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

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

March 11, 2012

Directory Assisted RBridge Edge
draft-dunbar-trill-directory-assisted-edge-05.txt

Status of this Memo

Copyright Notice

Abstract

RBridge edge nodes currently learn the mapping between MAC addresses and their corresponding RBridge edge nodes by observing the data packets traversed through. When ingress RBridge receives a data packet with its destination address (MAC&VLAN) unknown, the data packet is flooded across RBridge domain. 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 of using directory assisted RBridge edge to improve TRILL network scalability 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 3
2. Terminology 3
3. Impact on RBridge domain of massive number of hosts in Data Center 4
3.1. Issues of Flooding Based Learning in Data Centers 4
3.2. Some Examples 5
4. Benefits of Directory Assisted RBridge Edge in DC Environment 7
5. Generic operation of Directory Assistance..... 8
5.1. Information in Directory Servers for TRILL..... 8
5.2. Push Model 8
5.3. Pull model: 10
6. Conclusion and Recommendation..... 11
7. Manageability Considerations..... 11
8. Security Considerations..... 11
9. IANA Considerations 11
10. Acknowledgments 11
11. References 12

Authors' Addresses 12
Intellectual Property Statement..... 13
Disclaimer of Validity 13

1. Introduction

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

1. Data centers, especially Internet or multi-tenant data centers, tend to have large number of hosts with a wide variety of applications.
2. Topology is based on racks and rows.
   Hosts assignment to Servers, Racks, and Rows is orchestrated by Server/VM Management system, not at random.
3. Rapid workload shifting in data centers can accelerate the frequency of one physical server being re-loaded with different applications. Sometimes, applications re-loaded to one physical server at different time 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, or to aggregate VMs onto fewer servers when demand is light.

Both 3) and 4) above can lead to hosts in one subnet being placed under different locations (racks or rows) or one rack having hosts 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 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. Impact on RBridge domain of massive number of hosts in Data Center

3.1. Issues of Flooding Based Learning in Data Centers

It is common for Data Center networks to have multiple tiers of switches, e.g. 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 are high port density switches. It is not uncommon to see aggregation switches interconnecting hundreds of ToR switches.

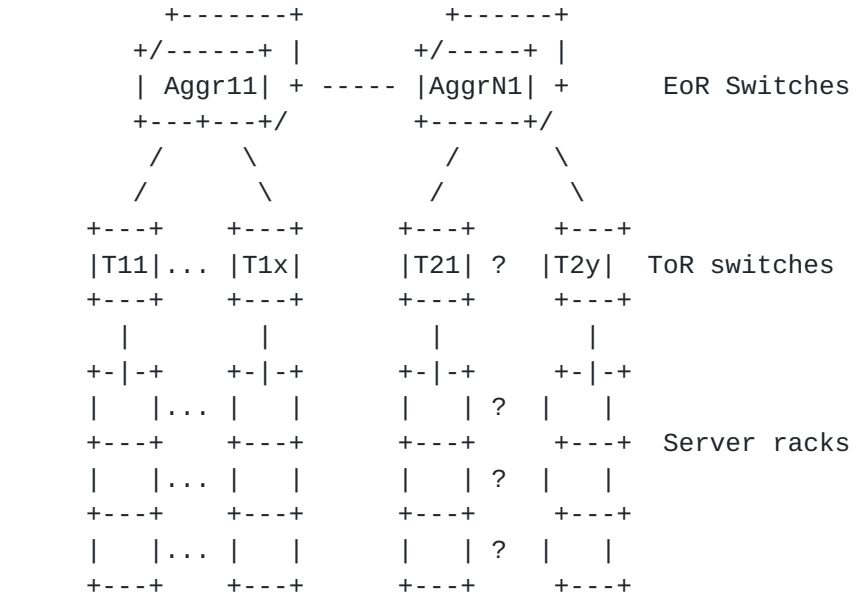

Figure 1: Typical Data Center Network Design

When TRILL is deployed in a data center with large number of hosts, with the possibility of hosts in one subnet/VLAN being placed under multiple edge RBridges and each edge RBridge having hosts from different subnets/VLANs, the following problems will occur:

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

Some edge RBridge ports being blocked for user traffic when there are more than one RBridge ports connected to one bridged LAN. When there are multiple RBridge ports connected to a bridged LAN, only one, i.e. the AF port, can forward/receive traffic for that bridged LAN (i.e. 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 (RBridge Edges), which is very common in data center environment, some links can't be fully utilized.

