Internet Engineering Task Force (IETF)
Request for Comments: 7956
Category: Standards Track
ISSN: 2070-1721

W. Hao
Y. Li
Huawei
A. Qu
MediaTec
M. Durrani
Equinix Inc.
P. Sivamurugan
IP Infusion

September 2016

Transparent Interconnection of Lots of Links (TRILL) Distributed Layer 3 Gateway

Abstract

The base TRILL (Transparent Interconnection of Lots of Links) protocol provides optimal pair-wise data frame forwarding for Layer 2 intra-subnet traffic but not for Layer 3 inter-subnet traffic. A centralized gateway solution is typically used for Layer 3 inter-subnet traffic forwarding but has the following issues:

1. Sub-optimum forwarding paths for inter-subnet traffic.

2. A centralized gateway that may need to support a very large number of gateway interfaces in a Data Center, one per tenant per Data Label used by that tenant, to provide interconnect functionality for all the Layer 2 Virtual Networks in a TRILL campus.

3. A traffic bottleneck at the gateway.

This document specifies an optional TRILL distributed gateway solution that resolves these centralized gateway issues.

Status of This Memo

This is an Internet Standards Track document.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Further information on Internet Standards is available in Section 2 of RFC 7841.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc7956.

Copyright Notice

Table of Contents

1. Introduction

The TRILL (Transparent Interconnection of Lots of Links) protocol [RFC6325] [RFC7780] provides a solution for least-cost transparent routing in multi-hop networks with arbitrary topologies and link technologies, using IS-IS [IS-IS] [RFC7176] link-state routing and a hop count. TRILL switches are sometimes called RBridges (Routing Bridges).

The base TRILL protocol provides optimal unicast forwarding for Layer 2 intra-subnet traffic but not for Layer 3 inter-subnet traffic, where "subnet" means a different IP address prefix and, typically, a different Data Label (VLAN or FGL (Fine-Grained Label)). This document specifies a TRILL-based distributed Layer 3 gateway solution that provides optimal unicast forwarding for Layer 3 inter-subnet traffic. With distributed gateway support, an edge RBridge provides routing based on the Layer 2 identity (address and Virtual Network (VN, i.e., Data Label)) among End Stations (ESs) that belong to the same subnet and also provides routing based on the Layer 3 identity among ESs that belong to different subnets of the same routing domain. An edge RBridge supporting this feature needs to provide routing instances and Layer 3 gateway interfaces for locally connected ESs. Such routing instances provide IP address isolation between tenants. In the TRILL distributed Layer 3 gateway solution, inter-subnet traffic can be fully spread over edge RBridges, so there is no single bottleneck.

1.1. Document Organization

This document is organized as follows: Section 3 gives a simplified example and also a more detailed problem statement. Section 4 gives the Layer 3 traffic forwarding model. Section 5 provides a distributed gateway solution overview. Section 6 gives a detailed distributed gateway solution example. Section 7 describes the TRILL protocol extensions needed to support this distributed gateway solution.

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 RFC 2119 [RFC2119].

The terms and acronyms in [RFC6325] are used, with the following additions:

AGG: Aggregation switch.

ARP: Address Resolution Protocol [RFC826].

Campus: The name for a network using the TRILL protocol in the same sense that a "bridged LAN" is the name for a network using bridging. In TRILL, the word "campus" has no academic implication.

COR: Core switch.

Data Label: VLAN or FGL [RFC7172].

DC: Data Center.

Edge RBridge: An RBridge that connects to one or more ESs without any intervening RBridges.

ES: End Station. A Virtual Machine or physical server, whose address is either the destination or source of a data frame.

FGL: Fine-Grained Label [RFC7172].

Gateway interface: A Layer 3 virtual interface that terminates Layer 2 forwarding and forwards IP traffic to the destination using IP forwarding rules. Incoming traffic from a physical port on a gateway will be distributed to its virtual gateway interface based on the Data Label (VLAN or FGL).

Inner.MacDA: The inner MAC destination address in a TRILL Data packet [RFC6325].

Inner.MacSA: The inner MAC source address in a TRILL Data packet [RFC6325].

Inner.VLAN: The inner VLAN tag in a TRILL Data packet payload [RFC6325].

L2: Layer 2.

L3: IP Layer 3.

LSP: Link State PDU

ND: IPv6’s Neighbor Discovery [RFC4861].

ToR: Top of Rack.

VN: Virtual Network. In a TRILL campus, a unique 12-bit VLAN ID or a 24-bit FGL [RFC7172] identifies each VN.

VRF: Virtual Routing and Forwarding. In IP-based computer networks, VRF technology supports multiple instances of routing tables existing within the same router at the same time.

3. Simplified Example and Problem Statement

There is normally a Data Label (VLAN or FGL) associated with each IP subnet. For traffic within a subnet -- that is, IP traffic to another ES in the same Data Label attached to the TRILL campus -- the ES just ARPs for the MAC address of the destination ES’s IP. It then uses this MAC address for traffic to that destination. TRILL routes the ingressed TRILL Data packets to the destination’s edge RBridge based on the egress nickname for that destination MAC address and Data Label. This is the regular TRILL base protocol [RFC6325] process.

If two ESs of the same tenant are on different subnets and need to communicate with each other, their packets are typically forwarded to an IP L3 gateway that performs L3 routing and, if necessary, changes the Data Label. Either a centralized L3 gateway solution or the distributed L3 gateway solution specified in this document can be used for inter-subnet traffic forwarding.

