''Push Model'' is that RBridge edge's MAC&VLAN<->RBridgeEdge mapping table will have more entries than it really needs.

One simple step for an RBridge to reduce the number of mapping entries pushed down from directory is to prune out entries belonging to VIDs which are not enabled on its bridged LANs ports. For example, if only {vid#1, vid#2, vid#3} are enabled on bridged LANs connected to an RBridge edge ports, only MAC&VLAN<->RBridgeEdge entries for those three VIDs need to be pushed down to the RBridge edge.

However, under the situations when hosts/VMs attached to one RBridge edge rarely communicate with hosts under the same VLAN attached to different RBridge, the normal process of RBridge edge's cache aging would have removed those MAC&VLAN entries from the RBridge's cache. But it can be difficult for Directory Servers to predict the communication patterns among hosts within one VLAN.

Therefore, even with VLAN pruning, it is likely that the Directory Servers will push down more the MAC&VLAN entries to RBridge Edges than the normal cache aging approach.

3.2. Aggregated entries to push down

Using Table 2 requires one entry per host/VM. When directory pushes down the entire mapping to an edge RBridge for the very first time, there usually are many entries. To minimize the amount of data pushed down, summarization should be considered, e.g. with one edge RBridge Nickname being associated with all attached hosts' MAC addresses and VLANs as shown below:

Nickname1	VID-1	MAC1, MAC2, MACn
	VID-2	MAC1, MAC2, MACn
	....	MAC1, MAC2, MACn
Nickname2	VID-1	MAC1, MAC2, MACn
	VID-2	MAC1, MAC2, MACn
	...	MAC1, MAC2, MACn
-------	------	----------
	....	MAC1, MAC2, MACn

Table 1: Summarized table pushed down from directory

Whenever there is any change in MAC&VLAN <-> RBridgeEdge mapping, which can be triggered by hosts being added, moved, or de-commissioned, an incremental update can be sent to the RBridge edges which are impacted by the change.

3.3. Messages to trigger pushing from directory

In push down model, it is necessary to have a message for RBridge node to request directory server(s) to start pushing down the mapping entries. This message should at least include the number of VLANs enabled on the RBridge edge ports, so that directory server doesn't need to push down the entire mapping entries for all the hosts in the data center.

RBridge node can use this message to get mapping entries when it is initialized or restarted.

4. Pull model of Directory Assisted RBridge Edge in DC Environment

Under this model, ''RBridge'' pulls the MAC&VLAN<->RBridgeEdge mapping entry from the directory server when needed.

RBridge edge node can send data frames with unknown DA to Directory Servers, or intercept all ARP/ND requests and forward them to the Directory Server(s).

The reply from the Directory Server can be the standard ARP/ND reply with an extra field showing the RBridge egress node's Nickname, as depicted in Table 2. RBridge ingress node can cache the mapping.

If there is no response from the directory server, the RBridge edge node can drop the packet.

RBridge edge can age out MAC&VLAN entries if they haven't been used for a certain period of time. Therefore, each RBridge edge will only keep the entries which are frequently used, i.e. mapping table size can be smaller.

The following table shows how target RBridge nickname can be attached to a standard ARP Reply when replying to an ARP request forwarded by ingress RBridge edge.

