TRILL

Weiguo Hao
Yizhou Li
Tao Han
Huawei

Internet Draft
Intended status: Standards Track
Expires: June 2014

December 11,2013

Centralized Replication for BUM traffic in active-active edge connection

draft-hao-trill-centralized-replication-00.txt

Status of this Memo

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 June 11, 2014.

Copyright Notice

Abstract

In TRILL active-active access scenario, RPF check failure issue may occur when pseduonode nickname mechanism in [TRILLPN] is used. This draft is to solve RPF check failure issue through centralized replication for BUM traffic solution. The basic idea is that all ingress RBs send BUM traffic to a centralized node through unicast TRILL encapsulation, the centralized node decapsulates the unicast TRILL packet. Then the centralized node searches its multicast forwarding table to replicates a copy of the BUM traffic to destination RBs through TRILL unicast encapsulation. Through centralized replication solution, only unicast forwarding behavior is required between edge RB and centralized RB, so no RPF check function

is required on any device in TRILL campus through centralized replication solution and no RPF check failure issue occurs.

Table of Contents

1. Introduction

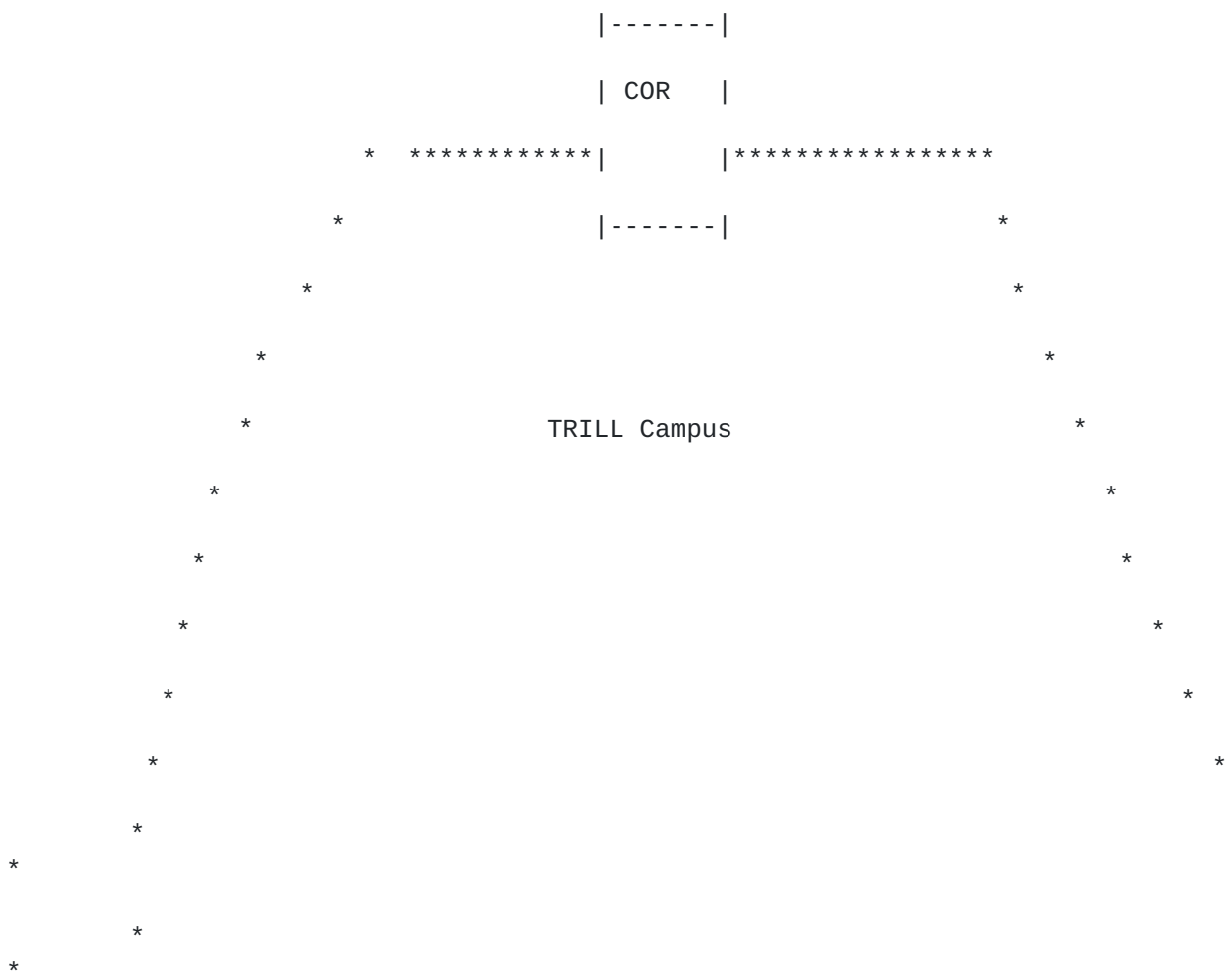
The IETF TRILL (Transparent Interconnection of Lots of Links) [RFC6325] protocol provides loop free and per hop based multipath data forwarding with minimum configuration. TRILL uses IS-IS [RFC6165] [RFC6326bis] as its control plane routing protocol and defines a TRILL specific header for user data.

