TRILL Working Group
INTERNET-DRAFT
Intended Status: Informational

Yizhou Li
Donald Eastlake
Weiguo Hao
Huawei Technologies
Radia Perlman
Intel Labs
Jon Hudson
Brocade
Hongjun Zhai
ZTE

Expires: July 7, 2014
January 3, 2014

Problems and Requirements of Active-Active connection at the TRILL Edge

draft-yizhou-trill-active-active-connection-prob-01

Abstract

The IETF TRILL (Transparent Interconnection of Lots of Links) protocol provides support for flow level multi-pathing with rapid failover for both unicast and multi-destination traffic in networks with arbitrary topology and link technology between TRILL switches. Active-active at the TRILL edge is the extension, in so far as practical, of these characteristics to end stations that are multiply connected to a TRILL campus. This informational document discusses the high level problems and requirements when providing active-active connection at the TRILL edge.

Status of this Memo

Copyright and License Notice

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. In a TRILL campus, communications between TRILL switches can

(1) use multiple parallel links and/or paths,

(2) load spread over different links and/or paths at a fine grained flow level through equal cost multipathing of unicast traffic and multiple distribution trees for multi-destination traffic, and

(3) rapidly re-configure to accommodate link or node failures or additions.

Active-active connection is the extension, to the extent practical, of similar load spreading and robustness to the connections between end stations and the TRILL campus. Such end stations may have multiple ports and will be connected, directly or via bridges, to multiple edge TRILL switches. It must be possible, except in some failure conditions, to load spread end station traffic at the flow level across links to such multiple edge TRILL switches and rapidly re-configure to accommodate topology changes.

1.1 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].

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

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

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.

TRILL switch - an alternative term for an RBridge.

2. Target Scenario

The TRILL appointed forwarder [RFC6325] [RFC6327bis] [RFC6439] mechanism provides per VLAN active-standby traffic spreading and loop avoidance at the same time. One and only one appointed RBridge can ingress/egress native frames into/from TRILL campus for a given VLAN among all edge RBridges connecting a legacy network to TRILL campus. This is true whether the legacy network is a simple point-to-point link or a complex bridged LAN or anything inbetween. By carefully selecting different RBridge as appointed forwarder for different set of VLANs, load spreading over different edge RBidges across different VLANs can be achieved.

This section presents a typical scenario of active-active connections to TRILL campus via multiple edge RBridges where current TRILL appointed forwarder mechanism is not applicable.

The appointed forwarder mechanism [RFC6439] requires each of the edge RBridges to exchange TRILL IS-IS Hello packets from their access ports. As figure 1 shows, when multiple access links of multiple edge RBridges are bundled as an MC-LAG (Multi-Chassis Link Aggregation Group), Hello messages sent by RB1 via access port to CE1 will not be forwarded to RB2 by CE1. RB2 (and other members of MC-LAG1) will not see that Hello from RB1. Every member RBridge of MC-LAG1 thinks of itself as appointed forwarder on MC-LAG1 link for all VLANs and will ingress/egress frames for all VLANs. Hence appointed forwarder mechanism is not applicable in such active-active scenario.

```
                ----------------------
               |                      |
               |   TRILL Campus       |
               |                      |
                ----------------------
                    |      |    |
               -----       |     --------
              |             |             |
          +------+      +------+      +------+
          |      |      |      |      |      |
          |(RB1) |      |(RB2) |      | (RBk)|
          +------+      +------+      +------+
            |..|          |..|          |..|
            |  +-----+    |  |          |  |
            |   +---|-----|--|----------+  |
            | +-|---|-----+  +-----------+ |
   MC-      | | |   +------------------+ | |
   LAG1--->(| | |)                    (| | |) <---MC-LAG2
          +-------+    .  .  .       +-------+
          | CE1   |                  | CEn   |
          |       |                  |       |
          +-------+                  +-------+
```

Active-Active connection is useful when we want to achieve the following requirements.

