INTERNET-DRAFT
Intended status: Proposed Standard

Linda Dunbar
Donald Eastlake
Huawei
Radia Perlman
EMC

Expires: June 27, 2017
December 28, 2016

Directory Assisted TRILL Encapsulation

<draft-ietf-trill-directory-assisted-encap-04.txt>

Abstract

This draft describes how data center networks can benefit from non-RBridge nodes performing TRILL encapsulation with assistance from a directory service.

Status of This Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Distribution of this document is unlimited. Comments should be sent to the authors or the TRILL working group mailing list: trill@ietf.org

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/1id-abstracts.html. The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

Table of Contents

1. Introduction..3
2. Conventions Used in This Document.....................4
3. Directory Assistance to Non-RBridge...................5
4. Source Nickname in Encapsulation by Non-RBridge Nodes...8

5. Benefits of Non-RBridge Performing TRILL Encapsulation..9
5.1. Avoid Nickname Exhaustion Issue.....................9
5.2. Reduce MAC Tables for Switches on Bridged LANs........9

6. Manageability Considerations.........................11
7. Security Considerations..............................11
8. IANA Considerations..................................12

Normative References....................................13
Informative References..................................13
Acknowledgments...13

Authors' Addresses......................................14

1. Introduction

This document describes how data center networks can benefit from non-RBridge nodes performing TRILL encapsulation with assistance from directory service and specifies a method for them to do so.

[RFC7067] and [Directory] describe the framework and methods for RBridge edge to get MAC&VLAN<->RBridgeEdge mapping from a directory service in data center environments instead of flooding unknown DAs across TRILL domain. If it has the needed directory information, any node, even a non-RBridge node, can perform the TRILL encapsulation. This draft is to describe the benefits and a scheme for non-RBridge nodes performing TRILL encapsulation.

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 [RFC2119].

AF: Appointed Forwarder RBridge port [RFC6439]

Bridge: IEEE 802.1Q compliant device. In this draft, Bridge is used interchangeably with Layer 2 switch.

DA: Destination Address

Host: Application running on a physical server or a virtual machine. A host usually has at least one IP address and at least one MAC address.

SA: Source Address

TRILL-EN: TRILL Encapsulating node. It is a node that only performs the TRILL encapsulation but doesn't participate in RBridge's IS-IS routing.

VM: Virtual Machines

3. Directory Assistance to Non-RBridge

With directory assistance [RFC7067] [Directory], a non-RBridge can be informed if a packet needs to be forwarded across the RBridge domain and the corresponding egress RBridge. Suppose the RBridge domain boundary starts at network switches (not virtual switches embedded on servers), a directory can assist Virtual Switches embedded on servers to encapsulate with a proper TRILL header by providing the nickname of the egress RBridge edge to which the destination is attached. The other information needed to encapsulate can be either learned by listening to TRILL Hellos, which will indicate the MAC address and nickname of appropriate edge RBridges, or by configuration.

If a destination is not shown as attached to one or more other RBridge edge nodes, based on the directory, the non-RBridge node can forward the data frames natively, i.e. not encapsulating with any TRILL header.

```
       \              +-------+         +------+ TRILL Domain/
        \           +/------+ |       +/-----+ |           /
         \          | Aggr11| + ----- |AggrN1| +           /
          \         +---+---+/        +------+/           /
           \         /     \            /      \         /
            \       /       \          /        \       /
             \   +---+    +---+      +---+     +---+   /
              \- |T11|... |T1x|      |T21| ..  |T2y|---
                 +---+    +---+      +---+     +---+
                   |        |          |         |
                 +-|-+    +-|-+      +-|-+     +-|-+
                 |   |... | V |      | V | ..  | V |<- vSwitch
                 +---+    +---+      +---+     +---+
                 |   |... | V |      | V | ..  | V |
                 +---+    +---+      +---+     +---+
                 |   |... | V |      | V | ..  | V |
                 +---+    +---+      +---+     +---+
```

Figure 1. TRILL domain in typical Data Center Network

When a TRILL encapsulated data packet reaches the ingress RBridge, the ingress RBridge simply forwards the pre-encapsulated packet to the RBridge that is specified by the egress nickname field of the TRILL header of the data frame. When the ingress RBridge receives a native Ethernet frame, it handles it as usual and may drop it if it has complete directory information indicating that the target is not attached to the TRILL campus.

In such an environment with complete directory information, the ingress RBridge doesn't flood or forward the received data frames when the DA in the Ethernet data frames is unknown.

