TRILL working group L. Dunbar
Internet Draft D. Eastlake
Intended status: Standard Track Huawei
Expires: Sept 2014 Radia Perlman
Intel
I. Gashinsky
Yahoo
January 11, 2014

Directory Assisted TRILL Encapsulation
draft-dunbar-trill-directory-assisted-encap-05.txt

Status of this Memo

Copyright Notice

Abstract

This draft describes how data center network can benefit from non-RBridge nodes performing TRILL encapsulation with assistance from directory service.

Conventions used 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 1. .

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

Table of Contents

1. Introduction

This draft describes how data center network can benefit from non-RBridge nodes performing TRILL encapsulation with assistance from directory service.

[RFC7067] describes the framework for RBridge edge to get MAC&VLAN<->RBridgeEdge mapping from a directory service in data center environment instead of flooding unknown DAs across TRILL domain. When directory is used, any node, even non-RBridge node, can perform the TRILL encapsulation. This draft is to demonstrate the benefits of non-RBridge nodes performing TRILL encapsulation.

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

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

VM: Virtual Machines

3. Directory Assistance to Non-RBridge

With directory assistance [RFC7067], a non-RBridge can determine if a packet needs to be forwarded across the RBridge domain. Suppose the RBridge domain boundary starts at network switches (i.e. not virtual switches embedded on servers), a directory can assist Virtual Switches embedded on servers to encapsulate proper TRILL header by providing the information of the egress RBridge edge to which the target is attached. If a target is not attached to other RBridge edge nodes based on the directory [RFC7067], the non-RBridge node can forward the data frames natively, i.e. not encapsulating any TRILL header.

```
          \              +--------+        +-------+ TRILL Domain/
           \           +/------+ |       +/-----+ |            /
            \          | Aggr11| + ----- |AggrN1| +           /
             \         +---+---+/        +------+/           /
              \         /     \            /      \         /
               \       /       \          /        \       /
                \   +---+    +---+      +---+     +---+   /
                 \- |T11|... |T1x|      |T21| ?  |T2y|---
                    +---+    +---+      +---+     +---+
                      |        |          |         |
                    +-|-+    +-|-+      +-|-+     +-|-+
                    |   |... | V |      | V | ..  | V |<-VSwitch
                    +---+    +---+      +---+     +---+
                    |   |... | V |      | V | ..  | V |
                    +---+    +---+      +---+     +---+
                    |   |... | V |      | V | ..  | V |
                    +---+    +---+      +---+     +---+
```

Figure 1: TRILL domain in typical Data Center Network

When a TRILL encapsulated data packet reaches the ingress RBridge, the ingress RBridge can simply forward the pre-encapsulated packet to the RBridge that is specified in the DA field of the TRILL header of the data frame. When the ingress RBridge receives a native Ethernet frame, it only forward the data frame to the directly attached bridged LAN.

Under this environment, the ingress RBridge doesn't flood or send the received Ethernet data frames to TRILL domain when the DA in the Ethernet data frames is unknown or instructed by the directory not to be sent across TRILL domain. Under this scheme, for an RBridge with multiple ports connected to a bridged LAN, data frames received from TRILL domain, decapsulated and forwarded to the bridged LAN via one port, and flooded back to the RBridge via another port, won't be encapsulated again and forwarded back TRILL domain.

That means there is no need to worry about AF ports and all RBridge edge ports connected to one bridged LAN can receive and forward pre-encapsulated traffic, which greatly improves the overall network utilization.

Note: [RFC6325] Section 4.6.2 Bullet 8 specifies that an RBridge port can be configured to accept TRILL encapsulated frames from a neighbor that is not an RBridge.

When data frames do not need to be sent across RBridge domain, they are switched by all nodes/ports per IEEE802.1Q and RBridge edge will not encapsulate and forward those data frames across RBridge domain.

When a pre-encapsulated TRILL frame arrives at an RBridge whose nickname matches with the destination nickname in the TRILL header, the processing is exactly same as normal, i.e. it decapsulates the received TRILL frame and forwards the decapsulated Ethernet frame to the target attached to its edge ports. If the DA of the decapsulated Ethernet frame is not in the egress RBridge's FDB, the egress RBridge can flood the decapsulated Ethernet frame to all hosts attached.

