INTERNET-DRAFT
Intended Status: Proposed Standard
Expires: August 2, 2012

Mingui Zhang
Huawei

January 30, 2012

Reverse Path Forwarding Check under Multiple Topology TRILL
draft-zhang-trill-multi-topo-rpfc-00.txt

Abstract

Multi-homing (RBridge Aggregation) is a promising approach to increase the reliability and access bandwidth of TRILL edge. Active-active forwarding in multi-homing allows multiple RBridges forward data frames for VLAN-x on a LAN link, which creates the possibility that multicast frames from a specific ingress RBridge may arrive at multiple incoming ports of a remote RBridge. This violates the Reverse Path Forwarding Check and multicast frames arrives at unexpected incoming ports will be discarded by this RBridge. This document makes use of multiple topology TRILL to solve this problem. Multiple topology TRILL provides physical separation of traffic from different members of aggregation. Multicast frames from aggregation members comply with the Reverse Path Forwarding Check per topology.

Table of Contents

1. Introduction

With the link state routing of IS-IS (Intermediate System to Intermediate System), TRILL provides a solution of least cost forwarding of data frames to replace the Spanning Tree Protocol (STP) running in traditional bridge networks.

RBridge Aggregation provides active-active multi-homing at the edge of TRILL [RBAgg]. It increases the access bandwidth and reliability of TRILL edge but creates the possibility that multiple RBridges ingress/egress data frames for end-stations from VLAN-x on a LAN link. A typical use of RBridges Aggregation is to represent a LAN link with a single virtual RBridge. RBridges participating the aggregation ingress/egress data frames on behalf of this virtual RBridge using a pseudonode nickname.

Reverse Path Forwarding Check (RPFC) is used by TRILL to suppress forwarding loops of multicast frames. Based on a Distribution Tree (DT), a multicast frame from a specific ingress RBridge arrives at a single expected link of an RBridge. RBridges MUST drop multicast frames that fail the RPFC [RFC6325]. When multiple RBridges ingress multicast frames for end-stations from VLAN-x on a LAN link simultaneously, it can not guarantee that these frames always arrive at the expected link of a remote RBridge.

Multiple Topology (MT) TRILL provides a physical separation of traffic [RFC5120] [MTc] [MTd]. An MT aware RBridge can participate data forwarding in multiple topologies at the same time. This feature is utilized in this document to resolve the issue that active-active multi-homing may fail RPFC. Each RBridge of the aggregation uses an individual topology to ingress/egress data frames for the target LAN link. Since distribution trees are calculated per topology by MT aware RBridges [MTd], multicast frames will be forwarded along these distribution tress separately, which helps the arriving multicast frames pass RPFC. To be backward compatible, the solution provided in this draft does not require all RBridges in a campus to upgrade to support multiple topology TRILL. Legacy RBridges that do not support multiple topology TRILL can inter-operate with the MT aware RBridges participating the RBridge Aggregation.

This document focus on solving the RPFC issues caused by active-active multi-homing. Other issues of multi-homing, such as failure recovery and load balance, are in the scope of RBridge Aggregation [RBagg]. One advantage of the adoption of multiple topology TRILL is that approaches for failure recovery developed for multiple topology routing ([RFC5714]) can be reused in RBridge Aggregation without the reinvention of the wheel.

1.1. Content

Section 2 explains why active-active multi-homing may cause trouble in Reverse Path Forwarding Check of TRILL.

Section 3 describes the approach of configuration for the edge RBridges to achieve RBridge Aggregation through multi-topology TRILL.

Backward compatibility is an essential requirement for the inter-operation between legacy RBridges and RBridges participating in aggregation. Section 4 describes solutions for three incremental deployment scenarios.

1.2. 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].

