TRILL

Weiguo Hao
Yizhou Li
Donald Eastlake
Huawei
Radia Perlman
Intel Labs

Internet Draft

Intended status: Standards Track
Expires: August 2014

February 14, 2014

TRILL anycast Layer 3 Gateway
draft-hao-trill-anycast-gw-00.txt

Status of this Memo

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/ietf/1id-abstracts.txt

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

This Internet-Draft will expire on August 14, 2014.

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

This draft mainly describes centralized anycast layer 3 gateway solution in TRILL campus. Comparing to traditional VRRP based active-standby layer 3 gateway solution, this solution can achieve better load balancing and scalability. Anycast nickname, anycast gateway IP and MAC are introduced. It can ensure inter-subnet traffic forwarding in flow-based load balancing mode among all physical layer 3 gateways. To avoid sending duplicated ARP reply message to the end system, ARP master gateway election mechanism is introduced. The election algorithm is described in this draft.

Table of Contents

1. Introduction

In a TRILL based multi-tenancy data center network (DCN), each tenant normally owns one routing domain (RD) which may consist of one or more IP subnets. It is a common practice that one layer 2 virtual network (VN) maps to a unique IP subnet. Layer 2 virtual network in a TRILL campus is identified by a 12-bit VLAN ID or 24-bit Fine Grained Label [FGL].

All the inter-subnet communication or inter VN communication need to pass through an L3 GW. Different subnets in one tenant are usually allowed to communicate with each other freely. Gateway plays an important role in both such west-to-east traffic and traditional north-to-south traffic.

Figure 1 shows a typical data center network topology. Multiple core switches serve as the layer 3 gateways. All the network nodes are RBridges running TRILL protocol. Gateway functions co-exist with traditional RBridge functions at the GW switch. There are several ways to organize the gateways. A traditional way is to use VRRP based gateways which is explained in section 3. However it has the issue of scalability and efficiency. In order to avoid single point of failure and achieve better load balancing, anycast gateway group can be used.The key idea of anycast gateway is to make multiple physical gateways share the same gateway IP and MAC address for single virtual network(VN).

