TRILL Working Group H. Zhai
Internet-Draft ZTE
Intended status: Standards Track T. Senevirathne
Expires: December 31, 2013 Cisco Systems
R. Perlman
Intel Labs
D. Eastlake 3rd
M. Zhang
Huawei
F. Hu
ZTE
June 29, 2013

RBridge: Pseudo-Nickname
draft-hu-trill-pseudonode-nickname-05

Abstract

RBridges provide provide end-station services to their attached end stations. To avoid potential frame duplication and loops, the rule that only one edge RBridge is allowed to be the frame forwarder of a VLAN on a shared LAN segment is employed by base TRILL protocol, even though there are multiple RBridges attached to that segment. However, in some application scenarios, for example an end station is multi-homed to multiple RBridges, there is a need to improve the resiliency and increase the available network bandwidth of the connection. This means all those RBriges attached to the end station can act as the frame forwarders of a specific VLAN. This kind of active-active connection violates the above rule. The violation may bring some additional problems, such as the flip-flopping of the nickname-MAC correspondences for such end stations in remote RBridges' forwarding tables, frame dropping because of failure of Reverse Path Forwarding Check(RPF Check) on RBridges, etc. The RPF Check problem has been addressed in [CMT]. This document proposes the concept of Virtual RBridge, along with the pseudo-nickname configuration for this Virtual RBridge, to address the above problems in accompany with [CMT].

Status of This Memo

This Internet-Draft will expire on December 31, 2013.

Copyright Notice

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. The design supports VLANs and optimization of the distribution of multi-destination frames based on VLANs and IP derived multicast groups. Devices that implement TRILL are called RBridges.

In TRILL protocol, RBridges are identified by nicknames (16-bits). At the edge of TRILL network, some RBridges connect to legacy networks on one side and connect to the TRILL network on the other side. These RBridges are called edge RBridges. For the connectivity between the two types of network, edge RBridges provide frame forwarding service to end stations located in legacy networks. When receiving a native frame from such a local end station S, the service edge RBridge RB1 encapsulates the frame in a TRILL header, addressing the packet to RBridge RBx to which the destination end station D is attached. The TRILL header contains an "ingress RBridge nickname" field (filled with RB1's nickname), an "egress RBridge nickname" field (filled with RBx's nickname), and a hop count. On receiving such a frame, RBx removes the TRILL header and forwards it in native form to D. Meanwhile, based on the de-capsulation of that frame, RBx learns the { ingress RBridge nickname, source MAC address, VLAN ID } triplet. Edge RBridges maintain such triplets in their forwarding tables for the future forwarding of native frames.

Due to failures, reconfiguration and other network dynamics, service edge RBridge for S may change over from RB1 to another edge RBridge. In this event, remote traffic addressed to S will be still forwarded to RB1 by remote RBridge RBx before perceiving this change, and then the traffic gets dropped at RB1, causing traffic disruption. Furthermore, to improve resiliency and maximize the available network bandwidth, an end station typically is multi-homed to several edge RBridges and treats all the uplink links as a link bundle. In this scenario, all those edge RBridges work in an active-active load sharing model to provide end-station services for an end station even in same a VLAN. When remote RBridge RB2 receives different frames,

which are originated by such an end station S and ingressed into TRILL campus by different such edge RBridge, flip-flopping of ingress RBridge nickname for MAC of S will be observed by RBx during de-capsulating such frames. This flip-flopping will cause disorder of different frames in traffic, worsening the traffic disruption.

In this document, concept of Virtual RBridge group, together with its Pseudo-nickname, is introduced to address the above issues. For a member RBridge in such a group, it uses the pseudo-nickname of this group, instead of its own device nickname, as ingress RBridge nickname when encapsulating a frame to its TRILL form with a TRILL header. So, in a RBridge Group, even if there are more than one RBridge providing end-station services for a end station or the service RBridge changes over from one member RBridge to another in same set of VLANs, the ingress RBridge nickname for the MAC of this end station will still remain unchanged in remote RBridges' forwarding tables. In this document, the concept of a Virtual RBridge group, together with its Pseudo-nickname, is introduced to address the rest of above issues. 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 into TRILL campus. So, in such a RBridge Group, even if there are more than one RBridge providing frame forwarding service for an end station or the service RBridge changes over from one to another member RBridge in same a group, the ingress RBridge nickname associated to this end station's MAC address(es) needs not be changed in remote RBridges' forwarding tables.

