TRILL working group
Internet Draft
Category: Informational

L. Dunbar
D. Eastlake
Huawei
Padia Perlman
Intel
Igor Gashinsky
Yahoo

Expires: December 2013

July 7, 2012

TRILL Edge Directory Assistance Framework

draft-ietf-trill-directory-framework-00

Status of this Memo

Copyright Notice

Abstract

Edge RBridges currently learn the mapping between MAC addresses and their egress RBridges by observing the data packets traversed through. When ingress RBridge receives a data frame with its destination address (MAC&VLAN) unknown, the data frame is flooded across the TRILL campus. When there are more than one RBridge ports connected to one bridged LAN, only one of them can be designated as AF port for forwarding/receiving traffic for each LAN, the rest have to be blocked for that LAN.

This draft describes the framework for using directory service to assist edge RBridges to improve TRILL network scalability in data center environment.

Conventions used in this document

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

Data center networks are different from enterprise campus networks in several ways, in particular:

1) Data centers, especially Internet or multi-tenant data centers tend to have a large number of end stations with a wide variety of applications.
2) Topology is usually based on racks and rows.
   - Guest OSs assignment to Servers, Racks, and Rows is orchestrated by a Server/VM Management system, not at random.
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 to one physical server at different times can belong to different subnets.
4) With server virtualization, there is an ever-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.

Both 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.

This draft describes why and how Data Center TRILL networks can be optimized by utilizing a directory assisted approach.

2. Terminology

AF Appointed Forwarder [RBridge-AF]

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

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

and at least one MAC address, which could be in DA or SA field of a data frame.

RBridge: A device implementing the TRILL protocol [RBridge]

RSTP: Rapid Spanning Tree Protocol

SA: Source Address

Station: A node, or a virtual node, with IP and/or MAC addresses, which could be in the DA or SA of a data frame.

STP: 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. Impact on RBridge Campus of Massive Number of stations in a DC

3.1. Issues of Flooding Based Learning in DCs

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.

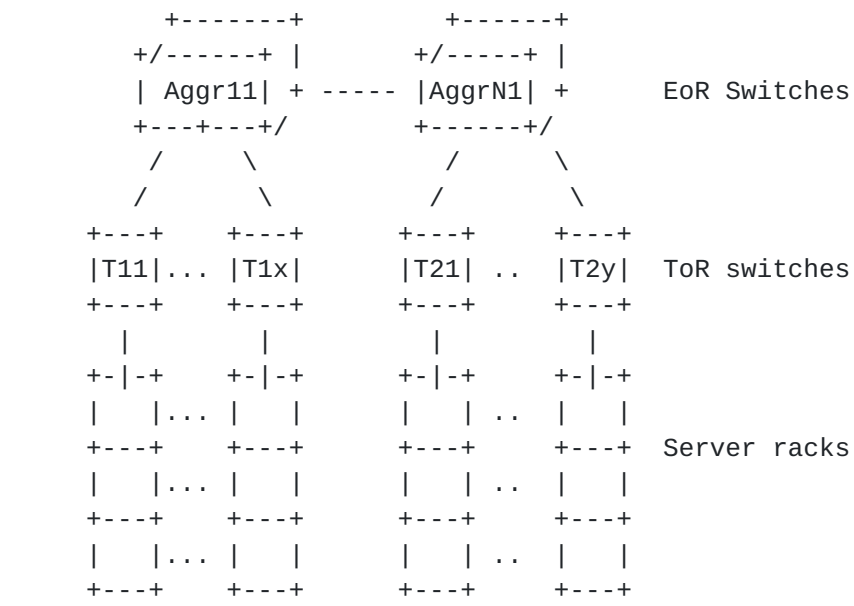

Figure 1: Typical Data Center Network Design

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 MAC table of edge RBridges RB1, due to RB1 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 RB1.
- Some edge RBridge ports being blocked for user traffic when there are more than one RBridge ports connected to an edge bridged LAN. When there are multiple RBridge ports connected to a bridged LAN, only one (the AF port) can forward/receive traffic for that bridged LAN or VLAN. The rest have to be blocked for forwarding/receiving traffic for that VLAN. When a rack has dual uplinks to two different ToR switches (or edge RBridges), some links may not be fully utilized.
- Packets being flooded across TRILL campus when their DAs are not in ingress RBridge's cache.
- In an environment where VMs migrates, there is higher chance of cached entries becoming invalid, causing traffic to be black holed by the egress RBridge. If VMs send out gratuitous ARP/ND or IEEE802.1Qbg's VDP upon arriving at new locations, the ingress nodes might not have the MAC entries for the newly arrived VMs, causing more unknown flooding.

3.2. Some 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 uplink to one ToR, there are 1600 edge RBridges. If each rack has dual uplinks to two ToR switches, then there will be 3200 edge RBridges

  In this scenario, the RBridge domain will have more than 1600 (or 3200) + 8*4 (or 8*8) nodes, which is quite a large IS-IS domain. Even though a mesh IS-IS domain can scale up to thousands of nodes, it is very 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 assumption 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.