Customer edge(CE) devices typically are multi-homed to several RBridges which form an edge group. All of the uplinks of CE is considered as an Multi-Chassis Link Aggregation (MC-LAG) bundle. An active-active flow-based load-sharing mechanism is implemented to achieve better load balancing and high reliability. A CE device can be a layer3 end system by itself or a bridge switch through which layer3 end systems are accessed to TRILL campus.

In active-active access scenario, pseduonode nickname solution can be used to avoid mac flip-flop on remote RBs. The basic idea is to represent all member links of the MC-LAG as a virtual RBridge with single pseduonode nickname. Any member RBridge of the MC-LAG should use this pseduonode nickname rather than its own nickname as ingress nickname when inject TRILL data frames. It solves the abovementioned

problems pretty well; however, it introduces another issue: packet drop due to RPF check.

This document proposes a centralized replication solution for broadcast, unknown unicast, multicast(BUM) traffic to solve the issue of packet drop due to RPF check. The basic idea is that all ingress RBs send BUM traffic to a centralized node through unicast TRILL encapsulation, the centralized node decapsulates the unicast TRILL packet. Then the centralized node searches its multicast forwarding table to replicates a copy of the BUM traffic to destination RBs through TRILL unicast encapsulation. Through centralized replication solution, only unicast forwarding behavior is required between edge RB and centralized RB, so no RPF check function is required on any device in TRILL campus through centralized replication solution and no RPF check failure issue occurs.