This document is organized as following: Section 2 is problem statement, which describes why virtual RBridge and its pseudo-nickname are required. Section 3 gives the concept of virtual RBridge. Section 5 describes the consideration for pseudo-nickname used in ingressing multi-destination frames. Section 6 covers processing of transit frame traffic when considering pseudo-nickname.

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

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].

1.2. Contributors

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.

2. Problem Statement

2.1. Appointed Forwarders on Shared Links

Even there are multiple RBridges on a shared link, together with end stations, only one RBridge is allowed to provide frame forwarding services in VLAN-x to the end stations to avoid possible frame duplication or loops in TRILL campus. The service RBridge is called VLAN-x Appointed Forwarder (AF).

However, AF for any set of VLANs on a shared link may change over from one RBridge to another, due to network dynamics such as failures and configuration changes. RBridges rely on LSPs to propagate these network dynamics. However, the propagation is time consuming and the network may take a considerable long time to converge. Before the network converges, remote RBridges may continue to forward traffic to the previous AF and the traffic is dropped at the previous egress RBridge, causing traffic disruption.

```
                    ...........................................
                   :                            TRILL Network  :
                   : +-----+      /\-/\-/\-/\-/\              :
              +------| RB1 |-----/              \              :
              |    : +-----+    /                \             :
           +-----+ :    |      /      Transit     \    +-----+ :
     S1 o--| SW1 | :    |     <      RBridges      >---| RBx |---o Sx
           +-----+ :    |      \      Campus      /    +-----+ :
              |    : +-----+    \                /             :
              +------| RB2 |-----\              /              :
                   : +-----+      \/\-/\-/\-/\-/              :
                   :                                           :
                    ...........................................
```

Figure 1 Multi-homed to TRILL Network

2.2. Multi-homed to TRILL Network

In order to improve the reliability of connection to a TRILL network, multi-homing technique may be employed by a legacy device which can be a switch or end host. Take Figure 1 as an example, switch SW1 multi-homed to the TRILL network by connecting to RB1 and RB2 with respective links. Then the end station S1 can continue to get frame forwarding service from the TRILL network even if one of its up-links (e.g., SW1-RB1) fails.

SW1 may treat the two links as a link bundle, so that the two links form active-active load sharing model instead of the previous active-standby model. That is to say, in Figure 1, two RBridges (i.e., RB1 and RB2) provides frame forwarding service to S1 simultaneously in a VLAN. As stated previously, simultaneous frame forwarding may result in frame duplication, loops and the flip-flopping of the ingress RBridge associated to the MAC of S1 in remote RBridges' (e.g., RBx) forwarding tables. The flip-flopping in turn causes packet disorder in reverse traffic and worsens the traffic disruption. Therefore, the concept of Virtual RBridge, together with its nickname, is introduced in the following section to fix these issues.

3. Concept of Virtual RBridge and Pseudo-nickname

A Virtual RBridge (RBv) represents a group of different end station service ports on different edge RBridges. After joining RBv, such a RBridge port is called a member port of RBv, and such a RBridge becomes a member RBridge of RBv. An RBv is identified by its virtual nickname in TRILL campus, and this nickname is also referred to as pseudo-nickname in this document.

After joining a RBv, a member RBridge will announce its connection to RBv by including the information of that RBv, e.g., the pseudo-nickname of RBv, in its self-originated LSP. From such LSPs, other RBridges that are not members of the RBv believe those member RBridges are connected to RBv.

When a native frame from an end station S1 is received from such a port, the member RBridge encapsulates the frame with the RBv's nickname, instead of its own nickname, as the ingress nickname. When the destination RBridges receive and de-capsulate this frame, they will learn that S1 is reachable through RBv.