  But bridged LANs are formed under the aggregation switches in this scenario. With aggregation switches being the RBridge edge nodes, multiple RBridge edge ports could be connected to one bridged LAN. To avoid potential loops, TRILL requires only one of multiple RBridge edge ports connected to each VLAN being designated as Appointed Forwarder (AF port), and other ports being blocked for native frames in that VLAN.

  There is also the possibility of loops on the bridged LAN attached to RBridge edge ports unless STP/RSTP is running. Running traditional Layer 2 STP/RSTP on the bridged LAN in this environment may be overkill because the topology among the ToR switches and aggregation switches is very simple.

In addition, 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 DC

In data center environment, applications placement to servers, racks, and rows 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 RBridge edge nodes won't need to flood data frames with unknown DA across the TRILL campus.

Avoiding unknown DA flooding to TRILL campus is especially valuable in data center environment because there is higher chance of an edge RBridge receiving packets with unknown DA 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 different IP/MAC addresses, it is more likely that the DA of data packets sent out from those VMs are unknown to their attached edge RBridges. In addition, gratuitous ARP (IPv4) or Unsolicited Neighbor Advertisement (IPv6) 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 DA 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 DA data frames to all ports of the edge RBridges in the frame's VLAN. Therefore, it is no longer necessary to designate one Appointed Forwarder among all

the RBridge Edge ports connected to a bridge LAN. All edge RBridge ports can forward/receive native traffic.

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

  When an RBridge receives a TRILL frame whose destination Nickname matches with its own, the normal procedure is for the RBridge to decapsulate the TRILL header and forward the decapsulated Ethernet frame to the directly attached bridged LAN. If the destination MAC is unknown, the normal Ethernet switch's flooding will occurs to the decapsulated Ethernet frame. With directory assistance, the egress RBridge can determine if DA in a frame matches with any end stations attached via the bridged LAN. Frames can be discarded if their DAs 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 which don't communicate with the end stations locally attached.

5. Generic operation of Directory Assistance

5.1. Information in Directory for Edge Bridges

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

[IP, MAC, attached RBridge nickname, {list of interested RBridges}]

The {list of interested RBridges} would get populated when an RBridge queries for information, or pushed down from management systems. The list is used to notify those RBridges whose connectivity to VMs changes due to VM migration or link failures.

There can be two different models for RBridge edge node to be assisted by Directory Service: Push Model and Pull Model.

5.2. Push Model

Under this model, Directory Server(s) push down the MAC&VLAN <-> Egress RBridge mapping for all the end stations which might communicate with end stations attached to an RBridge edge node.

Under this model, it is recommended that the ingress RBridge simply drops a data packet (instead of flooding to TRILL campus) if the packet's destination address can't be found in the MAC&VLAN<->Egress RBridge mapping table.

It may not be necessary for every edge RBridge to get the entire mapping table for all the end stations in a data center. 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 of only pushing down the mapping for the VLANs which have active end stations under an edge RBridge 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. Under the normal process of edge RBridge cache aging and unknown DA 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 major disadvantage of the Push Model.

In the Push Model, it is necessary to have a message for an RBridge node to request directory server(s) to start pushing down the mapping entries. This message 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 data center. An RBridge node can use this message to get mapping entries when it is initialized or restarted.

The detailed message format and hand-shake mechanism between RBridge and Directory Server(s) is beyond the scope of this framework draft.

When directory server needs to push down a very large number of entries to edge RBridges, summarization 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	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 <-> 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 described in a separate draft.

5.3. Pull Model

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 options to trigger the pulling process. For example, the RBridge edge node can send a pulling request whenever it receives an unknown DA, or the RBridge edge node can simply intercept all ARP/ND requests and forward them to the Directory Server(s) that has the information on where the target stations are located. The ingress RBridge can cache the mapping pulled down from the directory.

One advantage of the Pull Model is that edge RBridge can age out MAC&VLAN entries if they haven't been used for a certain period of time. Therefore, each edge RBridge will only keep the entries which are frequently used, so mapping table size can be smaller. Edge RBridges would query the Directory Server(s) for unknown DAs 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.

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

When the target doesn't exist in the subnet, routers generally re-send ARP/ND request a few more times before dropping the packets. Therefore, 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-MAC mapping from directory if target doesn't exist in the subnet.

A separate draft will describe the detailed messages and mechanism for edge RBridge to pull information from directory server(s).

6. Conclusion and Recommendation

The traditional RBridge learning approach of observing data plane can no longer keep pace with the ever growing number of end stations in Data center.

Therefore, we suggest TRILL consider directory assisted approach(es). This draft only describes the basic framework of using directory assisted approach for RBridge edge nodes. More complete mechanisms will be described in separate drafts.

7. Security Considerations

TBD

8. IANA Considerations

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

9. Acknowledgements

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

10. References

[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

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

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