TRILL WG
Internet-Draft
Intended status: Standards Track
Expires: November 25, 2015

Radia. Perlman
EMC Corporation
Fangwei. Hu
ZTE Corporation
Donald. Eastlake 3rd
Huawei technology
Kesava. Krupakaran
Dell
Ting. Liao
ZTE Corporation

May 24, 2015

TRILL Smart Endnodes
draft-ietf-trill-smart-endnodes-01.txt

Abstract

This draft addresses the problem of the size and freshness of the endnode learning table in edge RBridges, by allowing endnodes to volunteer for endnode learning and encapsulation/decapsulation. Such an endnode is known as a "smart endnode". Only the attached RBridge can distinguish a "smart endnode" from a "normal endnode". The smart endnode uses the nickname of the attached RBridge, so this solution does not consume extra nicknames. The solution also enables Fine Grained Label aware endnodes.

Table of Contents

1. Introduction

The IETF TRILL (Transparent Interconnection of Lots of Links) protocol [RFC6325] provides optimal pair-wise data frame forwarding without configuration, safe forwarding even during periods of temporary loops, and support for multipathing of both unicast and multicast traffic. TRILL accomplishes this by using IS-IS [IS-IS] [RFC7176] link state routing and encapsulating traffic using a header that includes a hop count. Devices that implement TRILL are called "RBridges" (Routing Bridges) or "TRILL Switches".

An RBridge that attaches to endnodes is called an "edge RBridge" or "edge TRILL Swtich", whereas one that exclusively forwards encapsulated frames is known as a "transit RBridge" or "transit TRILL Switch". An edge RBridge traditionally is the one that encapsulates a native Ethernet packet with a TRILL header, or that receives a

TRILL-encapsulated packet and decapsulates the TRILL header. To encapsulate efficiently, the edge RBridge must keep an "endnode table" consisting of (MAC,Data Label, TRILL egress switch nickname) sets, for those remote MAC addresses in Data Labels currently communicating with endnodes to which the edge RBridge is attached.

These table entries might be configured, received from ESADI [RFC7357], looked up in a directory [RFC7067], or learned from decapsulating received traffic. If the edge RBridge has attached endnodes communicating with many remote endnodes, this table could become large. Also, if one of the MAC addresses and Data Labels in the table has moved to a different remote TRILL switch, it might be difficult for the edge RBridge to notice this quickly, and because the edge RBridge is encapsulting to the incorrect egress RBridge, the traffic will get lost.

For these reasons, it is desirable for an endnode E (whether it is a server, hypervisor, or VM) to maintain the endnode table for remote endnodes that E is corresponding with. This eliminates the need for the edge RBridge RBx, to which E is connected, to know about those nodes (unless some non-smart endnode attached to RBx is also corresponding with those nodes), Once D is unreachable for E, which could be determined through ICMP messages or other techniques, the smart endnode should delete the entry of (MAC, Data Label, nickname). If D moves to a new place, E should attempt to acquire a fresh entry for D by flooding to D, examining updates to the ESADI link state database[RFC7357],or consulting a directory[RFC7067].

The mechanism in this draft is that E issue a Smart-Hello (even though E is just an endnode), indicating E's desire to act as a smart endnode, together with the set of MAC addresses and Data Labels that E owns, and whether E would like to receive ESADI packets. E learns from RBx's Smart-Hello, whether RBx is capable of having a smart endnode neighbor, what RBx's nickname is, and which trees RBx can use when RBx ingresses multi-destination frames. Although E transmits Smart-Hellos, E does not transmit or receive LSPs or E-L1FS FS-LSPs[I-D.ietf-trill-rfc7180bis].

RBx will accept already-encapsulated TRILL Data packets from E (perhaps verifying that the source MAC and Data Label is indeed one of the ones that E owns, that the ingress RBridge field is RBx's, and if the packet is an encapsulated multi-destination frame, the tree selected is one of the ones that RBx has claimed it will choose). When RBx receives (from the campus) a TRILL Data packet with RBx's nickname as egress, RBx checks whether the destination MAC address and Data Label in the inner packet is one of the MAC addresses and Data Labels that E owns, and if so, RBx forwards the packet onto E's port, keeping it encapsulated.