For a member RBridge, it MUST move out of a RBv and clear the RBv's information from its self-originated LSPs when it loses all of its member ports of the RBv, due to port failure, configuration, etc.

NOTE1: In the multi-homing scenario of same a RBv, it is RECOMMENDED that all devices multi-homed to that RBv SHOULD have operational links to all the member RBridges of that RBv unless one or more of the links failed or administratively down.

3.1. VLAN-x Appointed Forwarder for member interfaces in RBv

If member RBridges in RBv cannot see each others' Hellos on their member ports (e.g., in the multi-homing scenario), then each RBridge becomes Designated RBridge (DRB) for that port and appoints itself as AF for all VLANs.

On the other hand, if they can see each others' Hellos on the member ports in RBv (e.g., in the shared link scenario), the TRILL Hello protocol in [RFC6325] is used for DRB election and for VLAN-x AFs appointment on those ports. Then the DRB appoints different member ports as AFs for different sets of VLANs.

By using the AF framework specified in [RFC6325], a unified framework of RBv for both Multi-homed and shared LAN edge connectivity is provided in this document. It also allows:

- Detection and protection against mis-configuration at the edge, e.g., on the device SW1 the two interfaces are not configured as multi-homing then RB1 and RB2 work in an unexpected active-standby mode rather than expected active-active mode for S1 or

- Avoidance of loops in the event that S1 and S2 were connected by a native Ethernet Link. In this event, RB1's Hellos originated on link RB1-SW1 will be forwarded by S1 through the Ethernet Link to S2 then received by RB2, and vice versa. Therefore, RB1 and RB2 work in an active-standby mode for S1 (or S2) in any VLAN to avoid potential forwarding loops.

3.2. Announcing Pseudo-Nickname of RBv

Each member RBridge advertises the RBv's pseudo-nickname using the nickname sub-TLV [rfc6326bis], along with its regular nickname(s), in its LSPs. For a member RBridge, when its last member port is disconnected to RBv, it MUST leave from RBv and clear RBv's pseudo-nickname from its update LSPs.

RBv's pseudo-nickname is ignored when determining the distribution tree root for the campus. 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.

4. Acquision of RBv's Pseudo-nickname

In active-active connection scenario, a device is typically connected to multiple edge RBridges via a link bundle. From the perspective of the edge RBridges, the device can be identified by a globally unique identifier; and this identifier is called Link Aggregation Group Identifier (LAG-ID) in this document.

For an edge RBridge, if it has one or more operational ports through which a device multi-homed to it, it MUST announce that LAG-ID of the device to all other edge RBridges via RBridge Channel messages [RBChannel]. Based on the LAG-IDs received from other edge RBridges, edge RBridges can pick up, from TRILL campus, all the edge RBridges that can join same a RBv(See Section Section 4.1 for more details) and elect one of them as the Designated RBridge (DRB) for that RBv. That DRB is responsible for appointing an available pseudo-nickname for that RBv.

4.1. Picking up RBridges for different RBvs

```
                          ---------------------
                        /                        \
                       |       TRILL Campus        |
                        \                        /
                         -----------------------
                            |   |         |   |
                    +--------+   |         |   +---------+
                    |           |         |             |
               +--------+    +--------+    +--------+    +--------+
               |  RB1   |    |  RB2   |    |  RB3   |    |  RB4   |
               +--------+    +--------+    +--------+    +--------+
                 |    |        |  |          | | | |        |    |
                 | +--|-------+  | +--------|-+ | +--------|--+ |
                 | |  +---------+ | |        |   |          |  | |
                 | | +----------|-|-|-------+   | +-------+  | |
          LAG1->(| | |) LAG2->(| | |)   LAG3->(| |)  LAG4->(| |)
               +--------+    +--------+     +--------+     +--------+
               |  CE1   |    |  CE2   |     |  CE3   |     |  CE4   |
               +--------+    +--------+     +--------+     +--------+
```

Figure 2 Different LAGs to TRILL Campus

For each edge RBridge with available multi-homed devices connected, it MUST announce a list of LAG-IDs of all of those devices to all other edge RBridges via RBridge Channel message (See Section 8.1 for more details). Take Figure 2 as an example, RB1 and RB2 announce {LAG1, LAG2} in their lists respectively; RB3 announces {LAG1, LAG2, LAG3, LAG4}; and RB4 announces {LAG3, LAG4}, respectively.