Section 3.1 gives a simplified example in a TRILL campus with and without a distributed L3 gateway using VLAN Data Labels. Section 3.2 gives a detailed description of the issues related to using a centralized gateway (i.e., without a distributed L3 gateway). The remainder of this document, particularly Section 5, describes the distributed gateway solution in detail.

3.1. Simplified Example

Figure 1 depicts a TRILL DC network, where ToR switches are edge RBridges and the AGGs and CORs are non-edge RBridges.

```
                     -------                --------
                     | COR1|                | COR2 |
                     -------                --------
                        |                      |
                     -------                -------
                     |AGG1 |                |AGG2 |
                     -------                -------
                        |                      |
          -----------------------------------------------------
          |  -------------|------------------|----------------|
          |  |            |  |               |  |          |  |
        -------          -------           -------        -------
        | RB1 |          | RB2 |           | RB3 |        | RB4 |
        |ToR1 |          |ToR2 |           |ToR3 |        |ToR4 |
        -------          -------           -------        -------
         |    |           |    |            |    |         |    |
      -----  -----     -----  -----      -----  -----   -----  -----
      |ES1|  |ES2|     |ES3|  |ES4|      |ES5|  |ES6|   |ES7|  |ES8|
      -----  -----     -----  -----      -----  -----   -----  -----
```

Figure 1: A Typical TRILL DC Network

ES1 through ES8 belong to one tenant network, and the tenant has four subnets with each subnet corresponding to one VLAN (which indicates one individual L2 VN). Each ES's IP address, VLAN, and subnet are listed below:

ES	IP Address	Subnet	VLAN
ES1	192.0.2.2	192.0.2.0/24	10
ES2	198.51.100.2	198.51.100.0/24	11
ES3	192.0.2.3	192.0.2.0/24	10
ES4	198.51.100.3	198.51.100.0/24	11
ES5	203.0.113.2	203.0.113.0/25	12
ES6	203.0.113.130	203.0.113.128/25	13
ES7	203.0.113.3	203.0.113.0/25	12
ES8	203.0.113.131	203.0.113.128/25	13

Assume that a centralized gateway solution is used with both COR1 and COR2 acting as centralized gateways for redundancy in Figure 1. COR1 and COR2 each have four gateway interfaces for the four subnets in the tenant. In the centralized L3 gateway solution, all traffic within the tenant between different VLANs must go through the centralized L3 gateway device of COR1 or COR2, even if the traffic is between two ESs connected to the same edge RBridge, because only the L3 gateway can change the VLAN labeling of the traffic.

This is generally sub-optimal because the two ESs may be connected to the same ToR where L3 switching could have been performed locally. For example, in Figure 1 above, the unicast IP traffic between ES1 and ES2 has to go through a centralized gateway of COR1 or COR2. It can't be locally routed between them on ToR1. However, if an edge RBridge has the distributed gateway capabilities specified in this document, then it can still perform optimum L2 forwarding for intra-subnet traffic and, in addition, optimum L3 forwarding for inter-subnet traffic, thus delivering optimum forwarding for unicast packets in all important cases.

With a distributed L3 gateway, each edge RBridge acts as a default L3 gateway for local connecting ESs and has IP router capabilities to direct IP communications to other edge RBridges. Each edge RBridge only needs gateway interfaces for local connecting ESs, i.e., RB1 and RB2 need gateway interfaces only for VLAN 10 and VLAN 11 while RB3 and RB4 need gateway interfaces only for VLAN 12 and VLAN 13. No device needs to maintain gateway interfaces for all VLANs in the entire network. This will enhance scalability in terms of the number of tenants and subnets per tenant.

When each ES ARPs for its L3 gateway, that is, its IP router, the edge RBridge to which it is connected will respond with that RBridge's "gateway MAC". When the ES later sends IP traffic to the L3 gateway, which it does if the destination IP is outside of its subnet, the edge RBridge intercepts the IP packet because the destination MAC is its gateway MAC. That RBridge routes the IP packet using the routing instance associated with that tenant, handling it in one of three ways:

(1) ES1 communicates with ES2. The destination IP is connected to the same edge RBridge; the RBridge of ToR1 can simply transmit the IP packet out the right edge port in the destination VLAN.

(2) If the destination IP is located in an outside network, the edge RBridge encapsulates it as a TRILL Data packet and sends it to the actual TRILL campus edge RBridge connecting to an external IP router.

(3) ES1 communicates with ES4. The destination ES is connected to a different edge RBridge; the ingress RBridge ToR1 uses TRILL encapsulation to route the IP packet to the correct egress RBridge ToR2, using the egress RBridge's gateway MAC and an Inner.VLAN identifying the tenant. Finally, the egress RBridge terminates the TRILL encapsulation and routes the IP packet to the destination ES based on the routing instance for that tenant.

3.2. Problem Statement Summary

With FGL [RFC7172], in theory, up to 16 million L2 VNs can be supported in a TRILL campus. To support inter-subnet traffic, a very large number of L3 gateway interfaces could be needed on a centralized gateway, if each VN corresponds to a subnet and there are many tenants with many subnets per tenant. It is a big burden for the centralized gateway to support so many interfaces. In addition, all inter-subnet traffic will go through a centralized gateway that may become a traffic bottleneck.

