TRILL

Weiguo Hao
Yizhou Li

Internet Draft

Huawei Technologies

Intended status: Standards Track

July 12, 2013

Expires: January 2014

TRILL Integrated Routing and Bridging Solution
draft-hao-trill-irb-02.txt

Status of this Memo

This Internet-Draft will expire on July 12, 2013.

Copyright Notice

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.

Abstract

Currently, TRILL solution can only provide optimum unicast forwarding just for Layer2 traffic of intra-subnet forwarding, not for Layer3 traffic(inter-subnet forwarding). In this document, a TRILL Integrated Routing and Bridging (IRB) solution is introduced to provide optimum unicast forwarding not just for Layer 2 traffic (intra-subnet forwarding), but also for Layer 3 traffic (inter-subnet forwarding). In the TRILL IRB scenario, an edge RB MUST perform the bridging function for the End Systems that are on the same subnet and the IP routing for the End Systems that are on the different subnets of same tenant.ESADI extension can be used for synchronizing <MAC, IP> correspondence among edge RBridges. To reduce the number of ESADI session among edge RBridges, Management Data Label for ESADI is suggested to be used.

Table of Contents

1. Introduction

The IETF has standardized the TRILL (Transparent Interconnection of Lots of Links) protocol [RFC6325] that provides a solution for least cost transparent routing in multi-hop networks with arbitrary topologies and link technologies, using [IS-IS] [RFC6165] [RFC6326bis] link-state routing and a hop count. TRILL switches are sometimes called RBridges (Routing Bridges).

Currently, TRILL only provides optimum unicast forwarding for Layer 2 LAN traffic (intra-subnet forwarding), not for Layer 3 traffic (inter-subnet forwarding).

In this document, a TRILL Integrated Routing and Bridging (IRB) solution is introduced to provide optimum unicast forwarding not just for Layer 2 traffic (intra-subnet forwarding), but also for Layer 3 traffic (inter-subnet forwarding). In the TRILL IRB solution, the edge RBridge provides a per tenant virtual switching and routing instance with address isolation and Layer 3 tunnel encapsulation across the core. The edge RBridge supports bridging among end stations that belong to same subnet and routing among end stations that belongs to different subnets of same routing domain.

This document is organized as follows: Section 3 describes why an IRB solution is needed. Section 4 gives forwarding procedures. Section **5 describes TRILL protocol extensions to support TRILL IRB solution.**

Familiarity with [RFC6325] and [ESADI] is assumed in this document.

2. Conventions used in this document

In examples, "C:" and "S:" indicate lines sent by the client and server respectively.

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].

In this document, these words will appear with that interpretation only when in ALL CAPS. Lower case uses of these words are not to be interpreted as carrying RFC-2119 significance.

In this document, the characters ">>" preceding an indented line(s) indicates a compliance requirement statement using the key words listed above. This convention aids reviewers in quickly identifying or finding the explicit compliance requirements of this RFC.ARP:IPv4 Address Resolution Protocol [RFC826]

DC: Data Center

End Station: VM or physical server, whose address is either a destination or the source of a data frame.

IRB: Integrated Routing and Bridging

L2: Layer 2

L3: Layer 3

ND: IPv6's Neighbor Discovery [RFC4861]

RB: Router Bridge. RBs are switches that implement the TRILL protocol and combine the advantages of bridges and routers.

TRILL: Transparent Interconnection of Lots of Links. TRILL presented in [RFC6325] and other related documents, provides methods of utilizing all available paths for active forwarding, with minimum configuration. TRILL utilizes IS-IS (Intermediate System to Intermediate System) as its control plane and encapsulates native frames with a TRILL header.

VN: Virtual Network

VRF: Virtual Routing and Forwarding. In IP-based computer networks, Virtual Routing and Forwarding (VRF) is a technology that allows multiple instances of a routing table to co-exist within the same router at the same time.