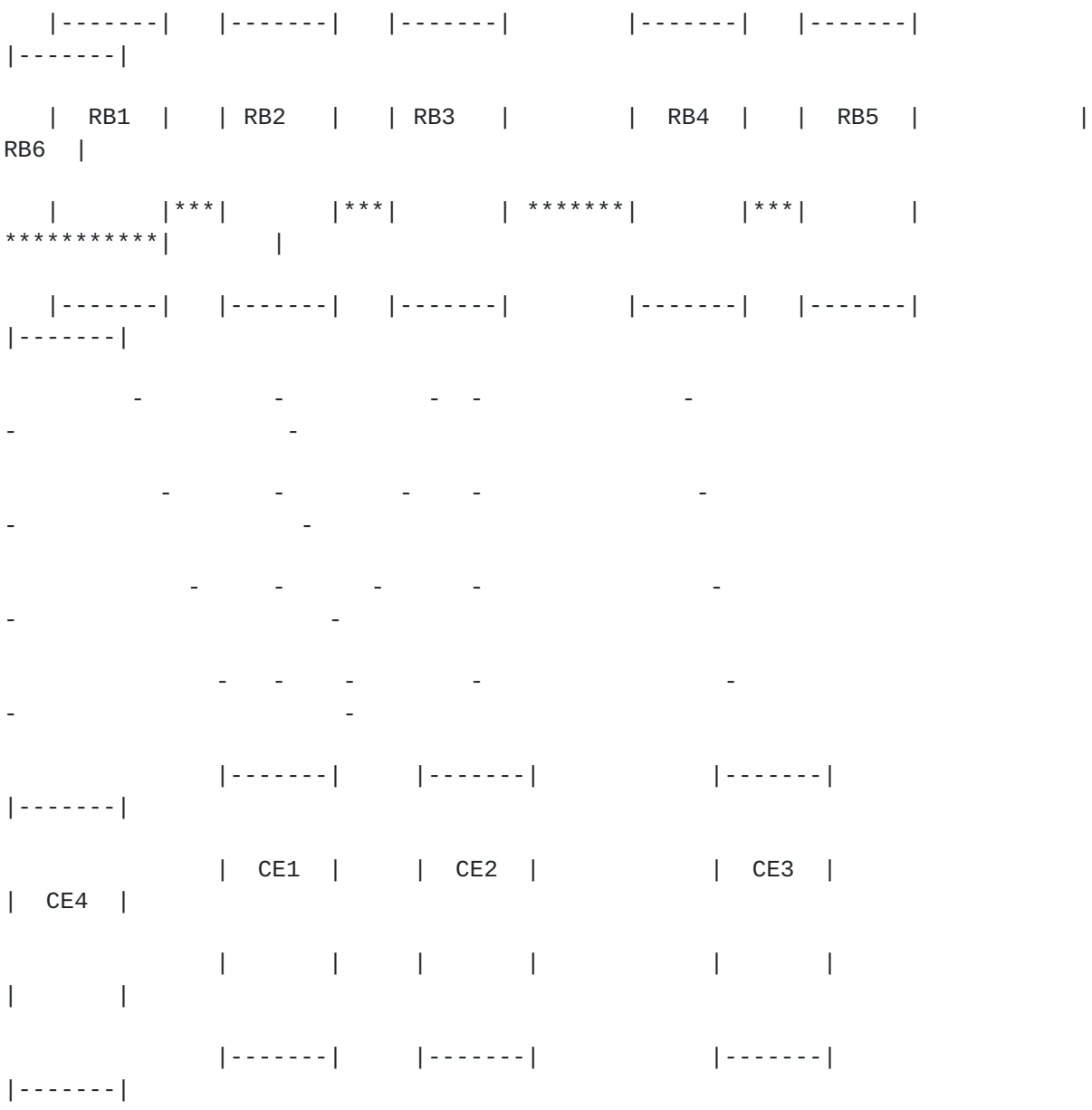


Figure 1 TRILL active-active access

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 acronyms and terminology in [RFC6325] is used herein with the following additions:

CE - customer equipment. Could be a bridge or end station.

3. Centralized Replication Solution Overview

When an edge RB receives BUM traffic from CE device, it acts as ingress RB and uses unicast TRILL encapsulation instead of multicast TRILL encapsulation to send the traffic to a centralized node.

The TRILL header of the unicast TRILL encapsulation contains an "ingress RBridge nickname" field and an "egress RBridge nickname" field. If ingress RB participates active-active access connection,

the ingress RBridge nickname should be set as pseduonode nickname to avoid mac flip-flop on remote RBs, otherwise the ingress nickname should be set as its own nickname. The egress RBridge nickname is the nickname of the centralized node.

When the centralized node receives the unicast TRILL encapsulated BUM traffic from ingress RB, the node decapsulates the packet. Then the centralized node searches its multicast forwarding table to replicate a copy of the BUM traffic to destination RBs through TRILL unicast encapsulation. Ingress nickname for the TRILL unicast encapsulation is unchangeable and is still the nickname of ingress RB. Egress nickname for each copy of TRILL unicast encapsulation is the nickname corresponds to each destination RB. The destinations are all edge RBs except the ingress RB itself. If ingress RB participates active-active access connection, the virtual RB that represent all member links of the MC-LAG acts as real ingress RB and the physical multi-homed edge RB acts as transit RB, the destinations RBs should exclude the virtual RB.

If destination RB participates active-active access connection, the egress nickname in TRILL unicast encapsulation should be set as pseduonode nickname to represent a MC-LAG corresponding a multi-homed destination CE device, the TRILL unicast encapsulation BUM traffic will reach destination CE device through any one of multi-homed edge RBs to realize flow-based loadbalancing.