Based on the LAG-IDs contained in these lists, each RBridge can know which set of RBridges each LAG is multi-homed to. For example, all the 4 RBridges know the information as follows:

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}

In the above table, there might be some LAGs that multi-homes only to one single RBridge due to mis-configuration or link failure, etc. Those LAGs are considered as invalid entries. Then each of the relative edge RBridges performs the following approach to pick up which valid LAGs can be served by same a RBv.

Step 1: Take all the valid LAGs that have their OE-flags (Occupying Exclusively a RBv) 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 mumber 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 the same RBridges as that of LAG_i and take the LAGs out of the table. Then appoint RBv_i as those LAGs' servicing RBv.

Step 5: Repeat Step 3-4 for the left LAGs in the table.

For the example given in Figure 2, 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 RBrdigs; 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 RBrdigs, 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 RBridge 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 in this RBv via RBridge Channel message (Refer Section 8.3 for more details).

If possible, the DRB SHOULD attempt to reuse the RBv's previous pseudo-nickname to avoid traffic disruption caused by pseudo-nickname changing. If there is no such a previous nickname available, the DRB will acquire a new available nickname from TRILL campus and announce it as the RBv's pseudo-nickname.

5. Distribution Trees for Member RBridges in RBv

In TRILL, RBridges use distribution trees to forward multi-destination frames. In the TRILL header of the multi-destination frames, the ingress nickname identifies the ingress RBridge and the egress nickname specifies the root of the chosen distribution tree. After receiving a multi-destination TRILL data frame, RBn performs Reverse Path Forwarding (RPF) check on the multi-destination frame to avoid temporary multicast loops during topology changes.

RPF specifies that a multi-destination TRILL data frame ingressed by an RBridge and forwarded on a distribution tree can only be received by RBn on an expected port. If the frame is not received from that port, it MUST be dropped.

However, member RBridges use RBv's pseudo-nickname other than their own nicknames as the ingress nickname when they forward unicast or non-unicast native frames. Therefore, when these TRILL data frames arrive at RBn, they will be treated as frames ingressed by the same RBridge, i.e., RBv. If they are multi-destination frames and the same distribution tree is chosen by different member RBridges to forward these frames, they may travel on the tree and arrive at RBn on different ports. Then the RPF check is violated, and some of the frames reaching the RBridge on unexpected ports will be dropped by RBn.

[CMT] proposes to assign different distribution trees for each member RBridge to fix the above RPF check issue, and makes use of the Affinity sub-TLV defined in [rfc6326bis] to achieve this kind of assignment.

This document supposes the approach proposed in [CMT] is supported by member RBridges of RBv.

To avoid duplication traffic being egressed through RBv to a multi-homing end-device, multi-destination TRILL traffic arriving at RBv on a tree (say Tx), only the Tx's Designated Forwarder is allowed to egress it to the device.

When a member RBridge joins in or leaves from a virtual RBridge group, the assignment of distribution trees may change. That change is beyond the scope of this document.

6. Frame Processing

Although, there are five types of Layer 2 frames in [RFC6325], e.g., native frame, TRILL data frame, TRILL control frames, etc., pseudo-nickname of RBv is only used for native frame and TRILL data frame in this specification.

6.1. Native Frames Ingressing

When RB1 receives a native frame on one of its valid member ports of RBv, it uses the pseudo-nickname of RBv, instead of its own nickname, as ingress nickname, if it is the appointed forwarder for the VLAN of the frame on the port. If the frame is not received on a member port, RB1 MUST NOT use RBv's pseudo-nickname as ingress nickname when doing TRILL-encapsulation on the frame. Otherwise, the reverse traffic may be forwarded to another member RBridge that does not connect to the link containing the destination, which may cause the traffic disruption.

