TRILL Working Group
INTERNET-DRAFT
Intended Status: Informational

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

Expires: November 13, 2014

May 12, 2014

Problem Statement and Goals for Active-Active TRILL Edge
draft-ietf-trill-active-active-connection-prob-03

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.

Table of Contents

1 Introduction 3
1.1 Terminology 3
2. Target Scenario 4
3. Problems in Active-Active at the TRILL Edge 6
3.1 Frame Duplications 7
3.2 Loop 7
3.2 Address Flip-Flop 7
3.3 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 9
8 References 9
8.1 Normative References 9
8.2 Informative References 10
Authors' Addresses 10

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] [RFC6326bis] 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) load spread 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 load spread end station traffic at the flow level 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]).

MC-LAG - Multi-Chassis Link Aggregation. Proprietary extensions to [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 RBridges to which at least one CE is multiply attached using MC-LAG. When multiple CEs attach to the exact

same set of edge RBridges, those edge RBridges can be considered as a single edge group. One RBridge can be in more than one edge group.

TRILL switch - an alternative term for an RBridge.

2. Target Scenario

The TRILL appointed forwarder [RFC6325] [RFC7177] [RFC6439] mechanism provides per Data Label active-standby traffic spreading and loop avoidance at the same time. One and only one appointed RBridge can ingress/egress native frames into/from TRILL campus for a given VLAN among all edge RBridges connecting a legacy network to 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 RBridges as appointed forwarder for different set of VLANs, load spreading over different edge RBidges across different Data Labels can be achieved.

This section presents a typical scenario of active-active connections to TRILL campus via multiple edge RBridges where the current TRILL appointed forwarder mechanism is not working as expected.

The appointed forwarder mechanism [RFC6439] requires each of the edge RBridges to exchange TRILL IS-IS Hello packets from their access ports. As Figure 1 shows, when multiple access links of multiple edge RBridges are bundled as an MC-LAG (Multi-Chassis Link Aggregation Group), Hello messages sent by RB1 via access port to CE1 will not be forwarded to RB2 by CE1. RB2 (and other members of MC-LAG1) will not see that Hello from RB1 via the MC-LAG. Every member RBridge of MC-LAG1 thinks of itself as appointed forwarder on an MC-LAG1 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.

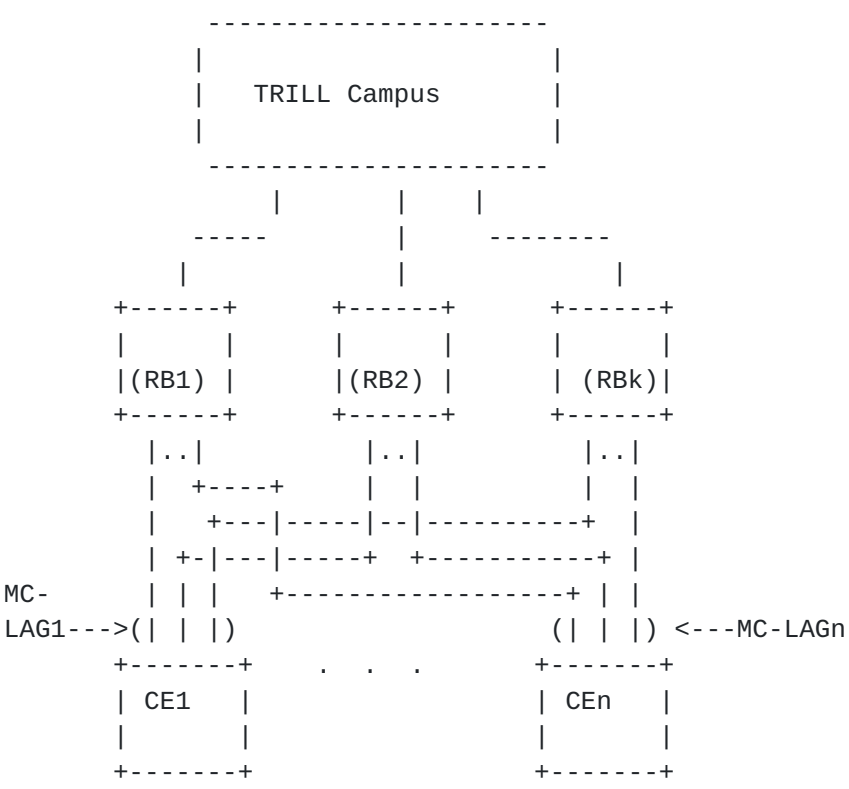

Figure 1 Active-Active connection to TRILL edge RBridges

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

- Flow rather than VLAN based load balancing is desired.