3. Problem statement

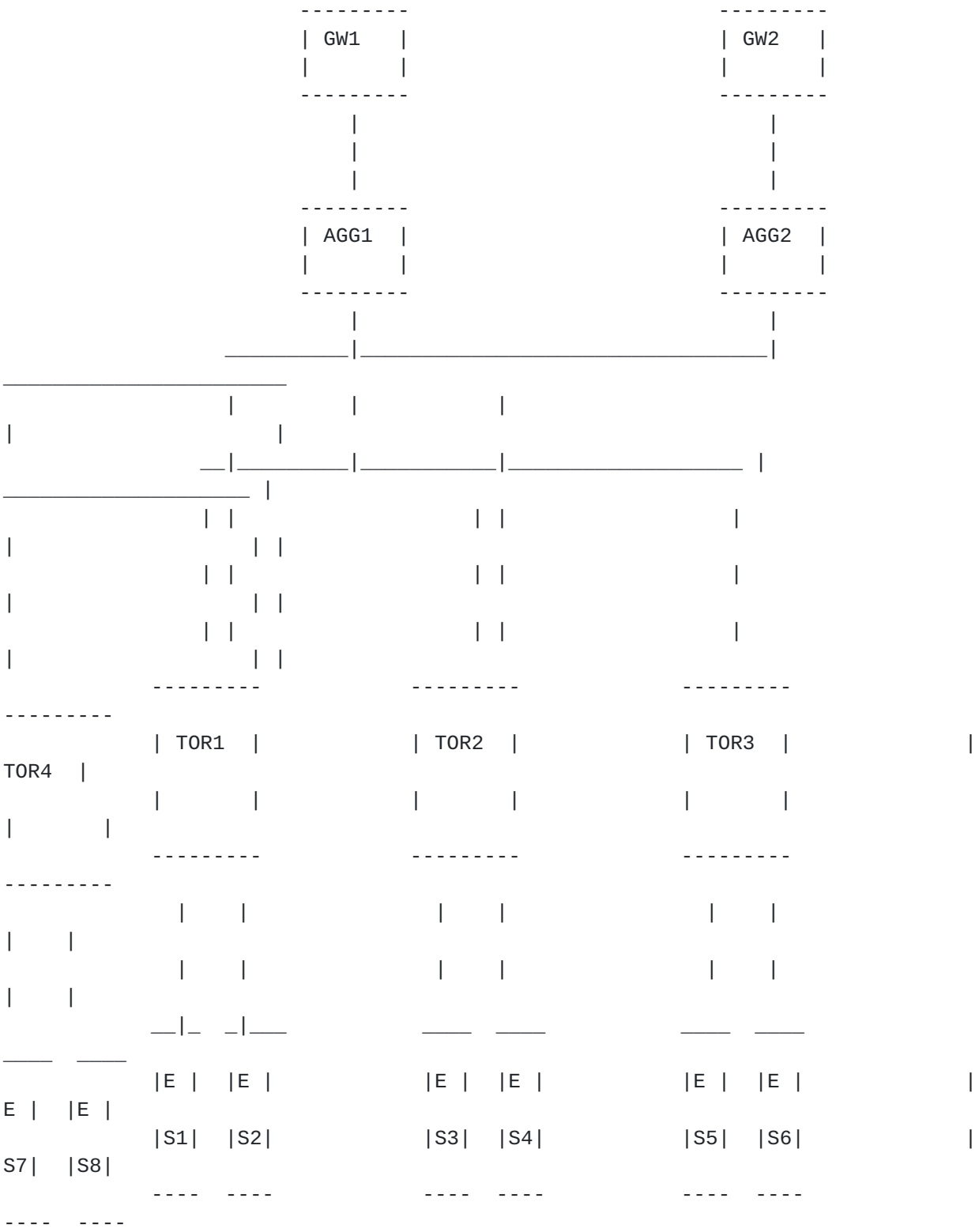

Figure 1 A typical DC network

Figure-1 depicts a Data Center Network (DCN) using TRILL where edge RB functionality resides in physical Top of Rack (ToR) switches.

Centralized gateway (GW) nodes are provided not only for north-south bound L3 forwarding but also for east-west bound inter-subnet L3 forwarding. If two end stations of same tenant are on two different subnets and need to communicate with each other, their packets need to be forwarded all the way to a centralized layer 3 GW so one of the GW devices can perform L3 forwarding. This is generally sub-optimal because the two end stations may be connected to the same TOR where L3 switching could have been performed locally. If an edge RB has IRB capability, then it can perform optimum L2 forwarding for intra-subnet traffic and optimum L3 forwarding for inter-subnet traffic, delivering optimum forwarding for unicast packets at all time.