Since a smart endnode can encapsulate TRILL Data frames, it can cause the Inner.Lable to be a Fine Grained Label [RFC7172], thus this method supports FGL aware endnodes.

2. Terminology

Edge RBridge: An RBridge providing endnode service on at least one of its ports. It is also named as edge TRILL Switch

Data Label: VLAN or FGL.

ESADI: End Station Address Distribution Information [RFC7357].

FGL: Fine Grained Label [RFC7172].

IS-IS: Intermediate System to Intermediate System [IS-IS].

RBridge: Routing Bridge, an alternative name for a TRILL switch.

Smart endnode: An endnode that has the capability specified in this document including learning and maintaining(MAC, Data Lable, Nickname) entries and encapsulating/decapsulating TRILL frame.

Transit RBridge: An RBridge exclusively forwards encapsulated frames. It is also named as transit RBridge.

TRILL: Transparent Interconnection of Lots of Links [RFC6325].

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

3. Smart-Hello Content

Suppose endnode E is attached to RBridge RBx. In order for E to act as a smart endnode, both E and RBx have to be signaled. The logical choice of frame to do this is Smart-Hello.

3.1. Edge RBridge's Smart-Hello

For smart endnode operation, RBx's Smart-Hello must contain the following information:

- RBridge's nickname. The nickname sub-TLV (Specified in section 2.3.2 in [RFC7176])could be reused here, and TLV 242 (IS-IS router capability) should be updated to be carried in Smart-Hello frame.

- Tree that RBx can use when ingressing multi-destination frames. The Tree Identifiers Sub-TLV (Specified in section 2.3.4 in [RFC7176]) could be reused here.

- Smart endnode neighbor list. The TRILL Neighbor TLV (Specified in section 2.5 in [RFC7176]) could be reused.

3.2. Smart Endnode's Smart-Hello

A new TLV (S-MAC TLV)is defined for smart endnode. If there are several VLANs/FGL Data Label for that smart endnode, the TLV could be filled several times in smart endnode's Smart-Hello.

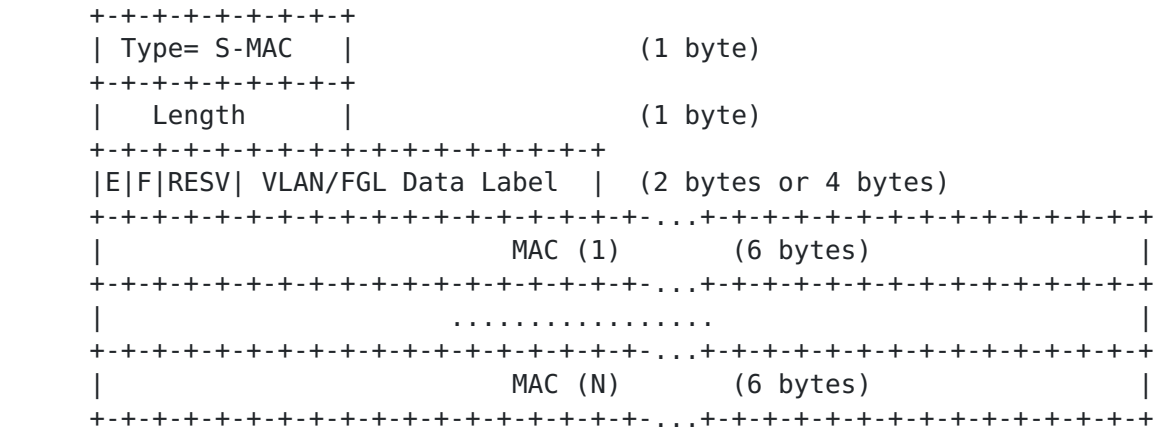

Figure 1 S-MAC TLV

- Type: S-MAC, the value is TBD.

- Length: Total number of bytes contained in the value field.

- E: one bit. If it sets to 1, which indicates that the endnode should receive ESADI frames.

- F: one bit. If it sets to 1, which indicates that the endnode supports FGL data label, otherwise, the VLAN/FGL Data Label [RFC7172] field is the VLAN ID.

- RESV: 2 bits or 6 bits, is reserved for the future use. If VLAN/FGL Data Label indicates the VLAN ID(or F flag sets to 0), the RESV field is 2 bits length, otherwise it is 6 bits.