The centralized gateway has the following issues:

1. Sub-optimum forwarding paths for inter-subnet traffic can occur due to the requirements to perform IP routing and possibly change Data Labels at a centralized gateway.

2. The centralized gateway may need to support a very large number of gateway interfaces -- in a DC, one per tenant per Data Label used by that tenant -- to provide interconnect functionality for all the L2 VNs in the TRILL campus.

3. There may be a traffic bottleneck at the centralized gateway.

A distributed gateway on edge RBridges addresses these issues. Through the distributed L3 gateway solution, the inter-subnet traffic is fully dispersed and is transmitted along optimal pair-wise forwarding paths, improving network efficiency.

4. Layer 3 Traffic Forwarding Model

In a DC network, each tenant has one or more L2 VNs, and, in normal cases, each tenant corresponds to one routing domain. Normally, each L2 VN uses a different Data Label and corresponds to one or more IP subnets.

Each L2 VN in a TRILL campus is identified by a unique 12-bit VLAN ID or 24-bit FGL [RFC7172]. Different routing domains may have overlapping address space but need distinct and separate routes. The ESs that belong to the same subnet communicate through L2 forwarding; ESs of the same tenant that belong to different subnets communicate through L3 routing.

Figure 2 depicts the model where there are n VRFs corresponding to n tenants, with each tenant having up to m segments/subnets (VNs).

```
              +---------------------------------------------+
              |                                             |
              |      +-----------+         +-----------+    |
              |      | Tenant n  |---------|  VRF n    |    |
              |   +------------+ |     +------------+  |    |
              |   |  +-----+   | |     |            |  |    |
              |   |  | VN1 |   | |     |            |  |    |
              |   |  +-----+   | |     |    VRF 1   |  |    |
              |   |     ..     +-------+            |  |    |
              |   |  +-----+   | |     |            |  |    |
              |   |  | VNm |   | |     |            |  |    |
              |   |  +-----+   | |     |            |  |    |
              |   |  Tenant1   |-+     |            |  |    |
              |   +------------+       |            |  |    |
              |   +------------+       +------------+       |
              |                                             |
              |               Edge RBridge                  |
              +---------------------------------------------+
```

Figure 2: Edge RBridge Model as Distributed Gateway

5. Distributed Gateway Solution Details

With the TRILL distributed gateway solution, an edge RBridge continues to perform routing based on the L2 MAC address for the ESs that are on the same subnet but performs IP routing for the ESs that are on the different subnets of the same tenant.

As the IP address space in different routing domains can overlap, VRF instances need to be created on each edge RBridge to isolate the IP forwarding process for different routing domains present on the edge RBridge. A Tenant ID unique across the TRILL campus identifies each routing domain. The network operator MUST configure the Tenant IDs on each edge RBridge for each routing domain consistently so that the same ID always refers to the same tenant. Otherwise, data might be delivered to the wrong tenant. If a routing domain spreads over multiple edge RBridges, routing information for the routing domain is synchronized among these edge RBridges through the link-state database to ensure reachability to all ESs in that routing domain. The routing information is, in effect, labeled with the Tenant ID to differentiate the routing domains.

From the data-plane perspective, all edge RBridges are connected to each other via one or more TRILL hops; however, they are always just a single IP hop away. When an ingress RBridge receives inter-subnet

IP traffic from a local ES whose destination MAC is the edge RBridge’s gateway MAC, that RBridge will perform Ethernet header termination. The tenant involved is determined by the VLAN of the traffic and the port on which it arrives. The edge RBridge looks up in its IP routing table for that tenant how to route the traffic to the IP next hop. If the destination ES is connected to a remote edge RBridge, the remote RBridge will be the IP next hop for traffic forwarding. For such inter-subnet traffic, the ingress RBridge will rewrite the original Ethernet header with the ingress RBridge’s gateway MAC address as the Inner.MacSA and the egress RBridge’s gateway MAC address as the Inner.MacDA and then perform TRILL encapsulation to the remote RBridge’s nickname, setting the inner Data Label to indicate the tenant involved. TRILL then routes it to the remote edge RBridge through the TRILL campus.

When that remote edge RBridge receives the traffic, it will decapsulate the TRILL Data packet and see that the inner destination MAC is its gateway MAC. It then terminates the inner Ethernet encapsulation and looks up the destination IP in the RBridge’s IP forwarding table for the tenant indicated by the inner Data Label, to route it to the destination ES.

Through this method, TRILL with distributed gateways provides optimum pair-wise data routing for inter-subnet traffic.

5.1. Local Routing Information

An ES can be locally connected to an edge RBridge through an L2 network (such as a point-to-point Ethernet link or a bridged LAN) or externally connected through an L3 IP network.

If the ES is connected to an edge RBridge through an L2 network, then the edge RBridge acts as an L3 gateway for the ES. A gateway interface is established on the edge RBridge for the connecting ES. Because the ESs in a subnet may be spread over multiple edge RBridges, each such edge RBridge that establishes its gateway interface for the subnet SHOULD share the same gateway MAC and gateway IP address configuration. Sharing the configuration and insuring configuration consistency can be done by local configuration and Network Configuration Protocol (NETCONF) / YANG models.

