TRILL Working Group
Internet Draft
Intended status: Standard Track
Expires: April 2013

L. Dunbar
D. Eastlake
Huawei
Radia Perlman
Intel
Igor Gashinsky
Yahoo
YiZhou Li
Huawei
October 22, 2012

Mechanisms for Directory Assisting TRILL
draft-dunbar-trill-scheme-for-directory-assist-03.txt

Status of this Memo

Copyright Notice

Abstract

This draft describes the mechanisms of using directory server(s) to assist TRILL (Transparent Interconnection of Lots of Links) edge switches in reducing ARP/ND and unknown unicast flooding across TRILL domain 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 switch'' 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 [RFC2119].

Table of Contents

1. Introduction

[TRILL-Directory-Framework] describes the framework for using directory servers to assist TRILL edge nodes to reduce multi-destination ARP/ND and unknow unicast flooding traffic, thus improving TRILL network scalability in data center environment. This draft describes the detailed mechanisms of using directory servers to assist RBridge edge nodes.

Although this document describes directory servers as being part of RBridges, they may be separate end-stations devices (i.e. standalone directory servers), or co-located with an RBridge.

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

DDS: Designated Directory Server for a specific VLAN or a group of VLANs

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.

NDSR: Non-Directory Server RBridge. An RBridge which is not directly connected or embedded with a Directory Server

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 IP&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 IP&MAC&VLAN<->RBridgeEdge mapping table.

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

The push scheme can be accomplished by using the [ESADI] protocol with some simplification or a similar protocol. It is important that it be VLAN scoped.

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.

To achieve this goal, RBridges need to subscribe directory services for the VLANs which they are interested in. Directory servers only send the directory information to an RBridge for the VLANs subscribed by the RBridge. This process eliminates unnecessary entries to be pushed down to RBridges.

RBridges uses the same mechanism as ESADI protocol to announce all the VLANs which they are interested in since ESADI is already VLAN scoped.

3.2. Messages to trigger pushing from directory

In push down model, it is necessary to have a way for RBridge node to request directory server(s) to start pushing down the mapping entries.

RBridges uses the same mechanism as ESADI protocol to announce, in the IS-IS link state database, all the VLANs which they are interested in. The difference from ESADI is that ''Request for Directory'' message is sent to the Push Directory Servers. All other RBridges who are not attached to any directory servers are not going to process this request.

3.3. Actions by Push Directory Servers

A Push Directory Server could be directly attached to an RBridge or embedded in an RBridge through which VLAN scoped directory contents are pushed to other RBridges.

A Push Directory Server could also be a standalone server which is capable of sending required LSPs to announce its ability and push the content to the subscribed RBridges. A standalone Push Directory is almost like a dummy RBridge node which participates in TRILL link state flooding, but doesn't perform RBridge's forwarding, encapsulating, or decapsulation of native Ethernet data frames.

Push Directory servers advertise their availability by turning on a flag bit in the Interested VLANs sub-TLV [rfc6326bis] in their LSP for the VLAN or VLANs for which they offer Push Directory services. If more than one Directory Server is advertising that it can provide Push Directory Service for a particular VLAN, only the Directory Server associated with the RBridge with the highest System ID on the ESADI pseudo-link [EASDI] should push the information for that VLAN. Other Push Directory servers for that VLAN (presumably present for backup) SHOULD NOT push their directory information to avoid unnecessary duplication.

The Directory Server with the highest System ID is called Designated Directory Server - DDS. Different ESADI pseudo-links [EASDI] for different VLANS could have different DDSs.

There is a reserved Multicast Address for all Push Directory Servers.

3.4. Applicable Components of ESADI used in Push Scheme

RBridges that are not associated with any Push Directory Servers should only participate in ESADI for getting the mapping information for the interested VLAN, but SHOULD NOT advertise any locally learned MAC attachment information into ESADI.

When a non-directory server RBridge detects that the information appears to be missing from the directory information, they can

advertise the information only to the Push Directory Servers by using the well-known multicast address for the Push Directory Servers. This behavior is different from ESADI protocol where the locally learned MAC attachment information is advertised to all RBridges who are interested in the VLANs.

ESADI only advertises the locally learned MAC address. But the directory needs to push down IP/MAC/VLAN and their directly attached RBridges.

[draft-eastlake-isis-ia-tlv] describes a proposed TLV to carry the VLAN scoped IP/MAC/VLAN information for all the hosts.