When all nodes attached to an ingress RBridge can pre-encapsulate with a TRILL header for traffic across the TRILL domain, the ingress RBridge don't need to encapsulate any native Ethernet frames to the TRILL domain. The attached nodes can be connected to multiple edge RBridges by having multiple ports or by an bridged LAN. Under this environment, there is no need to designate AF ports and all RBridge edge ports connected to one bridged LAN can receive and forward pre-encapsulated traffic, which can greatly improve the overall network utilization.

The TRILL base protocol specification [RFC6325] Section 4.6.2 Bullet 8 specifies that an RBridge port can be configured to accept TRILL encapsulated frames from a neighbor that is not an RBridge.

When a TRILL frame arrives at an RBridge whose nickname matches with the destination nickname in the TRILL header of the frame, the processing is exactly same as normal, i.e. as specified in [RFC6325] the RBridge decapsulates the received TRILL frame and forwards the decapsulated frame to the target attached to its edge ports. When the DA of the decapsulated Ethernet frame is not in the egress RBridge's local MAC attachment tables, the egress RBridge floods the decapsulated frame to all attached links in the frame's VLAN, or drops the frame (if the egress RBridge is configured with that policy).

We call a node that only performs the TRILL encapsulation but doesn't participate in RBridge's IS-IS routing a TRILL Encapsulating node (TRILL-EN). The TRILL Encapsulating Node can get the MAC&VLAN<->RBridgeEdge mapping table pulled from directory servers [Directory].

Upon receiving a native Ethernet frame, the TRILL-EN checks the MAC&VLAN<->RBridgeEdge mapping table, and perform the corresponding TRILL encapsulation if the entry is found in the mapping table. If the destination address and VLAN of the received Ethernet frame doesn't exist in the mapping table and there is no positive reply from pulling requests to a directory, the Ethernet frame is dropped or forwarded in native form to an edge RBridge.

```
       +------------+--------+---------+---------+--+-------+---+
       |OuterEtherHd|TRILL HD| InnerDA | InnerSA |..|Payload|FCS|
       +------------+--------+---------+---------+--+-------+---+
               ^
               |             |<Inner Ether Header>  |
               |
               |
               |      +--------+  TRILL    +-------+
               |      |  R1   |-----------|  R2  |  Decapsulate
               |      +---+---+  domain   +-------+  TRILL header
               |          |                   |
               +----------|                   |
                          |                   |
                       +-----+             +-----+
     Non-RBridge node:|T12  |             | T22 |
     Encapsulate TRILL+-----+             +-----+
     Header for data
     Frames to traverse
     TRILL domain.
```

Figure 2. Data frames from TRILL-EN

4. Source Nickname in Encapsulation by Non-RBridge Nodes

The TRILL header includes a Source RBridge's Nickname (ingress) and Destination RBridge's Nickname (egress). When a TRILL header is added by TRILL-EN, the Ingress RBridge edge node's nickname is used in the source address field. The TRILL-EN learns this nickname by listening to the TRILL IS-IS Hellos from the Ingress RBridge. Those Hellos have that nickname in a field in the Special VLANs and Flags Sub-TLV [RFC7176] contained in the Hello.

5. Benefits of Non-RBridge Performing TRILL Encapsulation

5.1. Avoid Nickname Exhaustion Issue

For a large Data Center with hundreds of thousands of virtualized servers, setting the TRILL boundary at the servers' virtual switches will create a TRILL domain with hundreds of thousands of RBridge nodes, which has issues of TRILL Nicknames exhaustion and challenges to IS-IS. On the other hand, setting TRILL boundary at aggregation switches that have many virtualized servers attached can limit the number of RBridge nodes in a TRILL domain, but introduce the issues of very large MAC&VLAN<->RBridgeEdge mapping table to be maintained by RBridge edge nodes and the necessity of enforcing AF ports.

Allowing Non-RBridge nodes to pre-encapsulate data frames with TRILL header makes it possible to have a TRILL domain with a reasonable number of RBridge nodes in a large data center. All the TRILL-ENs attached to one RBridge are represented by one TRILL nickname, which can avoid the Nickname exhaustion problem.

5.2. Reduce MAC Tables for Switches on Bridged LANs

When hosts in a VLAN (or subnet) span across multiple RBridge edge nodes and each RBridge edge has multiple VLANs enabled, the switches on the bridged LANs attached to the RBridge edge are exposed to all MAC addresses among all the VLANs enabled.