With a distributed gateway, the edge RBridge to which an ES is connected appears to be the local IP router on its link. As in any IP network, before the ES starts to send inter-subnet traffic, it acquires its gateway’s MAC through the ARP/ND process. Local connecting edge RBridges that support this distributed gateway feature always respond with the gateway MAC address when receiving ARP/ND requests for the gateway IP. Through the ARP/ND process, the edge RBridge can learn the IP and MAC correspondence of a local ES connected to the edge RBridge by L2 and then generate local IP routing entries for that ES in the corresponding routing domain.

To TRILL, an IP router connected to an edge RBridge looks like an ES. If a router/ES is located in an external IP network, it normally provides access to one or more IP prefixes. The router/ES SHOULD run an IP routing protocol with the connecting TRILL edge RBridge. The edge RBridge will learn the IP prefixes behind the router/ES through that IP routing protocol, and the RBridge will then generate local IP routing entries in the corresponding routing domain. If such a routing protocol is not run with the edge RBridge, then only the IP prefixes behind the router/ES that are explicitly configured on the edge RBridge will be accessible.

5.2. Local Routing Information Synchronization

When a routing instance is created on an edge RBridge, the Tenant ID, tenant Data Label (VLAN or FGL), and tenant gateway MAC that correspond to that instance are configured and MUST be globally advertised (see Section 7.1). The Tenant ID uniquely identifies that tenant throughout the campus. The tenant Data Label identifies that tenant at the edge RBridge. The tenant gateway MAC MAY identify that tenant, all tenants, or some subset of tenants at the edge RBridge.

When an ingress RBridge performs inter-subnet traffic TRILL encapsulation, the ingress RBridge uses the Data Label advertised by the egress RBridge as the inner VLAN or FGL and uses the tenant gateway MAC advertised by the egress RBridge as the Inner.MacDA. The egress RBridge relies on this tenant Data Label to find the local VRF instance for the IP forwarding process when receiving inter-subnet traffic from the TRILL campus. (The role of the tenant Data Label is akin to an MPLS VPN Label in an MPLS IP / MPLS VPN network.) Tenant Data Labels are independently allocated on each edge RBridge for each routing domain. An edge RBridge can use an access Data Label from a routing domain to act as the inter-subnet Data Label, or the edge RBridge can use a Data Label different from any access Data Labels to be a tenant Data Label. It is implementation dependent, and there is no restriction on this assignment of Data Labels.

The tenant gateway MAC differentiates inter-subnet L3 traffic from intra-subnet L2 traffic on the egress RBridge. Each tenant on an RBridge can use a different gateway MAC or the same tenant gateway MAC for inter-subnet traffic purposes. This is also implementation dependent, and there is no restriction on it.

When a local IP prefix is learned in a routing instance on an edge RBridge, the edge RBridge should advertise the IP prefix information for the routing instance so that other edge RBridges will generate IP routing entries. If the ESs in a VN are spread over multiple RBridges, these RBridges MUST advertise each local connecting ES's IP address in the VN to other RBridges. If the ESs in a VN are only connected to one edge RBridge, that RBridge only needs to advertise the subnet corresponding to the VN to other RBridges using host routes. A Tenant ID unique across the TRILL campus is also carried in the advertisement to differentiate IP prefixes between different tenants, because the IP address space of different tenants can overlap (see Sections 7.3 and 7.4).

If a tenant is deleted on an edge RBridge RB1, RB1 updates the local tenant Data Label, tenant gateway MAC, and related IP prefix information it is advertising to include only the rest of the tenants. It may take some time for these updates to reach all other RBridges, so during this period of time there may be transient route inconsistency among the edge RBridges. If there is traffic in flight during this time, it will be dropped at the egress RBridge due to local tenant deletion. When a stable state is reached, the traffic to the deleted tenant will be dropped by the ingress RBridge. Therefore, the transient route inconsistency won't cause issues other than wasting some network bandwidth.

If a new tenant is created and a previously used tenant Data Label is assigned to the new tenant immediately, this may cause a security policy violation for the traffic in flight, because when the egress RBridge receives traffic from the old tenant, it will forward it in the new tenant's routing instance and deliver it to the wrong destination. So, a tenant Data Label MUST NOT be reallocated until a reasonable amount of time -- for example, twice the IS-IS Holding Time generally in use in the TRILL campus -- has passed to allow any traffic in flight to be discarded.

When the ARP entry in an edge RBridge for an ES times out, it will trigger an edge RBridge LSP advertisement to other edge RBridges with the corresponding IP routing entry deleted. If the ES is an IP router, the edge RBridge also notifies other edge RBridges that they must delete the routing entries corresponding to the IP prefixes accessible through that IP router. During the IP prefix deleting process, if there is traffic in flight, the traffic will be discarded at the egress RBridge because there is no local IP routing entry to the destination.

If an edge RBridge changes its tenant gateway MAC, it will trigger an edge RBridge LSP advertisement to other edge RBridges, giving the new gateway MAC to be used as the Inner.MacDA for future traffic destined to the edge RBridge. During the gateway MAC changing process, if there is traffic in flight using the old gateway MAC as the Inner.MacDA, the traffic will be discarded or will be forwarded as L2 intra-subnet traffic on the edge RBridge. If the inter-subnet tenant Data Label is a unique Data Label that is different from any access Data Labels, when the edge RBridge receives the traffic whose Inner.MacDA is different from the local tenant gateway MAC, the traffic will be discarded. If the edge RBridge uses one of the access Data Labels as an inter-subnet tenant Data Label, the traffic will be forwarded as L2 intra-subnet traffic unless a special traffic-filtering policy is enforced on the edge RBridge.