3.5. 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/IP, MAC2/IP, ,, MACn/IP
	VID-2	MAC1/IP, MAC2/IP, ,, MACn/IP
	,,,,,	MAC1/IP, MAC2/IP, ,, MACn/IP
Nickname2	VID-1	MAC1/IP, MAC2/IP, ,, MACn/IP
	VID-2	MAC1/IP, MAC2/IP, ,, MACn/IP
	,,,,,	MAC1/IP, MAC2/IP, ,, MACn/IP
-------	------	----------
	,,,,,	MAC1/IP, MAC2/IP, ,, MACn/IP

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.

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

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

Pull Directory Servers for a particular VLAN are located by looking in the link state database for RBridges that advertise themselves by having the Pull Directory Server flag on in their Interested VLANs sub-TLV [rfc6326bis] for that VLAN. If multiple RBridges indicate that they are Pull Directory Servers for a particular VLAN, then pull requests can be sent to any of them.

Pull Directory requests are sent to the RBridge, or a dummy RBridge, whose LSP contains the Interested VLANs sub-TLV advertising that it is a Pull Directory server for the relevant VLAN. These requests are sent by enclosing them in an RBridge Channel [Channel] message using the Pull Directory channel protocol number (see Section 8). Responses are returned in an RBridge Channel message using the same protocol number.

The requests to Pull Directory Servers are derived from normal ARP [RFC826], ND [RFC4861], RARP [RFC903] messages intercepted by the RBridge, or data frame with unknown DA.

However, additional information is desired from the directory server response in this case, such as the nickname to which an end station (probably identified by IP address) is attached. For this purpose, extended ARP op codes are specified in Table 2.

For a Pull Request derived from an unknown data frame, the RBridge edge node can drop the data frame if there is no response from the directory server after X number of tries.

The requesting RBridge node can cache the mapping and age out MAC&IP&VLAN entries if the entries 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.

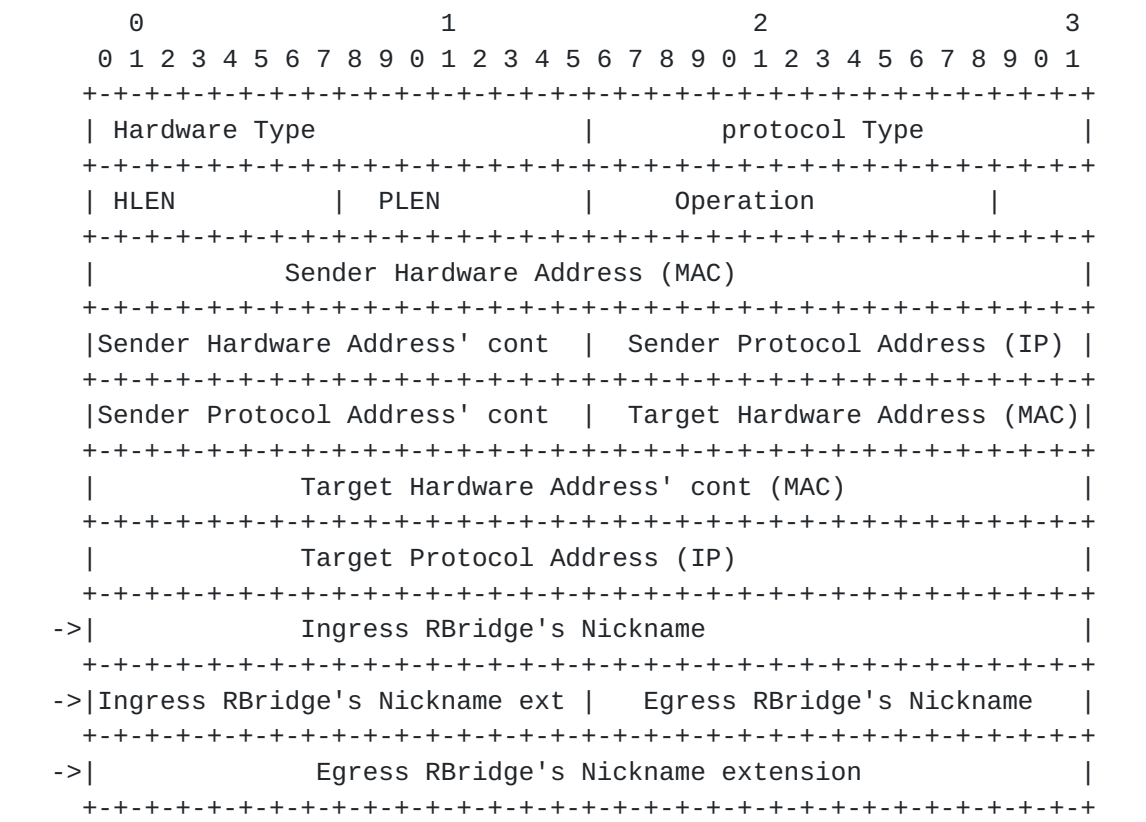

Table 2: Extended fields added to standard ARP reply

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.

