

TRILL Working Group
INTERNET-DRAFT
Intended Status: Informational

Yizhou Li
Weiguo Hao
Huawei Technologies
Radia Perlman
Intel Labs
Jon Hudson
Brocade
Hongjun Zhai
ZTE
July 29, 2014

Expires: January 30, 2015

**Problem Statement and Goals for Active-Active TRILL Edge
draft-ietf-trill-active-active-connection-prob-06**

Abstract

The IETF TRILL (Transparent Interconnection of Lots of Links) protocol provides support for flow level multi-pathing with rapid failover for both unicast and multi-destination traffic in networks with arbitrary topology. Active-active at the TRILL edge is the extension of these characteristics to end stations that are multiply connected to a TRILL campus. This informational document discusses the high level problems and goals when providing active-active connection at the TRILL edge.

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/lid-abstracts.html>

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

Copyright and License Notice

Copyright (c) 2014 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.

Table of Contents

- [1. Introduction](#) [3](#)
- [1.1 Terminology](#) [3](#)
- [2. Target Scenario](#) [4](#)
- [2.1 LAALP and Edge Group Characteristics](#) [6](#)
- [3. Problems in Active-Active at the TRILL Edge](#) [7](#)
- [3.1 Frame Duplications](#) [7](#)
- [3.2 Loop Back](#) [7](#)
- [3.3 Address Flip-Flop](#) [7](#)
- [3.4 Unsynchronized Information Among Member RBridges](#) [8](#)
- [4. High Level Requirements and Goals for Solutions](#) [8](#)
- [5. Security Considerations](#) [9](#)
- [6. IANA Considerations](#) [9](#)
- [7. Acknowledgments](#) [10](#)
- [8. References](#) [10](#)
- [8.1 Normative References](#) [10](#)
- [8.2 Informative References](#) [10](#)
- Authors' Addresses [11](#)

1. Introduction

The IETF TRILL (Transparent Interconnection of Lots of Links) [[RFC6325](#)] protocol provides loop free and per hop based multipath data forwarding with minimum configuration. TRILL uses [[IS-IS](#)] [[RFC6165](#)] [[RFC7176](#)] as its control plane routing protocol and defines a TRILL specific header for user data. In a TRILL campus, communications between TRILL switches can

- (1) use multiple parallel links and/or paths,
- (2) spread load over different links and/or paths at a fine grained flow level through equal cost multipathing of unicast traffic and multiple distribution trees for multi-destination traffic, and
- (3) rapidly re-configure to accommodate link or node failures or additions.

"Active-active" is the extension, to the extent practical, of similar load spreading and robustness to the connections between end stations and the TRILL campus. Such end stations may have multiple ports and will be connected, directly or via bridges, to multiple edge TRILL switches. It must be possible, except in some failure conditions, to spread end station traffic load at the granularity of flows across links to such multiple edge TRILL switches and rapidly re-configure to accommodate topology changes.

1.1 Terminology

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](#)].

The acronyms and terminology in [[RFC6325](#)] are used herein with the following additions:

CE - As in [[CMT](#)], Classic Ethernet device (end station or bridge). The device can be either physical or virtual equipment.

Data Label - VLAN or FGL (Fine Grained Label [[RFC7172](#)]).

LAALP - Local Active-Active Link Protocol. Any protocol similar to MC-LAG that runs in a distributed fashion on a CE, the links from that CE to a set of edge group Rbridges, and on those Rbridges.

MC-LAG - Multi-Chassis Link Aggregation. Proprietary extensions to IEEE Std 802.1AX-2011 [[802.1AX](#)] standard so that the aggregated links

can, at one end of the aggregation, attach to different switches.

Edge group - a group of edge R Bridges to which at least one CE is multiply attached using an LAALP. When multiple CEs attach to the exact same set of edge R Bridges, those edge R Bridges can be considered as a single edge group. An R Bridge can be in more than one edge group.

TRILL switch - an alternative term for an R Bridge.

2. Target Scenario

This section presents a typical scenario of active-active connections to a TRILL campus via multiple edge R Bridges where the current TRILL appointed forwarder mechanism does not work as expected.

The TRILL appointed forwarder mechanism [[RFC6439](#)] can handle fail over (active-standby), provides loop avoidance and, with administrative configuration, provides load spreading based on VLAN. One and only one appointed R Bridge can ingress/egress native frames into/from the TRILL campus for a given VLAN among all edge R Bridges connecting a legacy network to the TRILL campus. This is true whether the legacy network is a simple point-to-point link or a complex bridged LAN or anything in between. By carefully selecting different R Bridges as appointed forwarder for different sets of VLANs, load spreading over different edge R Bridges across different Data Labels can be achieved.

