TRILL Working Group
Internet-Draft
Intended status: Standards Track
Expires: November 21, 2014

H. Zhai
ZTE
T. Senevirathne
Cisco Systems
R. Perlman
Intel Labs
D. Eastlake 3rd
M. Zhang
Y. Li
Huawei

May 20, 2014

RBridge: Pseudo-Nickname for Active-active Access
draft-hu-trill-pseudonode-nickname-07

Abstract

The IETF TRILL (Transparent Interconnection of Lots of Links) protocol provides support for flow level multi-pathing for both unicast and multi-destination traffic in networks with arbitrary topology. Active-active access at the TRILL edge is the extension of these characteristics to end stations that are multiply connected to a TRILL campus. In this document, the edge RBridge group providing active-active access to such an end station can be represented as a Virtual RBridge. Based on the concept of Virtual RBridge along with its pseudo-nickname, this document facilitates the TRILL active-active access of such end stations.

Table of Contents

1. Introduction

The IETF TRILL protocol [RFC6325] provides optimal pair-wise data frame forwarding without configuration, safe forwarding even during periods of temporary loops, and support for multi-pathing of both unicast and multicast traffic. TRILL accomplishes this by using IS-IS [RFC1195] link state routing and encapsulating traffic using a header that includes a hop count. Devices that implement TRILL are called RBridges or TRILL switch.

In the current TRILL protocol, an end node can be attached to TRILL campus via a point-to-point link or a shared link (such as a Local Area Network (LAN) segment). Although there might be more than one edge RBridge on a shared link, to avoid potential forwarding loops, one and only one of the edge RBridges is permitted to provide forwarding service for end station traffic in each VLAN (Virtual LAN). That RBridge is referred as to Appointed Forwarder (AF) for that VLAN on the link [RFC6325] [RFC6439]. However, in some practical deployments, to increase the access bandwidth and reliability, an end station might be multiply connected to several edge RBridges and treat all of the uplinks as a Multi-Chassis Link Aggregation (MC-LAG) bundle. In this case, it's required that traffic can be ingressed/egressed into/from TRILL campus by any of such RBridges for each given VLAN. These RBridges constitutes an Active-Active Edge (AAE) RBridge group for the end station.

Traffic with the same VLAN and source MAC address might be sent by such an end station to any member RBridge in the AAE group, and then is ingressed into TRILL campus by the RBridge. When some egress RBridge receives TRILL data packets from different ingress RBridges but with same VLAN and source MAC address, it learns different VLAN and MAC to nickname address correspondences continuously when decapsulating those frames. This issue is known as the "MAC flip-flopping" issue, which makes most TRILL switches behave badly and causes the returning traffic to reach the destination via different paths resulting in persistent re-ordering of the frames. In addition to this issue, other issues such as duplication egressing and loop of multi-destination frames may be encountered by the end stations multiply connected to the member RBridges of such an AAE group [AAProb].

In this document, edge RBridge group, which can be represented as a Virtual RBridge (RBv) and assigned a pseudo-nickname, is used to address the AAE issues in the scope of TRILL. For a member RBridge of such a group, it uses the pseudo-nickname, instead of its own

nickname as the ingress RBridge nickname when ingressing frames received on the related MC-LAG links.

The main body of this document is organized as follows: Section 2 gives an overview of the TRILL active-active access issues and the reason virtual RBridge is used to resolve the issues. Section 3 gives the concept of virtual RBridge and its pseudo-nickname. Section 4 describes how some edge RBridges to constitute a virtual RBridge (RBv) automatically and get a pseudo-nickname for the RBv. Section 5 discusses how to protect multi-destination traffic against disruption due to Reverse Forwarding Path (RPF) check failure, duplication copies and forwarding loop, etc. Section 6 covers the special processing of native frames and TRILL data frames at member RBridges of a RBv (also referred as to an edge RBridge group); followed by Section 7 which describes the MAC information synchronization among the member RBridges of a RBv. Section 8 discusses the protection of downlink failure at a member RBridge; and Section 9 gives the necessary TLV extension for edge RBridge group.

Familiarity with [RFC6325] is assumed in this document.

1.1. Terminology and Acronyms

This document uses the acronyms defined in [RFC6325] and the following additional acronym:

AF - Appointed Forwarder

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

AAE - Active-active edge RBridge group, a group of edge RBridges to which at least one CE is multiply attached using MC-LAG. AAE is also referred to as edge group or Virtual RBridge in this document.

RBv - Virtual RBridge, an alias of active-active edge RBridge group in this document.

vDRB - The Designated RBridge in a RBv. It is responsible to appoint a pseudo-nickname for the RBv.

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 [RFC2119].

When used in lower case, these words convey their typical use in common language, and are not to be interpreted as described in [RFC2119].

2. Overview

To minimize impact during failures and maximize available access bandwidth, end stations (referred to as CEs in this document) may be multiply connected to TRILL campus via multiple edge RBridges. Figure 1 shows such a typical deployment scenario, where CE1 attaches to RB1, RB2, ... RBk and treats all of the uplinks as a Multi-Chassis Link Aggregation (MC-LAG) bundle. Then RB1, RB2, ... RBk constitute an Active-active Edge (AAE) RBridge group for CE1 in this MC-LAG. Even if a member RBridge or an uplink fails, CE1 can still get frame forwarding service from TRILL campus if there are still member RBridges and uplinks available in the AAE group. Furthermore, CE1 can make flow-based load balancing across the available member links of the MC-LAG bundle in the AAE group when it communicates with other end stations across TRILL campus.

