Internet Working Group
Internet Draft
Category: Standards Track

Ali Sajassi
Samer Salam
Cisco

Aldrin Issac
Bloomberg

Nabil Bitar
Verizon

Sam Aldrin
Huawei

Expires: April 1, 2015

October 1, 2014

TRILL-EVPN
draft-ietf-l2vpn-trill-evpn-02

Status of this Memo

Copyright and License Notice

Abstract

This document discusses how Ethernet VPN (E-VPN) technology is used to interconnect TRILL [TRILL] networks over an MPLS/IP network, with two key characteristics: C-MAC address transparency on the hand-off point and control-plane isolation among the interconnected TRILL networks.

Conventions

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.

Table of Contents

1. Introduction

[E-VPN] introduces a solution for multipoint L2VPN services, with advanced multi-homing capabilities, using BGP for distributing customer/client MAC address reach-ability information over the core MPLS/IP network. [TRILL] defines a solution for optimal forwarding of Ethernet frames with support for multipathing of unicast and multicast traffic, using IS-IS control-plane and associated tunneling encapsulation that includes a hop count. In this document, we discuss how Ethernet VPN (E-VPN) technology can be used to interconnect TRILL [TRILL] networks over an MPLS/IP network, while guaranteeing two key characteristics: C-MAC address transparency on the hand-off point and control-plane isolation among the interconnected TRILL networks. The resulting solution is referred to as TRILL-EVPN.

2. Contributors

In addition to the authors listed above, the following individuals also contributed to this document.

Keyur Patel Cisco
Tissa Senevirathne Cisco

3. Terminology

CE: Customer Edge
C-MAC: Customer/Client MAC Address
DHD: Dual-homed Device
DHN: Dual-homed Network
E-VPN: Ethernet VPN
LACP: Link Aggregation Control Protocol
LSM: Label Switched Multicast
MDT: Multicast Delivery Tree
MES: MPLS Edge Switch
MP2MP: Multipoint to Multipoint
P2MP: Point to Multipoint
P2P: Point to Point
PoA: Point of Attachment
PW: Pseudowire
TRILL: Transparent Interconnect of Lots of Links

4. Requirements

4.1. C-MAC Address Transparency on the Hand-off Point

[TRILL] addresses the problem of MAC address table scalability on

intermediate switching devices by introducing a tunneling technology and confining C-MAC address learning/forwarding to the edge of the TRILL network. When TRILL networks are interconnected over an MPLS/IP network, it is required to maintain C-MAC address transparency on the hand-off point and the edge (i.e. MES) of the MPLS network. Otherwise, the MPLS edge nodes may suffer from MAC address table space exhaustion given that they would need to learn the C-MAC addresses from all interconnected TRILL networks.

TRILL-EVPN supports seamless interconnect with TRILL while guaranteeing C-MAC address transparency on the MES nodes.

4.2. Control Plane Isolation among TRILL Networks

It is required to maintain control-plane isolation among the various TRILL networks being interconnected over the MPLS/IP network. This ensures the following characteristics:

- scalability of the IS-IS control plane in large deployments. In TRILL, all nodes must calculate the shared multicast trees, so as the number of interconnected cloud networks scales, this places a burden on the RBridges, especially if roots are to be located in every site to ensure optimality of the data-path.

- fault domain localization, where link or node failures in one site do not trigger SPF re-computations in remote sites.

Interconnect solutions which extend the IS-IS control-plane over the MPLS network, as an overlay, do not meet this requirement. TRILL-EVPN provides control-plane isolation between interconnected TRILL networks by terminating the TRILL IS-IS at the MPLS edge nodes.

5. Solution Overview

TRILL-EVPN enables seamless connectivity of TRILL networks over an MPLS/IP core while ensuring control-plane separation among these networks, and maintaining C-MAC address transparency on the MES nodes.

Every TRILL network that is connected to the MPLS core runs an independent instance of the IS-IS control-plane. Each MES participates in the TRILL IS-IS control plane of its local site. The MES peers, in IS-IS protocol, with the RBridges internal to the site, but does not terminate the TRILL data-plane encapsulation. So, from a control-plane viewpoint, the MES appears as an edge RBridge; whereas, from a data-plane viewpoint, the MES appears as a core RBridge to the TRILL network. The MES nodes encapsulate TRILL frames with MPLS in the imposition path, and de-capsulate them in the disposition path.

