INTERNET-DRAFT Mingui Zhang
Intended Status: Proposed Standard Donald Eastlake
Expires: February 23, 2014 Huawei
August 22, 2013

Problem Statement: TRILL Active/Active Edge
draft-zhang-trill-aggregation-04.txt

Abstract

This document specifies TRILL active/active edge which allows multiple RBridges concurrently forward data frames of the same VLAN on links bundled by a Multi-Chassis Link Aggregation Group. With this kind of connection, end nodes may increase the bandwidth and reliability of the access at the edge of TRILL campuses. It's required that no loop or duplication is caused by this new connection type. Besides this basic requirement, this document outlines other potential issues associated with TRILL active/active edge and investigates how these issues may be addressed.

Status of this Memo

Copyright and License Notice

Table of Contents

1. Introduction

TRILL makes use of the ISIS link state routing to provide least cost paths between TRILL switches (a.k.a. Routing Bridge, RBridge). When a multi-access LAN link connects end-stations to multiple RBridges, a single RBridge has to be appointed as the frame forwarder for each VLAN-x on this LAN link. Other RBridges MAY be appointed as frame forwarders for other VLANs but MUST be inhibited from forwarding frames for the same VLAN-x on this LAN link [RFC6349].

An MC-LAG can also be used to connect end-stations to multiple RBridges. There are two possible scenarios: (a) an end-station is connected to multiple RBridges by an MC-LAG directly; (b) end-stations are attached to a bridge and this bridge uses an MC-LAG to connect multiple RBridges. An MC-LAG may choose any component link to forward frames and never forwards between them. Therefore, it requires the up-connected RBridges to provide active/active attachment instead of the active/standby mode adopted in the Appointed Forwarder mechanism [RFC6349]. This kind of attachment allows end nodes increase the bandwidth and reliability of their access to the TRILL campus via MC-LAG.

Similar as a LAN link, an MC-LAG can be represented by a pseudonode. All member RBridges should report their adjacencies to this pseudonode using LSPs. In this way, RBridges attached to the same MC-LAG forms an active/active edge group. Other RBridges in the campus communicate with this pseudonode using forwarding paths computed according to ISIS link state routing. No additional add-on characteristics are required.

The baseline requirement is that the active/active edge MUST provide frame forwarding without causing loops or duplications to TRILL campus and the end node. In order to work properly, the TRILL active/active edge has to conduct several other issues. The purpose of this document is to outline these issues while specific solutions to address them are to be explored in the future as building blocks of the whole TRILL active/active edge mechanism.

The rest of this document is organized as follows. Section 2 gives acronyms and terminology. Section 3 provides an overview. Section 4 specifies the frame processing behaviors of member RBridges. Section 5 describes how pseudonode is set up. Section 6 explains the MAC sharing among member RBridges. Section 7 describes the self-healing issue. Section 8 investigates how to go through Reverse Path Forwarding Check without packet loss.

2. Acronyms and Terminology

2.1. Acronyms

MC-LAG: Multi-Chassis Link Aggregation Group
ISIS: Intermediate System to Intermediate System
TRILL: TRansparent Interconnection of Lots of Links
AF: Appointed Forwarder
DT: Distribution Tree
RPFC: Reverse Path Forwarding Check

2.2. 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].

In this document, the term "end node" means the end station or bridge connected to the TRILL active/active edge by MC-LAG.

Familiarity with [RFC6325], [RFC6327], and [RFC6349] is assumed in this document. As in [RFC6325], in this document the word "link" means a "bridged LAN", unless otherwise qualified.

3. Overview

If an end node (end station or bridge) uses an MC-LAG to connect multiple edge RBridges, it's expected that all these RBridges can ingress and egress frames for the end node. In contrast, if multiple RBridges are connected to a LAN link, only one of them can be appointed as the frame forwarder for each VLAN-x [RFC6349], as illustrated in Figure 2.1 (a). Other RBridges will be inhibited from ingressing and egressing frames for VLAN-x.

```
             +-----+                   +-----+
             | RBi |                   | RBi |(Remote RBridge)
             +-----+                   +-----+
           /\/\/\/\/\/\              /\/\/\/\/\/\
          /   Transit  \            /   Transit  \
         <    RBridges  >          <    RBridges  >
          \            /            \            /
           \/\/\/\/\/\/              \/\/\/\/\/\/
            |        |                |        |
         +-----+  +-----+          +-----+  +-----+
         | RB1 |--| RB2 |          | RB1 |--| RB2 |(Active/Active Edge)
         +-----+  +-----+          +-----+  +-----+
            AF\    /                     \   /
               +---+                    *******
               |LAN|                    * RBv * (Virtual RBridge)
               +---+                    *******
                                          | |(MC-LAG)
                                         +---+
                                         | E |
                                         +---+
         (a) Appointed Forwarder   (b) Active/Active Edge
```

