TRILL working group
Internet Draft
Intended status: Standard Track
Expires: October 2014

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

April 8, 2014

Directory Assisted TRILL Encapsulation
draft-dunbar-trill-directory-assisted-encap-07.txt

Status of this Memo

Abstract

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

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 a non-RBridge node, can perform the TRILL encapsulation. This draft is to describe the benefits and the scheme of non-RBridge nodes performing TRILL encapsulation.

2. 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 [RFC2119].

In this document, these words will appear with that interpretation only when in ALL CAPS. Lower case uses of these words are not to be interpreted as carrying RFC-2119 significance.

AF Appointed Forwarder RBridge port [RFC6439]

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

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.

TRILL-EN: TRILL Encapsulating node. It is a node that only performs the TRILL encapsulation but doesn't participate in RBridge's IS-IS routing.

VM: Virtual Machines

3. Directory Assistance to Non-RBridge

With directory assistance [RFC7067], a non-RBridge can be informed if a packet needs to be forwarded across the RBridge domain and the corresponding egress RBridge. Suppose the RBridge domain boundary starts at network switches (not virtual switches embedded on servers), a directory can assist Virtual Switches embedded on servers to encapsulate with a proper TRILL header by providing the nickname of the egress RBridge edge to which the target is attached. The other information needed to encapsulate can be either learned by listening to TRILL Hellos, which will indicate the MAC address and nickname of appropriate edge RBridges, or by configuration.

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.

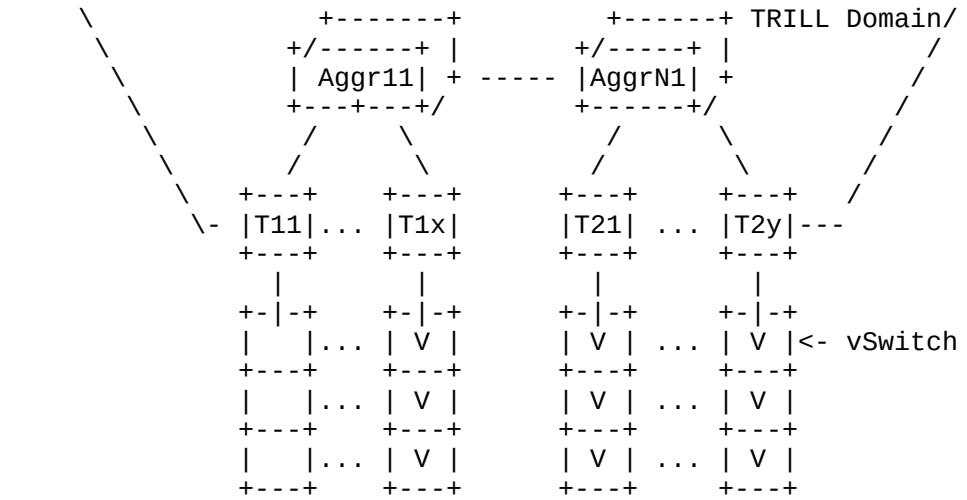

Figure 1 TRILL domain in typical Data Center Network

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 by the egress nickname field of the TRILL header of the data frame. When the ingress RBridge receives a native Ethernet frame, it handles it as usual and may drop it if it has complete directory information indicating that the target is not attached to the TRILL campus.

In this environment with complete directory information, 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.

When all attached nodes to ingress RBridge can pre-encapsulate TRILL header for traffic across the TRILL domain, the ingress RBridge don't need to encapsulate any native Ethernet frames to the TRILL domain. All native Ethernet frames are switched by the attached bridged LAN per IEEE802.1Q. Under this environment, there is no need to designate AF ports and all RBridge edge ports connected to one bridged LAN can receive and forward pre-encapsulated traffic, which can greatly improve 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 a TRILL frame arrives at an RBridge whose nickname matches with the destination nickname in the TRILL header of the frame, the processing is exactly same as normal, i.e. the RBridge decapsulates the received TRILL frame and forwards the decapsulated Ethernet frame to the target attached to its edge ports. When the DA of the decapsulated Ethernet frame is not in the egress RBridge's local MAC attachment tables, the egress RBridge can flood the decapsulated Ethernet frame to all hosts attached or drop the frame (if the egress RBridge is configured with the policy).

We call a node that only performs the TRILL encapsulation but doesn't participate in RBridge's IS-IS routing a TRILL Encapsulating node (TRILL-EN). The TRILL Encapsulating Node can get the MAC&VLAN<->RBridgeEdge mapping table pulled from directory servers [RFC7067].

Editor's note: RFC7067 has defined Push and Pull model for edge nodes to get directory mapping information. While Pull Model is relative simple for TRILL-EN to implement, Pushing requires some reliable flooding mechanism, like the one used by IS-IS, between the edge RBridge and the TRILL encapsulating node. Something like an extension to ES-IS might be needed.

Upon receiving a native Ethernet frame, the TRILL-EN 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 dropped or forwarded per IEEE802.1Q.

