Internet Engineering Task Force (IETF)
Request for Comments: 8380
Category: Standards Track
ISSN: 2070-1721

L. Dunbar
D. Eastlake 3rd
Huawei
R. Perlman
Dell/EMC

May 2018

Directory-Assisted
Transparent Interconnection of Lots of Links (TRILL) Encapsulation

Abstract

This document describes how data center networks can benefit from non-RBridge nodes performing TRILL (Transparent Interconnection of Lots of Links) encapsulation with assistance from a directory service.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

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

[RFC7067] and [RFC8171] describe the framework and methods for edge RBridges to get (MAC and VLAN) <-> Edge RBridge mapping from a directory service instead of flooding unknown destination MAC addresses across a TRILL domain. If it has the needed directory information, any node, even a non-RBridge node, can perform the TRILL data packet encapsulation. This document describes the benefits of 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", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

AF:	Appointed Forwarder RBridge port [RFC8139].
Bridge:	A device compliant with IEEE 802.1Q. In this document, Bridge is used interchangeably with Layer 2 switch.
DA:	Destination Address.
ES-IS:	End System to Intermediate System [RFC8171].

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

IS-IS: Intermediate System to Intermediate System [RFC7176].

SA: Source Address.

TRILL-EN: TRILL Encapsulating Node. A node that performs the TRILL encapsulation but doesn’t participate in an RBridge’s IS-IS routing.

VM: Virtual Machine.

3. Directory Assistance to Non-RBridge

With directory assistance [RFC7067] [RFC8171], a non-RBridge node can learn if a data packet needs to be forwarded across the RBridge domain and, if so, the corresponding egress RBridge.

Suppose the RBridge domain boundary starts at network switches (not virtual switches embedded on servers). (See Figure 1 for a high-level diagram of a typical data center network.) 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 learned either by listening to TRILL ES-IS and/or IS-IS Hellos [RFC7176] [RFC8171], which will indicate the MAC address and nickname of appropriate local edge RBridges, or by configuration.

If it is not known whether a destination is attached to one or more edge RBridges, based on the directory, the non-RBridge node can forward the data frames natively, i.e., not encapsulating with any TRILL header. Or, if the directory is known to be complete, the non-RBridge node can discard such data frames.

```
          \           +-----------+       +-----------+            /
           \        +/----------+ |     +/----------+ |  TRILL    /
            \       |Aggregation| |     |Aggregation| | Domain   /
             \      |     11    | + --- |     N1    | +         /
              \     +-----------+/      +-----------+/         /
               \         /     \            /      \          /
                \       /       \          /        \        /
         Top-    \   +---+    +---+      +---+     +---+    /
         of- -->  \- |T11|... |T1x|      |T21| ..  |T2y|---/
         Rack        +---+    +---+      +---+     +---+
         Switches      |        |          |         |
                     +-|-+    +-|-+      +-|-+     +-|-+
                     |   |... | V |      | V | ..  | V | <- vSwitch
                     +---+    +---+      +---+     +---+
                     |   |... | V |      | V | ..  | V |
                     +---+    +---+      +---+     +---+
                     |   |... | V |      | V | ..  | V |
                     +---+    +---+      +---+     +---+
```

Figure 1: TRILL Domain in a Typical Data Center Network

When a TRILL-encapsulated data packet reaches the ingress RBridge, that RBridge simply performs the usual TRILL processing and forwards the pre-encapsulated packet to the RBridge that is specified by the egress nickname field of the TRILL header. When an ingress RBridge receives a native Ethernet frame in an environment with complete directory information, the ingress RBridge doesn’t flood or forward the received data frames when the destination MAC address in the Ethernet data frames is unknown.

When all end nodes attached to an ingress RBridge pre-encapsulate with a TRILL header for traffic across the TRILL domain, the ingress RBridge doesn’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 through a bridged LAN. All RBridge edge ports connected to one bridged LAN can receive and forward pre-encapsulated traffic; this can greatly improve the overall network utilization. However, it is still necessary to, for example, designate AF ports to be sure that multi-destination packets from the TRILL campus are only egressed through one RBridge.