If the above native frame is ingressed by RB1 as a multi-destination TRILL data frame, e.g., its destination is unknown to RB1 or it is non-unicast frame, RB1 can only choose one of its assigned distribution trees for RBv to distribute the TRILL-encapsulated frame [CMT]. If not so, the multi-destination TRILL data frame will fail RPF check on another RBridge and be dropped.

Furthermore, for such a frame, its source MAC address information ({ VLAN, Outer.MacSA, port }) is learned by default if its source address is unicast. Then the learned information is shared with other member RBridges of RBv (See Appendix A for more details for the information sharing).

6.2. 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 own nicknames, the frame is processed as defined in in [RFC6325].

If the egress nickname is RBv's pseudo-nickname and RBn is a member RBridge of RBv, RBn is responsible to learn the source MAC address. If the learned { Inner.MacSA, Inner.VLAN ID, ingress nickname } triplet is a new one or it updates a previously learned one, this triplet SHOULD be shared with other member RBridges within the RBv (See Appendix A for more details for the triplet sharing).

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 address cache, and one of the three following cases occurs:

- o If the destination end station identified by the Inner.MacDA and Inner.VLAN ID is on a local link to RBv, this frame is egressed onto that link if RBn is the Inner.VLAN AF on this link.

- o Else if RBn can reach the destination through another member RBridge RBk, it tunnels the frame to RBx [ClearCorrect] by re-encapsulating the native frame into a unicast TRILL data frame. RBn uses RBk's own nickname, instead of RBv's pseudo-nickname as the egress nickname for the re-encapsulation, and remains the ingress nickname remains unchanged. 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 own nickname rather than the RBv!_s pseudo-nickname, RBk will process it as Section 4.6.2.4 in [RFC6325], and will not re-forward it to another RBridge.

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

6.2.2. Multi-Destination TRILL Data Frames

If RBn is the AF for the Inner.VLAN, the source MAC address is learned. If the learned { Inner.MacSA, Inner.VLAN ID, ingress nickname } triplet is a new one or updates a previously learned one, this triplet SHOULD be shared among the members RBridges within the virtual RBridge group (See Appendix A for more details for the triplet sharing).

Then a copy of the frame is de-capsulated into its native form. Before the native frame is sent out of the ports on which RBn is appointed forwarder for the Inner.VLAN, the following extra check is performed for each member port of RBv:

o Frames MUST only be forwarded out on member ports of RBv where RBn is the Designated Forwarder for the Tree Tx on which the frame was received.

7. Member Link Failure in RBv

As shown in Figure 3, suppose the link RB2-CE1 fails. Both unicast frames and multi-destination frames cannot be sent from RB2 to CE1. Section 7.1 discusses the failure protection for unicast frames egressing.

```
                              ------------------
                            /                    \
                           |     TRILL Campus     |
                            \                     /
                             ---------------------
                                 |     |     |
                             +---+     |     +----+
                             |         |          |
                         +------+   +------+   +------+
                         | RB1  |   | RB2  |   | RB3  |
                        o|oooooo|ooo|oooooo|ooo|oooooo|o
                       o +------+   +------+   +------+ o
                       o  |   |  \|/  |  |      |   |   o
                       o  | +-|-- B --+  +------|-+ |   o
                       o  | | |  /|\    +-------+ | |   o
                        oo|o|o|ooooooooo|ooooooooo|o|ooo
                          | | +---------|-------+ | |
                          | | +---------+       | | |
                         (| | |)<--LAG1        (| | |)<--LAG2
                        +-------+              +-------+
                        |  CE1  |              |  CE2  |
                        +-------+              +-------+
```

              B - Failed Link or Link bundle

                   Figure 3 Member Link Failure in LAG1

7.1. Link Protection for Unicast Frame Egressing

When the link CE1-RB2 fails, RB2 loses its direct connection to CE1. The MAC entry through the failed link to CE1 is removed from RB2's local forwarding table immediately. Another MAC entry through another member RBridge (say RB1) that has local link to CE1 is installed into RB2's forwarding table only if RB2 is still a member RBridge of RBv. Then when the TRILL data frames to CE1 are delivered to RB2, they can be re-encapsulated (ingress nickname remains unchanged and egress nickname is replaced with RB1's nickname) by RB2 and forwarded based on the above installed MAC entry. The member RBridge who receives the redirected frames will egress them to CE1.

