Internet Engineering Task Force (IETF)
Request for Comments: 7379
Category: Informational
ISSN: 2070-1721

Y. Li
W. Hao
Huawei Technologies
R. Perlman
EMC
J. Hudson
Brocade
H. Zhai
JIT

October 2014

Problem Statement and Goals for Active-Active Connection at the Transparent Interconnection of Lots of Links (TRILL) Edge

Abstract

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

Status of This Memo

Copyright 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 [IS-IS] [RFC6165] [RFC7176] 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) spread load 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 reconfigure to accommodate link or node failures or additions.

To the degree practical, "active-active" is the extension 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 spread end-station traffic load at the granularity of flows across links to such multiple edge TRILL switches and rapidly reconfigure 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 the RBridges base protocol [RFC6325] are used herein with the following additions:

CE: Customer Equipment (end station or bridge).

Data Label: VLAN or FGL (Fine-Grained Label [RFC7172]).

LAALP: Local Active-Active Link Protocol. Any protocol similar to MC-LAG that runs in a distributed fashion on a CE, on the links from that CE to a set of edge group RBridges, and on those RBridges.

MC-LAG: Multi-Chassis Link Aggregation. Proprietary extensions to IEEE Std 802.1AX-2011 [802.1AX] so that the aggregated links can, at one end of the aggregation, attach to different switches.

Edge group: a group of edge RBridges to which at least one CE is multiply attached using an LAALP. When multiple CEs attach to the exact same set of edge RBridges, those edge RBridges can be considered as a single edge group. An RBridge can be in more than one edge group.

RBridge: Routing Bridge. An alternative name for a TRILL switch.

TRILL: Transparent Interconnection of Lots of Links.

TRILL switch: a device that implements the TRILL protocol; an alternative term for an RBridge.

2. Target Scenario

This section presents a typical scenario of active-active connection to a TRILL campus via multiple edge RBridges where the current TRILL Appointed Forwarder mechanism does not work as expected.

The TRILL Appointed Forwarder mechanism [RFC6439] can handle failover (active-standby), provides loop avoidance, and, with administrative configuration, provides load spreading based on VLAN. One and only one appointed RBridge can ingress/egress native frames into/from the TRILL campus for a given VLAN among all edge RBridges connecting a legacy network to the TRILL campus. This is true whether the legacy network is a simple point-to-point link or a complex bridged LAN or anything in between. By carefully selecting different RBridges as Appointed Forwarder for different sets of VLANs, load spreading over different edge RBridges across different Data Labels can be achieved.

The Appointed Forwarder mechanism [RFC6439] requires all of the edge group RBridges to exchange TRILL IS-IS Hello packets through their access ports. As Figure 1 shows, when multiple access links of multiple edge RBridges are connected to a CE by an LAALP, Hello messages sent by RB1 via access port to CE1 will not be forwarded to RB2 by CE1. RB2 (and other members of LAALP1) will not see that Hello from RB1 via the LAALP1. Every member RBridge of LAALP1 thinks of itself as Appointed Forwarder on an LAALP1 link for all VLANs and will ingress/egress frames. Hence, the Appointed Forwarder mechanism cannot provide active-active or even active-standby service across the edge group in such a scenario.

```
                   ----------------------
                  |                      |
                  |   TRILL Campus       |
                  |                      |
                   ----------------------
                       |       |    |
                  -----        |     --------
                 |             |             |
             +------+      +------+      +------+
             |      |      |      |      |      |
             |(RB1) |      |(RB2) |      | (RBk)|
             +------+      +------+      +------+
               |..|          |..|          |..|
               |  +----+     |  |          |  |
               |   +---|-----|--|----------+  |
               | +-|---|-----+  +-----------+ |
               | | |   +------------------+ | |
    LAALP1--->(| | |)                    (| | |) <---LAALPn
             +-------+    .  .  .       +-------+
             | CE1   |                  | CEn   |
             |       |                  |       |
             +-------+                  +-------+
```

Figure 1: Active-Active Connection to TRILL Edge RBridges

Active-active connection is useful when we want to achieve the following two goals:

- Flow-based rather than VLAN-based load balancing is desired.