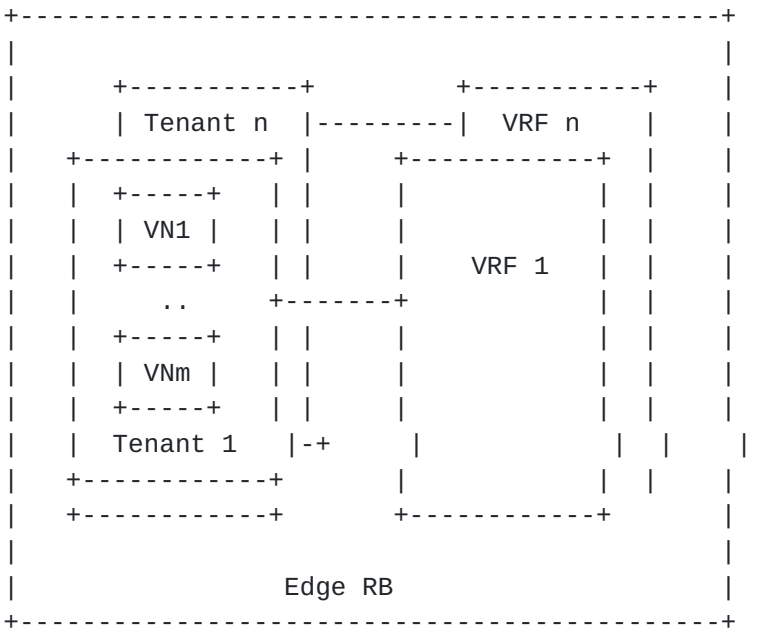

Figure 2 Edge RB Model as default GW

In a data center network (DCN), each tenant may include one or more IP subnets. Each IP subnet corresponds to one layer 2 virtual network and in normal cases each tenant corresponds to one routing domain (RD). One layer 2 virtual network (VN) maps to a unique IP subnet within a VRF context. Each layer 2 virtual network in a TRILL campus is identified by a unique 12-bit VLAN ID or 24-bit Fine Grained Label [FGL]. Different routing domains should have overlapping address space, distinct and separate routes. The end systems that belongs to the same subnet communicate through L2 forwarding, end systems of same tenant that belongs to different subnet communicate through L3 forwarding.

The above figure 2 depicts the model where there are N VRFs corresponding to N tenants with each tenant having up to M segments/subnets (virtual network).

4. Requirements of edge RB acts as default GW

In the TRILL IRB scenario, an edge RB MUST perform the bridging function for the End Systems that are on the same subnet and the IP routing for the End Systems that are on the different subnets of same tenant. For L3 traffic edge RB must act as default GW for connected end systems that belongs to each routing domain.

Each GW should establish a gateway interface and VRF for each routing domain. Each L2 VN maps to a unique IP subnet within a VRF context. Because the end systems in each routing domain may spread over multiple edge RBs, all these edge RBs should act as default GWs and have same gateway IP and MAC address for the connected end systems that belong to same routing domain. The default GW must satisfy following requirements:

1, Support <MAC, IP> correspondence learning on each default GW. An edge RBridge can learn IP/MAC correspondence of locally attached end stations by inspecting the ARP message or other data frame. An end system uses the ARP/ND protocol to discover other end system MAC addresses if they are on the same subnet; An end system sends a packet to a known gateway if the destination of the packet is on different subnet from the sender end system and the end system uses ARP/ND protocol to find the gateway MAC address. When the default GW receives ARP/ND request packet from an access link, if destination IP in the packet is equals to the IP address of the default GW, it returns an ARP reply with self MAC and IP mapping information. After the end system acquires the MAC address of the GW, it will send unicast IP packets to destination end systems with destination MAC equals to the MAC of default GW, the default GW will perform L2 termination and find routing table entry with destination IP to perform L3 forwarding for the unicast packet.

2, Support <MAC, IP> correspondence synchronization for each routing domain among default GWs.

For each tenant, there may be multiple L2 VNs and the end systems in each L2 VN may spread over multiple edge RBs. These edge RBs can only acquire the ARP/ND table for locally attached end systems. To support inter-subnet communication between locally attached end stations and remote end stations, the edge RB should have <MAC, IP, L2 VNID > mapping information for all remote end stations that are attached to all other edge RBs.

So all edge RBs should synchronize ARP/ND table (i.e. MAC, IP, L2 VNID for each local attached end systems) of local attached end systems to all other edge RBs that have the same routing domain. Similarly, when a ARP/ND table in an edge RB ages, the edge RB should flood the ARP/ND table flush event to all other RBs.

After ARP/ND table synchronization is finished, all edge RBs keep all ARP/ND tables and install an IP forwarding table for all end

systems in each VRF. After that, these edge RBs can support inter-subnet L3 forwarding for all end systems in each routing domain.

3, Support L2 forwarding for intra-subnet traffic and L3 forwarding for inter-subnet traffic on each default GW.

When ingress edge RB receives packets from local attached end station, the RB performs following process:

1. The RB will check the destination MAC, if the destination MAC equals to default GW's MAC, the GW will perform L3 forwarding process. Otherwise, the RB will perform L2 forwarding process and jump to step 4.