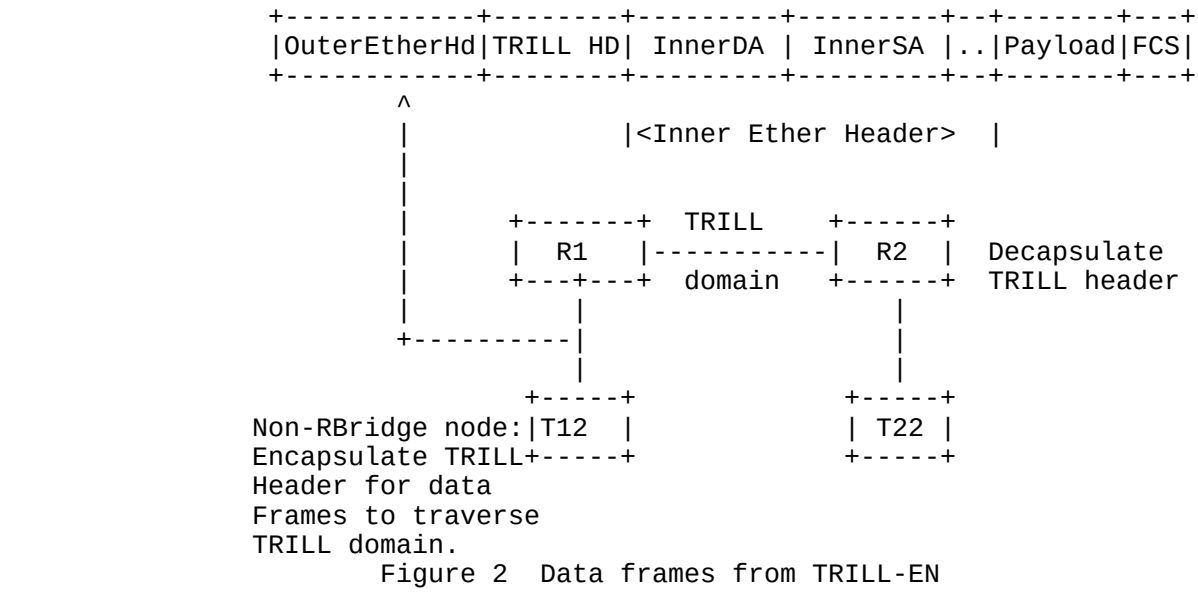

Figure 2 Data frames from TRILL-EN

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 TRILL-EN, the Ingress RBridge edge node's nickname is used in the source address field.

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-ENs attached to one RBridge are represented by one TRILL nickname, which can avoid the Nickname exhaustion problem.

5.2. Reduce MAC Tables 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.

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.

But the traffic from nodes attached to other RBridges is decapsulated and has the true source and destination MACs. To prevent local bridges from learning remote hosts' MACs and adding to their MAC tables, one simple way is to disable learning on local bridges. The local bridges can be pre-installed with MAC addresses of local hosts with the assistance of directory. The local bridges can always send frames with unknown Destination to the ingress RBridge. In an environment where end stations are VMs embedded in a server, the amount of remote MAC addresses could be very large. If it is not feasible to disable learning and pre-install MAC tables for local bridges, one effective method to minimize local bridges' MAC table size is to use the server's MAC address to hide MAC addresses of the attached VMs. I.e. the server acting as an edge node using its own MAC address in the Source Address field of the packets originated from a host (or VM) embedded. When the Ethernet frame arrives at the target edge node (the server), the target edge node can send the packet to the corresponding destination host based on the packet's IP address. Very often, the target edge node communicates with the embedded VMs via a layer 2 virtual switch. Under this case, the target edge node can construct the proper Ethernet header with the assistance from directory. The information from directory includes the proper host IP to MAC mapping information.

6. Conclusion and Recommendation

When directory information is available, nodes outside TRILL domain become capable of encapsulating TRILL header for data frames destined for remote RBridges that are 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

It requires directory assistance to make it possible for a non-TRILL node to pre-encapsulate packets destined towards remote RBridges.

8. Security Considerations

Pull Directory queries and responses are transmitted as RBridge-to-RBridge or native RBridge Channel messages. Such messages can besecured as specified in [ChannelTunnel].

For general TRILL security considerations, see [RFC6325].

9. IANA Considerations

This document requires no IANA actions. RFC Editor: Please remove this section before publication.

10. References

10.1. Normative References

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

[RFC6325] Perlman, et, al, "Routing Bridges (RBridges): Base Protocol Specification", RFC6325, July 2011

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

10.2. Informative References

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

[ChannelTunnel] - D. Eastlake, Y. Li, "TRILL: RBridge Channel Tunnel Protocol", draft-eastlake-trill-channel-tunnel, work in progress.

11. Acknowledgments

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

Authors' Addresses

Linda Dunbar
Huawei Technologies
5340 Legacy Drive, Suite 175
Plano, TX 75024, USA
Phone: (469) 277 5840
Email: linda.dunbar@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