- VLAN/FGL Data Label: This carries a 12-bits VLAN identifier or 24-bits FGL Data Label that is valid for all subsequent MAC addresses in this TLV, or the value zero if no VLAN/FGL data label is specified.

- MAC(i): This is the 48-bit MAC address reachable in the Data Label given from the IS that is announcing this TLV.

4. Frame Processing

4.1. Frame Processing for Smart Endnode

Smart endnode E does not issue or receive LSPs or E-L1FS FS-LSPs or calculate topology. E does the following:

- E maintains an endnode table of (MAC, Data Label, nickname) entries of end nodes with which the smart endnode is communicating. Entries in this table are populated the same way that an edge RBridge populates the entries in its table:
  - learning from (source, ingress) on packets it decapsulates.
  - from ESADI[RFC7357].
  - by querying a directory [RFC7067].
  - by having some entries configured.
- When E wishes to transmit to unicast destination D, if (D, nickname) is in E's endnode table, E encapsulates with ingress nickname=RBx, egress nickname as indicated in D's table entry. If D is unknown, D either queries a directory or encapsulates the packet as a multi-destination frame, using one of the trees that RBx has specified in RBx's Smart-Hello.
- When E wishes to transmit to a multicast or broadcast destination, E encapsulates the packet using one of the trees that RBx has specified.

The smart endnode E need not send Smart-Hellos as frequently as normal RBridges. These Smart-Hellos could be periodically unicast to the Appointed Forwarder RBx. In case RBx crashes and restarts, or the DRB changes and E receives the Smart-Hello without mentioning E, E SHOULD send a Smart-Hello immediately. If RBx is AF for any of the VLANs that E claims, RBx MUST list E in its Smart-Hellos as a smart endnode neighbor.

4.2. Frame Processing for Edge RBridge

The attached RBridge RBx does the following:

- If receiving an encapsulated unicast data frame from a port with a smart endnode, with RBx's nickname as ingress, RBx forwards the

frame to the specified egress nickname, as with any encapsulated frame. However, RBx MAY filter the encapsulation frame based on the inner source MAC and Data Label as specified for the smart endnode. If the MAC (or Data Label) are not among the expected set of the smart endnode, the frame would be dropped by the edge RBridge.

- o If receiving an mulit-destination TRILL Data packet from a port with smart endnode, RBridge RBx forwards the TRILL encapsulation to the TRILL campus based on the distribution tree. If there are some normal endnodes (i.e, non-smart endnode) attached to RBridge RBx, RBx should decapsulates the frame and sends the native frame to these ports.

- o When RBx receives a multicast frame from a remote RBridge, and the exit ports includes hybrid endnodes, it should send two copies of mulicast frames, one as native and the other as TRILL encapsulated frame. When smart endnode receives the encapsulated frame, it learns the remote (MAC, Data Label, Nickname) set, A smart endnodes ignores any native data frames. The normal endnode receives the native frame and learns the remote MAC address and ignore the native frame. This transit solution may bring some complex for the edge RBridge and waste network bandwidth resource, so it is recommended to avoid the hybrid endnodes scenario by attaching the smart endnodes and non-smart endnodes to different ports when deployed. Another solution is that if there are one or more endnodes on a link, the non-smart endnodes are ignored on a link; but we can configure a port to support mixed links. The RBx only sends TRILL encapsulated frame to the link in this situation.

5. Multi-homing

It is supposed that endnode E is attached to the TRILL campus in two places: to RBridges RB1 and RB2. There are two ways for this to work:

(1) E can choose either RB1 or RB2's nickname, when encapsulating a frame, whether the encapsulated frame is sent via RB1 or RB2. If E wants to do active-active load splitting, and uses RB1's nickname when forwarding through RB1, and RB2's nickname when forwarding through RB2, which will cause the flip-floping of the endnode table entry in the remote RBridges (or smart endnodes). One solution is to set a multi-homing bit in the RESV field of the TRILL data Frame. When remote RBs or smart endnodes receive the data frame with the multi-homed bit set, the MAC entry (E, RB1's nickname) and (E, RB2's nickname) will be coexist as two entries for that MAC address. Another solution is to extend the ESADI protocol to distribute multiple attachments of a MAC

address of a multi-homing group.(Please refer to the option C in section 4 of [I-D.ietf-trill-aa-multi-attach] for details).