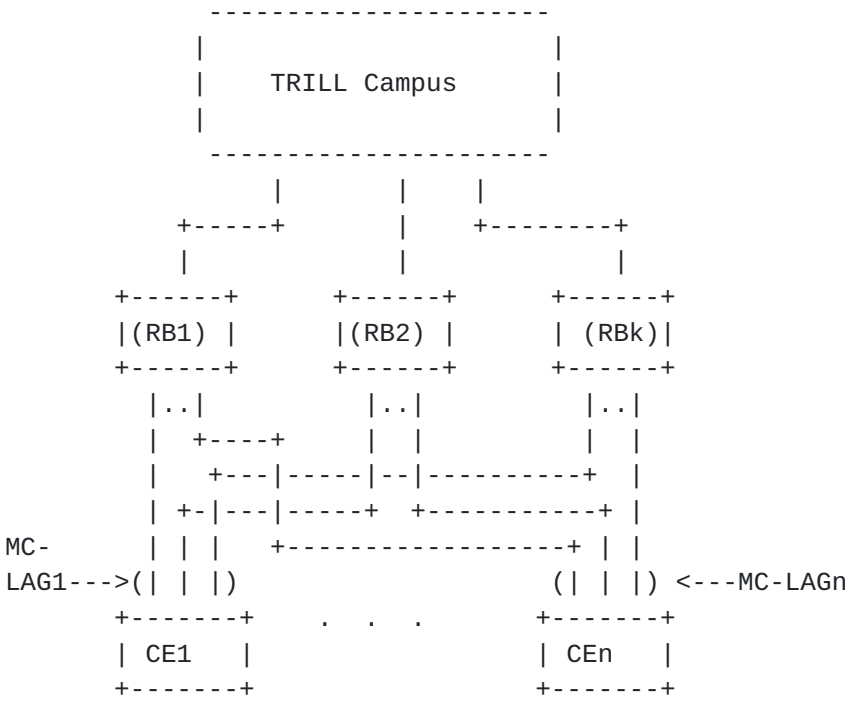


Figure 1 Active-Active connection to edge RBridges

By design, an MC-LAG does not forward packets received on one of its member ports to other member ports. As a result, the hello messages sent by one member RBridge (say RB1) via port to CE1 will not be forwarded to other member RBridges by CE1. That is to say, member RBridges will not see each other's hellos via the MC-LAG. So every member RBridge of MC-LAG1 thinks of itself as appointed forwarder for all VLANs enabled on an MC-LAG1 link and can ingress/egress frames simultaneously in these VLANs. The simultaneous flow-based

ingressing/egressing may cause some problems in some cases. For example, simultaneous egressing of multi-destination traffic by multiple member RBridges will result in frame duplication at CE1 (see Section 3.1 of [AAProb]); simultaneous ingressing of frames originated by CE1 for different flows in same VLAN will also result in MAC flip-flopping at remote egress RBridges (see Section 3.3 of [AAProb]). The flip-flopping in turn causes packet disorder in reverse traffic and worsens the traffic disruption.

Since the fact is true that edge RBridges learn Data Label and MAC to nickname address correspondences by default via decapsulating TRILL data frames (see Section 4.8.1 of [RFC6325]), the MAC flip-flopping issue SHOULD be solved based on the assume that the default learning is enabled at edge RBridges. So in this document, Virtual RBridge, together with its pseudo-nickname is introduced to fix these issues.

3. Edge Group and its Pseudo-nickname

A Virtual RBridge (RBv) represents a group of edge RBridges to which at least one CE is multiply attached using MC-LAG. More exactly, it represents a group of end station service ports on the edge RBridges and the end station service provided to the CE(s) on these ports, through which the CE(s) is multiply attached to TRILL campus using MC-LAG(s). Such end station service ports are called RBv ports; in contrast, other access ports at edge RBridges are called regular access ports in this document. RBv ports are always MC-LAG connecting ports, but not vice versa (see Section 4.1). For an edge RBridge, if one or more of its end station service ports are ports of a RBv, that RBridge is a member RBridge of the RBv.

For the convenience of description, a Virtual RBridge is also referred to as an edge group in this document. In TRILL campus, a RBv is identified by its pseudo-nickname which is different to RBridge's regular nickname(s). A RBv has one and only one pseudo-nickname. Each member RBridge (say RB1, RB2,..., RBk) of a RBv advertises the pseudo-nickname of the RBv using the Nickname sub-TLV in its TRILL IS-IS LSP (Link State PDU), [RFC7176], along with their regular nickname(s). Then from these LSPs, other RBridges outside the edge group know that the RBv is reachable through RB1 to RBk.

A member RBridge (say RBi) loses its membership from a RBv when its last port of that RBv becomes unavailable due to failure, configuration, etc. Then RBi removes that RBv's pseudo-nickname from its LSPs and updates them to the TRILL campus. From those updated LSPs, other RBridges know that their path(s) to RBv is not through RBi now.

When member RBridges receive native frames from their ports of a RBv and decide to ingress the frames into TRILL campus, they use the RBv's pseudo-nickname instead of their own regular nicknames as the ingress nickname in encapsulating TRILL headers on the frames. So when these frames arrive at an egress RBridge, even they are originated by same an end station in same a VLAN and ingressed by different member RBridges, no address flip-flopping is observed on the egress RBridge via decapsulating these frames. Kindly note when a member RBridge of an edge group ingresses a frame from a non-RBv port, it still use its own nickname as the ingress nickname.

Since RBv is not a physical node and no TRILL frames are forwarded between its ports via a local MC-LAG, pseudo-node LSP(s) MUST NOT be created for an RBv. RBv cannot act as root when constructing distribution trees for multi-cast traffic and its pseudo-nickname is ignored when determining the distribution tree root for TRILL campus [CMT]. So the tree root priority of RBv's nickname SHOULD be set to 0, and this nickname SHOULD NOT be listed in the s nicknames by the RBridge holding the highest priority tree root nickname.

NOTE 1: In order to reduce the consumption of nicknames, especially in large TRILL campus with lots of RBridges and/or active-active accesses, when multiple CEs attach to the exact same set of edge RBridges via MC-LAGs, those edge RBridges should be considered as a single edge group and represented by a single RBv with a pseudo-nickname.

4. Member RBridges Auto-Discovery

Edge RBridges connected by CE(s) via MC-LAG(s) can automatically discover each other with minimal to no configuration through exchange of the MC-LAG(s) information.

From the perspective of edge RBridges, a CE that connects to edge RBridges via a MC-LAG can be identified by the globally unique ID of the MC-LAG (i.e., the LAG-ID). On each of such edge RBridges, the access port to such a CE is associated with a LAG_ID for the CE. A MC-LAG is considered valid on an edge RBridge only if the RBridge still has operational down-link to that MC-LAG. For such an edge RBridge, it advertises a list of LAG-IDs for all the valid local MC-LAGs to other edge RBridges via TRILL IS-IS LSPs. Based on the LAG-IDs advertised by other edge RBridges, each RBridge can know which edge RBridges could constitute an edge group (See Section 4.1 for more details). Then one RBridge is elected from the group for the duty to allocate an available nickname from TRILL campus as the pseudo-nickname for the group (See Section 4.2 for more details).

4.1. Discovering Member RBridge for an Edge Group

Take Figure 2 as an example, where CE1 and CE2 multiply attach to RB1, RB2 and RB3 via MC-LAG1 and MC-LAG2 respectively; CE3 and CE4 attach to RB3 and RB4 via MC-LAG3 and MC-LAG4 respectively. Assume all the down-links are operational, and MC-LAG3 is configured to occupy a Virtual RBridge by itself.