- More rapid failure recovery is desired.

The current Appointed Forwarder mechanism relies on the TRILL Hello timer expiration to detect the unreachability of another edge RBridge connecting to the same local link. Then, reappointing the forwarder for specific VLANs may be required. Such procedures take time on the scale of seconds although this can be improved with TRILL use of Bidirectional Forwarding Detection (BFD) [RFC7175]. Active-active connection usually has a faster built-in mechanism for member node and/or link failure detection. Faster detection of failures minimizes the frame loss and recovery time.

Today, LAALP is usually a proprietary facility whose implementation varies by vendor. So, to be sure the LAALP operates successfully across a group of edge RBridges, those edge RBridges will almost always have to be from the same vendor. In the case where the LAALP is an MC-LAG, the CE normally implements the logic described in IEEE Std 802.1AX-2011 [802.1AX], so proprietary elements would only be at the end of the edge group. There is also a revision of IEEE Std 802.1AX-2011 [802.1AX] underway (802.1X-REV) to remove the restriction in IEEE Std 802.1AX-2011 [802.1AX] that there be one box at each end of the aggregation. So it is possible that, in the future, an LAALP could be implemented through such a revised IEEE Std 802.1AX-2011 [802.1AX] with standards-conformant logic at the ends of both the CE and edge group. In order to have a common understanding of active-active connection scenarios, the assumptions in Section 2.1 are made about the characteristics of the LAALP and edge group of RBridges.

2.1. LAALP and Edge Group Characteristics

For a CE connecting to multiple edge RBridges via an LAALP (active-active connection), the following characteristics apply:

a) The LAALP will deliver a frame from an end node to TRILL at exactly one edge group RBridge.

b) The LAALP will never forward frames it receives from one uplink to another.

c) The LAALP will attempt to send all frames for a given flow on the same uplink. To do this, it has some unknown rule for which frames get sent to which uplinks (typically based on a simple hash function of Layer 2 through 4 header fields).

d) Frames are accepted from any of the uplinks and passed down to end nodes (if any exist).

e) The LAALP cannot be assumed to send useful control information to the uplink such as "this is the set of other RBridges to which this CE is attached" or "these are all the MAC addresses attached".

For an edge group of RBridges to which a CE is multiply attached with an LAALP:

a) Any two RBridges in the edge group are reachable from each other via the TRILL campus.

b) Each RBridge in the edge group knows an ID for each LAALP instance multiply attached to that group. The ID will be consistent across the edge group and globally unique across the TRILL campus. For example, if CE1 attaches to RB1, RB2, ... RBn using an LAALP, then each of the RBs will know, for the port to CE1, that it has a label such as "LAALP1".

c) Each RB in the edge group can be configured with the set of acceptable VLANs for the ports to any CE. The acceptable VLANs configured for those ports should include all the VLANs the CE has joined and be consistent for all the member RBridges of the edge group.

d) When an RBridge fails, all the other RBridges that have formed an LAALP instance with it learn of the failure in a timely fashion.

e) When a downlink of an edge group RBridge to an LAALP instance fails, that RBridge and all the other RBridges participating in the LAALP instance, including that downlink, learn of the failure in a timely fashion.

f) The RBridges in the edge group have a mechanism to exchange information with each other, information such as the set of CEs they are connecting to or the IDs of the LAALP instances their downlinks are part of.

Other than the applicable characteristics above, the internals of an LAALP are out of the scope of TRILL.

3. Problems in Active-Active Connection at the TRILL Edge

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

3.1. Frame Duplications

When a remote RBridge ingresses a multi-destination TRILL data packet in VLAN x, all edge group RBridges of LAALP1 will receive the frame if any local CE1 joins VLAN x. As each of them thinks it is the Appointed Forwarder for VLAN x, without changes made for active-active connection support, they would all forward the frame to CE1. The bad consequence is that CE1 receives multiple copies of that multi-destination frame from the remote end-host source.

