INTERNET-DRAFT
Intended Status: Informational
Linda Dunbar
Donald Eastlake
Huawei
Radia Perlman
Intel
Igor Gashinsky
Yahoo
Expires: August 22, 2012
February 23, 2013

TRILL (Transparent Interconnection of Lots of Links): Edge Directory Assistance Framework

<draft-ietf-trill-directory-framework-04.txt>

Abstract

Edge RBridges currently learn the mapping between MAC addresses and their egress RBridges by observing the data packets they ingress or egress or by the TRILL ESADI protocol. When an ingress RBridge receives a data frame whose destination address (MAC&VLAN) that RBridge does not know, the data frame is flooded within the VLAN across the TRILL campus.

This document describes the framework for using directory services to assist edge RBridges in reducing multi-destination frames, particularly unknown unicast frames flooding, and ARP/ND, thus improving TRILL (Transparent Interconnection of Lots of Links) network scalability.

Status of This Memo

Table of Contents

1. Introduction

Edge RBridges (devices implementing [RFC6325], also known as TRILL Switches) currently learn the mapping between destination MAC addresses and their egress RBridges by observing data packets or by the ESADI (End Station Address Distribution Information) protocol. When an ingress RBridge receives a data frame for a destination address (MAC&VLAN) that RBridge does not know, the data frame is flooded within that VLAN across the TRILL campus.

This document describes a framework for using directory services to assist edge RBridges by reducing multi-destination frames, particularly ARP [RFC826], ND [RFC4861], and unknown unicast, improving TRILL network scalability in environments where a directory can be available, such as data centers.

Data center networks differ from enterprise campus networks in several ways that make them attractive for the use of directory assistance, in particular:

1. Data centers, especially Internet and/or multi-tenant data centers tend to have a large number of end stations with a wide variety of applications.
2. Topology is often based on racks and rows. Furthermore, guest operating system assignment to Servers, Racks, and Rows is orchestrated by a Server/VM (virtual machine) Management system, not done at random. So the information necessary for a directory is normally available.
3. Rapid workload shifting in data centers can accelerate the frequency of the physical servers being re-loaded with different applications. Sometimes, the applications loaded into one physical server at different times can belong to different subnets. When a VM is moved to a new location or a server is loaded with a new application with different IP/MAC addresses, it is more likely that the destination address of data packets sent out from those VMs are unknown to their attached edge RBridges.
4. With server virtualization, there is an increasing trend to dynamically create or delete VMs when demand for resource changes, to move VMs from overloaded servers to less loaded servers, or to aggregate VMs onto fewer servers when demand is light. This results in the more common occurrence of multiple subnets on the same port at the same time and a higher change rate for VMs than for physical servers.

Both items 3 and 4 above can lead to applications in one subnet being placed in different locations (racks or rows) or one rack having applications belonging to different subnets.

2. Terminology

The terms "Subnet" and "VLAN" are used interchangeably in this document because it is common to map one subnet to one VLAN.

Bridge: IEEE Std 802.1Q-2011 compliant device [802.1Q]. In this document, Bridge is used interchangeably with Layer 2 switch.

EoR: End of Row switches in data center. Also known as aggregation switches.

End Station: Guest OS running on a physical server or on a virtual machine. An end station in this document has at least one IP address and at least one MAC address.

IS-IS: Intermediate System to Intermediate System. TRILL uses IS-IS [IS-IS] [RFC6326].

RBridge: "Routing Bridge", an alternative name for a TRILL switch.

Station: A node, or a virtual node, with IP and/or MAC addresses.

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

TRILL: Transparent Interconnection of Lots of Links [RFC6325]

TRILL switch: A device implementing the TRILL protocol [RFC6325]

VM: Virtual Machine

3. Impact of Massive Number of End Stations

This section discusses the impact of a massive number of end stations in a TRILL campus using Data Centers as an example.

3.1 Issues of Flooding Based Learning in Data Centers

It is common for Data Center networks to have multiple tiers of switches, for example, one or two Access Switches for each server rack (ToR), aggregation switches for some rows (or EoR switches), and some core switches to interconnect the aggregation switches. Many aggregation switches deployed in data centers have high port density. It is not uncommon to see aggregation switches interconnecting hundreds of ToR switches.

```
                    +-------+         +------+
                  +/------+ |       +/-----+ |
                  | Aggr11| + ----- |AggrN1| +    EoR Switches
                  +---+---+/        +------+/
                   /     \            /      \
                  /       \          /        \
               +---+    +---+      +---+     +---+
               |T11|... |T1x|      |T21| ..  |T2y| ToR switches
               +---+    +---+      +---+     +---+
                 |        |          |         |
               +-|-+    +-|-+      +-|-+     +-|-+
               |   |... |   |      |   | ..  |   |
               +---+    +---+      +---+     +---+Server racks
               |   |... |   |      |   | ..  |   |
               +---+    +---+      +---+     +---+
               |   |... |   |      |   | ..  |   |
               +---+    +---+      +---+     +---+
```