Item 8 of Section 4.6.2 of the TRILL base protocol specification [RFC6325] 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 the destination nickname in the TRILL header of the frame, the processing is exactly as normal: 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 destination MAC address 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, as specified herein, only performs TRILL encapsulation, but doesn’t participate in RBridge’s IS-IS routing, a TRILL Encapsulating Node (TRILL-EN). The TRILL Encapsulating Node can pull (MAC and VLAN) <-> Edge RBridge mapping from directory servers [RFC8171]. In order to do this, a TRILL-EN MUST support TRILL ES-IS [RFC8171].

Upon receiving or locally generating a native Ethernet frame, the TRILL-EN checks the (MAC and VLAN) <-> Edge RBridge mapping and performs the corresponding TRILL encapsulation if the mapping entry is found as shown in Figure 2. If the destination MAC address and VLAN of the received Ethernet frame doesn’t exist in the mapping table and there is no positive reply from pull requests to a directory, the Ethernet frame is dropped or is forwarded in native form to an edge RBridge, depending on the TRILL-EN configuration.

```
       +------------+--------+---------+---------+--+-------+---+
       |OuterEtherHd|TRILL HD| InnerDA | InnerSA |..|Payload|FCS|
       +------------+--------+---------+---------+--+-------+---+
               |
               |             |<Inner Ether Header>  |
               |
               |
               |       +-------+  TRILL    +------+
               |       |  RB1  |---------->|  RB2 |  Decapsulate
               |       +-------+  domain   +------+  TRILL header
               v           ^                   |
               +---------->|                   |
                           |                   V
                        +--------+         +--------+
      Non-RBridge node: |TRILL-EN|         |TRILL-EN|
      Encapsulate TRILL |   1    |         |   2    |
      Header for data   +--------+         +--------+
      Frames to traverse TRILL domain.
```

Figure 2: Data Frames from a 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 to a data packet by a TRILL-EN, the ingress RBridge nickname field in the TRILL header is set to a nickname of the AF for the data packet’s VLAN. The TRILL-EN determines the AF by snooping on IS-IS Hellos from the edge RBridges on the link with the TRILL-EN in the same way that the RBridges on the link determine the AF [RFC8139]. A TRILL-EN is free to send the encapsulated data frame to any of the edge RBridges on its link.

5. Benefits of a Non-RBridge Performing TRILL Encapsulation

This section summarizes the benefits of having a non-RBridge node perform 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; this could lead to TRILL nickname exhaustion and challenges to IS-IS. On the other hand, setting the TRILL boundary at aggregation switches that have many virtualized servers attached can limit the number of RBridge nodes in a TRILL domain, but introduces the issue of having very large (MAC and VLAN) <-> Edge RBridge mapping tables that need to be maintained by edge RBridges.

Allowing non-RBridge nodes to pre-encapsulate data frames with TRILL headers 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 can be 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 edge RBridges and each edge RBridge has multiple VLANs enabled, the switches on the bridged LANs attached to the edge RBridges 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 remote edge RBridges, the outer MAC destination address of those TRILL-encapsulated data frames will be the MAC address of a local RBridge edge, i.e., the ingress RBridge. 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 TRILL 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 would be a problem, is to disable this data-plane learning on local bridges. With the assistance of a directory, the local bridges can be pre-configured with MAC addresses of local hosts. The local bridges can always send frames with unknown destination MAC addresses 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 and all important traffic is IP, 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. That is, the server acting as an edge node uses its own MAC address in the source MAC address field of the packets originated from a host (or VM). When the Ethernet frame arrives at the target edge node (the egress), 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 mapping of host IP to MAC.

6. Manageability Considerations

Directory assistance [RFC8171] is required to make it possible for a non-TRILL node to pre-encapsulate packets destined towards remote RBridges. TRILL-ENs have the same configuration options as any pull directory client. See Section 4 of [RFC8171].