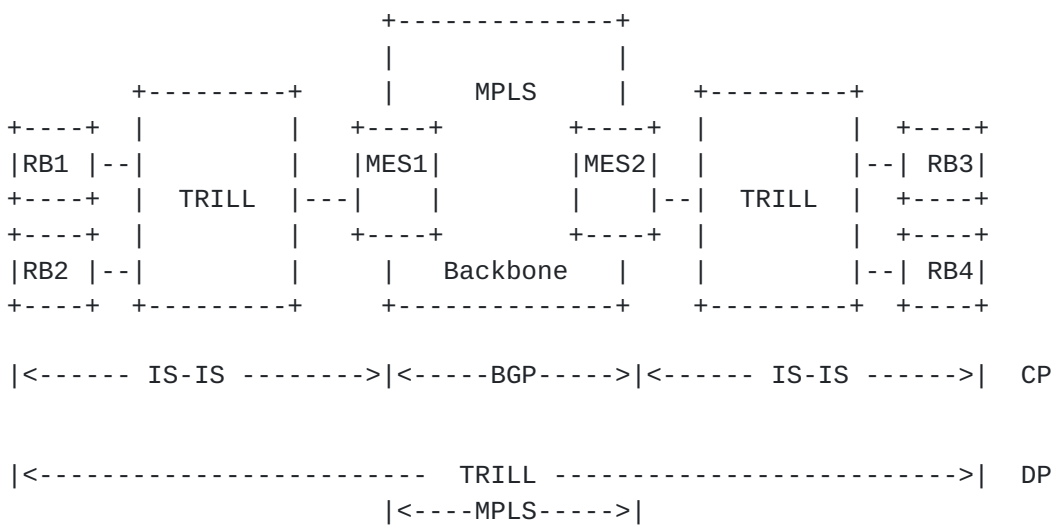

Figure 1: Interconnecting TRILL Networks with TRILL-EVPN

5.1. TRILL Nickname Assignment

In TRILL, edge RBridges build forwarding tables that associate remote C-MAC addresses with remote edge RBridge nicknames via data-path learning (except if the optional ESADI function is in use). When different TRILL networks are interconnected over an MPLS/IP network using a seamless hand-off, the edge RBridges (corresponding to the ingress and egress RBridges of particular traffic flows) may very well reside in different TRILL networks. Therefore, in order to guarantee correct connectivity, the TRILL Nicknames must be globally unique across all the interconnected TRILL islands in a given EVI. This can be achieved, for instance, by using a hierarchical Nickname assignment paradigm, and encoding a Site ID in the high-order bits of the Nickname:

Nickname = [Site ID : Rbridge ID]

The Site ID uniquely identifies a TRILL network, whereas the RBridge ID portion of the Nickname has local significance to a TRILL site, and can be reused in different sites to designate different RBridges. However, the fully qualified Nickname is globally unique in the entire domain of interconnected TRILL networks for a given EVI.

It is worth noting here that this hierarchical Nickname encoding scheme guarantees that Nickname collisions do not occur between different TRILL islands. Therefore, there is no need to define TRILL Nickname collision detection/resolution mechanisms to operate across

separate TRILL islands interconnected via TRILL-EVPN.

Another point to note is that there are proposals to achieve per-site Nickname significance; however, these proposals either require C-MAC learning on the border RBridge (i.e. violate the C-MAC address transparency requirement), or require a completely new encapsulation and associated data-path for TRILL [TRILL-MULTILEVEL].