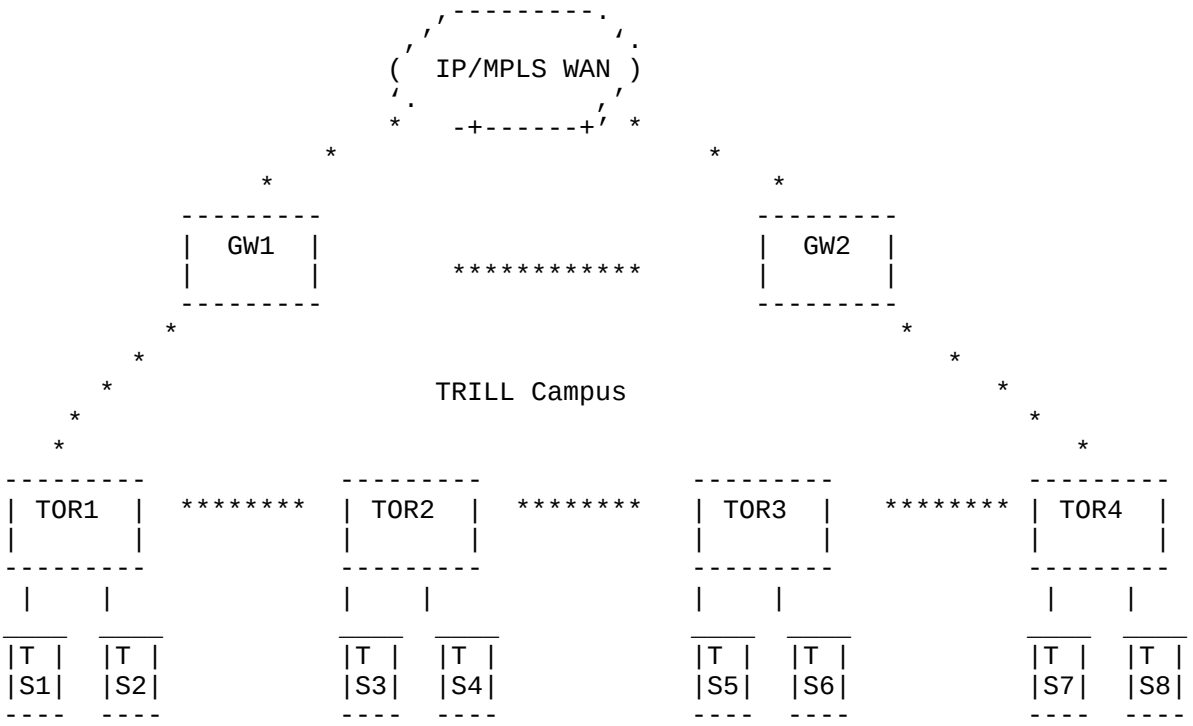


Figure 1 Centralized layer 3 gateway in TRILL campus

For inter-subnet layer 3 traffic, centralized layer 3 gateway is normally used and put at the boundary of TRILL network and the external IP network. In figure 1 above, GW1 and GW2 are integrated devices of layer 3 gateway and TRILL RB function. TRILL protocol runs on TOR and GW devices. West-to-east IP traffic among different VNs and north-to-south IP traffic between TRILL network and external IP network both pass through the layer 3 gateway. When the gateway receives the unicast TRILL encapsulated traffic from one layer 2 VN, it removes the TRILL encapsulation header. If destination MAC in inner Ethernet header is gateway's MAC, the gateway removes inner Ethernet header. Then the gateway looks up local IP forwarding table. If destination IP belongs to another VN in TRILL campus, the gateway will encapsulate the frame in TRILL format and send to the destination.

To eliminate the single point of gateway failure and to enhance the reliability, multiple layer 3 gateways are deployed. These gateways can work in active-standby mode or active-active mode. In active-standby mode, for each VN only one gateway acts as master and is

responsible for IP traffic forwarding between VNs. Network bandwidth usage is inefficient with such deployment. In a cloud computing data center, it is estimated that about 70% of traffic is east-west traffic which requires a non-blocking forwarding for line-speed traffic transmission between servers.

For inter-subnet layer 3 traffic, multiple centralized layer 3 gateways working in flow-based active-active mode will enhance the network efficiency. In this draft, such anycast layer 3 gateway solution for TRILL campus is illustrated. Anycast nickname, anycast gateway IP and MAC address are introduced. Anycast gateway IP and MAC address are set on each layer 3 gateway for each VN to terminate Ethernet traffic. Anycast nickname also is shared by multiple gateways, the TRILL traffic with anycast nickname as egress nickname could go to any one of the gateways by the natural support of ECMP from TRILL protocol, so flow-based load balancing among physical gateways will be achieved. Comparing to traditional VRRP based active-standby layer 3 gateway, anycast gateway can achieve better load balancing and scalability.

This document is organized as follows: Section 3 describes VRRP based gateway solution and its disadvantage. Section 4 gives anycast gateway solution overview. Section 5 describes ARP handling process. Section 6 describes data traffic forwarding. Section 7 describes TRILL protocol extension.

Familiarity with [RFC6325] is assumed in this document.

2. Conventions used in this document

ARP - Address Resolution Protocol.

ES - End Station.

VN - Virtual Network. In TRILL network, each VN can be identified by a 12 bit VLAN ID or a 24 bit Fine Grained Label.

3. VRRP based gateways

Assuming in figure 1 above, COR1 and COR2 are centralized gateway in active-standby mode. TRILL protocol runs on TOR and GW device. ES is end station. ES1,ES3,ES5 and ES7 belong to VLAN1. ES2,ES4,ES6 and ES8 belong to VLAN2.