If there are multiple nicknames owned by an edge RBridge, the edge RBridge can also specify one nickname as the egress nickname for inter-subnet traffic forwarding. A NickFlags APPsub-TLV with the SE flag set can be used for this purpose. If the edge RBridge doesn’t specify a nickname for this purpose, the ingress RBridge can use any one of the nicknames owned by the egress as the egress nickname for inter-subnet traffic forwarding.

TRILL Extended Level 1 Flooding Scope (E-L1FS) FS-LSP [RFC7780] APPsub-TLVs are used for IP routing information synchronization in each routing domain among edge RBridges. Based on the synchronized information from other edge RBridges, each edge RBridge generates routing entries in each routing domain for remote IP addresses and subnets.

Through this solution, the intra-subnet forwarding and inter-subnet IP routing functions are integrated, and network management and deployment are simplified.

5.3. Active-Active Access

TRILL active-active service provides ESs with flow-level load balance and resilience against link failures at the edge of TRILL campuses, as described in [RFC7379].

If an ES is connected to two TRILL RBridges, say RB1 and RB2, in active-active mode, RB1 and RB2 MUST both be configured to act as a distributed L3 gateway for the ES in order to use a distributed gateway. RB1 and RB2 each learn the ES’s IP address through the ARP/ND process, and then they announce the IP address to the TRILL campus independently. The remote ingress RBridge will generate an IP routing entry corresponding to the IP address with two IP next hops of RB1 and RB2. When the ingress RBridge receives inter-subnet traffic from a local access network, the ingress RBridge selects RB1 or RB2 as the IP next hop based on least cost or, if costs are equal, the local load-balancing algorithm. The traffic will then be transmitted to the selected next-hop destination RB1 or RB2 through the TRILL campus.

5.4. Data Traffic Forwarding Process

After ES1, connected by L2 in VLAN-x, acquires its gateway’s MAC, it can start inter-subnet data traffic transmission to ES2 in VLAN-y.

When the edge RBridge attached to ES1 receives inter-subnet traffic from ES1, that RBridge performs L2 header termination; then, using the local VRF corresponding to VLAN-x, it performs the IP routing process in that VRF.

If destination ES2 is attached to the same edge RBridge, the traffic will be locally forwarded to ES2 by that RBridge. Compared to the centralized gateway solution, the forwarding path is optimal, and a traffic detour through the centralized gateway is avoided.

If ES2 is attached to a remote edge RBridge, the remote edge RBridge is the IP next hop, and the inter-subnet traffic is forwarded to the IP next hop through TRILL encapsulation. If there are multiple equal-cost shortest paths between the ingress RBridge and the egress RBridge, all these paths can be used for inter-subnet traffic forwarding, so load-spreading can be achieved for inter-subnet traffic.

When the remote RBridge receives the inter-subnet TRILL-encapsulated traffic, the RBridge decapsulates these TRILL packets and checks the Inner.MacDA. If that MAC address is the local gateway MAC corresponding to the inner label (VLAN or FGL), the inner label will be used to find the corresponding local VRF; the IP routing process in that VRF will then be performed, and the traffic will be locally forwarded to destination ES2.

In summary, this solution avoids traffic detours through a central gateway. Both inter-subnet and intra-subnet traffic can be forwarded along pair-wise shortest paths, and network bandwidth is conserved.

6. Distributed Layer 3 Gateway Process Example

This section gives a detailed description of a distributed L3 gateway solution example for IPv4 and IPv6.

In Figure 3, RB1 and RB2 support the distribution gateway function, ES1 connects to RB1, and ES2 connects to RB2. ES1 and ES2 belong to Tenant1 but are in different subnets.

```
---------             ---------
|  RB3  |             |  RB4  |
---------             ---------
#   *                     #  *
#   **************************
###########################  *
#                            *
#                            *
#                            *
---------              ---------
|  RB1  |              |  RB2  |
---------              ---------
   |                       |
 -----                   -----
 |ES1|                   |ES2|
 -----                   -----
```

Figure 3: Distributed Gateway Scenario

For IPv4, the IP address, VLAN, and subnet information of ES1 and ES2 are as follows:

ES	Tenant	IP Address	Subnet	VLAN
ES1	Tenant1	192.0.2.2	192.0.2.0/24	10
ES2	Tenant1	198.51.100.2	198.51.100.0/24	20

Figure 4: IPv4 ES Information

For IPv6, the IP address, VLAN, and subnet information of ES1 and ES2 are as follows:

ES	Tenant	IP Address	Subnet	VLAN
ES1	Tenant1	2001:db8:0:1::2	2001:db8:0:1::0/64	10
ES2	Tenant1	2001:db8:0:2::2	2001:db8:0:2::0/64	20

Figure 5: IPv6 ES Information

The nickname, VRF, tenant Label, and tenant gateway MAC for Tenant1 on RB1 and RB2 are as follows:

RB	Nickname	Tenant	VRF	Tenant Label	Gateway MAC
RB1	nick1	Tenant1	VRF1	100	MAC1
RB2	nick2	Tenant1	VRF2	100	MAC2

Figure 6: RBridge Information

6.1. Control-Plane Process

RB1 advertises the following local routing information to the TRILL campus:

Tenant ID: 1

Tenant gateway MAC: MAC1

Tenant Label for Tenant1: VLAN 100

IPv4 prefix for Tenant1: 192.0.2.0/24

IPv6 prefix for Tenant1: 2001:db8:0:1::0/64

RB2 announces the following local routing information to the TRILL campus:

Tenant ID: 1

Tenant gateway MAC: MAC2

Tenant Label for Tenant1: VLAN 100

IPv4 prefix for Tenant1: 198.51.100.0/24

IPv6 prefix for Tenant1: 2001:db8:0:2::0/64

Relying on the routing information from RB2, remote routing entries on RB1 are generated as follows:

Prefix/Mask	Inner.MacDA	Inner VLAN	Egress Nickname
198.51.100.0/24	MAC2	100	nick2
2001:db8:0:2::0/64	MAC2	100	nick2

Figure 7: Tenant1 Remote Routing Table on RB1

Similarly, relying on the routing information from RB1, remote routing entries on RB2 are generated as follows:

Prefix/Mask	Inner.MacDA	Inner VLAN	Egress Nickname
192.0.2.0/24	MAC1	100	nick1
2001:db8:0:1::0/64	MAC1	100	nick1

Figure 8: Tenant1 Remote Routing Table on RB2

6.2. Data-Plane Process

Assuming that ES1 sends unicast inter-subnet traffic to ES2, the traffic forwarding process is as follows:

1. ES1 sends unicast inter-subnet traffic to RB1 with RB1’s gateway’s MAC as the destination MAC and the VLAN as VLAN 10.

2. Ingress RBridge (RB1) forwarding process:

   RB1 checks the destination MAC. If the destination MAC equals the local gateway MAC, the gateway function will terminate the L2 header and perform L3 routing.

   RB1 looks up IP routing table information by destination IP and Tenant ID to get IP next-hop information, which includes the egress RBridge’s gateway MAC (MAC2), tenant Label (VLAN 100), and egress nickname (nick2). Using this information, RB1 will perform inner Ethernet header encapsulation and TRILL encapsulation. RB1 will use MAC2 as the Inner.MacDA, MAC1 (RB1’s own gateway MAC) as the Inner.MacSA, VLAN 100 as the Inner.VLAN, nick2 as the egress nickname, and nick1 as the ingress nickname.

   RB1 looks up TRILL forwarding information by egress nickname and sends the traffic to the TRILL next hop as per [RFC6325]. The traffic will be sent to RB3 or RB4 as a result of load-balancing.

Assuming that the traffic is forwarded to RB3, the following occurs:

3. Transit RBridge (RB3) forwarding process:

   RB3 looks up TRILL forwarding information by egress nickname and forwards the traffic to RB2 as per [RFC6325].

4. Egress RBridge forwarding process:

   As the egress nickname is RB2’s own nickname, RB2 performs TRILL decapsulation. It then checks the Inner.MacDA and, because that MAC is equal to the local gateway MAC, performs inner Ethernet header termination. Using the inner VLAN, RB2 finds the local corresponding VRF and looks up the packet’s destination IP address in the VRF’s IP routing table. The traffic is then locally forwarded to ES2 with VLAN 20.

7. TRILL Protocol Extensions

If an edge RBridge RB1 participates in the distributed gateway function, it announces its tenant gateway MAC and tenant Data Label to the TRILL campus through the tenant Label and gateway MAC APPsub-TLV. It should announce its local IPv4 and IPv6 prefixes through the IPv4 Prefix APPsub-TLV and the IPv6 Prefix APPsub-TLV, respectively. If RB1 has multiple nicknames, it can announce one nickname for use by the distributed gateway, by using the NickFlags APPsub-TLV with the SE flag set to 1.

The remote ingress RBridges belonging to the same routing domain use this information to generate IP routing entries in that routing domain. These RBridges use the nickname, tenant gateway MAC, and tenant Label of RB1 to perform inter-subnet TRILL encapsulation when they receive inter-subnet traffic from a local ES. The nickname is used as the egress nickname, the tenant gateway MAC is used as the Inner.MacDA, and the tenant Data Label is used as the Inner.Label. The following APPsub-TLVs MUST be included in a TRILL GENINFO TLV in E-L1FS FS-LSPs [RFC7780].

7.1. The Tenant Label and Gateway MAC APPsub-TLV

```
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
      |   Type                        | (2 bytes)
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
      |   Length                      | (2 bytes)
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
      |                     Tenant ID   (4 bytes)                     |
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
      | Resv1 |     Label1            | (2 bytes)
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
      | Resv2 |     Label2            | (2 bytes)
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+....-+-+-+-+-+-+-+-+-+
      |            Tenant Gateway MAC   (6 bytes)                     |
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+....-+-+-+-+-+-+-+-+-+
```

- o Type: Set to the TENANT-GWMAC-LABEL sub-TLV type (7). 2 bytes, because this APPsub-TLV appears in an extended TLV [RFC7356].

- o Length: If the Label1 field is used to represent a VLAN, the value of the Length field is 12. If the Label1 and Label2 fields are used to represent an FGL, the value of the Length field is 14.

- o Tenant ID: This identifies a Tenant ID unique across the TRILL campus.

- Resv1: 4 bits that MUST be sent as zero and ignored on receipt.

- Label1: If the value of the Length field is 12, it identifies a tenant Label corresponding to a VLAN ID. If the value of the Length field is 14, it identifies the higher 12 bits of a tenant Label corresponding to an FGL.

- Resv2: 4 bits that MUST be sent as zero and ignored on receipt. Only present if the Length field is 14.