We call a node that only performs the TRILL encapsulation but doesn't participate in RBridge's IS-IS routing a ''TRILL Encapsulating node'' or ''Simplified RBridge''. The TRILL Encapsulating Node gets the MAC&VLAN<->RBridgeEdge mapping table pushed down or pulled from directory servers [RFC7067]. Upon receiving a native Ethernet frame, the TRILL Encapsulating Node checks the MAC&VLAN<->RBridgeEdge mapping table, and perform the corresponding TRILL encapsulation if the entry is found in the mapping table. If the destination address and VLAN of the received Ethernet frame doesn't exist in the mapping table and no positive reply from pulling request to a directory, the Ethernet frame is forwarded per IEEE802.1Q.

```
       +------------+--------+---------+---------+--+-------+----+
       |OuterEtherHd|TRILL HD| InnerDA | InnerSA |..|Payload| FCS|
       +------------+--------+---------+---------+--+-------+----+
               ^
               |             |<Inner Ether Header>  |
               |
               |
               |      +-------+  TRILL    +------+
               |      |  R1   |-----------|  R2  |  Decapsulate
   TRILL
               |      +---+---+  domain   +------+  header
               |          |                   |
               +----------|                   |
                          |                   |
                       +-----+             +-----+
      Non-RBridge node:|T12  |             | T22 |
      Encapsulate TRILL+-----+             +-----+
      Header for data
      Frames to traverse
      TRILL domain.
```

4. Source Nickname in Frames Encapsulated by Non-RBridge Nodes

The TRILL header includes a Source RBridge's Nickname (ingress) and Destination RBridge's Nickname (egress). When a TRILL header is added by a non-RBridge node, using the Ingress RBridge edge node's nickname in the source address field will make the ingress RBridge node receive TRILL frames with its own nickname in the frames' source address field, which can be confusing.

To avoid confusion of edge RBridges receiving TRILL encapsulated frames with their own nickname in the frames' source address field from neighboring non-RBridge nodes, a new nickname can be given to an RBridge edge node, e.g. Phantom Nickname, to represent all the TRILL Encapsulating Nodes attached to the RBridge edge node.

When the Phantom Nickname is used in the Source Address field of a TRILL frame, it is understood that the TRILL encapsulation is actually done by a non-RBridge node which is attached to an edge port of an RBridge Ingress node.

5. Benefits of Non-RBridge encapsulating TRILL header

5.1. Avoid Nickname Exhaustion Issue

For a large Data Center with hundreds of thousands of virtualized servers, setting TRILL boundary at the servers' virtual switches will create a TRILL domain with hundreds of thousands of RBridge nodes, which has issues of TRILL Nicknames exhaustion and challenges to IS-IS. Setting TRILL boundary at aggregation switches that have many virtualized servers attached can limit the number of RBridge nodes in a TRILL domain, but introduce the issues of very large MAC&VLAN<->RBridgeEdge mapping table to be maintained by RBridge edge nodes and the necessity of enforcing AF ports.

Allowing Non-RBridge nodes to pre-encapsulate data frames with TRILL header makes it possible to have a TRILL domain with reasonable number of RBridge nodes in a large data center. All the TRILL encapsulating nodes attached to one RBridge are represented by one TRILL nickname, i.e. Phantom Nickname, which avoids the Nickname exhaustion problem.

5.2. Reduce FDB size for switches on Bridged LANs

When hosts in a VLAN (or subnet) span across multiple RBridge edge nodes and each RBridge edge has multiple VLANs enabled, the switches on the bridged LANs attached to the RBridge edge are exposed to all MAC addresses among all the VLANs enabled.

For example, for an Access switch with 40 physical servers attached, where each server has 100 VMs, there are 4000 hosts under the Access Switch. If indeed hosts/VMs can be moved anywhere, the worst case for the Access Switch is when all those 4000 VMs belong to different VLANs, i.e. the access switch has 4000 VLANs enabled. If each VLAN has 200 hosts, this access switch's MAC table potentially has 200*4000 = 800,000 entries.

However, if the virtual switches on server pre-encapsulate the data frames towards hosts attached to other RBridge Edge nodes with TRILL header, the outer MAC DA of those TRILL encapsulated data frames will be the MAC address of the local RBridge edge, i.e. the ingress RBridge. Therefore, the switches on the local bridged LAN don't need to keep the MAC entries for remote hosts attached to other RBridge edges.

There are multiple ways for local switches to avoid adding remote hosts' MAC to their FDB. One simple way is by disabling learning on source addresses. The local switches can be pre-installed with MAC addresses of local hosts with the assistance of directory.

6. Conclusion and Recommendation

When directory service is available, nodes outside TRILL domain become capable of encapsulating TRILL header for data frames destined for remote RBridges that is not on the same bridged LAN. The non-RBridge encapsulation approach is especially useful when there are a large number of servers in a data center equipped with hypervisor-based virtual switches. It is relatively easy for virtual switches, which are usually software based, to get directory assistance and perform network address encapsulation.

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

[RFC7067] Dunbar, et, al ''Directory Assistance Problem and High-Level Design Proposal'', RFC7067, Nov, 2013

[RFC6325] Perlman, et, al ''RBridge: Base Protocol Specification'', RFC6325, July, 2011

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

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

Authors' Addresses

Linda Dunbar
Huawei Technologies
1700 Alma Drive, Suite 500
Plano, TX 75075, USA
Phone: (972) 543 5849
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 Liability

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

Acknowledgment

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