TRILL WG
Internet-Draft
Intended status: Standards Track
Expires: April 4, 2014

Radia Perlman
Intel Labs
Fangwei Hu
ZTE Corporation
Donald Eastlake 3rd
Huawei technology
Kesava Vijaya Krupakaran
Dell
Ting Liao
ZTE Corporation

Oct 2013

TRILL Smart Endnodes
draft-perlman-trill-smart-endnodes-02.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.

Table of Contents

1. Introduction

The IETF TRILL (Transparent Interconnection of Lots of Links) protocol implemented by devices called RBridges (Routing Bridges, [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([RFC1195]) ([RFC6165]) ([I-D.ietf-isis-rfc6326bis])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", whereas one that exclusively forwards encapsulated frames is known as a "transit RBridge". 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 removes the TRILL header. To encapsulate, the edge RBridge must keep an "endnode table" consisting of (MAC, TRILL egress switch nickname) pairs, for those MAC addresses currently communicating with endnodes to which the edge RBridge is attached.

These table entries might be configured, received from ESADI ([I-D.ietf-trill-esadi]), looked up in a directory([I-D.ietf-trill-directory-framework]), or learned from received traffic. If the edge RBridge has many attached endnodes, this table could become large. Also, if one of the MAC addresses in the table has moved to a different switch, it might be difficult for the edge RBridge to notice this quickly, and because the edge RBridge is tunneling 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 nodes that E is corresponding with. This eliminates the need for the attached RBridge R to know about those nodes (unless some non-smart endnode attached to R is also corresponding with those nodes), and it enables E to immediately discard an entry of (D, egress nickname), if E cannot talk to D. Then E can attempt to acquire a fresh entry for D by flooding to D, listening for ESADI,or consulting a directory.

The mechanism in this draft is that E issue a TRILL-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 that E owns, and whether E would like to receive ESADI frame. E learns from R’s Hello, whether R is capable of having a smart endnode neighbor, what R’s nickname is, and which trees R can use when R ingresses

multidestination frames. Although E transmits TRILL-Hellos, E does not transmit or receive LSPs.

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

2. TRILL-Hello Content

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

2.1. Edge RBridge's TRILL-Hello

For smart endnode operation, R's TRILL-Hello must contain the following information:

- RBridge's nickname. The nickname sub-TLV (Specified in section 2.3.2 in [I-D.ietf-isis-rfc6326bis])could be reused here, and TLV 242 (ISIS router capability) should be updated to be carried in TRILL-Hello frame.

- Tree roots that R can use when ingressing multidestination frames. The Tree Identifiers Sub-TLV(Specified in section 2.3.4 in [I-D.ietf-isis-rfc6326bis]) could be reused here.

- Smart endnode neighbor list. The TRILL Neighbor TLV(Specified in section 2.5 in [I-D.ietf-isis-rfc6326bis]) could be reused.

2.2. Smart Endnode's TRILL-Hello

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

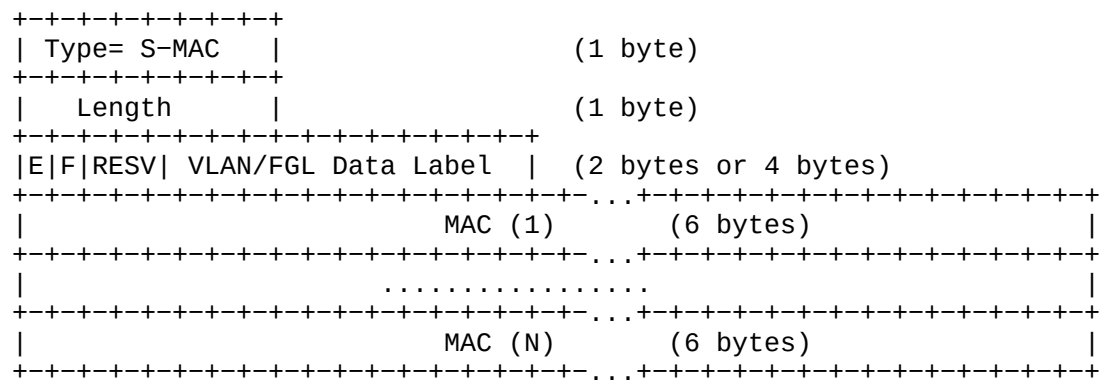

Figure 1 S-MAC TLV

- 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 could receive ESADI frame.
- F: one bit. If it sets to 1, which indicats that the endnode supports FGL data label, otherwise, the VLAN/FGL Data Label (([I-D.ietf-trill-fine-labeling])) 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 from the IS that is announcing this TLV.

