Weiguo Hao
Yizhou Li
Donald Eastlake
Huawei

Internet Draft
Intended status: Informational
February 14,2014
Expires: August 2014

Analysis of Active-Active connection solutions
draft-hao-trill-analysis-active-active-01.txt

Status of this Memo

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document.

Abstract

Draft [TRILL-Active-PS] lists basic problems which any active-active solutions should address, these problems include frame duplications, loop, MAC address flip-flop and unsynchronized information among member RBridges. For each problem, there may be multiple ways to deal with it. Some solutions solve all or most of the problems listed, and at the same time introduces extra issues. This draft tries to analyze and compare the different solutions for each of the issues, gives a brief summary on the pros and cons, and/or the applicable scenarios.

Table of Contents

1. Introduction

The IETF TRILL (Transparent Interconnection of Lots of Links) [RFC6325] protocol provides loop free and per hop based multipath data forwarding with minimum configuration. TRILL uses IS-IS [RFC6165] [RFC6326bis] as its control plane routing protocol and defines a TRILL specific header for user data.

Customer edge(CE) devices typically are multi-homed to several RBridges. All of the uplinks of a CE are considered as an Multi-Chassis Link Aggregation (MC-LAG) bundle. An edge group is the group of edge RBridges that a CE is multi-homed to in active-active mode. An edge group corresponds to an MC-LAG. One RB can be in more than one edge group. An active-active flow-based load-sharing mechanism is desirable to achieve better load balancing and high reliability. A CE device can be a layer3 end system by itself or a bridge switch through which layer3 end systems are accessed to TRILL campus.

Draft [TRILL-Active-PS] lists the following problems which any active-active solution should address:

```
                                   +------+
                                   | CEx  |
                                   +------+
                                       |
                                    +------+
                                   |(RBx) |
                                   +------+
                                       |
                               -------------------
                              /                   \

                             |                     |
                             |   TRILL Campus      |
                             |                     |
                              \                   /
                               --------------------
                                    |     |    |
                            --------      |    --------
                             |            |             |
                          +------+      +------+      +------+
                          |(RB1) |      |(RB2) |      | (RBk)|
                          +------+      +------+      +------+
                           |              |   |          |
                           |   __--------_|   |------    |
                           |   |LAG1          LAG2   |   |
                          +------+                  +------+
                          |  CE1 |                  | CE2  |
                          +------+                  +------+
                Figure 1 TRILL Active-Active Access Scenario
```

1. Frame duplications

2. Loop

3. Address flip-flop

4. Unsynchronized information among member RBridges

For each problem, there may be multiple ways to deal with it. And some solutions solve all or most of the problems listed, and at the same time introduces extra issues. This draft tries to analyze and compare the different solutions for each of the issue, gives a brief summary on the pros and cons, and/or the applicable scenarios. The co-authors believe such analysis is helpful to design a more completed solution in future.

2. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT","SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

The acronyms and terminology in [RFC6325] is used herein with the following additions:

BUM - Broadcast, Unknown unicast, and Multicast.

CE - Customer equipment. Could be a bridge or end station or a hypervisor.

CMT - Coordinated Multicast Trees [CMT].

Edge group - a group of edge RBs to which at least one CE is multiply attached. One RB can be in more than one edge group.

LACP - Link Aggregation Control Protocol.

LAG - Link Aggregation, as specified in [8021AX].

3. Frame duplications

Frame duplication may occur when a remote host sends multi-destination frame to a local CE which has an active-active connection to the TRILL campus.

To avoid local CE receiving multiple copies from a remote RBridge, the designated forwarder (DF) mechanism should be supported. DF allows only one port in one RB of MC-LAG to forward multicast traffic from TRILL campus to local access side for each VLAN. The basic idea of DF is to elect one RBridge per VLAN from an edge group to be responsible for egressing the multicast traffic.