Figure 1: Typical Data Center Network Design

The following problems could occur when TRILL is deployed in a data center with large number of end stations and the end stations in one subnet/VLAN could be placed under multiple edge RBridges:

- Unnecessary filling of slots in the MAC address learning table of edge RBridges, e.g. RBridge T11, due to T11 receiving broadcast / multicast traffic (e.g. ARP/ND, cluster multicast, etc.) from end stations under other edge RBridges that are not actually communicating with any end stations attached to T11.

- Packets being flooded across TRILL campus when their destination MAC addresses are not in ingress RBridge's MAC address to egress

RBridge cache.

- In an environment where VMs migrates, there is higher chance of cached information becoming invalid, causing traffic to be black-holed by the ingress RBridge, that is, persistently sent to the wrong egress RBridge. If VMs do not flood gratuitous ARP/ND or VDP [802.1Qbg] messages upon arriving at new locations, the ingress nodes might not have MAC entries for the MAC of the newly arrived VMs, causing unknown address flooding.

3.2 Two Examples

Consider a data center with 1600 server racks. Each server rack has at least one ToR switch. The ToR switches are further divided into 8 groups, with each group being connected by a set of aggregation switches. There could be 4 to 8 aggregation switches in each set to achieve load sharing for traffic to/from server racks. If TRILL is deployed in this data center environment, let's consider the following two scenarios for the TRILL campus boundary:

- Scenario #1: TRILL campus boundary starts at ToR switches:

  If each server rack has one ToR, there are 1600 edge RBridges. If each rack has two ToR switches, then there will be 3200 edge RBridges

  In this scenario, the TRILL domain will have more than 1600 (or 3200) + 8*4 (or 8*8) nodes, which is a large IS-IS domain. Even though a mesh IS-IS domain can scale up to thousands of nodes, it is challenging for aggregation switches to handle IS-IS link state advertisement among hundreds of parallel ports.

- Scenario #2: TRILL campus boundary starts at the aggregation switches:

  With the same assumptions as before, the number of nodes in the TRILL campus will be less than 100, and aggregation switches don't have to handle IS-IS link state advisements among hundreds of parallel ports.

  However, the number of MAC&VLAN<->Egress RBridge Mapping entries to be learned and managed by RBridge edge node can be very large. In the example above, each edge RBridge has 200 edge ports facing the ToR switches. If each ToR has 40 downstream ports facing servers and each server has 10 VMs, there could be 200*40*10 = 80000 end stations attached. If all those end stations belong to 1600 VLANs (i.e. 50 per VLAN) and each VLAN has 200 end stations, then under the worst-case scenario, the total number of MAC&VLAN entries to be learned by the edge

RBridge can be 1600*200=320000, which is very large.

4. Benefits of Directory Assisted Edge RBridge

In some environments, particularly data centers, the assignment of applications to servers, including rack and row selection, is orchestrated by Server (or VM) Management System(s). That is, there is a database or multiple databases (distributed model) that have the knowledge of where each application is placed. If the application location information can be fed to RBridge edge nodes, in some form of Directory Service, then there is much less chance of RBridge edge nodes receiving unknown MAC destination address, therefore less chance of flooding.

Avoiding unknown unicast address flooding to the TRILL campus is especially valuable in the data center environment because there is a higher chance of an edge RBridge receiving packets with unknown unicast destination address and broadcast / multicast messages due to VM migration and servers being loaded with different applications. When a VM is moved to a new location or a server is loaded with a new application with a different IP/MAC addresses, it is more likely that the destination address of data packets sent out from those VMs are unknown to their attached edge RBridges. In addition, gratuitous ARP (IPv4, [RFC826]) or Unsolicited Neighbor Advertisement (IPv6, [RFC4861]) sent out from those newly migrated or activated VMs have to be flooded to other edge RBridges that have VMs in the same subnets.

The benefits of using directory assistance include:

- Avoid flooding unknown unicast destination address across TRILL campus. The Directory enforced MAC&VLAN <-> Egress RBridge mapping table can determine if a data packet needs to be forwarded across TRILL campus.