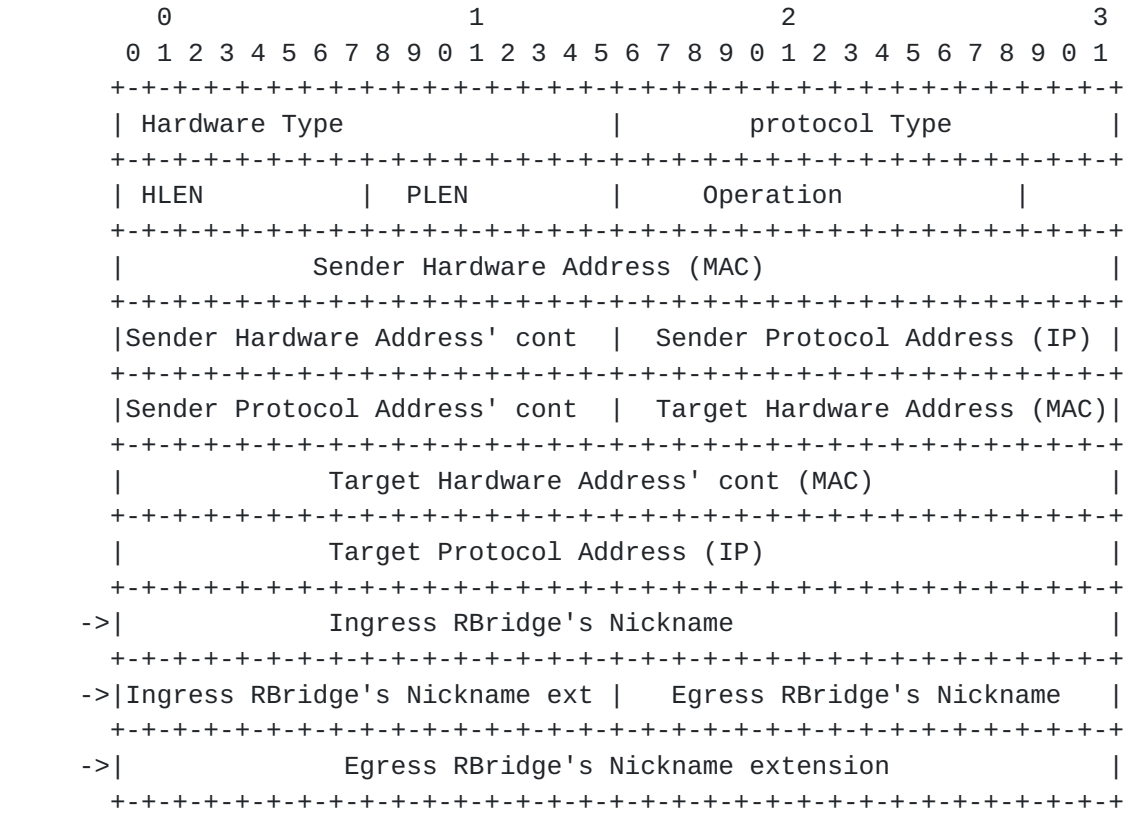

Table 2: Extended fields added to standard ARP reply

The original ARP reply format consists of the first 28 octets shown in this table. The last 12 octets in this table marked by ''->'' are extended fields to indicate the Ingress RBridge to which originating host is attached and the Egress RBridge to which the target host is attached. More bits are reserved for RBridge nicknames in case

multiple levels of nicknames are needed in the future for large data centers.

There are 16 bits for Operation type field in ARP message. IANA has assigned 0~25 for various purposes and leave 26~65534 unassigned [http://www.iana.org/assignments/arp-parameters/arp-parameters.xml]. If this approach is taken, a new ARP Operation code has to be assigned by IANA.

5. Push-Pull Hybrid Model

For some edge nodes which have great number of VIDs enabled, managing the MAC&VLAN <-> RBridgeEdge mapping for hosts under all those VIDs can be challenge. This is especially true for Data Center gateway nodes, which need to maintain majority of VIDs if not all.

For those RBridge Edge nodes, hybrid model should be considered. I.e. Push model are used for some VIDs, and pull model are used for other VIDs. It can be operator's decision (i.e. by configuration) on which VIDs' mapping entries are pushed down from directory and which VIDs' mapping entries are pulled.

For example, in a data center when hosts in specific VIDs (vid#1, vid#2, ? vid#100)communicate regularly with external peers, the mapping entries for those 100 VIDs should be pushed down to the data center gateway routers. For hosts in other VIDs which only communicate with external peers once a day (or once a few days) for management interface, the mapping entries for those VIDs should be pulled down from directory whenever the needs come up.

6. To be continued

This draft only describes the high level view of the mechanism on how to use directory to assist RBridge edge. More details will be added.

7. Manageability Considerations

TBD.

8. Security Considerations

TBD.

9. IANA Considerations

There are 16 bits for Operation type field. IANA has assigned 0~25 for various purposes and leave 26~65534 unassigned [http://www.iana.org/assignments/arp-parameters/arp-parameters.xml]. If this approach is taken, a new ARP Operation code has to be assigned by IANA.

10. Acknowledgments

This document was prepared using 2-Word-v2.0.template.dot.

11. References

[Directory-assisted-RBridge] Dunbar, et, al ''Directory Assisted RBridge Edge'', <draft-dunbar-trill-directory-assisted-edge>, March 2012,

[RBridges] Perlman, et, al ''RBridge: Base Protocol Specification'', <draft-ietf-trill-rbridge-protocol-16.txt>, March, 2010

[RBridges-AF] Perlman, et, al ''RBridges: Appointed Forwarders'', <draft-ietf-trill-rbridge-af-02.txt>, April 2011

[ARMD-Problem] Dunbar, et,al, ''Address Resolution for Large Data Center Problem Statement'', Oct 2010.

[ARP reduction] Shah, et. al., "ARP Broadcast Reduction for Large Data Centers", Oct 2010

Authors' Addresses

Linda Dunbar
Huawei Technologies
5430 Legacy Drive, Suite #175
Plano, TX 75024, USA
Phone: (469) 277 5840
Email: ldunbar@huawei.com

Donald Eastlake
Huawei Technologies
155 Beaver Street
Milford, MA 01757 USA
Phone: 1-508-333-2270
Email: d3e3e3@gmail.com

Radia Perlman
Intel Labs
2200 Mission College Blvd.
Santa Clara, CA 95054-1549 USA
Phone: +1-408-765-8080
Email: Radia@alum.mit.edu

Igor Gashinsky
Yahoo
45 West 18th Street 6th floor
New York, NY 10011
Email: igor@yahoo-inc.com

Intellectual Property Statement

users of this specification can be obtained from the IETF on-line IPR repository at http://www.ietf.org/ipr

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement any standard or specification contained in an IETF Document. Please address the information to the IETF at ietf-ipr@ietf.org.

Disclaimer of Validity

All IETF Documents and the information contained therein are provided on an "AS IS" basis and THE CONTRIBUTOR, THE ORGANIZATION HE/SHE REPRESENTS OR IS SPONSORED BY (IF ANY), THE INTERNET SOCIETY, THE IETF TRUST AND THE INTERNET ENGINEERING TASK FORCE DISCLAIM ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION THEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Acknowledgment

Funding for the RFC Editor function is currently provided by the Internet Society.