When the failure recovers, RB2 will be aware that it can reach CE1 by observing CE1's native frames. Then RB2 installs the MAC entry for link CE1-RB2.

8. TLV Extensions for RBv

8.1. LAG Membership (LM) Sub-TLV

We propose to use LM sub-TLV to advertise the state of the RBridges' LAG membership. There are following 3 different events, as follows:

o Membership Add

o Membership Withdrawal

o Membership Refresh

```
+-+-+-+-+-+-+-+-+
| Type= LM      |                  (1 byte)
+-+-+-+-+-+-+-+-+
|   Length      |                  (1 byte)
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   RBv Nickname                |  (2 bytes)
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| RESV      |OC |                  (1 byte)
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                LAG-ID (1)     |  (10 bytes)
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
.                               .
.                               .
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                LAG-ID (n)     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 4 Edge Membership advertisement sub-TLV

where each LAG_ID record is of the following form:

```
+-+-+-+-+-+-+-+-+
|     RESV    |OE|                  (1 byte)
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Virtual Network ID(VNID)      |  (3 bytes)
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|              LAG ID            |  (6 bytes)
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

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

o Length (1 byte): Defines the length of this sub-TLV which should be greater than 3.

o RBv Nickname (2 bytes): 2 byte nickname of the RBv. By default, this field is zero. Otherwise, it indicates the pseudo-nickname that the originator of the TLV considers the RBv has used, which providing information for the DRB to reuse the RBv's previous pseudo-nickname.

o RESV (6 bits): Transmitted as zero and ignored on receipt.

o OE(1 bit): an flag indicating whether the end-device identified by the combination of the VNID and LAG-ID needs to Occupy a RBv exclusively or can share a RBv with other end-devices; 1 for occupying exclusively, and 0 for sharing. By default, it is set to 0.

o OC (2 bits): Define the operation code.

  * 00: Add (LAG-IDs in this sub-TLV are new and will trigger the process of picking up member RBridge for a RBv and the Designated Forwarder election on the relative edge RBridges).

  * 01: Withdrawal (LAG-IDs in this sub-TLV do not have an active links from the announcing RBridge for RBv, the process of picking up member RBridge for a RBv and Designated Forwarder election MUST be triggered on the relative edge RBridges).

  * 10: Refresh (LAG-IDs in this sub-TLV are being refreshed and no state change from the perspective of the announcing RBridge).

  * 11: Reserved and currently unused.

o VNID(24 bits): an identifier of an Virtual Network where the end-device populated. By default, this field is set to zero.

o LAG-ID (2 bytes): an unsigned positive integer that uniquely identifies an end device multi-homed to the RBv. This ID along with the VNIT is globally meaningful in the scope of the TRILL campus. For convenience, this ID can be one of the MAC addresses of the end-device..

When receiving such a sub-TLV, if the RBridge has no membership for the listed LAGs in the RBv, it ignores the sub-TLV. If it has the membership, receiving such a sub-TLV where the operation code is 00 or 01 will triggers it to re-calculate the Designated Forwarder on each tree for the listed LAGs.

8.2. PN-RBv sub-TLV

The DRB employs PN-RBv sub-TLV to announce the RBv's pseudo-nickname, along with all the LAGs serviced by this RBv, to other relative edge RBridges.