2. The RB will find IP forwarding table by destination IP to get the MAC and VN ID of destination end station.

3. The RB will modify source MAC, destination MAC and VN ID of the packet. Source MAC is modified to GW's MAC, destination MAC is modified to destination end station's MAC, VN ID is modified to destination end station's VNID.

4. The RB will perform L2 forwarding process by destination MAC in destination L2 VN ID and will get remote nickname by finding MAC table entry in destination L2 VN. Then it performs TRILL encapsulation and goes through optimal TRILL forwarding to the egress RB. After decapsulation at the egress RB, the packet will reach to destination end station.

So when edge RBs support default GW function, optimum unicast forwarding will be performed not just for L2 traffic (intra-subnet forwarding), but also for L3 traffic (inter-subnet forwarding).In the TRILL IRB solution, edge RBridges are connected to each other via one or multiple RBridge hops, however they are always a single IP hop away.

5. Protocol extension to support <MAC, IP> correspondence synchronization

Edge RBs that belong to same routing domain should synchronize their ARP/ND tables with each other. One routing domain may include multiple subnets and each subnet maps to a L2 VN ID. A possible solution to synchronize ARP/ND tables among edge RBs was described by [ESADI].

ESADI is a Data Label scoped way for RBridges to announce and learn end station MAC addresses. There is a separate ESADI instance for

each Data Label (VLAN or FGL). ESADI protocol can be extended to announce and learn end station ARP/ND tables amongst all edge RBs for each routing domain where edge RB acts as a default GW for local attached end stations.

The Interface Addresses APPsub-TLV is used to indicate that a set of addresses on the same end-station interface and to associate that interface with the TRILL switch by which the interface is reachable. The TLV supports multiple address families and can be used to declare MAC and IPV4/IPV6 correspondence on each edge RBridge to TRILL campus.

When an edge RBridge learns IP/MAC correspondence of a locally attached end station 1 by inspecting the ARP message or other data frame, it will use Interface Addresses APPsub-TLV and flood such information to all other edge RBs belonging to same routing domain. Edge RBs in the same routing domain must establish ESADI sessions for each layer 2 network beforehand. When an edge RBridge receives Interface Addresses APPsub-TLV, it retrieves IPv4 and MAC mapping information of the end station and install it to its IP routing table in the corresponding VRF. After that, the end stations attached to the receiving edge RBridges can communicate to end station 1 through layer 2 and layer 3 forwarding procedures.

6. Management Data Label for ESADI

As ESADI is a Data Label(VLAN or FGL) scoped solution, each edge RBridge needs to establish ESADI session for each L2 VN in a routing domain. Therefore the number of ESADI session is huge and is a big burden for each RBridge's CPU. So we suggest a management Data Label for ESADI to be used for this purpose.

Every RBridge should be configured with a globally unique management Data Label. RBridges establishes ESADI session using this management Data Label. In extreme case, we can use one management ESADI session for all routing domains. With this approach CPU consumption can be greatly reduced on every RBridge. The correspondence of management Data Label and L2 VNs can be statically configured on every RBridge.The operator must make sure the configuration consistency for all RBridges. A new TLV is suggested to be defined in ESADI to synchronize ARP/ND tables for multiple L2 VN in one ESADI session.

7. Security Considerations

This document adds no additional security risks to IS-IS, nor does it provide any additional security for IS-IS.

See [RFC6325] for general TRILL Security Consideration.

8. IANA Considerations

See [DIRECTORY] for IANA allocation and registry considerations.

9. References

[1] [RFC6325] Perlman, R., Eastlake 3rd, D., Dutt, D., Gai, S., and A. Ghanwani, "Routing Bridges (RBridges): Base Protocol Specification", RFC 6325, July 2011.

[2] [rfc6326bis] - Eastlake, D., Banerjee, A., Dutt, D., Perlman, R., and A. Ghanwani, "TRILL Use of IS-IS", draft-ietf-isisrfc6326bis-00.txt, work in progress.

[3] [ESADI] - Zhai, H., F. Hu, R. Perlman, D. Eastlake, J. Hudson, "TRILL(Transparent Interconnection of Lots of Links): The ESADI (End Station Address Distribution Information) Protocol", draft-ietf-trill-esadi-02.txt, work in progress.

[4] [DIRECTORY] - L.Dunar., D. Eastlake, " TRILL: Directory Assistance Mechanisms", draft-dunbar-trill-scheme-for-directory-assist-04.txt, work in progress.

9.1. Informative References

[1] [RFC6165] Banerjee,A., Ward, D., Dutt, D.,

, "Extensions to IS-IS for Layer-2 Systems", RFC 6165, April 2011.

10. Acknowledgments

The authors wish to acknowledge the important contributions of Donald Eastlake, Zhenbin Li.

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