Each RB in an edge group elects a DF using same algorithm which guarantees the same RB elected as DF per MC-LAG per VLAN. The RB that is elected as a DF for a given VLAN will forward multi-destination traffic in the egress direction towards the CE. All non-DF RBs drop multi-destination traffic in the egress direction towards the CE. All edge RBs, including DF and non-DF, can ingress the traffic to TRILL campus as usual.[draft-hao-trill-dup-avoidance-active-active-00] describes the detail DF mechanism and TRILL protocol extension for DF election.

4. Loop

If a CE sends a broadcast, unknown unicast, or multicast (BUM) packet to DF RB, it will forward that packet to all or subset of the other RBs including the non-DF RBs. Because non-DF RBs don't egress BUM frame to local access side, in this case the frame won't loop back to the CE.

If a CE sends a BUM packet to one of the non-DF (Designated Forwarder) RBs, say RB1, then RB1 will forward that packet to all or subset of the other RBs including the DF RB for that MC-LAG. In this case the frame will loop back to the CE and traffic split-horizon filtering mechanism should be used to avoid looping back among RBridges in a edge group.

Split-horizon mechanism relies on ingress nickname to check if a packet's egress port belongs to a same MC-LAG with the packet's incoming port to TRILL campus.

4.1. Independent nickname allocation

Each ingress RBridge allocates a unique nickname for each MC-LAG independently. It is not required that the nickname provisioned on all involved edge RBridges remains the same for one corresponding MC-LAG.

When the ingress RBridge receives a BUM frame from a local CE, it uses the nickname as ingress nickname for TRILL tunnel encapsulation and sends the frame to other RBridge(s).

When an egress RBridge receives a multicast frame from the TRILL campus, it checks the ingress nickname in the TRILL header and filters out the frame on all local interfaces connected to the same CE. Each egress RBridge should track the nickname(s) associated with the other RBridge(s) with which it has a shared multi-homed LAG. The solution has limited nickname allocation scalability issue, because each RBridge needs allocate per nickname per MC-LAG.

4.2. Consistent nickname allocation

Edge RBridges forming an MC-LAG in an edge group are assigned a globally unique pseudo-nickname. If multiple MC-LAGs exist, edge BRridges for each individual MC-LAG should be assigned such a pseudo-nickname. It should be guaranteed that pseudo-nickname provisioned on all involving edge RBridges remains the same for one corresponding MC-LAG.

When a ingress RBridge receives traffic from a active-active accessed CE, it performs TRILL encapsulation with the pseudo-nickname as ingress nickname. When the traffic comes to each egress RBridge, the egress RBridge checks the ingress nickname in TRILL header and filters out the frame on all local interfaces connected to the same CE. Each egress RBridge relies on the pseudo-nickname to filter out the frame on all local interfaces connected to the same CE.

4.3. Comparison

Solution	Independent Allocation	Consistent Allocation
Nickname consumption	High	Normal
Scalability	Low	High

5. Address flip-flop

MAC learning in TRILL can be performed either in data plane or control plane. When a local host h1 attaches to multiple edge RBridges, learning at the remote host for h1 may have MAC flip-flop problem. There are different ways to avoid this for data plane learning and control plane learning scenarios.

5.1. Data plane learning mode

For data plane learning mode, to avoid mac address flip-flop on remote RBs, a pseudo-nickname [TRILLPN] solution was proposed. The basic idea is to represent all member links of the MC-LAG as a virtual RBridge with single pseudo-nickname. Any member RBridge of the MC-LAG should use this pseudo-nickname rather than its own nickname as ingress nickname when inject TRILL data frames. It solves the above mentioned problems pretty well; however, it introduces another issue: packet drop due to RPF check. To overcome the RPF check failure issue, three solutions have been proposed.

5.1.1. CMT

CMT [CMT] solution allows edge RBridges to specify different distribution trees to forward BUM traffic from a connecting CE device by using a new IS-IS Affinity sub-TLV. Remote RBridges calculate their forwarding tables and derive the RPF for distribution trees based on the distribution tree association advertisements.

In this solution, it’s required to establish multiple distribution trees in a TRILL campus, i.e. if a CE is active-active accessed to 4 edge RBridges, at least 4 distribution trees are required. No hardware upgrade is needed for RBridges in the TRILL campus, only software upgrade is needed.