- Flow rather than VLAN based load balancing is desired.

- More rapid failure recovery is desired. Current appointed forwarder mechanism relies on the Hello timer expiration to detect the unreachability of another edge RBridge connecting to the same local Ethernet link. Then re-appointing the forwarder for specific VLANs may be required. Such procedures takes time in the scale of seconds. Active-Active connection usually has faster built-in mechanism for member node and/or link failure detection. Faster detection of failure would minimize the frame loss and recovery time.

MC-LAG implementation varies by vendor. In order to have common understanding of active-active connection scenarios, the following assumptions are held regardless of the implementation.

For CE connecting to multiple edge RBs via active-active connection:
a) it will forward packets from endnodes to exactly one up-link
b) it will never forward packets from one up-link to another
c) it will attempt to send all packets for a given flow on the same uplink
d) packets are accepted from any of the uplinks and passed down endnodes (if exist)
e) it has no pre-determined rules for which packets get sent to which uplinks (such as certain VLANs go on certain uplinks, or certain source addresses go on certain uplinks)
f) it cannot be assumed to give useful control information to the up-link (such as "set of other RBridges CE is attached", or "all the MAC addresses attached".

For edge group to which CE is multiply attached:
a) Any two RBs in the edge group are reachable to each other
b) Each RB in the edge group is configured with a name for each down-link to an CE multiply attached to that group. The names will be consistent across the edge group. For instance, if CE1 attaches to RB1, RB2 to RBn, then each of RBs will have been configured, for the port to CE1, that it is labeled "MC-LAG1"
c) The RBs in the edge group have existing mechanisms to exchange states and information with each other, including the set of CEs they are connecting to or name of MC-LAGs their down-links have joined
d) Each RB in the edge group can be configured with the set of acceptable VLANs (or fine-grained labels) for the ports to any CE. The acceptable VLANs configured for those port should include all the VLANs the CE has joined and be consistent for all the member RB.
e) When a RB fails, all the other RBs having formed any MC-LAG with it know the information timely

f) When a down-link of a RB fails , all the other RBs having formed any MC-LAG with that down-link know the information timely

3. Problems in active-active connection at the edge

This section presents the problems need to be addressed in active-active connection scenarios. Topology in Figure 1 is used in the following sub-sections as the example scenario for illustration purpose.

3.1 Frame duplications

When a remote RBridge sends a TRILL encapsulated multi-destination frame of VLAN x, all member RBridges of MC-LAG1 will receive the frame if local CE1 joins VLAN x. As each of them thinks it is the appointed forwarder for all VLANs, they would all forward the frame to CE1. The consequence is CE1 receives multiple copies of that multi-destination frame from the remote end host.

It should be noted frame duplication may only happen in multi-destination frame forwarding. Unicast forwarding does not have this issue.

3.2 Loop

As shown in Figure 1, CE1 may send a native multi-destination frame to TRILL campus via a member of MC-LAG1 (say RB1). This frame will be TRILL encapsulated and then forwarded through the campus to another member (say RB2) of the same MC-LAG. In this case, RB2 will decapsulate the frame and forward it. The frame loops back to CE1.

3.2 Address flip-flop

Consider RB1 and RB2 using their own nickname as source nickname to ingress data frame into a TRILL campus. As shown by Figure 1, CE1 may send a data frame with the same source VLAN/MAC address to any member RB of MC-LAG1. If the egress RBridge receives TRILL packet from different ingress RBridge RBridges but with same same source VLAN/MAC address, it learns different address correspondence from the data frames. Address correspondence may keep flip-flopping among nicknames of the member RBridges of the MC-LAG for the same VLAN/MAC address in the same VLAN.

Some TRILL switches may behave badly under these circumstances and, for example, interpret this as a severe network problem. It may also cause the returning traffic to go through the different paths to reach the destination resulting in persistent re-ordering of the frames.