To differentiate the unicast TRILL encapsulation BUM traffic from normal unicast TRILL traffic, a special nickname on centralized node should be used exclusively for centralized replication. Only when centralized node receives unicast TRILL encapsulation traffic with egress nickname equivalent to the special nickname, the node does TRILL decapsulaton and performs corresponding replication procedures. The centralized nodes should announce its special use nickname to all TRILL campus.

In default mode BUM traffic on edge RB is forwarded along distribution tree established by TRILL base protocol. If at least a centralized node exists in TRILL campus, ingress RB can use the centralized replication mode instead of default mode to forward BUM traffic to TRILL campus.

4. Centralized Replication Forwarding Process

As described in figure1, CE1 is multi-homed to RB1,RB2 and RB3 in active-active mode, CE2 is single homed to RB3, CE3 is multi-homed to RB4, RB5 in active-active mode, CE4 is single homed to RB6. CE1,CE2,CE3 and CE4 belong to same CE-VLAN. COR is a centralized

replication node which can forward all BUM traffic in the TRILL campus. The pseduonode nickname corresponds to CE1 is p-nickname1. The pseduonode nickname corresponds to CE3 is p-nickname2. The nicknames of RB1 to RB6 are nickname1 to nickname6. The nickname on COR for centralized replication is S-nickname, the nickname on COR for normal unicast TRILL forwarding is N-nickname.

The BUM traffic forwarding process from CE1 to CE2,CE3 and CE4 is as follows:

1. CE1 sends BUM traffic to RB2.

2. RB2 sends the BUM traffic to COR device through unicast TRILL encapsulation. Ingress nickname is set as p-nickname1, egress nickname is set as S-nickname.

3. COR decapsulates unicast TRILL encapsulation. Then the centralized node searches its multicast forwarding table to replicate a copy of the BUM traffic to CE2,CE3 and CE4 through unicast TRILL encapsulation. The egress nickname of the traffic to CE2 is set as nickname3, the egress nickname to CE3 is set as p-nickname2, while the egress nickname of the traffic to CE4 is set as nickname6. Ingress nickname is p-nickname1.

4. RB3 receives the unicast TRILL encapsulation BUM traffic from COR. It decapsulates the unicast TRILL packet. Because ingress nickname of p-nickname1 is equivalent to the nickname of local MC-LAG connecting CE1, RB3 doesn't forward the packet to CE2 to avoid loop. RB3 only forwards the packet to CE3.

5. RB4(or RB5) receives the unicast TRILL encapsulation BUM traffic from COR. It decapsulates the unicast TRILL packet, then forwards the packet to CE2. The MAC of CE1 associated with p-nickname1 is learned on RB4(or RB5).

RB6 receives the unicast TRILL encapsulation BUM traffic from COR1. It decapsulates the unicast TRILL packet, then forwards the packet to CE3. The MAC of CE1 associated with p-nickname1 is learned on RB6.

5. Multicast forwarding table generation on the centralized node

There are two kinds of multicast forwarding tables of pruned and un-pruned on centralized node. Through un-pruned multicast forwarding table, all edge RBs are connected through the centralized node, BUM traffic from any one ingress RB is sent to all other edge RBs exclude ingress RB, some egress RB maybe receive extra BUM traffic if the RB has no local BUM traffic receiver.

To save the link bandwidth in TRILL campus, pruned multicast forwarding table should be used. The multicast forwarding table that connecting all edge RBs should be pruned based on CE-VLAN, centralized node relies on the pruned multicast table to replicate and forward BUM traffic to each destination RB that have the same VLAN attached. Similarly, the multicast forwarding table also can be pruned based on FGL or multicast group.

Centralized node makes a local decison on whether to generate a pruned or un-pruned multicast forwarding table.There is only one forwarding item in un-pruned multicast forwarding table. No key is required for un-pruned multicast forwarding table. There are multiple multicast forwarding items in pruned multicast forwarding table and each item corresponds to a CE-VLAN(or FGL/Multicast group). For pruned multicast forwarding table, the key is CE-VLAN(or FGL/Multicast group) which is obtained from original BUM traffic.

Centralized node relies on TRILL based protocol to acquire edge RB information and generate multicast forwarding tables. Each edge RBridge specifies the VLAN it is interested in the Interested VLANs and Spanning Tree Roots (INT-VLAN) sub-TLV [RFC6326],while the Multicast Group it is interested is specified in The Group Address TLV [RFC6326]. Centralized node can learn all access VLAN and multicast group information from edge RBs.