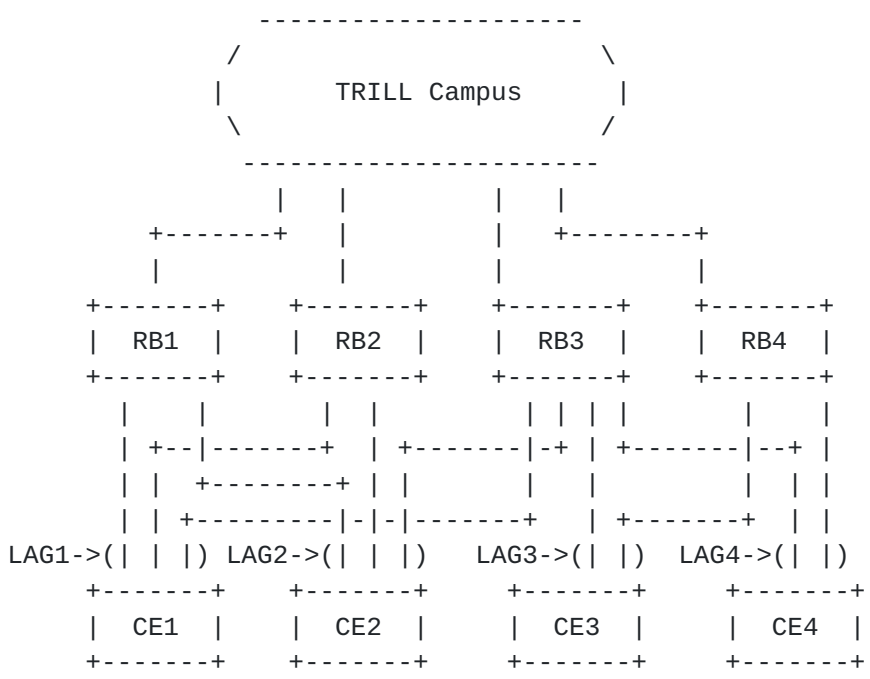


Figure 2 Different LAGs to TRILL Campus

RB1 and RB2 advertise {LAG1, LAG2} in the LAG Membership sub-TLV (see Section 9.1 for more details) via their TRILL IS-IS LSPs respectively; RB3 announces {LAG1, LAG2, LAG3, LAG4}; and RB4 announces {LAG3, LAG4}, respectively.

An edge RBridge is called a MC-LAG related RBridge if it has at least one MC-LAG configured on access port. On receipt of the LAG Membership sub-TLVs, RBn ignores them if it is not a MC-LAG related RBridge; otherwise, RBn SHOULD use the MC-LAG information contained in the sub-TLVs, along with its own LAG Membership sub-TLVs to decide which RBv(s) it should join and which edge RBridges constitute each of such RBvs. Based on the information received, each of the 4 RBridge knows the following information:

LAG-ID	OE-flag	Set of multi-homed RBridges
LAG1	0	{RB1, RB2, RB3}
LAG2	0	{RB1, RB2, RB3}
LAG3	1	{RB3, RB4}
LAG4	0	{RB3, RB4}

Where the E-flag indicates that LAG3 needs to occupy an edge group exclusively, even other MC-LAGs (for example MC-LAG4) multiply attaches to the same set of edge RBridges as its. By default, this flag is set zero. For an MC-LAG, this flag is considered 1 only if one edge RBridge advertises it as one (see Section 9.1).

In the above table, there might be some LAGs that attach to single RBridge per one due to mis-configuration or link failure, etc. Those LAGs are considered as invalid entries. Then each of the MC-LAG related edge RBridges performs the following approach to decide which valid LAGs can be served by an RBv.

Step 1: Take all the valid LAGs that have their E-flags (occupying a Virtual RBridge Exclusively) set 1 out of the table and create a RBv per such LAG.

Step 2: Sort the left valid LAGs in the table in descending order based on the number of RBriges in their associated set of multi-homed RBridges.

Step 3: Take the valid LAG (say LAG_i) with the maximum set of RBridges, say S_i, out of the table and create a new RBv (Say RBv_i) for it.

Step 4: Walk through the remainder valid LAGs in the table one by one, pick up all the valid LAGs that their sets of multi-homed RBridges contain same the RBridges as that of LAG_i and take them out of the table. Then appoint RBv_i as the servicing RBv for those LAGs.

Step 5: Repeat Step 3-4 for the left LAGs until all the valid entries in the table has be associated with a RBv.

After performing the above steps, all the 4 RBridges know that LAG3 is served by a RBv, say RBv1, which has RB3 and RB4 as member RBrdges; LAG1 and LAG2 are served by another RBv, say RBv2, which has RB1, RB2 and RB3 as member RBridges; and LAG4 is served by RBv3, which has RB3 and RB4 as member RBridges, shown as follows:

RBv	Serving LAGs	Member RBridges
RBv1	{LAG3}	{RB3, RB4}
RBv2	{LAG1, LAG2}	{RB1, RB2, RB3}
RBv3	{LAG4}	{RB3, RB4}

In each RBv, one of its member RBridges is elected as DRB. The winner is the member RBridge with the maximum device nickname in this RBv. Then this DRB picks up an available nickname as this RBv's pseudo-nickname and announce it to all other member RBridges via TRILL IS-IS LSPs (Refer Section 9.2 for more details).

4.2. Appointing Pseudo-nickname for RBv

As described in Section 3, in TRILL campus, a RBv is identified by its pseudo-nickname. In an edge group (i.e., RBv), one member RBridge is elected for the duty to appoint pseudo-nickname for this RBv; this RBridge is called Designated RBridge of the RBv (vDRB) in this document. The winner is the RBridge with maximum System ID in the group. Then based on its TRILL IS-IS link state database and the potential pseudo-nickname(s) reported in the LAG Membership sub-TLVs by other member RBridges of this RBv (see Section 9.1 for more details), the vDRB acquires an available nickname as the pseudo-nickname for this RBv and advertizes it to the other RBridges via its TRILL IS-IS LSPs (see Section 9.2). If a nickname is neither assigned as regular nickname to any RBridge nor as a pseudo-nickname to any other RBv, it is an available nickname. On receipt of the pseudo-nickname advertised by the vDRB, the other RBridges of that group associate it with the MC-LAGs served by the RBv, and then download the association to their data plane.