5.1.2. Centralized replication

Ingress RB participating in active-active connection sends BUM traffic to one of a distribution tree root node through unicast TRILL encapsulation. The distribution tree root node acts as centralized replication node. When the distribution tree root node receives unicast TRILL encapsulation BUM traffic from the ingress RB, it decapsulates the unicast TRILL packet. Then it replicates and forwards the BUM traffic to all other destination RBs through the distribution tree established per TRILL base protocol. [draft-hao-trill-centralized-replication-00] describes the detail centralized replication solution. Through the centralized replication solution, only unicast forwarding behavior is required between edge RB and distribution tree root RB, so no RPF check function is required along the path between ingress RB and distribution tree node.

When the ingress RBridge receives BUM traffic from an active-active accessing CE device, the traffic will be injected to TRILL campus through TRILL encapsulation. Then it is replicated and forwarded to other CE devices through TRILL distribution tree, even when the receiver CE is connected to the same RBridge as the sender CE. To avoid duplicated traffic on receiver CE, ingress RBridge can’t locally replicate and forward the BUM traffic to other connecting CE when it receives BUM traffic from an active-active sender CE, i.e. the access port of the ingress RBridge should be isolated from other local access ports.

In this solution, it’s required to consume more network bandwidth between ingress RB and distribution tree root node than CMT solution.

Both hardware and software upgrade are required on edge RBs participating in active-active connection and the distribution tree root node. This solution doesn't require multiple distribution trees in TRILL campus, so it has better scalability than CMT.

5.1.3. Tunneling among edge RBs

This solution allows only a selected edge RBridge in an edge group participating in active-active access to be responsible for forwarding BUM traffic from connecting CE to TRILL campus along distribution tree per TRILL base protocol. All other edge RBridges in the virtual RBridge send BUM traffic from connecting CE to the selected edge RBridge through unicast TRILL encapsulation. When the selected edge RBridge receives TRILL traffic from other RBs in a same virtual RBridge, the selected RB decapsulates the unicast TRILL packet. Then it forwards the BUM traffic to trill campus along distribution tree established per TRILL protocol.

Similar to the solution of centralized replication, to avoid duplicated traffic on receiver CE, the access port of ingress RBridge connecting to an active-active accessing sender CE should be isolated from other local access ports.

In this solution, it's required to consume more network bandwidth among edge RBs. Both hardware and software upgrade are required on edge RBs participating active-active connection. This solution doesn't require multiple distribution trees in TRILL campus, so it has better scalability than CMT.

5.1.4. Comparison

Solution	CMT	Centralized replication	Tunneling among edge RBs
Scalability	Medium	High	High
Network bandwidth consumption	Low	High	High

Software upgrade	All RBs	root and edge nodes	root and edge nodes
Hardware upgrade	No	root and edge nodes	root and edge nodes

5.2. Control plane learning mode

If a CE device is multi-homed to multiple edge RBs in active-active mode, each edge RB should announce the MAC of its attached end systems to all other RBs through ESADI-like control protocol. Remote RBriges will learn the MAC association with different ingress RB nicknames and generate multiple MAC forwarding entries in ECMP mode. All edge RBs should disable the data plane MAC learning function. MAC to nickname association should be learned only through the control plane.

Pseudo-nickname mechanism was basically designed to avoid MAC address learning flip-flop when a MAC address could be learnt to more than one RBridge. With control plane MAC leaning, pseudo-nickname is not required since multiple mac to nickname entries can be leaned for the same MAC. The problem of RPF check failure for multicast frame caused by pseudo-nickname mechanism is not an issue here.

In the control plane MAC learning solution, if an edge RB participating TRILL active-active access receives BUM traffic from connecting CE device, it uses its own nickname as ingress nickname instead of pseudo-nickname to ingress data frame into a TRILL campus.

This method requires hardware and software changes.