The appointed forwarder mechanism [[RFC6439](#)] requires all of the edge group R Bridges to exchange TRILL IS-IS Hello packets through their access ports. As Figure 1 shows, when multiple access links of multiple edge R Bridges are connected to a CE by an LAALP, Hello messages sent by RB1 via access port to CE1 will not be forwarded to RB2 by CE1. RB2 (and other members of LAALP1) will not see that Hello from RB1 via the LAALP1. Every member R Bridge of LAALP1 thinks of itself as appointed forwarder on an LAALP1 link for all VLANs and will ingress/egress frames. Hence the appointed forwarder mechanism cannot provide active-active or even active-standby service across the edge group in such a scenario.

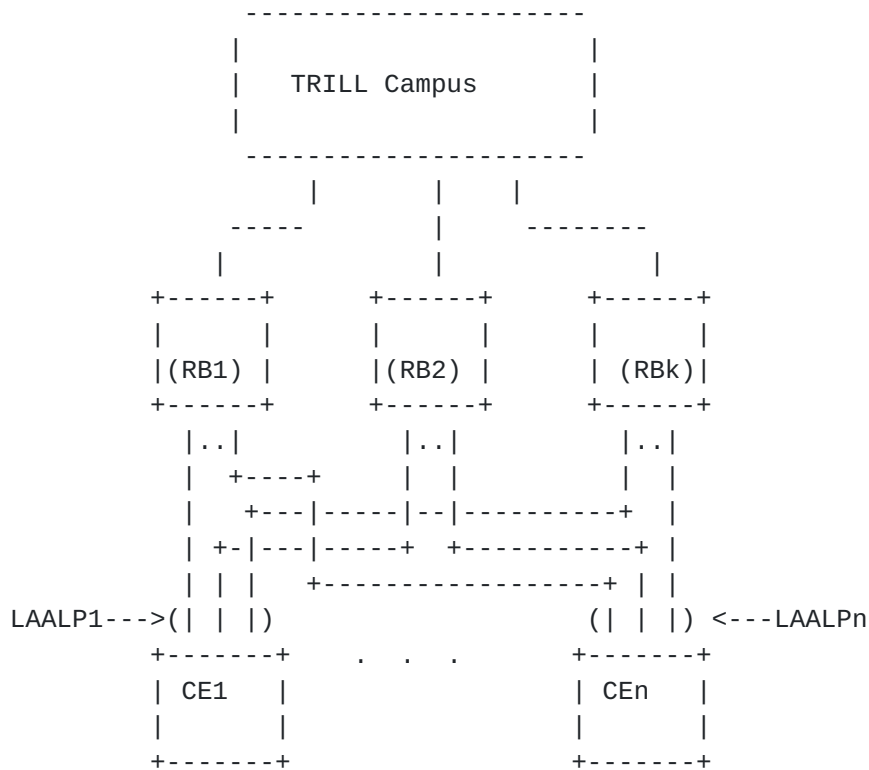


Figure 1 Active-Active connection to TRILL edge RBRidges

Active-Active connection is useful when we want to achieve the following two goals:

- Flow rather than VLAN based load balancing is desired.
- More rapid failure recovery is desired. The current appointed forwarder mechanism relies on the TRILL Hello timer expiration to detect the unreachability of another edge RBridge connecting to the same local link. Then re-appointing the forwarder for specific VLANs may be required. Such procedures take time on the scale of seconds although this can be improved with TRILL use of BFD [RFC7175]. Active-Active connection usually has a faster built-in mechanism for member node and/or link failure detection. Faster detection of failures minimizes the frame loss and recovery time.

LAALP is usually a proprietary facility whose implementation varies by vendor. So, to be sure the LAALP operations successfully across a group of edge RBRidges, those edge RBRidges will almost always have to be from the same vendor. In order to have a common understanding of active-active connection scenarios, the assumptions in [Section 2.1](#) are made about the characteristics of the LAALP and edge group of RBRidges.

2.1 LAALP and Edge Group Characteristics

For a CE connecting to multiple edge R Bridges via an LAALP (active-active connection), the following characteristics apply:

- a) The LAALP will deliver a frame from an endnode to TRILL at exactly one edge group R Bridge.
- b) The LAALP will never forward frames it receives from one up-link to another.
- c) The LAALP will attempt to send all frames for a given flow on the same uplink. To do this, it has some unknown rule for which frames get sent to which uplinks (typically based on a simple hash function of Layer 2 through 4 header fields).
- d) Frames are accepted from any of the uplinks and passed down to endnodes (if any exist).
- e) The LAALP cannot be assumed to send useful control information to the up-link such as "this is the set of other R Bridges to which this CE is attached", or "these are all the MAC addresses attached".