5.2. TRILL Nickname Advertisement Route

A new BGP route is defined to support the interconnection of TRILL networks over TRILL-EVPN: the TRILL Nickname Advertisement' route, encoded as follows:

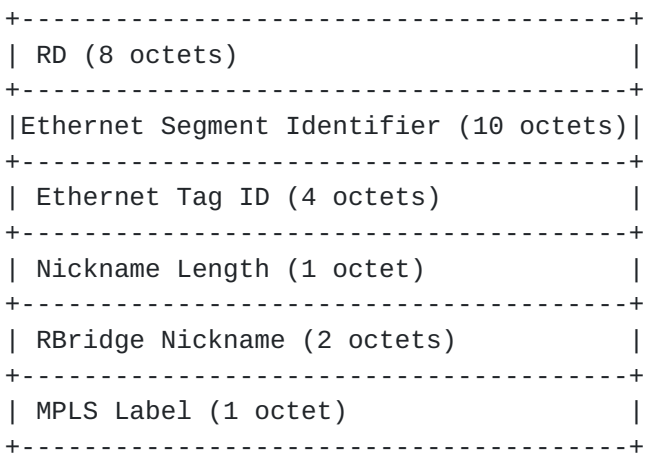

Figure 2: TRILL Nickname Advertisement Route

The MES uses this route to advertise the reachability of TRILL RBridge nicknames to other MES nodes in the EVI. The MPLS label advertised in this route is allocated on a per EVI basis and serves the purpose of identifying to the disposition MES that the MPLS-encapsulated packet holds an MPLS encapsulated TRILL frame.

5.3. Frame Format

The encapsulation for the transport of TRILL frames over MPLS is encoded as shown in the figure below:

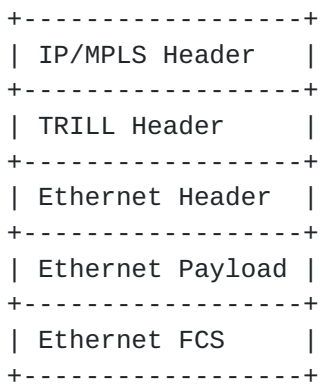

Figure 3: TRILL over MPLS Encapsulation

It is worth noting here that while it is possible to transport Ethernet encapsulated TRILL frames over MPLS, that approach unnecessarily wastes 16 bytes per packet. That approach further requires either the use of well-known MAC addresses or having the MES nodes advertise in BGP their device MAC addresses, in order to resolve the TRILL next-hop L2 adjacency. To that end, it is simpler and more efficient to transport TRILL natively over MPLS, and this is the reason why a new BGP route for TRILL Nickname advertisement is defined.

5.4. Unicast Forwarding

Every MES advertises in BGP the Nicknames of all RBridges local to its site in the TRILL Nickname Advertisement routes. Furthermore, the MES advertises in IS-IS, to the local island, the Rbridge nicknames of all remote switches in all the other TRILL islands that the MES has learned via BGP. This is required since TRILL [RFC6325] currently does not define the concept of default routes. However, if the concept of default routes is added to TRILL, then the MES can advertise itself as a border RBridge, and all the other Rbridges in the TRILL network would install a default route pointing to the MES. The default route would be used for all unknown destination Nicknames. This eliminates the need to redistribute Nicknames learnt via BGP into TRILL IS-IS.

Note that by having multiple MES nodes (connected to the same TRILL island) advertise routes to the same RBridge nickname, with equal BGP Local_Pref attribute, it is possible to perform active/active load-balancing to/from the MPLS core.

When a MES receives an Ethernet-encapsulated TRILL frame from the access side, it removes the Ethernet encapsulation (i.e. outer MAC header), and performs a lookup on the egress RBridge nickname in the TRILL header to identify the next-hop. If the lookup yields that the

next hop is a remote MES, the local MES would then encapsulate the TRILL frame with appropriate MPLS label stack. The label stack comprises of the VPN label (advertised by the remote MES), followed by an LSP/IGP label. From that point onwards, regular MPLS forwarding is applied.

On the disposition MES, assuming penultimate-hop-popping is employed, the MES receives the MPLS-encapsulated TRILL frame with a single label: the VPN label. The value of the label indicates to the disposition MES that this is a TRILL packet, so the label is popped, the TTL field (in the TRILL header) is reinitialized and normal TRILL processing is employed from this point onwards.

5.5. Handling Multicast

Each TRILL network independently builds its shared multicast trees. The number of these trees need not match in the different interconnected TRILL islands. In the MPLS/IP network, multiple options are available for the delivery of multicast traffic:

- Ingress replication
- LSM with Inclusive trees
- LSM with Aggregate Inclusive trees
- LSM with Selective trees
- LSM with Aggregate Selective trees

When LSM is used, the trees may be either P2MP or MP2MP.