(2) RB1 and RB2 might indicate, in their Smart-Hello, a virtual nickname that attached end nodes may use if they are multihomed to RB1 and RB2, separate from RB1 and RB2's nicknames (which they would also list in their Smart-Hello). This would be useful if there were many end nodes multihomed to the same set of RBridges. This would be analogous to a pseudonode nickname; return traffic would go via the shortest path from the source to the endnode, whether it is RB1 or RB2. If E loses connectivity to RB2, then E would revert to using RB1's nickname. In order to avoid RPF check issue for multi-destination frame, the affinity TLV [I-D.ietf-trill-cmt] is recommended to be used in this solution.

6. Security Considerations

For general TRILL Security Considerations, see [RFC6325].

7. Acknowledgements

The contributions of the following persons are gratefully acknowledged: Mingui Zhang, Weiguo Hao, Linda Dunbar and Andrew Qu.

8. IANA Considerations

IANA is requested to allocate a S-MAC TLV identifier. TLV 242(ISIS router capability) is required to updated to be carried by Smart-Hello frame.

9. Normative References

[I-D.ietf-trill-aa-multi-attach]
Zhang, M., Perlman, R., Zhai, H., Durrani, M., and S. Gupta, "TRILL Active-Active Edge Using Multiple MAC Attachments", draft-ietf-trill-aa-multi-attach-03 (work in progress), February 2015.

[I-D.ietf-trill-cmt]
Senevirathne, T., Pathangi, J., and J. Hudson, "Coordinated Multicast Trees (CMT) for TRILL", draft-ietf-trill-cmt-06 (work in progress), March 2015.

[I-D.ietf-trill-rfc7180bis]
Eastlake, D., Zhang, M., Perlman, R., Banerjee, A., Ghanwani, A., and S. Gupta, "TRILL: Clarifications, Corrections, and Updates", draft-ietf-trill-rfc7180bis-04 (work in progress), March 2015.

[IS-IS] ISO/IEC 10589:2002, Second Edition,, "Intermediate System to Intermediate System Intra-Domain Routing Exchange Protocol for use in Conjunction with the Protocol for Providing the Connectionless-mode Network Service (ISO 8473)", 2002.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

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

[RFC6325] Perlman, R., Eastlake, D., Dutt, D., Gai, S., and A. Ghanwani, "Routing Bridges (RBridges): Base Protocol Specification", RFC 6325, July 2011.

[RFC7067] Dunbar, L., Eastlake, D., Perlman, R., and I. Gashinsky, "Directory Assistance Problem and High-Level Design Proposal", RFC 7067, November 2013.

[RFC7172] Eastlake, D., Zhang, M., Agarwal, P., Perlman, R., and D. Dutt, "Transparent Interconnection of Lots of Links (TRILL): Fine-Grained Labeling", RFC 7172, May 2014.

[RFC7176] Eastlake, 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.

[RFC7357] Zhai, H., Hu, F., Perlman, R., Eastlake, D., and O. Stokes, "Transparent Interconnection of Lots of Links (TRILL): End Station Address Distribution Information (ESADI) Protocol", RFC 7357, September 2014.

Authors' Addresses

Radia Perlman
EMC Corporation
2010 156th Ave NE, suite #200
Bellevue, WA 98007
USA

Phone: +1-206-291-367
Email: radiaperlman@gmail.com

Fangwei Hu
ZTE Corporation
No.889 Bibo Rd
Shanghai 201203
China

Phone: +86 21 68896273
Email: hu.fangwei@zte.com.cn

Donald Eastlake,3rd
Huawei technology
155 Beaver Street
Milford, MA 01757
USA

Phone: +1-508-634-2066
Email: d3e3e3@gmail.com

Kesava Vijaya Krupakaran
Dell
Olympia Technology Park
Guindy Chennai 600 032
India

Phone: +91 44 4220 8496
Email: Kesava_Vijaya_Krupak@Dell.com

Ting Liao
ZTE Corporation
No.50 Ruanjian Ave.
Nanjing, Jiangsu 210012
China

Phone: +86 25 88014227
Email: liao.ting@zte.com.cn