3.3 Unsynchronized Information among member RBridges

A local Rbridge, say RB1 in MC-LAG1, may have learned a VLAN/MAC and nickname correspondence for a remote host h1 when h1 sends a packet to CE1. The returning traffic from CE1 may go to any other member RBridge of MC-LAG1, e.g., RB2. RB2 has no h1's VLAN/MAC and nickname correspondence stored. Therefore it has to do the flooding for unknown unicast. Such flooding is considered unnecessary since the returning traffic is always expected and RB1 had learned the address correspondence.

Synchronization on the VLAN/MAC and nickname correspondence information among member RBridges will reduce the unnecessary flooding.

Unsynchronized multicast group information causes problem too. The edge RBridge snoops the IGMP [RFC3376] join message from CE may not be the one receiving the multicast traffic for the joined group later. Therefore the multicast traffic can be dropped incorrectly.

TRILL[RFC6325] designed its multi-destination traffic forwarding with some specific mechanism, e.g., RPF checking, tree calculation, construction and selection, pruning, etc. Solutions of active-active connection to edge RBridges should carefully examine those features and make sure they work correctly.

4 High level requirements for solutions

Problems identified by section 3 should be solved in any solution used for active-active connection to RBridges. The requirements are summarized as follows,
a) Loop and frame duplication MUST be prevented
b) Learning of VLAN/MAC and nickname correspondence by a remote RBridge MUST not flip-flop between the local multiply attached edge RBridges
c) Member RBridges of a MC-LAG MUST be able to share the relevant TRILL specific information with each other

In addition, the following high level requirements should be fulfilled too.

Data plane:
1) all up-links of CE MUST be active. CE is free to choose any up-link to send packets
2) packets for a flow should stay in order
3) RPF check MUST work properly as per [RFC6325]
4) Single up-link failure on CE to an edge group MUST not cause persistent packet delivery failure from TRILL campus to CE

Control plane:
1) no requirement on information passed between edge RBs and CE
2) If there is any TRILL specific parameters required to be exchanged between RBridges in a edge group, e.g., nickname, solution SHOULD specify the mechanism to perform such exchange.

Configuration, incremental deployment and others:
1) Solution should provide minimal configuration
2) Solution should automatically detects misconfiguration of edge RBridge group
3) Solution should support incremental deployment, i.e. not require campus wide hardware upgrading for all RBridges
4) Solution should be able to support 4 active-active up-links on an multiply attached CE

5 Security Considerations

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

6 IANA Considerations

No IANA action is required. RFC Editor: please delete this section before publication.

6 References

5.1 Normative References

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

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

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

[RFC6327bis] Eastlake 3rd, D., R. Perlman, A. Ghanwani, H. Yang, and V. Manral, "TRILL: Adjacency", draft-ietf-trill-rfc6327bis, work in progress.

[RFC6439] Eastlake, D. et.al., "RBridge: Appointed Forwarder", RFC 6439, November 2011.

5.2 Informative References

[RFC3376] Cain, B., Deering, S., Kouvelas, I., Fenner, B., and A. Thyagarajan, "Internet Group Management Protocol, Version 3", RFC 3376, October 2002.

[TRILLPN] Zhai,H., et.al., "RBridge: Pseudonode Nickname", draft-hu-trill-pseudonode-nickname, Work in progress, November 2011.

[8021AX] IEEE, "Link Aggregation", 802.1AX-2008, 2008.

[8021Q] IEEE, "Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks", IEEE Std 802.1Q-2011, August, 2011

Authors' Addresses

Yizhou Li
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China

Phone: +86-25-56625375
EMail: liyizhou@huawei.com

Donald Eastlake
Huawei R&D USA
155 Beaver Street
Milford, MA 01757 USA

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

Weiguo Hao
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China

Phone: +86-25-56623144
EMail: haoweiguo@huawei.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

Jon Hudson
Brocade
130 Holger Way
San Jose, CA 95134 USA

Phone: +1-408-333-4062
jon.hudson@gmail.com

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

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