To reduce the traffic disruption caused by nickname changing, if possible, vDRB should attempt to reuse the pseudo-nickname recently used by the group when appointing nickname for the RBv. To help the vDRB to do so, each MC-LAG related RBridge advertises a potential pseudo-nickname for each of its MC-LAGs in its LAG Membership sub-TLV if it has used such one for that MC-LAG recently. Although it is up to the implementation of the vDRB as to how reusing pseudo-nickname, one suggestion is given as follows:

- If there are more than one available pseudo-nickname that are reported by all the member RBridges of some MC-LAGs in this RBv, the one that is reported by most of such MC-LAGs is chosen as the pseudo-nickname for this RBv. In the case that tie exists, the minimum one is chosen.

- Else, the minimum available potential pseudo-nickname is chosen.

If there is no available potential pseudo-nickname reported, the vDRB selects a nickname randomly from the apparently available nicknames as the pseudo-nickname for this RBv, based on its copy of the TRILL IS-IS link state.

Then the selected pseudo-nickname is announced by the vDRB to other member RBridges of this RBv in the PN-RBv sub-TLV (see Section 9.2) via TRILL IS-IS LSPs. After receiving the pseudo-nickname, other RBridges of that RBv associate the nickname with their ports of that RBv and download the association to their data plane.

5. Distribution Trees and Designated Forwarder

In a RBv, as each of the member RBridges thinks it is the appointed forwarder for VLAN x, without changes made for active-active connection support, they would all ingress/egress frames into/from TRILL campus for VLAN x. For multi-destination frames, more than one member RBridges ingress them may cause some of them suffering failure of Reverse Path Forwarding (RPF) Check on other RBridges; for a multi-destination traffic, more than one RBridges egress it may cause local CE(s) receiving duplication frames [AAProb]. Furthermore, in an edge group, a multi-destination frame sent by a CE (say CEi) may be ingressed into TRILL campus by one member RBridge, then another member RBridge will receive and egress it to CEi, which will result in loop of frame for CEi.

In the following sub-sections, the first two issues are discussed in Section 5.1 and Section 5.2, respectively; the third one is discussed in Section 5.3.

5.1. Different Trees for Different Member RBridges

In TRILL, RBridges use distribution trees to forward multi-destination frames. RPF Check along with other technical is used to avoid temporary multicast loops during topology changes (Section 4.5.2 of [RFC6325]). RPF check mechanism only allows a multi-destination frame ingressed by an RBridge RBi and forwarded on a distribution tree Tx to arrive at another RBridge RBn on an expected port. If arriving on other ports, the frame MUST be dropped.

To avoid address flip-flopping on remote RBridges, member RBridges use RBv's pseudo-nickname instead of their regular nicknames as ingress nickname to ingress native frames, including multicast frames. From the view of other RBridges, these frames appear as if they were ingressed by the RBv. When multicast frames of different flows are ingressed by different member RBridges of a RBv and forwarded along same a distribution tree, they will arrive at RBn

from different ports. Some of them will violate the RFC check principle at RBn and be dropped, which may result in traffic disruption.

In a RBv, if different member RBridge uses different distribution trees to ingress multi-destination frames, the RFC check violence issue can be fixed. Coordinated Multicast Trees (CMT) proposes such an approach, and makes use of the Affinity sub-TLV defined in [RFC7176] to tell other RBridges which trees a member RBridge (say RBi) may choose when ingressing multi-destination frames, then all RBridges calculate RFC check information for RBi on those trees [CMT].

In this document, the approach proposed in [CMT] is used to fix the RFC check violence issue, please refer to [CMT] for more details of the approach.

5.2. Designated Forwarder for Member RBridges

Take Figure 3 as an example, where CE1 and CE2 are served by a RBv, which has RB1 and RB2 as member RBridges. In VLAN x, the three CEs can communicate with each other.

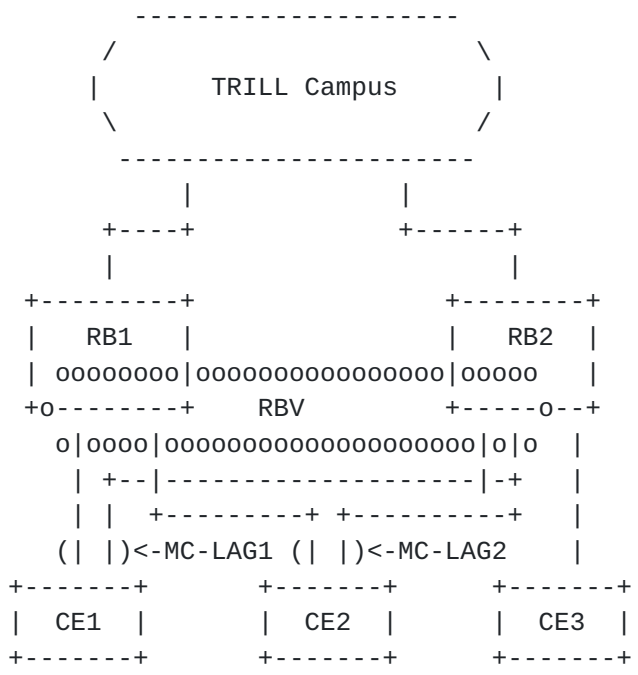


Figure 3 A Topology with Multi-homes and Single-homed CEs

When a remote RBridge (say RBn) sends a multi-destination TRILL Data packet in VLAN x, both RB1 and RB2 will receive it. As each of them

thinks it is the appointed forwarder for VLAN x, without changes made for active-active connection support, they would both forward the frame to CE1/CE2. As a result, CE1/CE2 would receive duplication copies of the frame through this RBv.

In another case, assume CE3 is single-homed to RB2. When it transmits a native multi-destination frame onto link CE3-RB2 in VLAN x, the frame can be locally replicated to the ports to CE1/CE2, and also TRILL encapsulated and ingressed into TRILL campus. When the encapsulated frame arrives at RB1 across the TRILL campus, it will be egressed to CE1/CE2 by RB1. Then CE1/CE2 receives duplicate copies from RB1 and RB2.

In this document, Designated Forwarder (DF) of VLAN is introduced to avoid the duplicate copies. The basic idea of DF is to elect one RBridge per VLAN from a RBv for the duty to forward multi-destination native traffic and/or egress multi-destination TRILL data traffic to the CEs served by the RBv.

Note that DF has an effect only on the forwarding/egressing of multi-destination traffic, no effect on the ingressing of frames or forwarding/egressing of unicast frames. Furthermore, DF check is performed only on RBv ports, not on regular access ports.

Each RBridge in a RBv elects a DF using same algorithm which guarantees the same RBridge elected as DF per VLAN.