The MES nodes are responsible for stitching the TRILL multicast trees, on the access side, to the ingress replication tunnels or LSM trees in the MPLS/IP core. The stitching must ensure that the following characteristics are maintained at all times:

1. Avoiding Packet Duplication: In the case where the TRILL network is multi-homed to multiple MES nodes, if all of the MES nodes forward the same multicast frame, then packet duplication would arise. This applies to both multicast traffic from site to core as well as from core to site.

2. Avoiding Forwarding Loops: In the case of TRILL network multi-homing, the solution must ensure that a multicast frame forwarded by a given MES to the MPLS core is not forwarded back by another MES (in the same TRILL network) to the TRILL network of origin. The same applies for traffic in the core to site direction.

3. Pacifying TRILL RPF Checks: For multicast traffic originating from a different TRILL network, the RPF checks must be performed against the disposition MES (i.e. the MES on which the traffic ingress into

the destination TRILL network).

There are two approaches by which the above operation can be guaranteed: one offers per-source load-balancing while the other offers per-flow load-balancing.

5.5.1. Multicast Stitching with Per-Source Load Balancing

The MES nodes, connected to a multi-homed TRILL network, perform BGP DF election to decide which MES is responsible for forwarding multicast traffic from a given source RBridge. An MES would only forward multicast traffic from source RBridges for which it is the DF, in both the site to core as well as core to site directions. This solves both the issue of avoiding packet duplication as well as the issue of avoiding forwarding loops.

In addition, the MES node advertises in IS-IS the nicknames of remote RBridges, learnt in BGP, for which it is the elected DF. This allows all RBridges in the local TRILL network to build the correct RPF state for these remote RBridge nicknames. Note that this results in all unicast traffic to a given remote RBridge being forwarded to the DF MES only (i.e. load-balancing of unicast traffic would not be possible in the site to core direction).

Alternatively, all MES nodes in a redundancy group can advertise the nicknames of all remote RBridges learnt in BGP. In addition, each MES advertises the Affinity sub-TLV, defined in [TRILL-CMT], on behalf of each of the remote RBridges for which it is the elected DF. This ensures that the RPF check state is set up correctly in the TRILL network, while allowing load-balancing of unicast traffic among the MES nodes.

In this approach, all MES nodes in a given redundancy group can forward and receive traffic on all TRILL trees.

5.5.2. Multicast Stitching with Per-VLAN Load Balancing

The MES nodes, connected to a multi-homed TRILL network, perform BGP DF election to decide which MES node is responsible for forwarding multicast traffic associated with a given VLAN. An MES would forward multicast traffic for a given VLAN only when it is the DF for this VLAN. This forwarding rule applies in both the site to core as well as core to site directions.

In addition, the MES nodes in the redundancy group partition among themselves the set of TRILL multicast trees so that each MES only sends traffic on a unique set of trees. This can be done using the RP Election Protocol as discussed in [TRILL-MULTILEVEL]. Alternatively,

the BGP DF election could be used for that. Each MES, then, advertises to the local TRILL network a Default Affinity sub-TLV, per [TRILL-MULTILEVEL], listing the trees that it will be using for multicast traffic originating from remote RBridges.

In this approach, each MES node in given TRILL network receives traffic from all TRILL trees but forwards traffic on only a dedicated subset of trees. Hence, the TRILL network must have at least as many multicast trees as the number of directly attached MES nodes.

5.5.3. Multicast Stitching with Per-Flow Load Balancing

This approach is similar to the per-VLAN load-balancing approach described above, with the difference being that the MES nodes perform the BGP DF election on a per-flow basis. The flow is identified by an N-Tuple comprising of Layer 2 and Layer 3 addresses in addition to Layer 4 ports. This can be done by treating the N-Tuple as a numeric value, and performing, for e.g., a modulo hash function against the number of PEs in the redundancy group in order to identify the index of the PE that is the DF for a given N-Tuple.