The Virtual Router Redundancy Protocol (VRRP) is designed to eliminate the single point of gateway failure. VRRP is an election protocol that dynamically assigns responsibility for a virtual

router to one of the VRRP routers on a layer 2 VN. Any of the virtual router's IP addresses on a LAN can then be used as the default first hop router by end-hosts. The layer 3 gateway of VRRP master is responsible for forwarding packets destined to the virtual router. If VRRP master fails, VRRP backup will take over.

VRRP based solution has the following issues:

1. Inefficient network bandwidth usage. Only the VRRP master gateway forwards the traffic. VRRP slave is idle most of the time.

2. Low scalability. VRRP session among physical layer 3 gateways should be established per layer 2 VN. Large number of layer 2 VN will cause heavy CPU workload for each layer 3 gateway.

4. Anycast layer 3 gateway

Multiple gateways share the same IP and MAC address for each VN. These IP and MAC address are called anycast IP and anycast MAC address respectively. Anycast IP is used as the default gateway IP address for all end hosts in the corresponding VN. Gateways always respond with the anycast MAC address when receiving ARP request for the anycast IP. As different VNs are allowed to have overlapping MAC address space, different anycast IP addresses can map to the same anycast MAC. That is to say, each VN should have a unique anycast gateway IP, however multiple anycast gateway IPs may map to the same anycast MAC. It is recommend to configure only one anycast MAC for all VNs on each gateway device for simplicity purpose. Each physical gateway performs layer 2 Ethernet traffic termination when the inner destination MAC of the incoming frame equal to its anycast MAC.

To support layer 3 traffic load-balancing among all gateways, besides each layer 3 gateway's own nickname, anycast nickname is introduced, multiple gateways share the same nickname. Each gateway announces anycast nickname through the Nickname Sub-Tlv specified in [RFC6326] to TRILL network and MUST ignore the nickname collision check as defined in basic TRILL protocol. The anycast nickname used by the gateway should be set to the highest priority. With such setting, in case some other RBridge tries to use the same nickname, the gateway can always win in the nickname conflicts.

Besides anycast nickname/IP/MAC, each physical gateway also has its own gateway IP and MAC for each VN and its own nickname.

The source MAC of ARP reply when responding to ARP request for anycast IP from ES is always the anycast MAC. Ingress nickname should be anycast nickname when the ARP reply message is a unicast

TRILL frame. For proactive ARP request from a gateway to ES, source MAC is the gateway's own MAC. In this case ingress nickname in TRILL header should be the gateway's own nickname. Edge nodes i.e. ToRs learn the consistent correspondence of anycast MAC and anycast nickname and correspondence of gateway's physical MAC and nickname through normal data plane learning mechanism.

An ES has no knowledge that MAC address it gets for a gateway is actually an address for anycast purpose. The ES operates in normal way. The ES acquires correspondence between anycast MAC and anycast IP through normal ARP procedures. When the ES tries to send traffic cross subnets, it will send the frame to the gateway first. The anycast MAC is used by the end system as destination MAC. As edge nodes, ToRs in this case, learn the consistent correspondence of anycast MAC and nickname for gateway beforehand, frame from the end host sending to the gateway could go to any one of the gateways by the natural support of ECMP from TRILL protocol. The workload is well spread over all the core switches. When one gateway fails, the rest could seamlessly take over the workload automatically without running any VRRP-like keepalive protocol in between.

It should not be allowed to telnet each physical gateway using the anycast IP address. The information exchange in a single telnet session may indeed go to the different physical gateways when the anycast gateway IP address is used for telnet. Consequently the state machine at the telnet initiator side may be in unpredictable and disordered states. To overcome this ,it is recommended to use gateway's own physical IP for telnet. ARP tables age independently on each physical gateways. A physical gateway should use its own MAC to send ARP request message to all ES belonging to a VN in proactive mode to acquire destination ES's ARP table. The source MAC of ARP request message should be the gateway's own MAC instead of anycast MAC, the destination ES uses the physical gateway's own MAC as destination MAC to send ARP reply message. Through this mode, the ARP reply message from destination ES can be ensured to reach the physical gateway. Inter-subnet traffic from gateway to ES can use either the gateway's own physical MAC or anycast MAC as source MAC.