1.3. Acronyms

IS-IS - Intermediate System to Intermediate System

TRILL - TRansparent Interconnection of Lots of Links

STP - Spanning Tree Protocol

MT - Multiple Topology

DT - Distribution Tree

LAG - Link Aggregation

RPFC - Reverse Path Forwarding Check

2. RPFC Issue in Active-Active Multi-homing

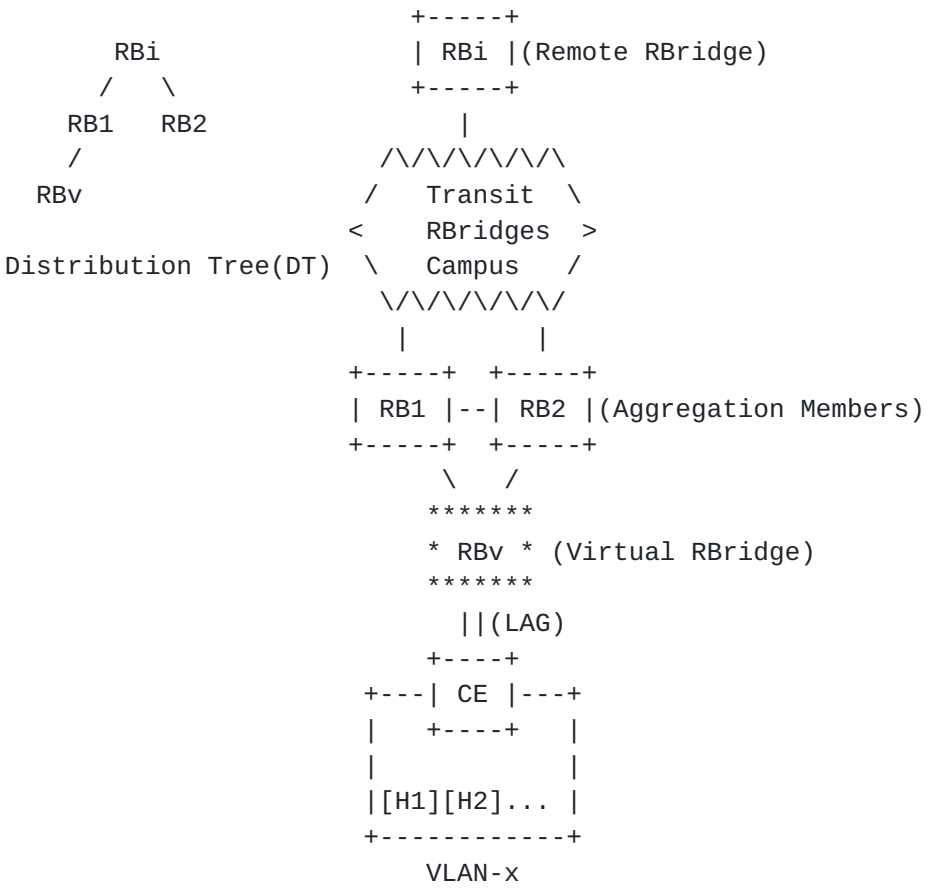

Figure 2.1: An Example Topology of RBridge Aggregation

RBridge Aggregation is first proposed in [RBagg]. RBridge Aggregation enables active-active multi-homing for LAN links [RBagg]. Several RBridges can ingress/egress data frames for end-stations of one VLAN on a LAN link, which increases the access bandwidth and reliability of TRILL edge.

Figure 2.1 shows an example topology of RBridge Aggregation and the distribution tree is shown on the left (Suppose the transit RBridge campus is null.). Based on the distributions tree, multicast frames from RBv to RBi is expected to be received at the port attaching to RB1.

Under RBridge Aggregation, RB2 can really ingress native data frames from the LAG links, therefore multicast frames from RBv to RBi may legally be received at the port attaching to RB2. These frames will be discarded according to the rule of Reverse Path Forwarding Check [RFC6325]. Active-active forwarding of multicast frames is the root cause of this issue. The rest of this document will make use of

multiple topology TRILL to solve this problem.

3. Multi-Topology for Aggregation

Documents [MTc] and [MTd] define the protocol extensions, data plane encoding and procedures to make use of the multiple topology routing supported by ISIS. Multiple topology routing provides physical traffic segregation to TRILL, which is utilized to solve the RPFC issue caused by RBridge Aggregation. RPFC will be done based on distribution trees which are calculated per topology abbreviation by MT aware RBridges.

Topology IDs are used to identify the aggregation members. If the number of available topologies is greater than the number of aggregation members, several topology IDs can be assigned to one aggregation member which can make use of these topologies to realize load-balancing. If available topologies are less than aggregation members, some of these members get no topology ID. These standby aggregation members can make use of the tunneling approach defined in Section 3.3 to redirect arriving data frames to other members for forwarding.