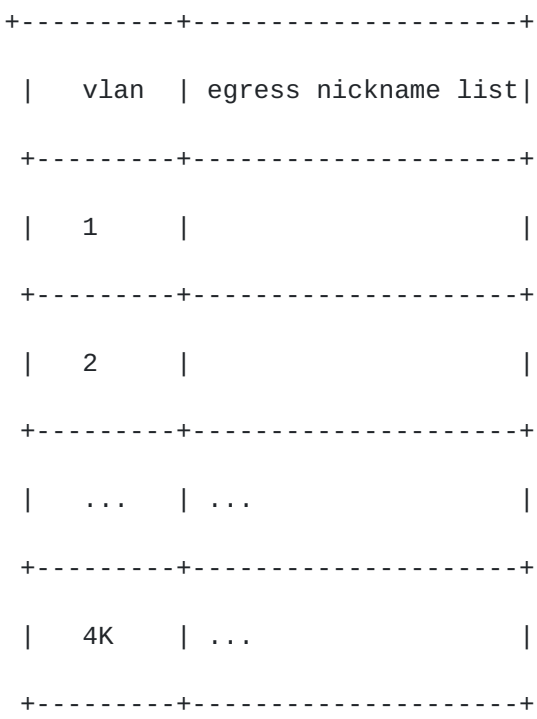


Figure 2 VLAN pruning multicast forwarding table on the centralized node

6. BUM traffic loadbalancing among multiple centralized nodes

To support unicast TRILL encapsulation traffic loadbalancing, multiple RBs can serve as centralized replication node and the traffic can be loadbalanced on these RBs in flow-based mode.

To support flow-based loadbalancing for BUM traffic between different centralized node, virtual centralized node mechanism should be introduced. The virtual centralized node is logically directly connected to both physical centralized node with equal link cost, BUM traffic special use nickname is attached to this virtual centralized node.

The egress nickname of unicast TRILL encapsulation for BUM traffic from ingress RB is the special use nickname attached to the virtual centralized node. The unicast TRILL encapsulation BUM traffic is transmitted in TRILL campus hop by hop until it reaches any one of the physical centralized node that logically connecting to the virtual centralized node. The physical centralized node will decapsulate the unicast TRILL encapsulation and performs centralized replication multicast forwarding procedures. Because ECMP of the unicast TRILL encapsulation BUM traffic is supported, so it can achieve better link bandwidth usage than VLAN-based(or FGL-based,etc) loadbalancing.

7. Link failure process

For member RBridges of a virtual RBridge occurs link (all member link of LAG) failure, the member RBridge should announce disconnection between this node and pseudonode, the member RBridge leaves the virtual RBridge. If a pseduonode nickname only represents a MC-LAG , because the member RBridge has left its corresponding virtual RBridge, so the unicast TRILL encapsulation BUM traffic with egress nickname equals to pseduonode nickname won't reach the member RBridge with failed access link, it can only reach normal member RBridges, the normal member RBridges then forward the original BUM traffic to local access link.

If a pseduonode nickname represents multiple MC-LAG, the unicast TRILL encapsulation BUM traffic with egress nickname equals to Pseduonode nickname can reach the member RBridge with failed access link, to forward the BUM traffic to local access CE device, the member RBridge should tunnel the BUM traffic to any one of remaining normal RBridges in the virtual RBridge in unicast TRILL encapsulation, destination port ID should be specified in the encapsulation in order

to destination RBridge can forward BUM traffic only to a local port connecting to corresponding CE. This solution is complex, so the method which a Pseduonode nickname only represents single MC-LAG is suggested!

8. Node Failure Process

If a member RBridges of a virtual RBridge occurs node failure, after TRILL network converges, the unicast TRILL encapsulation BUM traffic with egress nickname equals to pseduonode nickname won't reach the failed RBridge, it only reach any one of the remaining member RBridges of same virtual RBridge with the failed RBridge.

9. Enhanced Solution

Centralized replication solution can be deployed with TRILL distribution tree mechanism defined in TRILL base protocol at the same time. In this case, distribution tree root node acts as the centralized replication node.