Packets being flooded across RBridge domain 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 or re-flooded 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 to 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 to be deployed in this data center environment, let's consider following two scenarios for the TRILL domain boundary:

Scenario #1: TRILL domain 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 domain boundary starts at the aggregation switches:

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

But in this scenario, aggregation switches' downstream ports/links to ToR switches form the bridged LAN with links from ToR switches to servers. 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 one VLAN being designated as Appointed Forwarder (AF port) for forwarding native traffic across RBridge domain for that VLAN. That means other ports/links are blocked for native frames in that VLAN.

There is also 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<->RBridgeEdge 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 hosts attached. If all those hosts belong to 1600 VLANs (i.e. 50 per VLAN) and each VLAN has 200 hosts, then under the worst case scenario, the total number

of MAC&VLAN entries to be learned by the RBridge edge can be 1600*200=320000, which is very large.

4. Benefits of Directory Assisted RBridge Edge in DC Environment

In data center environment, applications placement to servers, racks, and rows is orchestrated by Server (or VM) Management System(s). I.e. there is a database or multiple ones (distributed model) which have the knowledge of where each host is located. If that host 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 RBridge domain.

Avoiding unknown DA flooding to RBridge domain is especially valuable in data center environment because there is higher chance of an RBridge edge 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 hosts are unknown to their attached RBridge edges. In addition, gratuitous ARP (IPv4) or Unsolicited Neighbor Advertisement (IPv6) sent out from those newly migrated or activated hosts have to be flooded to other RBridge edges which have hosts in the same subnets.

The benefits of using directory assistance include:

Avoid flooding unknown DA across RBridge domain. The Directory enforced MAC&VLAN <-> RBridgeEdge mapping table can determine if a data packet needs to be forwarded across RBridge domain.

When multiple RBridge edge ports are connected via bridged LAN to hosts (servers/VMs), a directory assisted RBridge edge won't need to flood unknown DA data frames to all ports of the RBridge edge. Under this circumstance, there is no chance for those data frames looping among multiple ports of RBridge edge. Therefore, it is no longer necessary to designate one Appointed Forwarder among all the RBridge Edge ports connected to a bridge LAN, which means that all RBridge ports can forward/receive traffic.

Reduce flooding 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 its directly attached bridged LAN. If the destination MAC is unknown, the decapsulated Ethernet frame is flooded in the LAN. With directory assistance, the RBridge edge can determine if DA in a frame matches with any hosts attached via the bridged LAN. Therefore, frames can be discarded if their DAs do not match.

Reduce the amount of MAC&VLAN <-> RBridgeEdge mapping maintained by RBridge edge. There is no need for an RBridge edge to keep the MAC entries for hosts which don't communicate with hosts attached to the RBridge edge.

5. Generic operation of Directory Assistance

5.1. Information in Directory Servers for TRILL

To achieve the benefits of directory service for TRILL, the corresponding directory server will need minimum 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 if VMs to RBridge's connectivity changes due to VMs 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 <-> 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.

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

However, it is inevitable that RBridge edge's MAC&VLAN<->RBridgeEdge mapping table will have more entries than they really need under the Push Model. When hosts attached to one RBridge Edge rarely communicate with hosts attached to different RBridge edges even though they are on the same VLAN, the normal process of RBridge edge's unknown DA flooding, learning and 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, it is likely that the Directory Servers will push down all the MAC&VLAN entries if there are hosts in the VLAN being attached to the RBridge Edge. This is a major disadvantage of push down model.

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 VLANs enabled on the RBridge, 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.

The detailed message format and hand-shake mechanism between RBridge and Directory Server(s) will be described in a separate draft because this draft only focuses on the framework of directory assisted Edge.

When directory pushes down the entire mapping to an edge RBridge for the very first time, there usually are many entries. To minimize the number of entries 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
	...	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. Therefore, something like 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, ''RBridge'' pulls the MAC&VLAN<->RBridgeEdge mapping entry from the directory server when needed. 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 RBridge edge node can simply intercept all ARP/ND requests and forward them to the Directory Server(s) that has the information on where each host is located. RBridge ingress node can cache the mapping pulled down from the directory.

One advantage of the Pull Model is that 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. RBridge edge would query the Directory Server(s) for unknown DAs in data frames or ARP/ND and cache the response. When hosts attached to one RBridge Edge rarely communicate with hosts attached to different RBridge edges even though they are on the same VLAN, the corresponding MAC&VLAN entries would be aged out from the RBridge's cache.

Some people are concerned of the performance with RBridge waiting for response from Directory Servers upon receiving a data frame with unknown DA. Actually this waiting practice is a common router behavior. Most deployed routers today do hold the packets and send an ARP/ND 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 RBridge edge 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 hosts 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. Manageability Considerations

TBD.

8. Security Considerations

TBD.

9. IANA Considerations

TBD

10. Acknowledgments

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

11. 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

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

Disclaimer of Validity

Acknowledgment

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