Table 3.1: A Sample Configuration for Aggregation

Aggregation Members	RBv's Nicknames	LAG Members
RB1	...001...	RB1-RBv
RB2	...010...	RB2-RBv
RB3	...011...	RB3-RBv
RB4	...100...	RB4-RBv

Since multiple topology TRILL identifies a topology using the ingress nickname [MTd], the topology assignment among aggregation members is embodied through the nickname configuration of RBv. Figure 3.1 shows a typical configuration of RBridge Aggregation with 4 members. Each aggregation member ingress native frames using one nickname of RBv. These frames will be confined to the topology as these nicknames indicate. For example, when RB1 ingress the native frames from the local link, it will use RBv001 as the ingress nickname and these frames will be forwarded in topology 1.

The rest of this section discusses multicast forwarding. For the

detail of unicast forwarding, one may refer to [RBagg].

3.1. Multicast Ingressing

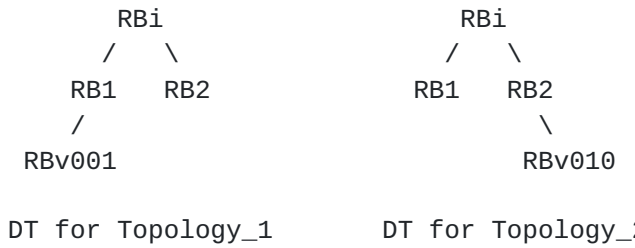

Figure 3.1: Sample Distribution Trees for Topology 1 and 2

LAG may use any of its links as the active link to send frames to a member of RBridge Aggregation. The receiver SHOULD encapsulate the native frames on behalf of RBv. Take Figure 3.1 as an example, RB1 and RB2 encapsulates native frame using RBv001 and RBv010 as their ingress nicknames respectively. If these frames are multicast frames, they will be forwarded according to the distribution trees calculated per topology. Since RBi calculates two different distributions trees for RBv001 and RBv010, multicast frames arriving at the ports attached to RB1 and RB2 can all pass the RPFC.

3.2. Multicast Egressing

Since distribution trees are built per topology, a multicast frame will be received by only one aggregation member. This member should egress the multicast frame to the local link on behalf of RBv. But remote RBridges is not aware that RBv actually does not exist. All aggregation members act as penultimate hops to RBv in the campus.

3.3. Address Flip-Flop Avoidance by Asymmetric Topologies

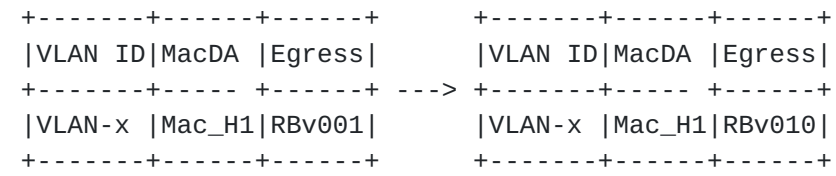

Figure 3.2: An Example of MAC Address Flip-Flop

In the above ingressing procedure, native data frames from one end station may be ingressed to the campus by different aggregation members. Current RBridges do not have the topology abbreviation as a separate column in their MAC tables. Therefore, when a remote RBridge receives multicast frames with the same source MAC address from different aggregation members, these multicast frames will create

only one entry in the MAC table of this remote RBridge.

As illustrated in Figure 3.2, when frames originated from H1 is sent to RBi from RB1, RBi will learn that the egress RBridge nickname for Mac_H1 is RBv001. Afterwards, if RB2 sends frames originated from H1 to RBi, the egress RBridge nickname will change to RBv010. It seems that the use of multiple topology TRILL brings a MAC address flip-flop issue. If RBv001 and RBv010 are regarded as two different egress RBridges and RBi prepares paths to them separately, it is possible RBi gets different forwarding paths. In other words, RBi will use different forwarding paths in different topologies for the data frames destined to the same end-station, which may cause packet disorder.

However, MT aware RBridges support asymmetric use of topologies [MTd]. In the above example, RBi can send data frames to Mac_H1 according to topology 1 even if it learns Mac_H1 from the data flow in topology 2. That is to say RBi can send return data frames to RBv001 all the time. In practical use, remote RBridges SHOULD adhere to a specific topology to send return data frames destined to a specific MAC address.