Unicast TRILL encapsulation for BUM traffic is only used for ingress RB participating active-active connection to avoid RPF check failure issue. When a centralized node receives unicast TRILL encapsulation BUM traffic from the ingress RB, it decapsulates the unicast TRILL packet. Then it replicates and forwards the BUM traffic to all other destination RBs through any one of a distribution tree established per TRILL base protocol.

Assuming the enhanced solution is used in the network of above figure1, the BUM traffic forwarding process from CE1 to CE2,CE3 and CE4 is as follows:

1. CE1 sends BUM traffic to RB2.

2. RB2 sends the BUM traffic to COR device through unicast TRILL encapsulation. Ingress nickname is set as p-nickname1, egress nickname is set as S-nickname.

3. COR device decapsulates the unicast TRILL packet. Then it selects a distribution tree to forward the packet to all other destination RBs. The egress nickname in the trill header is the nickname of distribution tree root.

4. RB3 receives multicast TRILL traffic from COR. It decapsulates the multicast TRILL packet. Because ingress nickname of p-nickname1 is equivalent to the nickname of local MC-LAG connecting CE1, RB3 doesn't forward the packet to CE1 to avoid loop. RB3 only forwards the packet to CE2.

5. RB3 receives multicast TRILL traffic from COR. It decapsulates the multicast TRILL packet. Then it forwards the packet to CE3. The MAC of CE1 associated with p-nickname1 is learned on RB4(or RB5).

RB6 receives multicast TRILL traffic from COR. It decapsulates the multicast TRILL packet. Then it forwards the packet to CE4. The MAC of CE1 associated with p-nickname1 is learned on RB6.

10. Network Migration Analysis

Centralized node should upgrade to support centralized replication process.

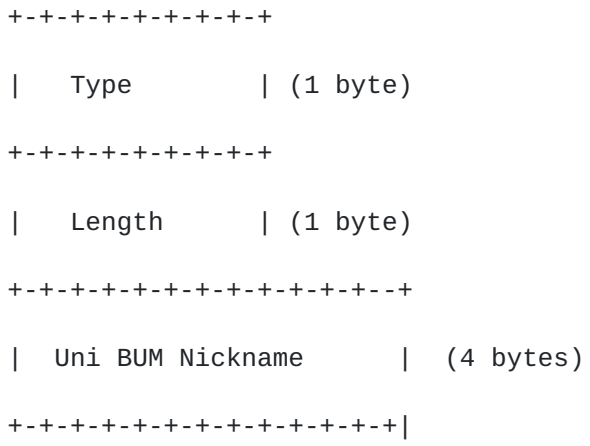
Edge RBs participating centralized replication process also should upgrade, although the forwarding process is similar to normal head-end replication process.

Transit nodes don't need upgrade.

11. TRILL protocol extension

The Unicast BUM Nickname TLV is introduced to announce its special use nickname for centralized replication by centralized node. It is carried in an LSP PDU. Ingress RBs rely on the TLV to learn the egress nickname of TRILL unicast encapsulation for BUM traffic.

11.1. The Unicast BUM Nickname sub-TLV



o Type: Router Capability sub-TLV type, TBD (Uni-BUM-VLANs).

o Length: indicates the length of Uni BUM Nickname field, it is a fixed value of 4.

o Uni BUM Nickname: The nickname is exclusively used for centralized replication solution purpose. Ingress RBs use the nickname as egress nickname in trill header of unicast TRILL encapsulation for BUM traffic.

12. Security Considerations

This draft does not introduce any extra security risks. For general TRILL Security Considerations, see [RFC6325].

13. IANA Considerations

TBD

14. References

14.1. 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.

14.2. Informative References

[4] [TRILLPN] Zhai,H., et.al., "RBridge: Pseduonode nickname", draft-hu-trill-pseudonode-nickname, Work in progress, November 2011.

[5] [TRILAA] Li,Y., et.al., "Problems of Active-Active connection at the TRILL Edge", draft-yizhou-trill-active-active-connection-prob-00, Work in progress, July 2013.

15. Acknowledgments

The authors wish to acknowledge the important contributions of Xiaomin Wu.

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

Tao Han
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China
Phone: +86-25-56623454
Email: billow.han@huawei.com