- Label2: This field has the lower 12 bits of the tenant Label corresponding to an FGL. Only present if the Length field is 14.

- Tenant Gateway MAC: This identifies the local gateway MAC corresponding to the Tenant ID. The remote ingress RBridges use the gateway MAC as the Inner.MacDA. The advertising TRILL RBridge uses the gateway MAC to differentiate L2 intra-subnet traffic and L3 inter-subnet traffic in the egress direction.

7.2. The SE Flag in the NickFlags APPsub-TLV

The NickFlags APPsub-TLV is specified in [RFC7780], where the IN flag is described. The SE Flag is assigned as follows:

```
+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
|   Nickname                                    |
+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
|IN|SE|         RESV                            |
+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
                 NICKFLAG RECORD
```

- SE: If the SE flag is set to 1, it indicates that the advertising RBridge suggests that the Nickname SHOULD be used as the Inter-Subnet Egress nickname for inter-subnet traffic forwarding. If the SE flag is set to 0, that Nickname SHOULD NOT be used for that purpose. The SE flag is ignored if the NickFlags APPsub-TLV is advertised by an RBridge that does not own the Nickname.

7.3. The IPv4 Prefix APPsub-TLV

```
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |   Type                          |                    (2 bytes)
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |   Total Length                  |                    (2 bytes)
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+...-+-+-+-+-+-+-+-+
     |                     Tenant ID                    | (4 bytes)
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+...-+-+-+-+-+-+-+-+
     |PrefixLength(1)|                                    (1 byte)
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+...-+-+-+-+-+-+-+-+
     |                     Prefix (1)                   | (variable)
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+...-+-+-+-+-+-+-+-+
     |     .....     |                                    (1 byte)
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+...-+-+-+-+-+-+-+-+
     |                    .....                         | (variable)
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+...-+-+-+-+-+-+-+-+
     |PrefixLength(N)|                                    (1 byte)
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+...-+-+-+-+-+-+-+-+
     |                     Prefix (N)                   | (variable)
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+...-+-+-+-+-+-+-+-+
```

- o Type: Set to the IPV4-PREFIX sub-TLV type (8). 2 bytes, because this APPsub-TLV appears in an extended TLV [RFC7356].

- o Total Length: This 2-byte unsigned integer indicates the total length of the Tenant ID, Prefix Length, and Prefix fields, in octets. A value of 0 indicates that no IPv4 prefix is being advertised.

- o Tenant ID: This identifies a Tenant ID unique across the TRILL campus.

- o Prefix Length: The Prefix Length field indicates the length, in bits, of the IPv4 address prefix. A length of 0 (i.e., the prefix itself is 0 octets) indicates a prefix that matches all IPv4 addresses.

- o Prefix: The Prefix field contains an IPv4 address prefix, followed by enough trailing bits to make the end of the field fall on an octet boundary. Note that the value of the trailing bits is irrelevant. For example, if the Prefix Length is 12, indicating 12 bits, then the Prefix is 2 octets and the low-order 4 bits of the Prefix are irrelevant.

7.4. The IPv6 Prefix APPsub-TLV

```
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |   Type                        |                    (2 bytes)
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |   Total Length                |                    (2 bytes)
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+...-+-+-+-+-+-+-+-+
     |                     Tenant ID                    | (4 bytes)
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+...-+-+-+-+-+-+-+-+
     |PrefixLength(1)|                                    (1 byte)
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+...-+-+-+-+-+-+-+-+
     |                     Prefix (1)                   | (variable)
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+...-+-+-+-+-+-+-+-+
     |     .....       |                                  (1 byte)
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+...-+-+-+-+-+-+-+-+
     |                    .....                         | (variable)
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+...-+-+-+-+-+-+-+-+
     |PrefixLength(N)|                                    (1 byte)
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+...-+-+-+-+-+-+-+-+
     |                     Prefix (N)                   | (variable)
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+...-+-+-+-+-+-+-+-+
```

- Type: Set to the IPV6-PREFIX sub-TLV type (9). 2 bytes, because this APPsub-TLV appears in an extended TLV [RFC7356].

- Total Length: This 2-byte unsigned integer indicates the total length of the Tenant ID, Prefix Length, and Prefix fields, in octets. A value of 0 indicates that no IPv6 prefix is being advertised.

- Tenant ID: This identifies a Tenant ID unique across the TRILL campus.

- Prefix Length: The Prefix Length field indicates the length, in bits, of the IPv6 address prefix. A length of 0 (i.e., the prefix itself is 0 octets) indicates a prefix that matches all IPv6 addresses.

- Prefix: The Prefix field contains an IPv6 address prefix, followed by enough trailing bits to make the end of the field fall on an octet boundary. Note that the value of the trailing bits is irrelevant. For example, if the Prefix Length is 100, indicating 100 bits, then the Prefix is 13 octets and the low-order 4 bits of the Prefix are irrelevant.

8. Security Considerations

Correct configuration of the participating edge RBridges is important to assure that data is not delivered to the wrong tenant, as such incorrect delivery would violate security constraints. IS-IS security [RFC5310] can be used to secure the information advertised by the edge RBridges in LSPs and FS-LSPs.

To avoid the mishandling of data in flight, see Section 5.2 for constraints on the reuse of a tenant Label and on tenant gateway MAC changes. Selecting tenant Labels and IDs in a pseudorandom fashion [RFC4086] can make it more difficult for an adversary to guess a tenant Label or ID that is in use.