Assuming there are k member RBridges in a RBv and each RBridge is referred to as RBi where 0 <= i < k-1, DF election algorithm per VLAN is as follows:

Step 1: All the RBridges are sorted in numerically ascending order by System ID such that RB0 < RB1 < ... < RBk-1. Each RBridge in the numerically sorted list is assigned a monotonically increasing number i such that; RB0=0, RB1=1 ..., RBi=i ..., RBk-1=k-1.

Step 2: For VLAN ID m, choose RBridge whose number equals (m mod k) as DF if any MC-LAG connecting to the RBv has been configured for VLAN m.

For a native multi-destination frame of VLAN x received, if RBi is an MC-LAG related RBridge, in addition to the local replication of the frame to regular access ports as per [RFC6325], it should also locally replicate the frame to the following RBv ports:

1) RBv ports associated with the same pseudo-nickname as that of the incoming port, no matter if RBi is the DF for the frame's VLAN on the outgoing ports;

2) RBv ports on which RBi is the DF for the frame's VLAN while they are associated with different pseudo-nickname(s) to that of the incoming port.

Furthermore, the frame MUST NOT be replicated back to the incoming port. For non MC-LAG related RBridges or for non RBv ports on an MC-LAG related RBridge, local replication is performed as per [RFC6325].

For a multi-destination TRILL data frame received, RBi MUST NOT egress it out of the RBv ports where it is not DF for the frame's Inner.VLAN. Otherwise, whether or not egressing it out of such ports is further subject to the filtering check result of the frame's ingress nickname on the ports (see Section 5.3).

5.3. Ingress Nickname Filtering

As shown in Figure 3, CE1 may send a multicast traffic in VLAN x to TRILL campus via a member RBridge of a RBv (say RB1). The traffic is then TRILL-encapsulated by RB1 and delivered through TRILL campus to multi-destination receivers. RB2 may receive the traffic, and egress it to CE1 if it is the DF for VLAN x on the port to MC-LAG1. Then the traffic loops back to CE1 (see Section 3.2 of [AAProb]).

To fix the above issue, the ingress nickname filtering check is introduced in this document. The idea of this check is to check the ingress nickname of a multi-destination TRILL data frame before egress a copy of it out of a port of a RBv. If the ingress nickname matches the pseudo-nickname of the RBv (associated with the port), the filtering check should be failed, and then the copy MUST NOT be egressed out of that RBv port and should be dropped. Otherwise, the copy is egressed out of that port if it has also passed other checks, such as the appointed forwarder check in Section 4.6.2.5 of [RFC6325] and the DF check in Section 5.2.

Note that ingress nickname filtering check has no effect on the multi-destination native frames received on access ports and replicated to other local ports (including RBv ports), since there is no ingress nickname associated with such frames. Furthermore, for the RBridge regular access ports, there is no pseudo-nickname associated with them; so no ingress nickname filtering check is required on those ports.

More details of data frame processing on RBv ports are given in the next section.

6. Data Frame Processing

Although there are five types of Layer 2 frames in TRILL basic protocol (see Section 1.4 of [RFC6325]), e.g., native frame, TRILL data frame, TRILL control frames, etc., RBv's pseudo-nickname is only used for native frame and TRILL data frame in this document.

6.1. Native Frames Ingressing

When RB1 receives a unicast native frame from one of its ports that enable end-station service, it processes the frame as described in Section 4.6.1.1 of [RFC6325] with the following exception.

- If the port is a RBv port, RB1 uses the RBv's pseudo-nickname, instead of one of its regular nickname(s) as the ingress nickname when doing TRILL encapsulation on the frame.

When RB1 receives a native BUM (Broadcast, Unknown unicast or Multicast) frame from one of its access ports (including regular access ports and RBv ports), it processes the frame as described in Section 4.6.1.2 of [RFC6325] with the following exceptions.

- If the incoming port is a RBv port, RB1 uses the RBv's pseudo-nickname, instead of one of its regular nickname(s) as the ingress nickname when doing TRILL encapsulation on the frame.

- For the copies of the frame replicated locally to RBv ports, there are two cases as follows:

  * If the outgoing port(s) is associated with the same pseudo-nickname as that of the incoming port, the copies are forwarded out of that outgoing port(s) after passing the appointed forwarder check for the frame's VLAN. That is to say, the copies are processed on such port(s) as Section 4.6.1.2 of [RFC6325].

  * Else, the Designated Forwarder (DF) check is further made on the outgoing ports for the frame's VLAN after the appointed forwarder check. The copies are dropped on the ports that failed the DF check (i.e., RB1 is not DF for the frame's VLAN on the ports); otherwise, the copies are forwarded out of the ports that pass the DF check (see Section 5.2).

For such a received frame, the MAC address information learned by observing it, together with the MC-LAG ID of the incoming port SHOULD be shared with other member RBridges in the group (see Section 7).

6.2. TRILL Data Frames Egressing

This section describes egress processing of the TRILL data frames received on a member RBridge (say RBn) of a RBv. Section 6.2.1 describes unicast TRILL data frames egress processing and Section 6.2.2 specifies the multi-destination TRILL data frames egressing.

6.2.1. Unicast TRILL Data Frames

When receiving a unicast TRILL data frame, RBn checks the egress nickname in the TRILL header of the frame. If the egress nickname is one of RBn's regular nicknames, the frame is processed as defined in Section 4.6.2.4 of [RFC6325].

If the egress nickname is the pseudo-nickname of one local RBv, RBn is responsible for learning the source MAC address. The learned {Inner.MacSA, Inner.VLAN ID, ingress nickname} triplet SHOULD be shared within the edge group (See Section 7).

Then the frame is de-capsulated to its native form. The Inner.MacDA and Inner.VLAN ID are looked up in RBn's local forwarding tables, and one of the three following cases may occur:

- If the destination end station identified by the Inner.MacDA and Inner.VLAN ID is on a local link, the native frame is sent onto that link.

- Else if RBn can reach the destination through another member RBridge RBk, it re-encapsulates the native frame into a unicast TRILL data frame and sends it to RBk. RBn uses RBk's regular nickname, instead of the pseudo-nickname as the egress nickname for the re-encapsulation, and the ingress nickname remains unchanged [RFC7180]. If the hop count value of the frame is too small for the frame to reach RBk safely, RBn SHOULD increase that value properly in doing the re-encapsulation. [NOTE: When receiving that re-encapsulated TRILL frame, as the egress nickname of the frame is RBk's regular nickname rather than the pseudo-nickname of a local edge group, RBk will process it as Section 4.6.2.4 of [RFC6325], and will not re-forward it to another RBridge.]

- Else, RBn does not know how to reach the destination; it sends the native frame out of all the local ports on which it is appointed forwarder for the Inner.VLAN.