3.4. Tunneling Approach

If available topologies are less than the aggregation members, there will be standby members who get no topology ID. These members can still ingress native frames from the LAG directly. But they should redirect them to other members through the following tunneling approach.

Suppose RB5 is a standby member of the aggregation. So it is not a parent of RBv on the distribution tree of any topology. Assume RB5 tunnels native frames from the LAG to RB1 which is the parent of RBv in topology 1. RB5 should ingress the native frame, fill its egress nickname as RB1 and fill its ingress nickname as the nickname of RBv001 which is used by RB1. Then RB5 sends this frame as a unicast frame to RB1. When RB1 receives this unicast frame, it can judge from its ingress nickname that this frame should be actually ingressed by RB1. Therefore, RB1 decapsulates this frame and re-capsulate it as if it is received from the LAG link RBv-RB1.

For the sake of load-balancing and resilience, it is recommended that standby RBridges tunnel their multicast frames evenly among those aggregation members who get topology IDs. The optimization of the tunneling configuration is out the scope of this document. Tunneling approach can also be used for any other purpose such as fail-over. However, in this document, tunneling is used only for redirecting ingress multicast frames to pass through the RPFC in TRILL.

4. Incremental Deployment

When RBridge Aggregation is put to use in a TRILL campus, it is probably that MT unaware RBridges have already been deployed in this campus. It is therefore necessary to enable the inter-operation of these two types of RBridges. On one hand, MT aware aggregation members MUST be backward compatible to those legacy MT unaware RBridges. On the other hand, legacy RBridges need not make any change in order to communicate with aggregation members.

With multi-topology TRILL, RBridge Aggregation can be incrementally deploy in an RBridge campus. This rest of this section provides approaches for three incremental deployment scenarios: (1) aggregation members need not to talk with MT unaware RBridges; (2) aggregation members need to communicate with MT unaware RBridges; (3) a combination of the above two scenarios.

4.1. Intra-Topology Communication

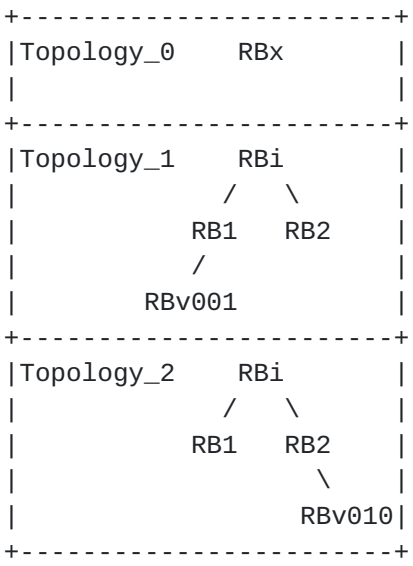

Figure 4.1: Aggregation Members Talk with MT aware RBridges Only

If MT aware aggregated RBridges do not talk with MT unaware RBridges, aggregation traffic can be confined to non-zero topologies. This kind of traffic segregation is achieved through multi-topology routing. As illustrated in Figure 4.1, when RB1 and RB2 forward multicast frames to RBi according to distribution trees for topology 1 and topology 2 respectively, the MT unaware RBx will not receive these frames from RBv. When RB1 and RB2 advertise LSPs in the base topology, they will not include their adjacencies to RBv001 and RBv010, therefore RBx will not be aware of RBv001 and RBv010. In particular, nickname RBv000 SHOULD be reserved and not used in aggregation configuration

so that even RBx can reach RBv000, they will not talk with each other.

4.2. Inter-Topology Communication

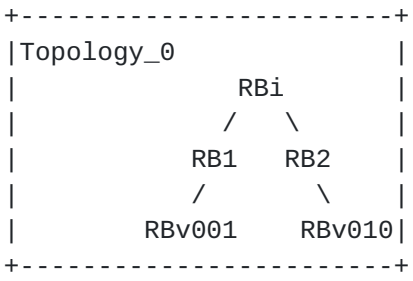

Figure 4.2: MT aware RBridges Need to Talk with MT unaware RBridges

If MT aware aggregated RBridges need to talk with MT unaware RBridges, the traffic segregation method in Section 4.1 can not be used again. Since MT unaware RBridges only understand the base topology, all aggregated RBridges advertise their connections to RBv in the base topology. Figure 4.2 illustrate this approach. Assume RBi is the MT unaware RBridge. RB1 and RB2 advertise their adjacencies to RBv, i.e., "RB1-RBv001" and "RB2-RBv010", in their LSPs. When RBi calculate the distribution tree, it should calculate as what is shown in Figure 4.2. RBv001 and RBv010 are regarded as two different RBridges by RBi.