Frame duplication may also occur when an ingress RBridge is non-remote -- say, ingress and egress are two RBridges belonging to the same edge group. Assume LAALP m connects to an edge group g, and the edge group g consists of RB1, RB2, and RB3. The multi-destination frames ingressed from a port not connected to LAALP m by RB1 can be locally replicated to other ports on RB1 and also TRILL encapsulated and forwarded to RB2 and RB3. CE1 will receive duplicate copies from RB1, RB2, and RB3.

Note that frame duplication is only a problem in multi-destination frame forwarding. Unicast forwarding does not have this issue as there is only ever one copy of the packet.

3.2. Loopback

As shown in Figure 1, CE1 may send a native multi-destination frame to the TRILL campus via a member of the LAALP1 edge group (say RB1). This frame will be TRILL encapsulated and then forwarded through the campus to the multi-destination receivers. Other members (say RB2) of the same LAALP edge group will receive this multicast packet as well. In this case, without changes made for active-active connection support, RB2 will decapsulate the frame and egress it. The frame loops back to CE1.

3.3. Address Flip-Flop

Consider RB1 and RB2 using their own nickname as ingress nickname for data into a TRILL campus. As shown in Figure 1, CE1 may send a data frame with the same VLAN and source Media Access Control (MAC) address to any member of the edge group LAALP1. If an egress RBridge receives TRILL data packets from different ingress RBridges but with the same source Data Label and MAC address, it learns different correspondences between a {Data Label and MAC address} and nickname when decapsulating the data frames. Address correspondence may keep flip-flopping among nicknames of the member RBridges of the LAALP for the same Data Label and MAC address. Existing hardware does not support data-plane learning of multiple nicknames for the same MAC address and Data Label -- when data-plane learning indicates attachment of the MAC address to a new nickname, it overwrites the old attachment nickname.

Implementers have stated that most current TRILL switch hardware, when doing data-plane learning, behaves badly under these circumstances and, for example, interprets address flip-flopping as a severe network problem. It may also cause the returning traffic to go through different paths to reach the destination, resulting in persistent reordering of the frames.

3.4. Unsynchronized Information among Member RBridges

A local RBridge, say RB1 connected to LAALP1, may have learned a correspondence between a {Data Label and MAC address} and nickname 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 LAALP1, for example, RB2. RB2 may not have h1’s correspondence between a {Data Label and MAC address} and nickname stored. Therefore, it has to do the flooding for unknown unicast addresses [RFC6325]. Such flooding is unnecessary since the returning traffic is almost always expected and RB1 had learned the address correspondence. It is desirable to avoid flooding; it imposes a greater burden on the network than known destination unicast traffic because the flooded traffic is sent over more links.

Synchronization of the correspondence between a {Data Label and MAC address} and nickname information among member RBridges will reduce such unnecessary flooding.

4. High-Level Requirements and Goals for Solutions

The problems identified in Section 3 should be solved in any solution for active-active connection to edge RBridges. The following high-level requirements and goals should be met.

Data plane:

1) All uplinks of a CE MUST be active: the LAALP is free to choose any uplink on which to send packets, and the CE is able to receive packets from any uplink of an edge group.

2) Loopback and frame duplication MUST be prevented.

3) Learning of correspondence between a {Data Label and MAC address} and nickname by a remote RBridge MUST NOT flip-flop between the local multiply attached edge RBridges.

4) Packets for a flow SHOULD stay in order.

5) The Reverse Path Forwarding Check MUST work properly as per the RBridges base protocol [RFC6325].

6) Single uplink failure on a CE to an edge group MUST NOT cause persistent packet delivery failure between a TRILL campus and CE.

Control plane:

1) No requirement for new information to be passed between edge RBridges and CEs or between edge RBridges and end nodes exists.

2) If there is any TRILL-specific information required to be exchanged between RBridges in an edge group, for example, Data Labels and MAC addresses binding to nicknames, a solution MUST specify the mechanism to perform such exchange unless this is handled internal to the LAALP.

3) RBridges SHOULD be able to discover other members in the same edge group by exchanging their LAALP attachment information.