In this approach, each MES node in given TRILL network receives traffic from all TRILL trees but forwards traffic on only a dedicated subset of trees. Hence, the TRILL network must have at least as many multicast trees as the number of directly attached MES nodes.

5.5.4. Multicast Stitching with Per-Tree Load Balancing

The MES nodes, connected to a multi-homed TRILL network, perform BGP DF election to decide which MES node is responsible for forwarding multicast traffic associated with a given TRILL multicast tree. An MES would forward multicast traffic with a given destination RBridge nickname only when it is the DF for this nickname. This forwarding rule applies in both the site to core as well as core to site directions. The outcome of the BGP DF election is then used to drive TRILL IS-IS advertisements: the MES advertises to the local TRILL network a Default Affinity sub-TLV, per [TRILL-MULTILEVEL], listing the trees for which it is the elected DF.

Note that on the egress MES, the destination RBridge Nickname in multicast frames identifies the multicast tree of the remote TRILL network from which the frame originated. If the TRILL tree identifiers are not coordinated between sites, then the egress Nickname has no meaning in the directly attached (destination) TRILL network. So, the MES needs to select a new tree (after the MPLS disposition) based on a hash function, and rewrite the frame with this new destination Nickname before forwarding the traffic. This may be necessary in certain deployments to ensure complete decoupling

between the TRILL sites connected to the MPLS core. On the other hand, if the TRILL tree identifiers are coordinated between sites, then the MES doesn't have to rewrite the destination nickname in the TRILL header, after the MPLS disposition.

In this approach, each MES node in a given redundancy group forwards and receives traffic on a disjoint set of TRILL trees. At a minimum, the TRILL network must have as many multicast trees as the number of directly attached MES nodes.

6. OAM Considerations

When TRILL networks are interconnected over MPLS networks, the IS-IS control plane is not extended across the MPLS network. As described in earlier sections, this will provide advantage in scaling the TRILL networks without any issues when changes happen within different TRILL networks.

TRILL OAM could be performed in TRILL networks, within the framework defined in [TRILL-OAM-FWRK]. When TRILL networks are interconnected, TRILL OAM frames just like TRILL data frames are transparently sent over the MPLS network. There are no changes required to perform TRILL OAM operations. MPLS OAM operations could be performed as defined in [MPLS-OAM] to ensure the working of MPLS network interconnecting the TRILL networks.

7. Acknowledgements

The authors would like to thank Sami Boutros and Dennis Cai for their valuable comments.

8. Security Considerations

There are no additional security aspects beyond those of VPLS/H-VPLS that need to be discussed here.

9. IANA Considerations

This document requires IANA to assign a new SAFI value for L2VPN_MAC SAFI.

10. Intellectual Property Considerations

This document is being submitted for use in IETF standards discussions.

11. Normative References

[TRILL] Perlman et al., "Routing Bridges (RBridges): Base Protocol Specification", RFC 6325, July, 2011.

12. Informative References

[EVPN-REQ] Sajassi et al., "Requirements for Ethernet VPN (E-VPN)", draft-ietf-l2vpn-evpn-req-04.txt, work in progress, July, 2013.

[E-VPN] Sajassi et al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-04.txt, work in progress, July, 2013.

[TRILL-CMT] Senevirathne et al., "Coordinated Multicast Trees for TRILL", draft-tissa-trill-cmt-01.txt, work in progress, January 2012.

[TRILL-MULTILEVEL] Senevirathne et al., "Default Nickname Based Approach for Multilevel TRILL", draft-tissa-trill-multilevel-02.txt, work in progress, February 2012.

[TRILL-OAM-FWRK] Salam et al., "TRILL OAM Framework", draft-ietf-trill-oam-framework-03, September, 2013.

[MPLS-OAM] Kompella & Swallow, "Detecting MPLS Data Plane Failures", RFC 4379, February, 2006.

13. Authors' Addresses

Ali Sajassi
Cisco
Email: sajassi@cisco.com

Samer Salam
Cisco
Email: ssalam@cisco.com

Nabil Bitar
Verizon Communications
Email : nabil.n.bitar@verizon.com

Aldrin Isaac
Bloomberg
Email: aisaac71@bloomberg.net

Sam Aldrin
Huawei
Email: sam.aldrin@gmail.com