For an edge group of R Bridges to which a CE is multiply attached with an LAALP:

- a) Any two R Bridges in the edge group are reachable from each other via the TRILL campus.
- b) Each R Bridge in the edge group knows an ID for each LAALP instance multiply attached to that group. The ID will be consistent across the edge group and globally unique across the TRILL campus. For example, if CE1 attaches to RB1, RB2, ... RBn using an LAALP, then each of RBs will know, for the port to CE1, that it is has some label such as "LAALP1"
- c) Each RB in the edge group can be configured with the set of acceptable VLANs for the ports to any CE. The acceptable VLANs configured for those ports should include all the VLANs the CE has joined and be consistent for all the member R Bridges of the edge group.
- d) When a R Bridge fails, all the other R Bridges having formed any LAALP instance with it know the information in a timely fashion.
- e) When a down-link of an edge group R Bridge to an LAALP instance fails, that R Bridge and all the other R Bridges participating in the LAALP instance including that down-link know of the failure in a timely fashion.
- f) The R Bridges in the edge group have some mechanism to exchange information with each other, including the set of CEs they are connecting to or the IDs of the LAALP instances their down-links are part of.

Other than the applicable characteristics above, the internals of an

LAALP are out of scope for TRILL.

3. Problems in Active-Active at the TRILL Edge

This section presents the problems that need to be addressed in active-active connection scenarios. The topology in Figure 1 is used in the following sub-sections as the example scenario for illustration purposes.

3.1 Frame Duplications

When a remote RBridge ingresses a multi-destination TRILL Data packet in VLAN x, all edge group RBridges of LAALP1 will receive the frame if any local CE1 joins VLAN x. As each of them thinks it is the appointed forwarder for VLAN x, without changes made for active-active connection support, they would all forward the frame to CE1. The bad consequence is that CE1 receives multiple copies of that multi-destination frame from the remote end host source.

Frame duplication may also occur when an ingress RBridge is non-remote, say ingress and egress are two RBridges belonging to the same edge group. Assume LAALP m connects to an edge group g and the edge group g consists of RB1, RB2 and RB3. The multi-destination frames ingressed from a port not connected to LAALP m by RB1 can be locally replicated to other ports on RB1 and also TRILL encapsulated and forwarded to RB2 and RB3. CE1 will receive duplicate copies from RB1, RB2 and RB3.

Note that frame duplication is only a problem in multi-destination frame forwarding. Unicast forwarding does not have this issue as there is only ever one copy of the packet.

3.2 Loop Back

As shown in Figure 1, CE1 may send a native multi-destination frame to the TRILL campus via a member of the LAALP1 edge group (say RB1). This frame will be TRILL encapsulated and then forwarded through the campus to the multi-destination receivers. Other members (say RB2) of the same LAALP edge group will receive this multicast packet as well. In this case, without changes made for active-active connection support, RB2 will decapsulate the frame and egress it. The frame loops back to CE1.

3.3 Address Flip-Flop

Consider RB1 and RB2 using their own nickname as ingress nickname for data into a TRILL campus. As shown by Figure 1, CE1 may send a data frame with the same VLAN and source MAC address to any member of the

edge group LAALP1. If some egress RBridge receives TRILL data packets from different ingress RBridges but with same source Data Label and MAC address, it learns different Data Label and MAC to nickname address correspondences when decapsulating the data frames. Address correspondence may keep flip-flopping among nicknames of the member RBridges of the LAALP for the same Data Label and MAC address. Existing hardware does not support data plane learning of multiple nicknames for the same MAC address and data label -- when data plane learning indicates attachment of the MAC to a new nickname, it overwrites the old attachment nickname.

Implementers have stated that most current TRILL switch hardware, when doing data plane learning, behaves badly under these circumstances and, for example, interpret address flip-flopping as a severe network problem. It may also cause the returning traffic to go through different paths to reach the destination resulting in persistent re-ordering of the frames.

3.4 Unsynchronized Information Among Member RBridges

A local RBridge, say RB1 connected to LAALP1, may have learned a Data Label and MAC to nickname correspondence for a remote host h1 when h1 sends a packet to CE1. The returning traffic from CE1 may go to any other member RBridge of LAALP1, for example RB2. RB2 may not have h1's Data Label and MAC to nickname correspondence stored. Therefore it has to do the flooding for unknown unicast [[RFC6325](#)]. Such flooding is unnecessary since the returning traffic is almost always expected and RB1 had learned the address correspondence. It is desirable to avoid flooding; it imposes a greater burden on the network than known destination unicast traffic because the flooded traffic is sent over more links.

Synchronization of the Data Label and MAC to nickname correspondence information among member RBridges will reduce such unnecessary flooding.