Hashing function is widely used in LAG for the purpose of load balancing. In a corner case, native data packets from one end-station may be mapped to any aggregated member. Similar as the example shown in Figure 3.3, the egress of Mac_H1 at RBi may change between RBv001 and RBv010 back and forth. When MAC address flip-flop happens, the MT unaware RBi is unable to use asymmetric topologies to send return TRILL data frames back to aggregated members. TRILL data frames destined to RBv001 and RBv010 may go through two different forwarding paths. Although this kind of MAC flip-flop is rare in real TRILL campus, it is recommended that the hashing function is configured to completely avoid it. The configuration of LAG should guarantee that native data frames from one end-station are mapped to only one aggregated member (one active link in the LAG). Destination IP/MAC address fields of a native data frame SHOULD not be used as the input of hashing function.

4.3. A Hybrid Scenario

It is allowable that some aggregated members report their connections to RBv in the base topology while others do not. For aggregated members which do not report the connections to RBv in the based

topology, they need tunnel multicast frames to those members who report their connections to RBv in the based topology in order to communicate with MT unaware RBridges. For example, RB1 and RB2 advertise their connections to RBv001 and RBv010 in topology 1 and 2, while RB0 advertises the adjacency to RBv000 in topology 0. Assume RBi is an MT unaware RBridge. The distribution tree calculated by RBi will include RBv000 while does not include RBv001 or RBv010. RB0 can talk with RBi directly on behalf of RBv000. When RB1 and RB2 communicates with MT aware RBridges, they can confine the traffic in topology 1 and 2. If RB1 and RB2 need to send TRILL data frames to MT unaware RBrdiges, such as RBi, they should redirect these frames to RB0 using the tunneling approach described in Section 3.3. RB0 will send these frames with RBv000 as their ingress nickname.

5. Security Considerations

This document raises no new security issues for IS-IS.

6. IANA Considerations

No new registry is requested to be assigned by IANA.

7. Acknowledgements

Discussions with authors and contributors of [Pseudo] and [CMT] provide a great help to the write up of this draft. This document is by no means to replace such kind of solutions used for RPFC relaxing. These solutions are designed for TRILL base topology and can be used in parallel in the same RBridge campus with the solution presented in this document.

8. References

8.1. Normative References

[RBAgg] M. Zhang, D. Eastlake, et al, "RBridge Aggregation", draft-zhang-trill-aggregation-01.txt, working in progress.

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

[MTc] Vishwas Manral, D. Eastlake, et al, "Multiple Topology Routing Extensions for Transparent Interconnection of Lots of Links (TRILL)", draft-manral-isis-trill-multi-topo-03.txt, working in progress.

[MTd] D. Eastlake, M. Zhang, et al, "Multiple Topology TRILL", draft-eastlake-trill-rbridge-multi-topo-02.txt, working in

progress.

8.2. Informative References

[RFC5120] Przygienda, T., Shen, N., and N. Sheth, "M-ISIS: Multi Topology (MT) Routing in Intermediate System to Intermediate Systems (IS-ISs)", RFC 5120, February 2008.

[RFC5714] Shand, M. and S. Bryant, "IP Fast Reroute Framework", RFC 5714, January 2010.

[Pseudo] H. Zhai, F. Hu, et atl, "RBridge: Pseudonode Nickname", draft-hu-trill-pseudonode-nickname-01.txt, working in progress.

[CMT] T. Senevirathne, J. Pathangi, et al, "Coordinated Multicast Trees (CMT)for TRILL", draft-tissa-trill-cmt-00.txt, working in progress.

Author's Addresses

Mingui Zhang
Huawei Technologies Co.,Ltd
Huawei Building, No.156 Beiqing Rd.
Z-park ,Shi-Chuang-Ke-Ji-Shi-Fan-Yuan,Hai-Dian District,
Beijing 100095 P.R. China

Email: zhangmingui@huawei.com