3. Frame Processing

3.1. Frame Processing for Smart Endnode

Smart endnode E does not issue LSPs, nor does it receive LSPs or calculate topology. E does the following:

- o E maintains an endnode table of (MAC, nickname) of end nodes with which the smart endnode is communicating. If E is attached to multiple VLANs (or FGL), there would be a separate (MAC, nickname) table for each VLAN/FGL that E is attached to. 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([I-D.ietf-trill-esadi]).
  - * by querying a directory([I-D.ietf-trill-directory-framework]).
  - * by having some entries configured.
- o When E wishes to transmit to unicast destination D, if (D, nickname) is in E's endnode table, E encapsulates with ingress nickname=R, egress nickname as indicated in D's table entry. If D is unknown, D either queries a directory or encapsulates the packet as a multidestination frame, using one of the trees that R has specified in R's TRILL-Hello.
- o When E wishes to transmit to a multicast destination, E encapsulates the packet using one of the trees that R has specified.

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

3.2. Frame Processing for Edge RBridge

The attached RBridge R does the following:

- o If receiving an encapsulated unicast data frame from a port with a smart endnode, with R's nickname as ingress, R forwards the frame to the specified egress nickname, as with any encapsulated frame. However, R MAY filter the encapsulation frame based on the inner source MAC and VLAN (or FGL) as specified for the smart endnode. If the MAC (or VLAN/FGL) are not among the expected set of the smart endnode, the frame would be dropped by the edge RBridge.
- o If receiving an mulitdestination data TRILL frame from a port with smart endnode, RBridge R 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 R, R should decapsulates the frame and sends the native frame to these ports.

- o When R receives a mulicast 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 address.

4. Multi-homing

Now suppose E is attached to the TRILL campus in two places: to RBridges R1 and R2. There are two ways for this to work:

(1) E can choose either R1 or R2's nickname, when encapsulating a frame, whether the encapsulated frame is sent via R1 or R2. If E wants to do active-active load splitting, and uses R1's nickname when forwarding through R1, and R2's nickname when forwarding through R2, which will cause the flip-floping of the endnode table entry in the remote RBridges(or smart endnodes). This issues could be solved by setting 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, R1's nickname) and (E, R2's nickname) will be coexist as two entries for that MAC address.

(2) R1 and R2 might indicate, in their Hello, a virtual nickname that attached end nodes may use if they are multihomed to R1 and R2, separate from R1 and R2's nicknames (which they would also list in their 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 R1 or R2. If E loses connectivity to R2, then E would revert to using R1'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.

5. Security Considerations

For general TRILL Security Considerations, see([RFC6325]).

6. Acknowledgements

7. 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 TRILL-Hello frame.

8. Normative References

[I-D.ietf-isis-rfc6326bis]
Eastlake, D., Senevirathne, T., Ghanwani, A., Dutt, D., and A. Banerjee, "Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS", draft-ietf-isis-rfc6326bis-01 (work in progress), April 2013.

[I-D.ietf-trill-cmt]
Senevirathne, T., Pathangi, J., and J. Hudson, "Coordinated Multicast Trees (CMT) for TRILL", draft-ietf-trill-cmt-02 (work in progress), October 2013.

[I-D.ietf-trill-directory-framework]
Dunbar, L., Eastlake, D., Perlman, R., and I. Gashinsky, "TRILL (Transparent Interconnection of Lots of Links): Edge Directory Assistance Framework", draft-ietf-trill-directory-framework-07 (work in progress), August 2013.

[I-D.ietf-trill-esadi]
Zhai, H., Hu, F., Perlman, R., Eastlake, D., and O. Stokes, "TRILL (Transparent Interconnection of Lots of Links): ESADI (End Station Address Distribution Information) Protocol", draft-ietf-trill-esadi-03 (work in progress), July 2013.

[I-D.ietf-trill-fine-labeling]
Eastlake, D., Zhang, M., Agarwal, P., Perlman, R., and D. Dutt, "TRILL (Transparent Interconnection of Lots of Links): Fine-Grained Labeling", draft-ietf-trill-fine-labeling-07 (work in progress), May 2013.

[RFC1195] Callon, R., "Use of OSI IS-IS for routing in TCP/IP and dual environments", RFC 1195, December 1990.

[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.

Authors' Addresses

Radia Perlman
Intel Labs
2200 Mission College Blvd.
Santa Clara, CA 95054-1549
USA

Phone: +1-408-765-8080
Email: Radia@alum.mit.edu

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