

TRILL working group  
Internet Draft  
Intended status: Standard Track  
Expires: February 2013

L. Dunbar  
D. Eastlake  
Huawei  
Radia Perlman  
Intel  
I. Gashinsky  
Yahoo  
August 21, 2012

**Directory Assisted TRILL Encapsulation**  
**draft-dunbar-trill-directory-assisted-encap-02.txt**

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of [BCP 78](#) and [BCP 79](#).

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at <http://www.ietf.org/ietf/1id-abstracts.txt>

The list of Internet-Draft Shadow Directories can be accessed at <http://www.ietf.org/shadow.html>

## Abstract

This draft describes how data center network can benefit from non-RBridge nodes performing TRILL encapsulation and how directory service can assist a non-RBridge node to encapsulate TRILL header.

## 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](#) 0.

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

<a href="#">1. Introduction</a>	<a href="#">2</a>
<a href="#">2. Terminology</a>	<a href="#">3</a>
<a href="#">3. Directory assistance on Non-RBridge</a>	<a href="#">4</a>
<a href="#">4. Source Nickname in frames encapsulated by non-RBridge nodes..</a>	<a href="#">6</a>
<a href="#">5. Conclusion and Recommendation</a>	<a href="#">6</a>
<a href="#">6. Manageability Considerations</a>	<a href="#">6</a>
<a href="#">7. Security Considerations</a>	<a href="#">6</a>
<a href="#">8. IANA Considerations</a>	<a href="#">6</a>
<a href="#">9. Acknowledgments</a>	<a href="#">7</a>
<a href="#">10. References</a>	<a href="#">7</a>
Authors' Addresses	<a href="#">7</a>
Other Provisions.....	<a href="#">8</a>

## [1. Introduction](#)

It is no longer uncommon for a data center to have thousands of server racks. Those thousands of server racks could be connected by multiple groups of aggregation switches, with each group connecting hundreds of ToR switches. For servers supporting virtualization, there is typically a virtual switch embedded in each physical server.

When TRILL is deployed in those data centers, there are issues no matter where the RBridge domain boundary starts. If RBridge domain boundary starts at aggregation switch level, the RBridge's IS-IS routing scales well, but there are problems with allowing only one (AF port) of multiple ports connected to a bridged LAN for forwarding traffic and requiring each RBridge edge to maintain a very large table of MAC&VLAN<-> RBridgeEdge mapping. If the RBridge domain boundary starts closer to hosts, e.g. at the virtual switches on servers, the number of MAC&VLAN<->Edge mapping is much smaller because each virtual switch only needs to maintain the mapping for remote hosts which actually communicate with the embedded VMs. But then, the number of nodes in RBridge IS-IS domain is very large, making it not scale well especially on aggregation switches which need to advertise link state over hundreds of ports.

[RBridge-Directory] introduces a method 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 [[RBridge-Directory](#)], a non-RBridge can determine if a packet should 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 RBridge edge to which the target is attached.

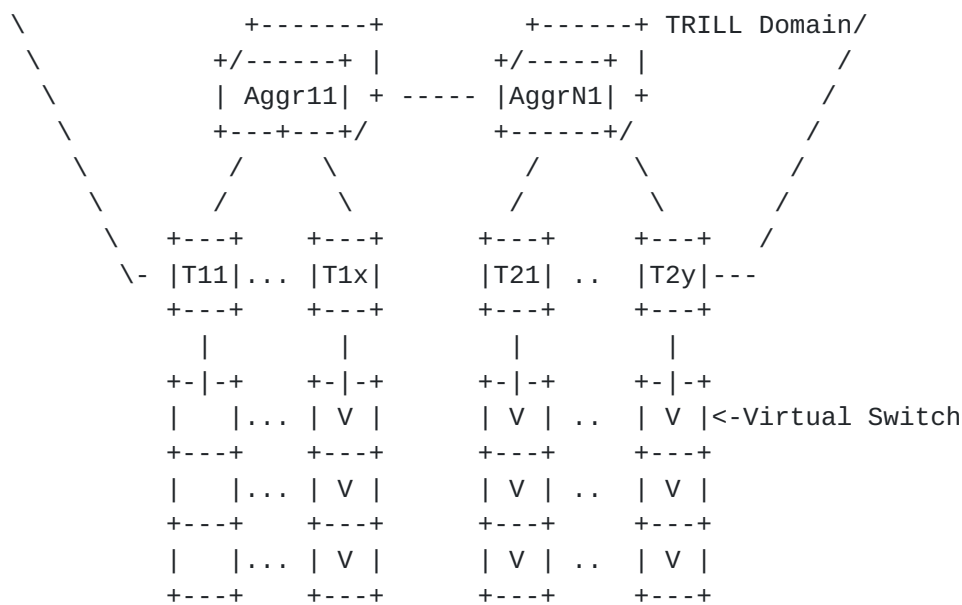


Figure 1: TRILL domain in typical Data Center Network

When a TRILL encapsulated data packet reaches an RBridge, the RBridge can simply forward the pre-encapsulated packet to the RBridge whose nickname is in the DA field of the TRILL header. By doing this, no ingress RBridge will receive a native frame with unknown DA, therefore, it won't need to flood received data packets to all other ports. 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.