Figure 2.1: TRILL Appointed Forwarder vs Active-Active Edge

As illustrated in Figure 2.1 (b), The end node 'E' are attached to both RB1 and RB2 using an MC-LAG. Each member RBridge can ingress and egress frames for the end node for VLAN-x. If each of them uses its own nickname as the ingress nickname, the remote RBridge may observe different locations for one MAC address at different time, which is referred as the "MAC move" problem in this document. The MAC move problem affects the path selection at the remote RBridge. Frames destined to the end node may go through different paths, which may cause frame disorder of a traffic flow.

In order to avoid the MAC move problem, each member RBridge should use a uniform nickname as the ingress nickname in TRILL data frame encapsulation. As shown in Figure 2.1 (b), member RBridges pretend there is an virtual RBridge connected to them, acting as the appointed forwarder of the end node. It is naturally to denote this virtual RBridge as a pseudonode. All RBridges connected to the MC-LAG forms adjacencies with the pseudonode. Other RBridges believe there is an RBridge RBv connecting RB1, RB2. Note that member RBridges SHOULD NOT announce they are VLAN-x Appointed Forwarder if VLAN-x is enabled on the MC-LAG.

Although the above example includes two edge RBridges, the TRILL active/active edge solution SHOULD support cases with more than two member RBridges.

4. Frame Processing

When the end node injects frames into the TRILL campus via a member RBridge, this RBridge encapsulates the native frames on behalf of the pseudonode. When frames are sent to the end node, the pseudonode is supposed to be the egress RBridge. It's REQUIRED that RBridges other than the active/active members are not aware of the active/active group and need not change their frame processing behavior.

Compared to the Appointed Forwarder mechanism, all active/active member RBridges are able to ingress and egress frames of VLAN-x on the same link. It is crucial to avoid loops and duplications in the frame processing.

4.1. Unicast Ingressing

Receiver RBridges encapsulate native frames using the nickname of the pseudonode as the ingress nickname. When these TRILL data frames arrive at the remote RBridge, the MAC addresses will be learnt from packet decapsulation. The remote RBridge will regard the pseudonode as the egress RBridge for these MAC addresses.

4.2. Unicast Egressing

As learnt in the MAC table, TRILL data frames from remote RBridges destined to the end node will be sent to the pseudonode rather than member RBridges. If member RBridges receive TRILL data frames whose egress RBridge is the pseudonode, they can judge that these frames should be egressed onto the MC-LAG.

However, member RBridges MUST NOT egress any TRILL data frames whose ingress RBridge is the pseudonode. Otherwise, loops will happen.

4.3. Multicast Ingressing

The end node chooses one component link of the MC-LAG to send multicast frames to member RBridges. Similar as the unicast ingressing, the receiver RBridge encapsulate the native frames using the nickname of the pseudonode as the ingress nickname.

Different member RBridges MUST NOT share the same Distribution Tree to ingress a multicast frame of a specific VLAN-x from the end node. Otherwise, some multicast frames may suffer from loss due to Reverse Path Forwarding Check. This issues is detailed in Section 8.

4.4. Multicast Egressing

Multicast frames sent along the VLAN-x Distribution Tree may reach

all member RBridges. However, only one of them can egress the multicast frames onto the MC-LAG. Otherwise, the end node will suffer from frame duplication. This requirement can be met if member RBridges calculate the Distribution Tree regarding the pseudonode as a normal RBridge. Then only one parent RBridge will be selected for the pseudonode. Other non-parent member RBridges MUST refrain from egressing multicast frames of VLAN-x onto the MC-LAG.

Similar as the unicast egressing, member RBridges MUST NOT egress any multicast frames whose ingress RBridge is the pseudonode.