  When multiple RBridge edge ports are connected via a bridged LAN to end stations (servers/VMs), a directory assisted edge RBridge won't need to flood unknown unicast destination data frames to all ports of the edge RBridges in the frame's VLAN when it ingresses a frame. It can depend on the directory to tell it where the destination is. When the directory doesn't have the needed information, the frames can be dropped or flooded depending on the policy configured.

- Reduce flooding of decapsulated Ethernet frames with unknown MAC destination address to a bridged LAN connected to RBridge edge ports.

  When an RBridge receives a TRILL data packet whose destination Nickname matches with its own, the normal procedure is for the

RBridge to decapsulate it and forward the decapsulated Ethernet frame to the directly attached bridged LAN. If the destination

MAC is unknown, the RBridge floods the decapsulated Ethernet frame out all ports in the fame's VLAN. With directory assistance, the egress RBridge can determine if the MAC destination address in a frame matches any end stations attached via the bridged LAN. Frames can be discarded if their destination addresses do not match.

- Reduce the amount of MAC&VLAN <-> Egress RBridge mapping maintained by edge RBridges. There is no need for an edge RBridge to keep MAC entries of remote end stations that don't communicate with the end stations locally attached.

- Eliminate ARP/ND being broadcasted or multi-casted through the TRILL core.

5. Generic operation of Directory Assistance

5.1 Information in Directory for Edge RBridges

To achieve the benefits of directory assistance for TRILL, the corresponding directory server entries will need, at a minimum, the following logical attributes:

[{IP, MAC/VLAN, {list of attached RBridge nicknames}, {list of interested RBridges}]

The {list of attached RBridges} are the edge RBridges to which the host (or VM) specified by the [IP or MAC/VLAN] in the entry is attached. The {list of interested RBridges} are the remote RBridges that might have attached hosts to communicate with the host in this entry.

When a host has multiple IP addresses, there will be multiple entries.

The {list of interested RBridges} could get populated when an RBridge queries for information, or pushed down from management systems. The list is used to notify those RBridges when the host (specified by the IP/MAC/VLAN) in the entry connectivity to its attached RBridges changes. An explicit list in the directory is not needed as long as the interested RBridges can be determined.

There are two different models for Directory assistance to edge RBridges: Push Model and Pull Model.

5.2 Push Model and Requirements

Under this model, Directory Server(s) push down the MAC&VLAN <-> Egress RBridge mapping for all the end stations that might communicate with end stations attached to an RBridge edge node. If the packet's destination address can't be found in the MAC&VLAN<->Egress RBridge table, the ingress RBridge could be configured to:

- simply drop a data packet,
- flood it to TRILL campus, or
- start the pull process to get information from directory server(s)

It may not be necessary for every edge RBridge to get the entire mapping table for all the end stations in a campus. There are many ways to narrow the full set down to a smaller set of remote end stations that communicate with end stations attached to an edge RBridge. A simple approach is to only pushing down the mapping for the VLANs that have active end stations under an edge RBridge. This approach can reduce the number of mapping entries being pushed down.

However, the Push Model usually will push down more entries of MAC&VLAN<->Egress RBridge mapping to edge RBridges than needed. Under the normal process of edge RBridge cache aging and unknown destination address flooding, rarely used mapping entries would have been removed. But it can be difficult for Directory Servers to predict the communication patterns among applications within one VLAN. Therefore, it is likely that the Directory Servers will push down all the MAC&VLAN entries if there are end stations in the VLAN being attached to the edge RBridge. This is a disadvantage of the Push Model compared with the Pull Model described below.

In the Push Model, it is necessary to have a way for an RBridge node to request directory server(s) to start pushing down the mapping entries. This method should at least include the VLANs enabled on the RBridge, so that directory server doesn't need to push down the entire mapping entries for all the end stations in the campus. An RBridge must be able to get mapping entries when it is initialized or restarted.

The Push Model's detailed method and any handshake mechanism between RBridge and Directory Server(s) is beyond the scope of this framework document.

When a directory server needs to push down a large number of entries to edge RBridges, efficient data organization should be considered. For example, with one edge RBridge Nickname being associated with all attached end stations' MAC addresses and VLANs as shown below:

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

Table 1: Summarized table pushed down from directory

Whenever there is any change in MAC&VLAN <-> Egress RBridge mapping, that can be triggered by end stations being added, moved, or de-commissioned, an incremental update can be sent to the edge RBridges which are impacted by the change. Therefore, something like a sequence number has to be maintained by directory servers and RBridges. Detailed mechanisms will be specified in a separate document.