It worth noting that the ''Egress RBridge's Nickname in Table 2 is the nickname of the ''Target RBridge'' to which the Target host is attached.

When Pull Directory Server is embedded in an RBridge, the Pull Request would have the ''Ingress RBridge Nickname'' in the TRILL Header. The ''Ingress RBridges Nickname'' field in the Pull Request is for future extension when directory server is not an RBridge.

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.

The mechanisms described above for Push and Pull Directory services make it easy to use Push for some VIDs and Pull for others. In fact, different RBridges can even be configured so that some use Push Directory services and some use Pull Directory services for the same VID if both Push and Pull Directory services are available for that VID. And there can be VIDs for which directory services are not used.

6. Manageability Considerations

TBD.

7. Security Considerations

For general TRILL security considerations, see [RFC6325].

8. IANA Considerations

There are 16 bits for ARP Operation type field [RFC826]. 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, IANA is requested to assign a new ARP Operation code for TRILL Directory Pull services.

IANA is request to allocate a new RBridge Channel protocol number for Pull Directory Services.

IANA is requested to allocate two currently reserved bits in the Interested VLANs field of the Interested VLANs and Spanning Tree Roots sub-TLV [rfc6326bis] to indicate directory servers and to create a sub-registry in the TRILL Parameters Registry, as follows:

Interested VLANs Flag Bits

Registration Procedures: IETF Review

Reference:

Bit	Mnemonic	Description	Reference
0	M4	IPv4 Multicast Router Attached	[rfc6326bis]
1	M6	IPv6 Multicast Router Attached	[rfc6326bis]
2	PS	Push Directory Server	This document
3	PL	Pull Directory Server	This document
16-19	-	available for allocation	[rfc6326bis]

9. Acknowledgments

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

10. References

[TRILL-Directory-Framework] Dunbar, et, al ''TRILL Edge Directory Assistance Framework'', <draft-ietf-trill-directory-framework-02>, work in progress,Oct 2012.

[RFC6325] Perlman, R., Eastlake 3rd, D., Dutt, D., Gai, S., and A. Ghanwani, "Routing Bridges (RBridges): Base Protocol Specification", RFC 6325, July 2011.

[RFC826] Plummer, D., "An Ethernet Address Resolution Protocol", RFC 826, November 1982.

[RFC903] Finlayson, R., Mann, T., Mogul, J., and M. Theimer, "A Reverse Address Resolution Protocol", STD 38, RFC 903, June 1984

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

[RFC4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, September

[RFC6325] Perlman, R., Eastlake 3rd, D., Dutt, D., Gai, S., and A. Ghanwani, "Routing Bridges (RBridges): Base Protocol Specification", RFC 6325, July 2011.

[RFC6439] Perlman, R., Eastlake, D., Li, Y., Banerjee, A., and F. Hu, "Routing Bridges (RBridges): Appointed Forwarders", RFC 6439, November 2011.

[rfc6326bis] Eastlake, D., Banerjee, A., Dutt, D., Perlman, R., and A. Ghanwani, ''TRILL Use of IS-IS'', draft-ietf-isis-rfc6326bis, work in progress

[ESADI] draft-ietf-trill-esadi, work in progress.

[InterfaceAddresses] ''Interface Addresses TLV'', draft-eastlake-isis-ia-tlv, work in progress.

[RFC6439] Perlman, R., Eastlake, D., Li, Y., Banerjee, A., and F. Hu, "Routing Bridges (RBridges): Appointed Forwarders", RFC 6439, November 2011.

[ARMD-Problem] Narten, et,al, ''Problem Statement for ARMD'', June 2012.

[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

YiZhou Li
Huawei
Email: liyizhou@huawei.com

Intellectual Property Statement

Disclaimer of Validity

Acknowledgment

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