7. Security Considerations

If the TRILL-ENs are not trusted, they can forge arbitrary ingress and egress nicknames in the TRILL Headers of the TRILL Data packets they construct. With data-plane learning, decapsulating a TRILL Data packet at an egress RBridge associates the inner source MAC address with the ingress nickname in the TRILL Header (assuming that MAC address is unicast). Thus, if those ingress nicknames are forged, incorrect learning will occur and future traffic destined for the inner source MAC will be sent to the wrong RBridge for egress. Because of this, an RBridge port should not be configured to accept encapsulated TRILL data frames on a link were it does not have an RBridge adjacency unless the end stations on that link are trusted.

As with any end station, TRILL-ENs can forge the outer MAC addresses of packets they send. (See Section 6 of [RFC6325].) Because they pre-encapsulate, they can also forge inner MAC addresses.

The pre-encapsulation performed by TRILL-ENs also means they can send data in any VLAN; this means they must be trusted in order to enforce a security policy based on VLANs. (See Section 6.1 of [RFC6325].)

Use of directory-assisted encapsulation by TRILL-ENs essentially involves those TRILL-ENs spoofing edge RBridges to which they are connected; this is another reason that TRILL-ENs should be trusted nodes. Such spoofing cannot cause persistently looping traffic because TRILL has a hop count in the TRILL header [RFC6325] so that, should there be a loop, a TRILL packet caught in that loop (i.e., an encapsulated frame) will be discarded. (In the potentially more dangerous case of multidestination packets (as compared with known unicast) where copies could multiply due to forks in the distribution tree, a Reverse Path Forwarding Check is also used [RFC6325] to discard packets that appear to be on the wrong link or when there is disagreement about the distribution tree.)

The mechanism described in this document requires a TRILL-EN to be aware of the MAC address(es) of the TRILL edge RBridge(s) to which the TRILL-EN is attached and the egress RBridge nickname from which the destination of the packets is reachable. With that information, TRILL-ENs can learn a substantial amount about the topology of the TRILL domain. Therefore, there could be a potential security risk when the TRILL-ENs are not trusted or are compromised.

If the path between the directory and a TRILL-EN is attacked, false mappings can be sent to the TRILL-EN causing packets from the TRILL-EN to be sent to wrong destinations, possibly violating security policy as to which end stations should receive what data. Therefore, a combination of authentication and encryption is RECOMMENDED between the directory and TRILL-EN. The entities involved will need to properly authenticate with each other, provide session encryption, maintain security patch levels, and configure their systems to allow minimal access and running processes to protect sensitive information.

For added security against the compromise of data due to its misdelivery for any reason, including the above, end-to-end encryption and authentication should be considered; that is, encryption and authentication from source end station to destination end station.

For Pull Directory and TRILL ES-IS security considerations, see [RFC8171].

8. IANA Considerations

This document has no IANA actions.

9. References

9.1. 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, <https://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, <https://www.rfc-editor.org/info/rfc6325>.

[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, <https://www.rfc-editor.org/info/rfc7176>.

[RFC8139] Eastlake 3rd, D., Li, Y., Umair, M., Banerjee, A., and F. Hu, "Transparent Interconnection of Lots of Links (TRILL): Appointed Forwarders", RFC 8139, DOI 10.17487/RFC8139, June 2017, <https://www.rfc-editor.org/info/rfc8139>.

[RFC8171] Eastlake 3rd, D., Dunbar, L., Perlman, R., and Y. Li, "Transparent Interconnection of Lots of Links (TRILL): Edge Directory Assistance Mechanisms", RFC 8171, DOI 10.17487/RFC8171, June 2017, <https://www.rfc-editor.org/info/rfc8171>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

9.2. Informative References

[RFC7067] Dunbar, L., Eastlake 3rd, D., Perlman, R., and I. Gashinsky, "Directory Assistance Problem and High-Level Design Proposal", RFC 7067, DOI 10.17487/RFC7067, November 2013, <https://www.rfc-editor.org/info/rfc7067>.

Acknowledgments

The following are thanked for their contributions:

Igor Gashinsky
Ben Nevin-Jenkins

Authors’ Addresses

Linda Dunbar
Huawei Technologies
5340 Legacy Drive, Suite 175
Plano, TX 75024
United States of America

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

Donald Eastlake 3rd
Huawei Technologies
155 Beaver Street
Milford, MA 01757
United States of America

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

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

Email: Radia@alum.mit.edu