5.3 Pull Model and Requirements

Under this model, an RBridge pulls the MAC&VLAN<->Egress RBridge mapping entry from the directory server when its cache doesn't have the entry. There are several possibilities to trigger the pulling process:

- The RBridge edge node can send a pull request whenever it receives an unknown MAC destination, or
- The RBridge edge node can intercept all ARP/ND requests and forward them or appropriate requests to the Directory Server(s) that has the information on where the target stations are located.
- The Pull Directory response could indicate that the address being queried is unknown or that the requestor is administratively prohibited from getting an informative response.

By using a Pull Directory, the frame with unknown MAC destination address doesn't have to be flooded across TRILL domain and the ARP/ND requests don't have to be broadcast or multicast across the TRILL domain.

The ingress RBridge can cache the response pulled down from the directory. The timer for cache should be short in an environment where VMs move frequently. The cache timer could be configured by management system or could be sent down along with the Pulled reply by the directory server(s). It is important that the cached information be kept consistant with the actual placement of addresses in the campus; therefore, there needs to be some mechanism by which RBridges that have pulled information that has not expired can be informed when that information changes or the like.

One advantage of the Pull Model is that edge RBridges can age out MAC&VLAN entries if they haven't been used for a certain configured period of time or a period of time provided by the Directory. Therefore, each edge RBridge will only keep the entries that are frequently used, so mapping table size will be smaller. Edge RBridges would query the Directory Server(s) for unknown MAC destination addresses in data frames or ARP/ND and cache the response. When end stations attached to remote edge RBridges rarely communicate with the locally attached end stations, the corresponding MAC&VLAN entries would be aged out from the RBridge's cache.

An RBridge waiting for response from Directory Servers upon receiving a data frame with an unknown destination address is similar to an L2/L3 boundary router waiting for ARP/ND response upon receiving an IP data packet whose destination IP is not in the router's IP/MAC cache table. Most deployed routers today do hold the packet and send ARP/ND requests to the target upon receiving a packet with destination IP not in its IP to MAC cache. When ARP/ND replies are received, the router will send the data packet to the target. This practice minimizes flooding when targets don't exist in the subnet.

When the target doesn't exist in the subnet, routers generally re-send an ARP/ND request a few more times before dropping the packets. So, the holding time by routers to wait for ARP/ND response can be longer than the time taken by the Pull Model to get IP to MAC mapping from a directory if target doesn't exist in the subnet.

For RBridges with mapping entries being pushed down from directory server, they can be configured to use Pull model for targets which don't exist in the mapping data pushed down.

A separate document will specify the detailed messages and mechanism for edge RBridges to pull information from directory server(s).

6. Recommendation

TRILL should provide a directory assisted approach. This document describes a basic framework of using a directory assisted approach for RBridge edge nodes. More detailed mechanisms will be described in a separate document or documents.

7. Security Considerations

Accurate mapping of IP addresses into MAC addresses and of MAC addresses to the RBridge from which they are reachable is important to the correct delivery of information. The security of specific directory assisted mechanisms will be discussed in the document or documents specifying those mechanisms.

For general TRILL security considerations, see [RFC6325].

8. IANA Considerations

This document requires no IANA actions. RFC Editor: please delete this section before publication.

9. Acknowledgements

Thanks for comments and review from the following:

David Black, Erik Nordmark

The document was prepared in raw nroff. All macros used were defined within the source file.

10. References

10.1 Normative References

As an Informational document, this draft has no Normative References.

10.2 Informative References

[802.1Q] - IEEE Std 802.1Q-2011, "IEEE Standard for Local and metropolitan area networks - Virtual Bridged Local Area Networks", May 2011.

[802.1Qbg] - IEEE Std 802.1Qbg-2012, ''Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks --- Edge Virtual Bridging'', July 2012.

[IS-IS] - ISO/IEC, "Intermediate system to Intermediate system routeing information exchange protocol for use in conjunction with the Protocol for providing the Connectionless-mode Network Service (ISO 8473)", ISO/IEC 10589:2002.

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

[RFC4861] - Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, September 2007.

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

[RFC6326] - Eastlake, D., Banerjee, A., Dutt, D., Perlman, R., and A. Ghanwani, "Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS", RFC 6326, July 2011.

Authors' Addresses

Linda Dunbar
Huawei Technologies
5430 Legacy Drive, Suite #175
Plano, TX 75024, USA
Phone: +1-469-277-5840
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 USA
Email: igor@yahoo-inc.com

Copyright, Disclaimer, and Additional IPR Provisions