Particularly sensitive data should be encrypted end-to-end -- that is, from the source ES to the destination ES. Since the TRILL campus is, for the most part, transparent to ES traffic, such ESs are free to use whatever end-to-end security protocol they would like.

For general TRILL security considerations, see [RFC6325].

9. Management Considerations

The configuration at each RBridge to support the distributed L3 gateway feature is visible, via the link-state database, to all other RBridges in the campus. Operations, Administration, and Maintenance (OAM) facilities for TRILL are primarily specified in [RFC7455] and [RFC7456].

10. IANA Considerations

IANA has assigned three APPsub-TLV type numbers that are lower than 255 in the "TRILL APPsub-TLV Types under IS-IS TLV 251 Application Identifier 1" registry. The registry has been updated as follows:

Type	Name	Reference
7	TENANT-GWMAC-LABEL	this document
8	IPV4-PREFIX	this document
9	IPV6-PREFIX	this document

IANA has assigned a flag bit in the NickFlags APPsub-TLV as described in Section 7.2 and updated the "NickFlags Bits" registry, created by [RFC7780], as follows:

Bit	Mnemonic	Description	Reference
1	SE	Inter-Subnet Egress	this document

11. References

11.1. Normative References

[IS-IS] International Organization for Standardization, "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 Standard 10589, 2002.

[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>.

[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, DOI 10.17487/RFC7172, 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, DOI 10.17487/RFC7176, May 2014, <http://www.rfc-editor.org/info/rfc7176>.

[RFC7356] Ginsberg, L., Previdi, S., and Y. Yang, "IS-IS Flooding Scope Link State PDUs (LSPs)", RFC 7356, DOI 10.17487/RFC7356, September 2014, <http://www.rfc-editor.org/info/rfc7356>.

[RFC7780] Eastlake 3rd, D., Zhang, M., Perlman, R., Banerjee, A., Ghanwani, A., and S. Gupta, "Transparent Interconnection of Lots of Links (TRILL): Clarifications, Corrections, and Updates", RFC 7780, DOI 10.17487/RFC7780, February 2016, <http://www.rfc-editor.org/info/rfc7780>.

11.2. Informative References

[RFC826] Plummer, D., "Ethernet Address Resolution Protocol: Or Converting Network Protocol Addresses to 48.bit Ethernet Address for Transmission on Ethernet Hardware", STD 37, RFC 826, DOI 10.17487/RFC0826, November 1982, <http://www.rfc-editor.org/info/rfc826>.

[RFC4086] Eastlake 3rd, D., Schiller, J., and S. Crocker, "Randomness Requirements for Security", BCP 106, RFC 4086, DOI 10.17487/RFC4086, June 2005, <http://www.rfc-editor.org/info/rfc4086>.

[RFC4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, DOI 10.17487/RFC4861, September 2007, <http://www.rfc-editor.org/info/rfc4861>.

[RFC5310] Bhatia, M., Manral, V., Li, T., Atkinson, R., White, R., and M. Fanto, "IS-IS Generic Cryptographic Authentication", RFC 5310, DOI 10.17487/RFC5310, February 2009, <http://www.rfc-editor.org/info/rfc5310>.

[RFC7379] Li, Y., Hao, W., Perlman, R., Hudson, J., and H. Zhai, "Problem Statement and Goals for Active-Active Connection at the Transparent Interconnection of Lots of Links (TRILL) Edge", RFC 7379, DOI 10.17487/RFC7379, October 2014, <http://www.rfc-editor.org/info/rfc7379>.

[RFC7455] Senevirathne, T., Finn, N., Salam, S., Kumar, D., Eastlake 3rd, D., Aldrin, S., and Y. Li, "Transparent Interconnection of Lots of Links (TRILL): Fault Management", RFC 7455, DOI 10.17487/RFC7455, March 2015, <http://www.rfc-editor.org/info/rfc7455>.

[RFC7456] Mizrahi, T., Senevirathne, T., Salam, S., Kumar, D., and D. Eastlake 3rd, "Loss and Delay Measurement in Transparent Interconnection of Lots of Links (TRILL)", RFC 7456, DOI 10.17487/RFC7456, March 2015, <http://www.rfc-editor.org/info/rfc7456>.

Acknowledgments

The authors wish to acknowledge the important contributions of Donald Eastlake, Gayle Noble, Mohammed Umair, Susan Hares, Guangrui Wu, Zhenbin Li, Zhibo Hu, Liang Xia, Scott Bradner, Stephen Farrell, Shawn Emery, Ben Campbell, Russ White, Kathleen Moriarty, Suresh Krishnan, Mirja Kuehlewind, and Francis Dupont.

Authors' Addresses

Weiguo Hao
Huawei Technologies
101 Software Avenue
Nanjing 210012
China

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

Yizhou Li
Huawei Technologies
101 Software Avenue
Nanjing 210012
China

Phone: +86-25-56625375
Email: liyizhou@huawei.com

Andrew Qu
MediaTec

Email: laodulaodu@gmail.com

Muhammad Durrani
Equinix Inc.

Email: mdurrani@equinix.com

Ponkarthick Sivamurugan
IP Infusion
RMZ Centennial
Mahadevapura Post
Bangalore 560048

Email: Ponkarthick.sivamurugan@ipinfusion.com