6.2.2. Multi-Destination TRILL Data Frames

When RB1 receives a multi-destination TRILL data frame, it checks and processes the frame as described in Section 4.6.2.5 of [RFC6325] with the following exception.

- o On each RBv port where RBn is the appointed forwarder for the frame's Inner.VLAN, the Designated Forwarder check (see Section 5.2) and the Ingress Nickname Filtering check (see Section 5.3) are further performed. For such an RBv port, if either the DF check or the filtering check fails, the frame MUST NOT be egressed out of that port. That is to say, 1) if the port is associated with the same pseudo-nickname as the ingress nickname of the frame, the frame SHOULD be discarded; or 2) if RBn is not the DF for the frame's Inner.VLAN on the port, the frame SHOULD also be discarded; otherwise, the frame can be egressed out of the port.

7. MAC Information Synchronization in Edge Group

An edge RBridge, say RB1 in MC-LAG1, may have learned a MAC&VLAN 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 MAC&VLAN 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. To avoid the unnecessary flooding, RB1 SHOULD share the correspondence with other RBridges of MC-LAG1. RB1 synchronizes the correspondence by using MAC-RI sub-TLV [RFC6165] in its ESADI LSPs [ESADI].

On the other hand, RB2 has learned the MAC&VLAN of CE1 when CE1 sends a packet to h1 through RB2. The returning traffic from h1 may go to RB1. RB1 may have not CE1's MAC&VLAN stored. Therefore it has to flood the traffic out of its all access ports where it is appointed forwarder for the VLAN (see Section 6.2.1). Such flooding is unnecessary since the returning traffic is almost always expected and RB2 had learned the CE1's MAC&VLAN information. To avoid that unnecessary flooding, RB2 SHOULD share the MAC&VLAN with other RBridges of MC-LAG1. RB2 synchronizes the MAC&VLAN by using MAC-RI-LAG sub-TLV (see Section 9.3) in its ESADI LSPs [ESADI].

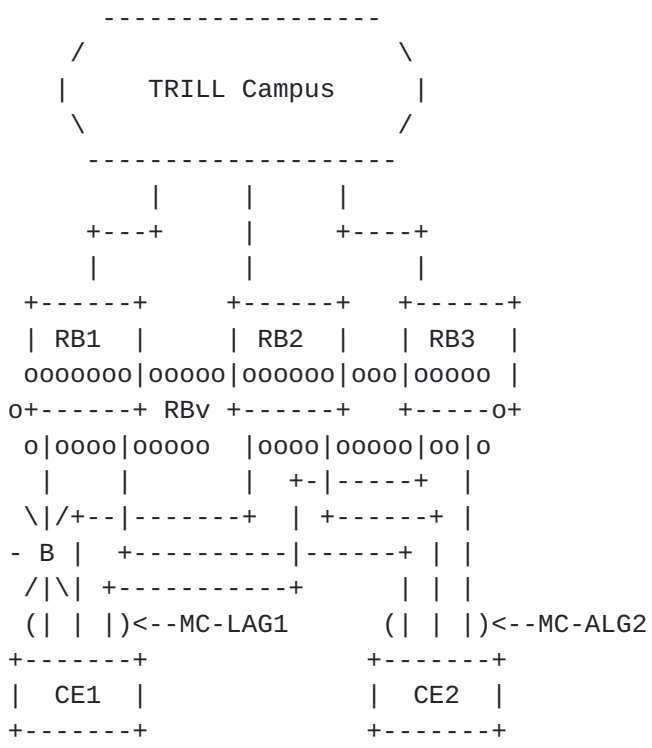
Besides the member RBridges, if there are other RBridges participating in ESADI for the VLAN and RB2 wants to distribute this MAC&VLAN to these ESADI RBridges, it SHOULD also put the MAC&VLAN into MAC-RI sub-TLV where the value of the Topology-id/Nickname field is pseudo-nickname of the RBv providing service for this MC-LAG. On

receipt of this sub-TLV, remote ESADI RBridges can learn that the MAC&VLAN is reachable through the pseudo-nickname.

8. Member Link Failure in RBv

As shown in Figure 4, suppose the link RB1-CE1 fails. Although a new RBv will be formed by RB2 and RB3 to provide active-active service for MC-LAG1 (see Section 5), the unicast traffic to CE1 might be still forwarded to RB1 before the remote RBridge learns CE1 is attached to the new RBv. That traffic might be disrupted by the link failure. Section 8.1 discusses the failure protection in this scenario.

Since the fact is true that multi-destination traffic can reach all member RBridges of a RBv and be egressed to CE1 by either RB2 or RB3 in the new RBv, the protection of down-link failure is not addressed in this document.



B - Failed Link or Link bundle

Figure 4 Member Link Failure in MC-LAG1

8.1. Link Protection for Unicast Frame Egressing

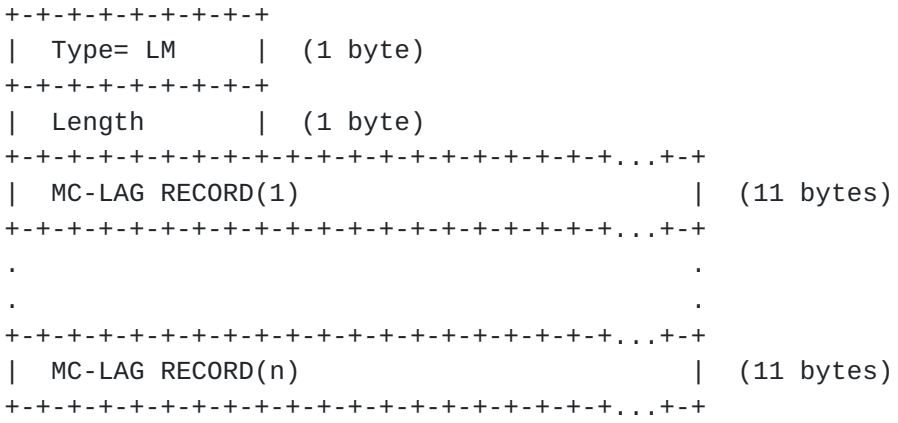
When the link CE1-RB1 fails, RB1 loses its direct connection to CE1. The MAC entry through the failed link to CE1 is removed from RB1's local forwarding table immediately. Another MAC entry learned from edge RBridge of the original RBv is installed into RB1's forwarding table (see Section 9.3). Then when a TRILL data frame to CE1 is delivered to RB1, it can be re-encapsulated (ingress nickname remains unchanged and egress nickname is replaced by RB2's nickname) and forwarded based on the above installed MAC entry (see the bullet 2 in Section 6.2.1). Then RB2 receives the frame egresses it to CE1.