6. Unsynchronized information among member RBridges

Synchronization mechanism should be provided to ensure information consistency among all edge RBridges in a edge group, such as MAC table, dynamic VLAN and multicast group, LACP configuration and state, DHCP snooping table, and etc. [draft-hao-trill-rb-syn-02] describes the detail synchronization requirements. Two synchronization solutions as follows are provided.

6.1. RBridge channel based communication protocol

RBridge channel based communication protocol among all RBridges in a edge group is introduced to implement synchronization. The communication protocol is restricted to RBridge nodes in each edge group, other RBridges in TRILL campus needn't involve. A new type of RBridge Channel message should be given by a Protocol field in the RBridge Channel Header to indicate synchronization information in the payload. RBridge channel message is forwarded through TRILL data plane. Transmission delay is relatively low.

6.2. TRILL LSP extension

TRILL LSP can be extended to implement synchronization among all edge RBridges. Synchronization information is conveyed through new TLVs or sub-TLVs in TRILL LSP. Because TRILL LSP is flooded to all RBridges in TRILL campus, so it may cause campus wide fluctuation. TRILL LSP is forwarded through control plane. Transmission delay is relatively high.

6.3. Comparison

Solution	RBridge channel based	TRILL LSP extension
Flooding scope	Edge group	Campus wide
Forwarding	Data plane	Control plane

7. Solution summary

Through the above analysis, a completed solution for active-active connection can be stitched together using mechanisms for each individual problem analyzed in this draft.

If there are multiple mechanisms for a single problem, any one can be picked up. For example, in MAC learning through data plane scenarios for address flip-flop problem, there are three mechanisms including CMT, centralized replication and tunneling among edge RBs to solve MAC address flip-flop problems. Any one out of three can be selected to combine with other mechanisms to form a whole solution. If there is only one mechanism for a single problem, then it is a mandatory part of the completed solution. For example, DF election mechanism is the only acceptable way to prevent frame duplication. Thus it is a mandatory part of the completed solution.

In summary, the whole solution for TRILL active-active connection is as follows.

Problem	Solution			
Frame duplication	DF election			
Loop	Data plane MAC learning			Control plane MAC learning
	CMT	Centralized replication	Tunneling among edge RBs	
Address flip-flop	Independant allocation			Consistent allocation
Unsynchronized information	RBridge channel based			LSP extension

8. Security Considerations

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

9. IANA Considerations

This document requires no IANA Actions. RFC Editor: Please remove this section before publication.

10. References

10.1. Normative References

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

[2] [RFC6325] Perlman, R., et.al. "RBridge: Base Protocol Specification", RFC 6325, July 2011.

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

10.2. Informative References

[4] [TRILAA] Li,Y., et.al., "Problems of Active-Active connection at the TRILL Edge", draft-yizhou-trill-active-active-connection-prob2, Work in progress, July 2013.

[5] [TRILLPN] Zhai,H., et.al., "RBridge: Pseudonode Nickname", draft-hu-trill-pseudonode-nickname, Work in progress, November

2011.

[6] [CMT] Senevirathne, T., Pathangi, J., and J. Hudson, "Coordinated Multicast Trees (CMT)for TRILL", draft-ietf-trill-cmt-01.txt Work in Progress, November 2012

[7] [RFCchannel] - D. Eastlake, V. Manral, L. Yizhou, S. Aldrin, D. Ward, "TRILL: RBridge Channel Support", draft-ietf-trill-rbridge-channel-08.txt, in RFC Edtior’s queue.

[8] [RFC6439] Perlman, R., Eastlake, D., Li, Y., Banerjee, A., and F. Hu, "Routing Bridges (RBridges): Appointed Forwarders", RFC 6439, November 2011.

Authors' Addresses

Weiguo Hao
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China
Phone: +86-25-56623144
Email: haoweiguo@huawei.com

Yizhou Li
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China
Phone: +86-25-56625375
Email: liyizhou@huawei.com

Donald Eastlake 3rd
Huawei Technologies
155 Beaver Street
Milford, MA 01757 USA
Phone: +1-508-333-2270
EMail: d3e3e3@gmail.com