5. DRB and Pseudonode

As we know, a DRB MAY give a pseudonode name to a LAN link, issue an LSP (Link State PDU) on behalf of the pseudonode, and issues CSNPs (Complete Sequence Number PDUs) on the LAN link [RFC6325]. Different from a LAN link, there is no HELLO exchanging on the MC-LAG. Thus, the DRB cannot be elected using HELLO protocol. Member RBridges MAY establish a dedicated RBridge Channel to discover each other and elect the DRB (DRB for active/active RBridge group, aDRB) to execute the above tasks: to assign the nickname and issue LSP and CSNPs. The member RBridge with the highest priority to be the tree root is a good choice.

Member RBridges SHOULD be able to discover each other to resolve misconfiguration and failures. Each member RBridge SHALL report their connection to the MC-LAG. The MAC address of the end node MAY be used to identify the MC-LAG to which the member RBridges are connected.

One RBridge may be connected to multiple MC-LAGs. It's probably that all these MC-LAGs share the same set of member RBridges. However, these MC-LAGs MUST NOT share the same pseudonode, otherwise it can cause the following issue.

- o Component Links from Different MC-LAGs Cannot be Distinguished: Assume member RBridge RBi is connected to multiple end nodes and these links are all advertised as a single ISIS link "RBi-RBv". Remote RBridges cannot distinguish these links connecting RBi and RBv. When one of these links fails, it becomes problematic. On one hand, if the failed link is not advertised as a down ISIS link, traffic sent from remote RBridges to RBv via the failed link will be trapped by blackholing. On the other hand, if the failed link is announced as a down ISIS link. Component links from other MC-LAGs will be disconnected mistakenly.

The right choice is to represent every MC-LAG as a unique pseudonode. In this way, the failure of a component link of an MC-LAG can be interpreted as an ISIS link failure. Thus the aDRB can issue a new

LSP on half of the pseudonode to trigger the link state update across the campus.

6. MAC Addresses Sharing

When a member RBridge learns a MAC address from the encapsulation or decapsulation of a TRILL data frame, it SHOULD share this learning among all member RBridges. Afterwards, a frame destined to this MAC address can be delivered to the MC-LAG or ingressed to the TRILL campus by any other member RBridge as a unicast native frame or TRILL data frame.

a) Northbound Sharing: When a remote RBridge chooses the path to send data frames to the end node, these frames may arrive at anyone of the member RBridges, given that member RBridges may be on the Equal Cost Multiple Paths from the remote RBridge to the pseudonode. If the MAC address from the end node was learnt and recorded by any member RBridge before. The receiver RBridge SHOULD have recorded this MAC (VLAN ID, MAC Address, Port Number) as well, so that the frame can be delivered as a known unicast to the end node. Therefore, local MAC addresses learnt from data frames sent by the end node (northbound) SHOULD be shared among member RBridges.

b) Southbound Sharing: The end node may choose any component link to inject a frame, which achieves load-balance on the MC-LAG. If the destination MAC address has been learnt by any member RBridge, the receiver RBridge SHOULD also hold that MAC record (VLAN ID, MAC Address, Egress RBridge Nickname). Thus the data frame need not be sent as a multicast frame (unknown unicast). Therefore, MAC addresses learnt from data frames sent by remote RBridges to the end node (southbound) should be shared as well.

When an RBridge learns a source MAC address from a data frame, it will record the VLAN ID, the source MAC address and location which can be the incoming port number or the ingress nickname. A MAC address shared by a peer RBridge is recorded as if it is locally learned. For example, when RB1 shares a MAC with RB2, RB2 should set the incoming port as its port attaching to the end node.

It is REQUIRED that all member RBridges set the same aging time for each MAC address. Every time a MAC address is learnt or updated, all member RBridges MUST update the record and reset its aging time. It's probably that data frames from one source MAC are received continuously. There is no problem to update the entry of this MAC locally. However, when this update is executed among multiple member RBridges, the intensive updates may consume a considerable bandwidth. Therefore, member RBridges need a communication channel to realize

the MAC sharing, which can be realized through the extension of ESADI or using a dedicated RBridge Channel [Channel].

7. Failures and Self-healing

Resilience is a major purpose that the active/active edge aims to achieve. From the side of the end node, the MC-LAG provides reliability of the access link. From the side of the member RBridges, the state change of the active/active edge caused by link or node failures is reflected by the update of LSPs of member RBridges. This provides self-healing of the active/active edge.

7.1. Link Failure

The failure of a component link of the MC-LAG link is translated into an ISIS link failure: if a member RBridge is disconnected from the end node, it will send out an LSP to announce that it is not connected to the pseudonode. This will trigger the update of forwarding tables of remote RBridges. Since other member RBridges have also reported the connection to the pseudonode, remote RBridges in the TRILL campus can send frames to the pseudonode via any other member RBridge. Therefore, the reach-ability to the end node is not broken by this link failure.