After the failure recovered, RB1 returns to the original RBv and learns that it can reach CE1 via link CE1-RB1 again by observing CE1's native frames or from the synchronization in Section 7, then it restores the MAC entry to its previous one.

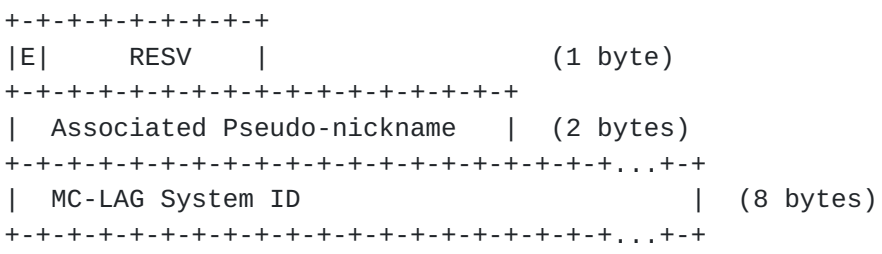
9. TLV Extensions for Edge RBridge Group

9.1. LAG Membership (LM) Sub-TLV

This TLV is used by edge RBridge to announce its associated MC-LAG information. It is defined as a sub-TLV of Multi-Topology-Aware Port Capability (MT-PORT-CAP) TLV [RFC6165]. It has the following format:



where each LAG_ID record is of the following form:



- LM (1 byte): Defines the type of Edge Membership sub-TLV, #TBD.

- Length (1 byte): 11*n bytes, where there are n MC-LAG Records.

- OE(1 bit): an flag indicating whether or not the MC-LAG wants to occupy an edge group Exclusively; 1 for occupying exclusively. By default, it is set to 0. This bit is used for edge RBridge group auto-discovery (see Section 4.1). For any one MC-LAG, the values of this flag might conflict in the LSPs advertised by different edge RBridges. In that case, the flag for this MC-LAG is considered as 1.

- RESV (7 bits): Transmitted as zero and ignored on receipt.

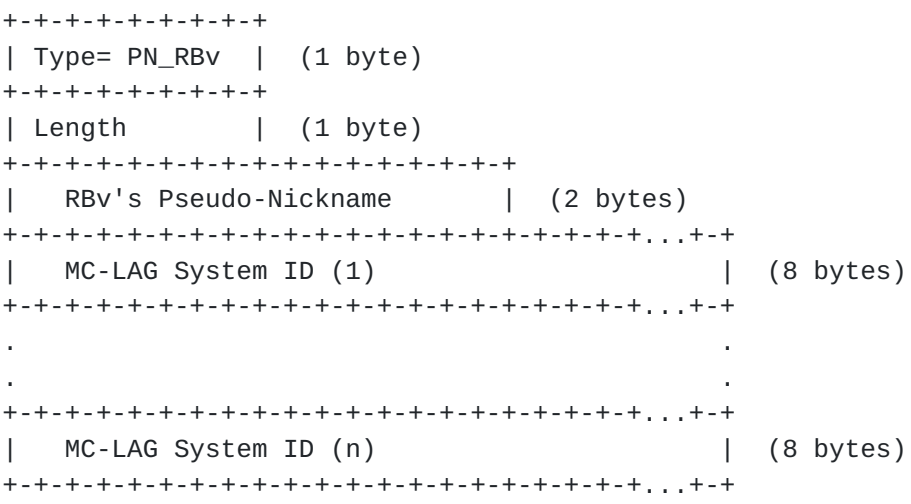
- Associated Pseudo-nickname (2 bytes): In a MC-LAG record, it suggests the pseudo-nickname of the edge group that the MC-LAG belongs to. If the MC-LAG does not belong to any edge group (RBv), this field MUST be set to zero. It is used by the originating RBridge to help the vDRB to reuse pseudo-nickname of an edge group (see Section 4.2).

- MC-LAG System ID (8 bytes): The System ID of the MC-LAG as specified in Section 6.3.2 in [IEEE802.1ax-rev]..

On receipt of such a sub-TLV, if RBn is not a MC-LAG related edge RBridge, it ignores the TLV but still stores the TLV in its database. Otherwise, the new copy of the sub-TLV is compared with its old copy stored in local database. If the fact is true that new MC-LAGs are found or old ones are withdrawn, and that those MC-LAGs are also configured on RBn, it triggers RBn to perform "Member RBridges Auto-Discovery" approach described in Section 4.1.

9.2. PN-RBv sub-TLV

PN-RBv sub-TLV is used by a Designated RBridge of a Virtual RBridge (vDRB) to appoint Pseudo-nickname for the MC-LAGs that are served by the edge group. It is defined as a sub-TLV of Multi-Topology-Aware Port Capability (MT-PORT-CAP) TLV [RFC6165]. It has the following format:



- PN_RBv (1 byte): Defines the type of this sub-TLV, #TBD.

- Length (1 byte): 2+8*n bytes, where there are n MC-LAG System IDs.

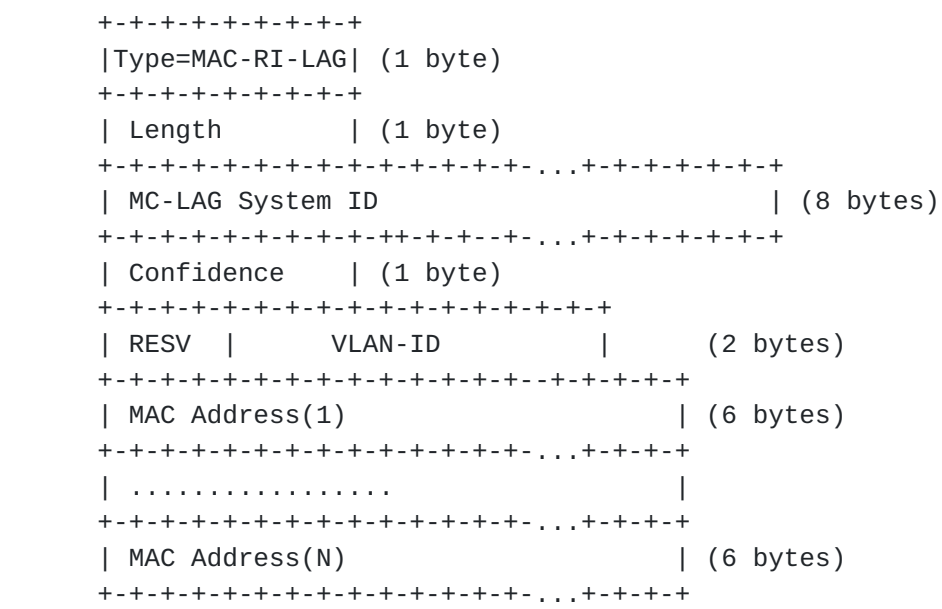
- RBv's Pseudo-Nickname (2 bytes): The appointed pseudo-nickname for the RBv that serves for the MC-LAGs listed in the following fields.