4.1. ARP Handling

Before an ES begins inter-subnet communication, it sends ARP request to ask the MAC address of the gateway. As the ES uses the anycast gateway IP as the target address, all physical layer 3 gateways could possibly respond it. To avoid duplicate ARP reply sending to the end system, only one physical gateway should be elected to respond. The physical gateway that responds to ARP request message

is called ARP master gateway. Assuming there are k physical gateways, the algorithm to elect ARP master gateway for each VN is as follows:

1. All physical gateways are ordered and numbered from 0 to k-1 in ascending order according to the 7-octet IS-IS ID.

2. For VN ID m, choose RB whose number equals (m mod k) as ARP master gateway.

The algorithm guarantees each VN has a consistent ARP master gateway. Only ARP master gateway sends ARP reply to an ES's ARP request for that VN. The rest gateways should ignore the ARP request.

Sender protocol address (SPA) and Sender hardware address (SHA) in the ARP reply message is set as anycast IP address and anycast MAC address. The ARP reply message is unicast TRILL encapsulated and sent to the ES. Ingress nickname should be anycast nickname. Egress nickname is set as the nickname of egress RB connecting to the ES.

As ES broadcasts ARP request message to TRILL campus, all physical gateways can learn the correspondence of <ES MAC, ES IP, VN ID, Ingress Nickname> from the frame. Gateways can use this information to generate IP forwarding table for that ES.

In summary, through the above ARP process:

1. Edge RBs i.e. TORs learn anycast MAC address associating with anycast nickname.

2. ES learns the anycast MAC address associating with anycast gateway IP.

All physical gateways learn the (ES MAC, ES IP and connected edge RB nickname) for all end systems. ARP tables age independently on each layer 3 gateway. To avoid the unnecessary flooding due to ARP table aging, the layer 3 gateway should send ARP detection message periodically in proactive mode to refresh the ARP table state. In this case, source MAC in inner Ethernet header and Sender hardware address (SHA) in the ARP request message is suggested to use the gateway's own MAC, ingress nickname is suggested to use the gateway's own nickname when it is unicast TRILL encapsulated. When the ES receives the ARP request message, ES returns unicast ARP reply message, destination MAC is the layer 3 gateway's own MAC. The message will only reach the layer 3 gateway. When the edge RB connecting the ES receives the ARP reply message, the edge RB will forward the packet to the ARP request sending layer 3 gateway.

4.2. Data traffic forwarding

After an ES acquires anycast MAC associated with anycast IP through above ARP handling process, it can start to send the inter-subnet IP traffic. Assuming ES1 tries to send data to ES4 in figure1. They belong to different subnet. The IP traffic forwarding process is as following:

1. ES1 sends unicast IP traffic to ES4. Destination IP is ES4’s IP address, destination MAC is anycast gateway’s MAC.

2. TOR1 receives the message from ES1. Because TOR1 has already learned anycast MAC address associating with anycast nickname through above ARP process, so it sends the packet with unicast TRILL encapsulation, egress nickname in TRILL header is anycast nickname. The TRILL data will reach one of the physical gateways through ECMP. Assuming the TRILL data reaches GW1.

3. GW1 receives the TRILL data from TOR1. It decapsulates the frame and get native packet. It looks up local IP forwarding table based on destination IP and tries to forward the packet to ES4. If entry of <ES4 MAC, ES4 IP, VLAN2, Nickname of TOR2> was stored on GW1, GW1 encapsulates the frame based on the information and sends it to the egress RB. The source MAC can be the gateway’s own MAC or anycast MAC. If the gateway’s own MAC is used as source MAC,ingress nickname of TRILL frame should be GW1’s own nickname. If anycast MAC is used, ingress nickname should be anycast nickname.(If the entry is not available on GW1, the gateway will send ARP Request message to ES4 proactively.)