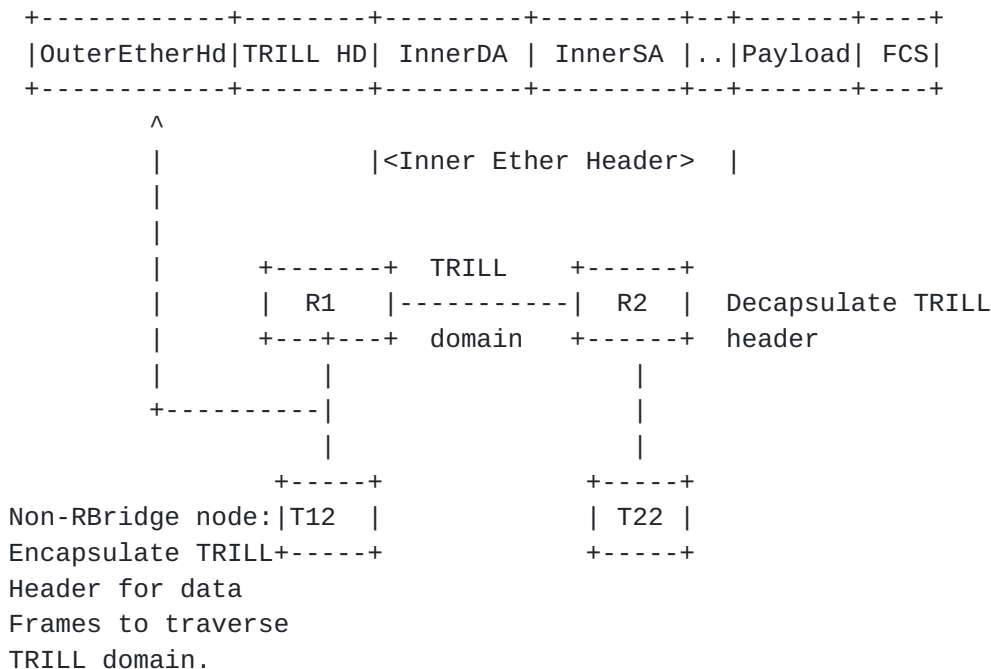
([\[RBridge\]](#) [Section 4.6.2](#) Bullet 8 specifies that an RBridge port can be configured to accept TRILL encapsulated frames from a neighbor that is not an RBridge.)



When data frames do not need to traverse RBridge domain, they are switched by all nodes/ports per IEEE802.1Q and RBridge edge will not encapsulate and forward native Ethernet 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 native frame from the received TRILL frame and forwards the decapsulated Ethernet frame to the host attached to its edge ports.

We call a node which 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 [[RBridge-Directory](#)]. 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, the Ethernet frame is forwarded per IEEE802.1Q.





#### **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. Conclusion and Recommendation**

As the number of hosts in data center gets large, the number of switches interconnecting them could increase to a point that TRILL no longer scales well. The situation will get worse as hypervisors on servers are equipped with virtual switches. Therefore, we suggest TRILL consider directory assisted non-RBridge encapsulation approach. The non-RBridge encapsulation approach is especially useful when there are many servers in a data center equipped with hypervisor-based virtual switches because it is relatively easy for virtual switches, which are usually software based, to get directory assistance and perform network address encapsulation.

#### **6. Manageability Considerations**

TBD.

#### **7. Security Considerations**

TBD.

#### **8. IANA Considerations**

TBD





## 9. Acknowledgments

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

## 10. References

[RBridge-Directory] Dunbar, et, al ''Directory Assisted RBridge Edge'', [draft-dunbar-trill-directory-assisted-edge](#), work in progress, Oct. 2011

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

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

## Acknowledgment

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

## Copyright and IPR Provisions

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to [BCP 78](#) and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License. The definitive version of an IETF Document is that published by, or under the auspices of, the IETF. Versions of IETF Documents that are published by third parties, including those that are translated into other languages, should not be considered to be definitive versions of IETF Documents. The definitive version of these Legal Provisions is that published by, or under the auspices of, the IETF. Versions of these Legal Provisions that are published by third parties, including those that are translated into other languages, should not be considered to be definitive versions of these Legal Provisions. For the avoidance of doubt, each Contributor to the IETF Standards Process licenses each Contribution that he or she makes as part of the IETF Standards Process to the IETF Trust pursuant to the provisions of [RFC 5378](#). No language to the contrary, or terms, conditions or rights that differ from or are inconsistent with the rights and licenses granted under [RFC 5378](#), shall have any effect and shall be null and void, whether published or posted by such Contributor, or included with or in such Contribution.