If the link connecting the aDRB and the end node fails, the link failure will trigger the election of aDRB. The new aDRB SHOULD reuse the nickname allocated to the pseudonode, which avoids changing the locations of MAC addresses from the end node learnt by remote RBridges.

The extreme case is that the last component link of the MC-LAG fails. Then the aDRB SHOULD update its LSPs to remove the pseudonode from the campus, which also destroys the whole active/active edge.

7.2. Node Failure

The node failure of member RBridges will also be reflected by LSP announcement. If the aDRB fails, a new aDRB will be elected and this new aDRB SHOULD reuse the nickname of the pseudonode allocated by the old aDRB.

8. Reverse Path Forwarding Check

Reverse Path Forwarding Check (RPFC) is used by TRILL to suppress forwarding loops of multicast frames [RFC6325]. For a specific Distribution Tree (DT), a multicast frame from a specific ingress RBridge can arrive at only one expected link of an RBridge. RBridges MUST drop multicast frames that fail the RPFC [RFC6325].

When multiple member RBridges ingress multicast frames for VLAN-x of the end node simultaneously, it can not guarantee that these frames always arrive at the expected link of at a remote RBridge. The following example explains this issue.

```
                                RBi
                               /   \
                             RB1   RB2
                             /
                           RBv
```

Figure 7.1: The Distribution Tree, root=RBi

Suppose a Distribution Tree of Figure 2.1 (b) is constructed as shown in Figure 7.1. For this Distributions Tree, multicast frames from RBv to RBi is expected to be received at the port attaching to RB1. With the active/active connection, RB2 can receive native data frames from the MC-LAG as well. If RB2 adopts the above Distribution Tree, multicast frames from RBv to RBi will be received at the port attaching to RB2. This brings the problem: these frames will be discarded according to the rule of RPFC.

```
                      RBx                RBy
                       |                  |
                      RBi                RBi
                     /   \              /   \
                   RB1   RB2          RB1   RB2
                   /                          \
                 RBv                           RBv
```

(a) DT, root=RBx   (b) DT, root=RBy

Figure 7.2: Assign an Unique Tree to each Member RBridge

One way to avoid the above issue is to leverage the feature that RBridges can compute multiple Distribution Trees. Be sure to assign an unique Distribution Tree to each member RBridge for multicast frame distribution. Identify these trees using their root RBridge nicknames. The example in Figure 7.2 illustrates this method, where RB1 and RB2 adopt two different Distribution Trees.

Active/active edge need to assign at least one Distribution Tree per component link of an MC-LAG, the maximally allowed number of component links depends on the number of Distribution Trees that all RBridges can compute. However, MC-LAGs of the best current practice have two component links, which are well supported by TRILL switches.

In [CMT], the Affinity TLV is used to achieve the above assignment of

Distribution Trees to member RBridges. It is REQUIRED that all RBridges in the campus are able to recognize the Affinity TLV and compute Distribution Trees as this TLV specified.

When there is a link or node failure in the active/active edge, the failed Distribution Tree should be re-allocated to a new member RBridge. It is RECOMMENDED that this re-allocation is incremental. In other words, other Distribution Trees not affected by the failure SHOULD be retained.

9. Security Considerations

This document raises no new security issues for ISIS.

10. IANA Considerations

This document requires no IANA actions. RFC Editor: please remove this section before publication.

11. References

11.1. Normative References

[RFC6325] R. Perlman, D. Eastlake, et al, "RBridges: Base Protocol Specification", RFC 6325, July 2011.

[RFC6349] R. Perlman, D. Eastlake, et al, "RBridges: Appointed Forwarders", RFC 6349, November 2011.

[Channel] D. Eastlake, V Manral, et al, "TRILL: RBridge Channel Support", draft-ietf-trill-rbridge-channel-08.txt, July 2012, working in progress.

[CMT] T. Senevirathne, J. Pathangi, et al, "Coordinated Multicast Trees (CMT)for TRILL", draft-ietf-trill-cmt-01.txt, November 2012, working in progress.

11.2. Informative References

None.

Author's Addresses

Mingui Zhang
Huawei Technologies
No.156 Beiqing Rd. Haidian District,
Beijing 100095 P.R. China

Email: zhangmingui@huawei.com

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

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