- MC-LAG System ID (8 bytes): The System ID of the MC-LAG as specified in Section 6.3.2 in [IEEE802.1ax-rev].

On receipt of such a sub-TLV, if RBn is not a MC-LAG related edge RBridge, it ignores the TLV but still stores the TLV in its database. Otherwise, if RBn is also a member RBridge of the RBv identified by the list of MC-LAGs, it associates the pseudo-nickname with the ports of these MC-LAGs and downloads the association onto data plane.

9.3. MAC-RI-LAG sub-TLV

MAC-RI-LAG sub-TLV is used by the member RBridges of an edge group to share MAC addresses learned from local RBv ports. It is defined as a sub-TLV of Multi-Topology-Aware Port Capability (MT-PORT-CAP) TLV [RFC6165]. It has the following format:



- PN_RBv (1 byte): Defines the type of this sub-TLV, #TBD.

- Length (1 byte): Total number of bytes contained in the value field given by 11 + 6*n bytes, where there are n MAC addresses.

- MC-LAG System ID (8 bytes): The System ID of the MC-LAG as specified in Section 6.3.2 in [IEEE802.1ax-rev]. This ID identifies the MC-LAG for all subsequent MAC addresses.

- Confidence (1 byte): This carries an 8-bit quantity indicating the confidence level in the MAC addresses being transported.

- RESV (4 bits): MUST be sent as zero and ignored on receipt.

- VLAN-ID (12 bits): This carries a 12-bit VLAN identifier that is valid for all subsequent MAC addresses in this TLV, or the value zero if no VLAN is specified. If this sub-TLV is used in the ESADI LSPs [ESADI], the VLAN ID MUST be sent as zero and ignored on receipt.

- MAC Address (6 bytes): This is the 48-bit MAC address learned from local attached end-station in the MC-LAG.

This TLV can be carried multiple times in a message and in multiple messages. On receipt of this sub-TLV, if the edge RBridge has not the MC-LAG configured, it ignores this sub-TLV. Otherwise, it parses this TLV and treats the MAC&VLANs as if it learned them on local operational port to the MC-LAG; else it learns the MAC&VLANs can be reached through the originator of this TLV if the port to the MC-LAGs is not operational.

10. OAM Frames

Attention must be paid when generating the OAM frames. When an OAM frame is generated with the ingress nickname of RBv, the originator RBridge's nickname MUST be included in the OAM message to ensure the response is returned to the originating member of the RBv group.

11. Configuration Consistency

It is important that VLAN membership of member ports of end switch SW1 is consistent across all of the member ports in the point-point scenario. Any inconsistencies in VLAN membership may result in packet loss or non-shortest paths.

Take Figure 1 for example, suppose RB1 configures VLAN1 and VLAN2 for the link CE1-RB1, while RB2 only configures VLAN1 for the CE1-RB2 link. Both RB1 and RB2 use the same ingress nickname RBv for all frames originating from CE1. Hence, a remote RBridge RBx will learn that CE1's MAC address in VLAN2 is originating from RBv. As a result, on the returning path, remote RBridge RBx may deliver VLAN2 traffic to RB2. However, RB2 does not have VLAN2 configured on CE1-RB2 link and hence the frame may be dropped or has to be redirected to RB1 if RB2 knows RB1 can reach CE1 in VLAN2.

12. Security Considerations

This draft does not introduce any extra security risks. For general TRILL Security Considerations, see [RFC6325]. For ESADI Security Considerations, see [ESADI].

13. IANA Considerations

IANA is requested to allocate code points for the 3 sub-TLV defined in Section 9.

14. Acknowledgements

We would like to thank Mingjiang Chen for his contributions to this document. Additionally, we would like to thank Erik Nordmark, Les Ginsberg, Ayan Banerjee, Dinesh Dutt, Anoop Ghanwani, Janardhanan Pathang, and Jon Hudson for their good questions and comments.

15. Contributing Authors

Weiguo Hao
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China

Phone: +86-25-56623144
Email: haoweiguo@huawei.com

16. References

16.1. Normative References

[AAProb] Li, Y., Hao, W., Perlman, R., Hudson, J., and H. Zhai, "Problem Statement and Goals for Active-Active TRILL Edge", draft-ietf-trill-active-active-connection-prob-01.txt Work in Progress, April 2014.

[CMT] Senevirathne, T., Pathangi, J., and J. Hudson, "Coordinated Multicast Trees (CMT)for TRILL", draft-ietf-trill-cmt-00.txt Work in Progress, April 2012.

[ESADI] Zhai, H., Hu, F., Perlman, R., Esstlake, D., and O. Stokes, "TRILL: ESADI (End Station Address Distribution Information) Protocol", draft-ietf-trill-esadi-07.txt Work in Progress, April 2014.

[RFC1195] Callon, R., "Use of OSI IS-IS for routing in TCP/IP and dual environments", RFC 1195, December 1990.

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

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

[RFC6325] Perlman, R., Eastlake, 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.

[RFC7176] Eastlake, 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.

[RFC7180] Eastlake, D., Zhang, M., Ghanwani, A., Manral, V., and A. Banerjee, "Transparent Interconnection of Lots of Links (TRILL): Clarifications, Corrections, and Updates", RFC 7180, May 2014.

16.2. Informative References

[IEEE802.1ax-rev] "Local and Metropolitan Area Networks - Link Aggregation", IEEE P802.1AX-REV/D3.2 , February 2014.

Authors' Addresses

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

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

Tissa Senevirathne
Cisco Systems
375 East Tasman Drive
San Jose, CA 95134
USA

Phone: +1-408-853-2291
Email: tsenevir@cisco.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

Donald Eastlake 3rd
Huawei
155 Beaver Street
Milford, MA 01757
USA

Phone: +1-508-333-2270
Email: d3e3e3@gmail.com

Mingui Zhang
Huawei
Huawei Building, No.156 Beiqing Rd.
Beijing, Beijing 100095
China

Email: zhangmingui@huawei.com

Yizhou Li
Huawei
101 Software Avenue,.
Nanjing, Nanjing 210012
China

Email: liyizhou@huawei.com