Configuration, incremental deployment, and others:

1) Solution SHOULD require minimal configuration.

2) Solution SHOULD automatically detect misconfiguration of edge RBridge group.

3) Solution SHOULD support incremental deployment, that is, not require campus-wide upgrading for all RBridges, only changes to the edge group RBridges.

4) Solution SHOULD be able to support from two up to at least four active-active uplinks on a multiply attached CE.

5) Solution SHOULD NOT assume there is a dedicated physical link between any two edge RBridges in an edge group.

5. Security Considerations

As an informational overview, this document does not introduce any extra security risks. Security risks introduced by a particular LAALP or other elements of solutions to the problems presented here will be discussed in the separate document(s) describing such LAALP or solutions.

End-station links in TRILL are Ethernet links, and consideration should be given to securing them with link security as described in IEEE Std 802.1AE-2006 [802.1AE] for the protection of end-station data and link-level control messages, including any LAALP control messages.

For general TRILL Security Considerations, see the RBridges base protocol [RFC6325].

6. References

6.1. Normative References

[IS-IS] ISO/IEC, "Information technology -- Telecommunications and information exchange between systems -- Intermediate System to Intermediate System intra-domain routeing information exchange protocol for use in conjunction with the protocol for providing the connectionless-mode network service (ISO 8473)", ISO/IEC 10589:2002, Second Edition, 2002.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997, <http://www.rfc-editor.org/info/rfc2119>.

[RFC6165] Banerjee, A. and D. Ward, "Extensions to IS-IS for Layer-2 Systems", RFC 6165, April 2011, <http://www.rfc-editor.org/info/rfc6165>.

[RFC6325] Perlman, R., Eastlake 3rd, D., Dutt, D., Gai, S., and A. Ghanwani, "Routing Bridges (RBridges): Base Protocol Specification", RFC 6325, July 2011, <http://www.rfc-editor.org/info/rfc6325>.

[RFC6439] Perlman, R., Eastlake, D., Li, Y., Banerjee, A., and F. Hu, "Routing Bridges (RBridges): Appointed Forwarders", RFC 6439, November 2011, <http://www.rfc-editor.org/info/rfc6439>.

[RFC7172] Eastlake 3rd, D., Zhang, M., Agarwal, P., Perlman, R., and D. Dutt, "Transparent Interconnection of Lots of Links (TRILL): Fine-Grained Labeling", RFC 7172, May 2014, <http://www.rfc-editor.org/info/rfc7172>.

[RFC7176] Eastlake 3rd, D., Senevirathne, T., Ghanwani, A., Dutt, D., and A. Banerjee, "Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS", RFC 7176, May 2014, <http://www.rfc-editor.org/info/rfc7176>.

6.2. Informative References

[802.1AE] IEEE, "IEEE Standard for Local and metropolitan area networks -- Media Access Control (MAC) Security", IEEE Std 802.1AE-2006, 18 August 2006.

[802.1AX] IEEE, "IEEE Standard for Local and metropolitan area networks -- Link Aggregration", IEEE Std 802.1AX-2008, 3 November 2008.

[RFC7175] Manral, V., Eastlake 3rd, D., Ward, D., and A. Banerjee, "Transparent Interconnection of Lots of Links (TRILL): Bidirectional Forwarding Detection (BFD) Support", RFC 7175, May 2014, <http://www.rfc-editor.org/info/rfc7175>.

Acknowledgments

Special acknowledgments to Donald Eastlake, Adrian Farrel, and Mingui Zhang for their valuable comments.

Authors' Addresses

Yizhou Li
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China

Phone: +86-25-56625409
EMail: liyizhou@huawei.com

Weiguo Hao
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China

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

Radia Perlman
EMC
2010 256th Avenue NE, #200
Bellevue, WA 98007
United States

EMail: Radia@alum.mit.edu

Jon Hudson
Brocade
130 Holger Way
San Jose, CA 95134
United States

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

Hongjun Zhai
JIT

EMail: honjun.zhai@tom.com