For example, for an Access switch with 40 physical servers attached, where each server has 100 VMs, there are 4000 hosts under the Access Switch. If indeed hosts/VMs can be moved anywhere, the worst case for the Access Switch is when all those 4000 VMs belong to different VLANs, i.e. the access switch has 4000 VLANs enabled. If each VLAN has 200 hosts, this access switch's MAC table potentially has 200*4000 = 800,000 entries.

If the virtual switches on servers pre-encapsulate the data frames destined for hosts attached to other RBridge Edge nodes, the outer MAC DA of those TRILL encapsulated data frames will be the MAC address of the local RBridge edge, i.e. the ingress RBridge. Therefore, the switches on the local bridged LAN don't need to keep the MAC entries for remote hosts attached to other edge RBridges.

But the traffic from nodes attached to other RBridges is decapsulated and has the true source and destination MACs. One simple way to prevent local bridges from learning remote hosts' MACs and adding to their MAC tables, if that is a problem, is to disable this data plane learning on local bridges. The local bridges can be pre-configured with MAC addresses of local hosts with the assistance of a directory. The local bridges can always send frames with unknown Destination to the ingress RBridge. In an environment where a large number of VMs are instantiated in one server, the number of remote MAC addresses could be very large. If it is not feasible to disable learning and pre- configure MAC tables for local bridges, one effective method to minimize local bridges' MAC table size is to use the server's MAC address to hide MAC addresses of the attached VMs. I.e. the server acting as an edge node uses its own MAC address in the Source Address field of the packets originated from a host (or VM) embedded. When the Ethernet frame arrives at the target edge node (the server), the target edge node can send the packet to the corresponding destination host based on the packet's IP address. Very often, the target edge node communicates with the embedded VMs via a layer 2 virtual switch. In this case, the target edge node can construct the proper Ethernet header with the assistance of the directory. The information from the directory includes the proper host IP to MAC mapping information.

6. Manageability Considerations

It requires directory assistance [Directory] to make it possible for a non-TRILL node to pre-encapsulate packets destined towards remote RBridges.

7. Security Considerations

TBD

For general TRILL security considerations, see [RFC6325].

8. IANA Considerations

This document requires no IANA actions. RFC Edtior: please remove this section before publication.

Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <http://www.rfc-editor.org/info/rfc2119>.

[RFC6325] Perlman, R., Eastlake 3rd, D., Dutt, D., Gai, S., and A. Ghanwani, "Routing Bridges (RBridges): Base Protocol Specification", RFC 6325, DOI 10.17487/RFC6325, 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, DOI 10.17487/RFC6439, November 2011, <http://www.rfc-editor.org/info/rfc6439>.

[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, DOI 10.17487/RFC7176, May 2014, <http://www.rfc-editor.org/info/rfc7176>.

[Directory] D. Eastlake, L. Dunbar, R. Perlman, Y. Li, "TRILL: Edge Directory Assist Mechanisms", draft-ietf-trill-directory-assist-mechanisms, work in progress.

Informative References

[RFC7067] Dunbar, et, al "Directory Assistance Problem and High-Level Design Proposal", RFC7067, November 2013.

Acknowledgments

The followiing are thanked for their contributions:

Igor Gashinsky

The document was prepared in raw nroff. All macros used were defined within the source file.

Authors' Addresses

Linda Dunbar
Huawei Technologies
5340 Legacy Drive, Suite 175
Plano, TX 75024, USA

Phone: +1-469-277-5840
Email: linda.dunbar@huawei.com

Donald Eastlake
Huawei Technologies
155 Beaver Street
Milford, MA 01757 USA

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

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

Email: Radia@alum.mit.edu

Copyright, Disclaimer, and Additional IPR Provisions

Copyright (c) 2016 IETF Trust and the persons identified as the document authors. All rights reserved.

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. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License. The definitive version of an IETF Document is that published by, or under the auspices of, the IETF. Versions of IETF Documents that are published by third parties, including those that are translated into other languages, should not be considered to be definitive versions of IETF Documents. The definitive version of these Legal Provisions is that published by, or under the auspices of, the IETF. Versions of these Legal Provisions that are published by third parties, including those that are translated into other languages, should not be considered to be definitive versions of these Legal Provisions. For the avoidance of doubt, each Contributor to the IETF Standards Process licenses each Contribution that he or she makes as part of the IETF Standards Process to the IETF Trust pursuant to the provisions of RFC 5378. No language to the contrary, or terms, conditions or rights that differ from or are inconsistent with the rights and licenses granted under RFC 5378, shall have any effect and shall be null and void, whether published or posted by such Contributor, or included with or in such Contribution.