4. TOR2 receives the TRILL data from GW1. It decapsulates the frame and forward the payload to ES4.

All layer 3 traffic will be processed in a flow-based load balancing mode among all physical gateways. Anycast gateway achieves better bandwidth utilization and scalability compared to VRRP-like mechanism.

5. Node failure

When one of the layer 3 gateways fails, after network convergence, the TRILL traffic to anycast nickname will only reach the remaining gateways. ARP master gateway will be re-elected among the remaining gateways. No VRRP-like protocol session among layer 3 gateways is required to detect the node failure. Network convergence relies purely on TRILL protocol.

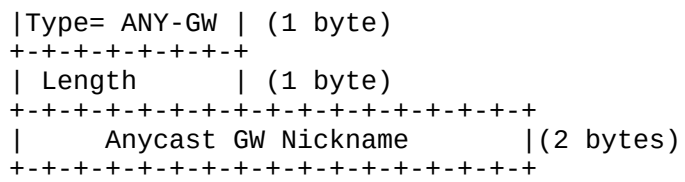
6. Anycast MAC aging on edge node

If anycast MAC aged on an edge node, when the edge node receives inter-subnet traffic from connecting ES, the edge node will flood the unicast traffic to TRILL campus as unknown unicast traffic. All physical gateways will receive the traffic, only one of the physical gateways should forward it, all others should drop it to avoid forwarding duplicated data to destination ES. The forwarding gateway is suggested to be same with ARP master device.

7. TRILL protocol extension

All layer 3 gateways should announce the anycast gateway TLV in LSP defined in section 6.1 to TRILL campus. Each gateway receiving the anycast gateway TLV from other RBs with the same anycast GW nickname thinks they are in one anycast gateway group. All the gateways should ensure the anycast nickname configuration consistency. If the anycast nickname is different from the local configured one, configuration error occurs and a network warning or SNMP trap should be sent to the network management system. Anycast nickname also is carried in the Nickname Sub-Tlv specified in [RFC6326], each gateway MUST ignore the nickname collision check for anycast nickname.

7.1. The Anycast Gateway TLV



o Type: TLV Type, TBD.

o Length: indicates the length of LAGID field, it is a fixed value of 1.

o Anycast GW Nickname: the nickname is shared by all the physical gateways in the anycast gateway group. All the inter-subnet traffic to the anycast gateways MUST use the nickname as egress nickname in TRILL header.

8. Security Considerations

The default value of anycast nickname priority should be set as highest value. If nickname on non-gateway and anycast nickname on gateways occurs collision, it can minimize the probability to modify anycast nickname.

9. IANA Considerations

TBD

10. Normative References

[1] [RFC6165] Banerjee, A. and D. Ward, "Extensions to IS-IS for Layer-2 Systems", RFC 6165, April 2011.

[2] [RFC6325] Perlman, R., et.al. "RBridge: Base Protocol Specification", RFC 6325, July 2011.

[3] [RFC6326bis] Eastlake, D., Banerjee, A., Dutt, D., Perlman, R., and A. Ghanwani, "TRILL Use of IS-IS", draft-eastlake-isis-rfc6326bis, work in progress.

11. Informative References

[4] [RFC 3768] R. Hinden, Ed., "Virtual Router Redundancy Protocol (VRRP)", RFC 3768, April 2004.

12. Acknowledgments

The authors wish to acknowledge the important contributions of Zhang Chengsong.

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

Donald E. Eastlake
Huawei Technologies
155 Beaver Street
Milford, MA 01757 USA
Phone: +1-508-333-2270
EMail: d3e3e3@gmail.com

Radia Perlman
Intel Labs
2200 Mission College Blvd.
Santa Clara, CA 95054-1549 USA
Phone: +1-408-765-8080
EMail: Radia@alum.mit.edu