The format of this sub-TLV is as follows, where the LAG-ID Record has the same format as the Record of LM Sub-TLV.

```
+-+-+-+-+-+-+-+-+
| Type= PN_RBv  |                  (1 byte)
+-+-+-+-+-+-+-+-+
|   Length      |                  (1 byte)
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   RBv's Pseudo-Nickname       |  (2 bytes)
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     RESV      |                  (1 byte)
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|         LAG-ID RECORD (1)     |  (10 bytes)
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
.                               .
.                               .
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|         LAG-ID RECORD (n)     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

LAG-ID RECORDs list all the end-devices to which the RBv identified by the pseudo-nickname provides services.

After receiving such a sub-TLV, if the receipt RBridge has membership for at least one of the listed LAGs and accepts the DRB membership of the originator of the TLV, it uses the RBv identified by the pseudo-nickname to service the end-devices identified by some of the listed LAGs and multi-homed to it. Otherwise, the received sub-TLV is ignored.

9. 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.

10. 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 SW1-RB1, while RB2 only configures VLAN1 for the SW1-RB2 link. Both RB1 and RB2 use the same ingress nickname RBv for all frames originating from S1. Hence, a remote RBridge RBx will learn that MAC addresses from S1 on VLAN2 are originating from RBv. As a result, on the returning path, RBx may deliver VLAN2 traffic to RB2. However, RB2 does not have VLAN2 configured on SW1-RB2 link and hence the frame may be dropped or has to be redirected to RB1 if RB2 knows RB1 can reach S1 in VLAN2.

11. IANA Considerations

TBD.

12. Security Considerations

TBD.

13. 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.

14. Normative References

[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.

[ClearCorrect]
Eastlake 3rd, D., Zhang, M., Ghanwani, A., Manral, V., and A. Banerjee, "TRILL: Clarifications, Corrections, and

Updates", draft-ietf-trill-clear-correct-06.txt In RFC Editting Queue, July 2012.

[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.

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

[rfc6326bis] Eastlake 3rd, D., Banerjee, A., Ghanwani, A., and R. Perlman, "Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS", draft-eastlake-isis-rfc6326bis-07.txt Work in Progress, March 2012.

Appendix A. Rationale for MAC Sharing among Member RBridges

With the introduction of virtual RBridge, MAC flip-flopping problem in LAN or LAG is resolved. However, in order to forward traffic effectively, member RBridges should share some of their learned MAC addresses with each other. For example, see Figure 5 shown below.

```
                 ..........................................
                :                           TRILL Network  :
             ^  : +-----+      /\-/\-/\-/\-/\              :
            +-----| RB1 |-----/              \             :
           /    : +-----+    /                \            :
          /     :    |      /      Transit     \    +-----+ :
     S1 o  RBv :    |     <      RBridges      >---| RBx |---o Sx
          \     :    |      \      Campus      /    +-----+ :
           \    : +-----+    \                /            :
            +-----| RB2 |-----\              /             :
             V  : +-----+      \/\-/\-/\-/\-/              :
                :                                          :
                 ..........................................
```

Figure 5 RBv in LAG scenario

Take Figure 5 as an example, the VLAN-x native frames from S1 to Sx will enter TRILL campus via one member RBridge of the RBv (say RB1). RB1 learns the location of S1 in VLAN-x. However, RBx may deliver the

reverse traffic to RB2 if it thinks the shortest path to RBv is through RB2. If RB2 has not learned the location of S1 in VLAN-x from the MAC sharing, RB2 has to transmit the reverse traffic to S1 as unknown unicast.

Thus, the learned MAC addresses of attached end stations on one member RBridge SHOULD be shared with rest of the member RBridges in the same RBv. With these information shared, when RB2 receives reverse frames, it can determine how to forward them to S1. For example, it can redirect them to RB1 if link RB2-S1 fails.

Since RBx always delivers the reverse traffic to RBv via RB2, RB2 egresses the traffic and learns the location of Sx. But RB1 will not know where Sx is, if RB2 does not share this information with RB1. As a result, RB1 has to treat the traffic from S1 to Sx as traffic with unknown destination and flood it in TRILL, which adds additional forwarding burden on the TRILL network.

Therefore, in addition to local attached end station MAC addresses, the learned remote MAC addresses should also be shared among all member RBridges of a RBv. With such information shared, RB1 can treat the traffic to Sx as known destination traffic and unicast it to RBx.

The design for above MAC sharing is currently beyond the scope of this document.

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

Fangwei Hu
ZTE
889 Bibo Road, Pudong District
Shanghai, Shanghai 201203
China

Phone: +86 21 68896273
Email: hu.fangwei@zte.com.cn