- More rapid failure recovery is desired. Current appointed forwarder mechanism relies on the Hello timer expiration to detect the unreachability of another edge RBridge connecting to the same local Ethernet 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 failure would minimize the frame loss and recovery time.

MC-LAG is a proprietary facility whose implementation varies by vendor. So, to be sure of MC-LAG operation across a group of edge RBridges, those edge RBridges will almost always be from the same vendor. In order to have a common understanding of active-active connection scenarios, the following assumptions are made:

For a CE connecting to multiple edge RBs via active-active

connection:
a) the CE will forward a packet from an endnode to exactly one up-link
b) the CE will never forward packets it receives from one up-link to another
c) the CE will attempt to send all packets for a given flow on the same uplink
d) packets are accepted from any of the uplinks and passed down to endnodes (if any exist)
e) the CE has some unknown rule for which packets get sent to which uplinks (typically based on a simple hash function of Layer 2 through 4 header fields)
f) the CE cannot be assumed to give useful control information to the up-link such as "this is the set of other RBridges to which this CE is attached", or "these are all the MAC addresses attached"

For an edge group to which a CE is multiply attached:
a) Any two RBs in the edge group are reachable from each other via TRILL campus.
b) Each RB in the edge group is configured with an ID for each down-link to a CE multiply attached to that group. The ID will be consistent across the edge group. For example, if CE1 attaches to RB1, RB2, ... RBn, then each of RBs will have been configured, for the port to CE1, that it is labeled "MC-LAG1"
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 RBridges of the edge group.
d) When a RBridge fails, all the other RBridges having formed any MC-LAG with it know the information in a timely fashion.
e) When a down-link of an edge group RBridge to an MC-LAG fails, all the other RBridges having formed any MC-LAG including that down-link know the information in a timely fashion
f) The RBs in the edge group have some mechanism to exchange the state and information with each other, including the set of CEs they are connecting to or ID of MC-LAGs their down-links have joined.

Some MC-LAG implementations corresponding to the assumptions d, e, and f above may only applicable to native Ethernet network or specific type of network. Then some extension for TRILL network may be required. It is not the assumption that RBridges in an edge group can naturally support them unless otherwise stated.

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 sends a multi-destination TRILL Data packet in VLAN x, all edge group RBridges of MC-LAG1 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.

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 MC-LAG m connects to an edge group g and the edge group g consists of RB1, RB2 and RB3. The multi-destination frame ingressed from a port not connected to MC-LAG 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.

It should be noted that frame duplication is only a problem in multi-destination frame forwarding. Unicast forwarding does not have this issue.

3.2 Loop

As shown in Figure 1, CE1 may send a native multi-destination frame to the TRILL campus via a member of the MC-LAG1 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 MC-LAG 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.

3.2 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 MC-LAG1. 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 MC-LAG for the same Data Label and MAC address.

Most TRILL switches behave badly under these circumstances and, for example, interpret this as a severe network problem. It may also cause the returning traffic to go through the different paths to reach the destination resulting in persistent re-ordering of the frames.

3.3 Unsynchronized Information Among Member RBridges

A local Rbridge, say RB1 in MC-LAG1, 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 MC-LAG1, 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. Such flooding is unnecessary since the returning traffic is almost always expected and RB1 had learned the address correspondence.

Synchronization on 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

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; CE is free to choose any up-link on which to send packets; CE is able to receive the packet from any up-link of an edge group.
2) Looping 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, solution SHOULD specify the mechanism to perform such exchange.
3) RBridges SHOULD be able to discover other members in the same edge group by exchanging their MC-LAG 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 line between any two of the edge RBridges in an edge group.

5 Security Considerations

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

[RFC6326bis] Eastlake, D., Banerjee, A., Dutt, D., Perlman, R., and A. Ghanwani, "TRILL Use of IS-IS", draft-eastlake-isis-rfc6326bis, work in progress.

[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.

[RFC7177] Eastlake 3rd, D., R. Perlman, A. Ghanwani, H. Yang, and V. Manral, "Transparent Interconnection of Lots of Links (TRILL): Adjacency", RFC7177, 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 Interconnetion of Lots of Links (TRILL): Bidirectional Forwarding Detection (BFD) Support", RFC7175, May 2014.

[802.1AX] IEEE, "Link Aggregration", 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