4. High Level Requirements and Goals for Solutions

The problems identified in [section 3](#) should be solved in any solution for active-active connection to edge RBridges. The following high-level requirements and goals should be met.

Data plane:

- 1) All up-links of CE MUST be active: the LAALP is free to choose any up-link on which to send packets and the CE is able to receive

packets from any up-link of an edge group.

- 2) Looping back and frame duplication MUST be prevented.
- 3) Learning of Data Label and MAC to nickname correspondence by a remote RBridge MUST NOT flip-flop between the local multiply attached edge RBridges.
- 4) Packets for a flow SHOULD stay in order.
- 5) The Reverse Path Forwarding Check MUST work properly as per [[RFC6325](#)].
- 6) Single up-link failure on CE to an edge group MUST NOT cause persistent packet delivery failure between TRILL campus and CE.

Control plane:

- 1) No requirement for new information to be passed between edge RBridges and CE or between edge RBridges and endnodes.
- 2) If there is any TRILL specific information required to be exchanged between RBridges in an edge group, for example data labels and MAC addresses binding to nicknames, a solution MUST specify the mechanism to perform such exchange unless this is handled internal to the LAALP.
- 3) RBridges SHOULD be able to discover other members in the same edge group by exchanging their LAALP attachment information.

Configuration, incremental deployment, and others:

- 1) Solution SHOULD require minimal configuration.
- 2) Solution SHOULD automatically detect misconfiguration of edge RBridge group.
- 3) Solution SHOULD support incremental deployment, that is, not require campus wide upgrading for all RBridges, only changes to the edge group RBridges.
- 4) Solution SHOULD be able to support from 2 up to at least 4 active-active up-links on a multiply attached CE.
- 5) Solution SHOULD NOT assume there is a dedicated physical link between any two of the edge RBridges in an edge group.

5. Security Considerations

As an informational overview, this draft does not introduce any extra security risks. Security risks introduced by any particular solutions to the problems presented here will be discussed in the separate document(s) describing such solutions. For general TRILL Security Considerations, see [[RFC6325](#)].

6. IANA Considerations

No IANA action is required. RFC Editor: please delete this section before publication.

7. Acknowledgments

Special acknowledgments to Donald Eastlake and Mingui Zhang for their valuable comments.

8. References

8.1 Normative References

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

- [IS-IS] ISO/IEC 10589:2002, Second Edition, "Intermediate System to Intermediate System Intra-Domain Routing Exchange Protocol for use in Conjunction with the Protocol for Providing the Connectionless-mode Network Service (ISO 8473)", 2002.

- [RFC6165] Banerjee, A. and D. Ward, "Extensions to IS-IS for Layer-2 Systems", [RFC 6165](#), April 2011.

- [RFC6325] Perlman, R., Eastlake 3rd, D., Dutt, D., Gai, S., and A. Ghanwani, "Routing Bridges (RBridges): Base Protocol Specification", [RFC 6325](#), July 2011

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

- [RFC7172] Eastlake, D., M. Zhang, P. Agarwal, R. Perlman, D. Dutt, "Transparent Interconnection of Lots of Links (TRILL): Fine-Grained Labeling", [RFC7172](#), May 2014.

- [RFC7176] Eastlake 3rd, D., Senevirathne, T., Ghanwani, A., Dutt, D., and A. Banerjee, "Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS", [RFC 7176](#), May 2014.

8.2 Informative References

- [CMT] Senevirathne, T., Pathangi, J., and J. Hudson, "Coordinated Multicast Trees (CMT)for TRILL", [draft-ietf-trill-cmt.txt](#), Work in Progress, April 2014.

- [RFC7175] Manral, V., D. Eastlake, D. Ward, A. Banerjee, "Transparent Interconnection of Lots of Links (TRILL): Bidirectional Forwarding Detection (BFD) Support", [RFC7175](#), May 2014.

[802.1AX] IEEE, "Link Aggregation", 802.1AX-2008, 2008.

[802.1Q] IEEE, "Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks", IEEE Std 802.1Q-2011, August, 2011

Authors' Addresses

Yizhou Li
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China

Phone: +86-25-56625409
EMail: liyizhou@huawei.com

Weiguo Hao
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China

Phone: +86-25-56623144
EMail: haoweiguo@huawei.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

Jon Hudson
Brocade
130 Holger Way
San Jose, CA 95134 USA

Phone: +1-408-333-4062
jon.hudson@gmail.com

Hongjun Zhai
ZTE
68 Zijinghua Road, Yuhuatai District
Nanjing, Jiangsu 210012
China

Phone: +86 25 52877345
Email: zhai.hongjun@zte.com.cn