L2VPN Working Group
Internet-draft
Intended Status: Proposed Standard
Expires: February 2013

Bhargav Bhikkaji
Balaji Venkat Venkataswami
Ramasubramani Mahadevan
Shivakumar Sundaram
Narayana Perumal Swamy
DELL-Force10
August 3, 2012

Connecting Disparate TRILL-based Data Center/PBB/Campus sites using BGP

draft-balaji-l2vpn-trill-over-ip-multi-level-02

Abstract

There is a need to connect (a) TRILL based data centers or (b) TRILL based networks which provide Provider Backbone like functionalities or (c) Campus TRILL based networks over the WAN using one or more ISPs that provide regular IP+GRE or IP+MPLS transport. A few solutions have been proposed as in [1] in the recent past that have not looked at the PB-like functionality. These solutions have not dealt with the details as to how these services could be provided such that multiple TRILL sites can be inter-connected with issues like nick-name collisons for unicast and multicast being taken care of. It has been found that with extensions to BGP the problem statement which we will define below can be handled. Both control plane and data plane operations can be driven into the solution to make it seamlessly look at the entire set of TRILL sites as a single entity which then can be viewed as one single Layer 2 cloud. MAC moves across TRILL sites and within TRILL sites can be realized. This document / proposal envisions the use of BGP-MAC-VPN vrfs both at the IP cloud PE devices and at the peripheral PEs within a TRILL site providing Provider Backbone like functionality. We deal in depth with the control plane and data plane particulars for unicast and multicast with nick-name election being taken care of as part of the solution.

In this version of the draft, we discuss how hierarchical MAC addresses can be doled out to the end stations thus reducing the size of the BGP-MAC-VPN VRFs in the IP+GRE or IP+MPLS edge devices. We also discuss how the MAC-Moves which involve changing the IP to MAC address associations where the IP addresses remain constant when VMs ot physical servers (without VMs) are removed from one part of the network and moved to another even between Trill Data Center sites.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the

Copyright and License Notice

Table of Contents

1 Introduction

There is a need to connect (a) TRILL based data centers or (b) TRILL based networks which provide Provider Backbone like functionalities or (c) Campus TRILL based networks over the WAN using one or more ISPs that provide regular IP+GRE or IP+MPLS transport. A few solutions have been proposed as in [1] in the recent past that have not looked at the Provider Backbone-like functionality. These solutions have not dealt with the details as to how these services could be provided such that multiple TRILL sites can be inter-connected with issues like nick-name collisions for unicast (multicast is still TBD) being taken care of. It has been found that with extensions to BGP the problem statement which we will define below can be well handled. Both control plane and data plane operations can be driven into the solution to make it seamlessly look at the entire set of TRILL sites as a single entity which then can be viewed as one single Layer 2 cloud. MAC moves across TRILL sites and within TRILL sites can be realized. This document / proposal envisions the use of BGP-MAC-VPN vrfs both at the IP cloud PE devices and at the peripheral PEs within a TRILL site providing Provider Backbone like functionality. We deal in depth with the control plane and data plane particulars for unicast (multicast is still TBD) with nick-name election being taken care of as part of the solution.

1.1 Acknowledgements

The authors would like to thank Janardhanan Pathangi, Anoop Ghanwani for their inputs for this proposal.

1.2 Terminology

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].

Legend :

U-PE / ARB : User-near PE device or Access Rbridge. U-PEs are edge devices in the Customer site or tier-2 site. This is a Rbridge with BGP capabilities. It has VRF instances for each tenant it is connected to in the case of Provider-Backbone functionality use-case.

U-Ps / CRB : Core Rbridges or core devices in the Customer site that do not directly interact with the Customer's Customer.

N-PE : Network Transport PE device. This is a device with RBridge capabilities in the non-core facing side. On the core facing side it is a Layer 3 device supporting IP+GRE and/or IP+MPLS. On the non-core

facing side it has support for VRFs one for each TRILL site that it connects to. It runs BGP to convey the BGP-MAC-VPN VRF routes to its peer N-PEs. It also supports IGP on the core facing side like OSPF or IS-IS for Layer 3 and supports IP+GRE and/or IP+MPLS if need be. A pseudo-interface representing the N-PE's connection to the Pseudo Level 2 area is provided at each N-PE and a forwarding adjacency is maintained between the near-end N-PE to its remote participating N-PEs pseudo-interface in the common Pseudo Level 2 area.

N-P : Network Transport core device. This device is IP and/or IP+MPLS core device that is part of the ISP / ISPs that provide the transport network that connect the disparate TRILL networks together.

1.2 Problem Statement

1.2.1 TRILL Data Centers requiring connectivity over WAN

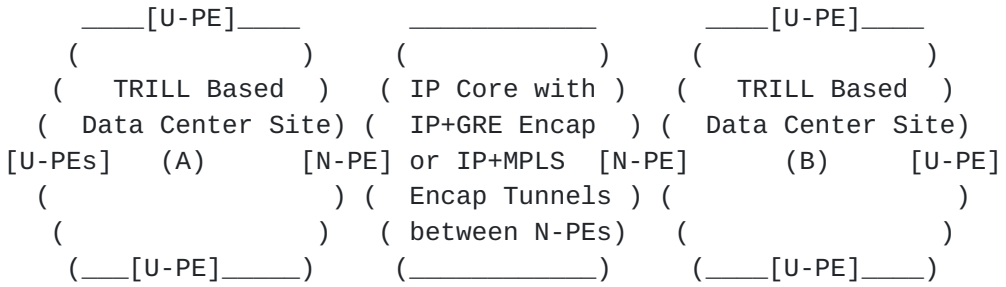


Figure 1.0 : TRILL based Data Center sites inter-connectivity.

o Providing Layer 2 extension capabilities amongst different disparate data centers running TRILL.

o Recognizing MAC Moves across data centers and within data centers to enjoin disparate sites to look and feel as one big Layer 2 cloud.

o Provide a solution agnostic to the technology used in the service provider network

o Provide a cost effective and simple solution to the above.

o Provide auto-configured tunnels instead of pre-configured ones in the transport network.

o Provide additional facilities as part of the transport network for eg., TE, QoS etc

o Routing and forwarding state is to be maintained at the network edges and not within the site or the core of the transport network.

This requires minimization of the state explosion required to provide this solution.

o So connectivity for end-customers is through U-PE onto N-PE onto remote-N-PE and onto remote U-PE.

1.2.2 Provider Backbone remote TRILL cloud requirements

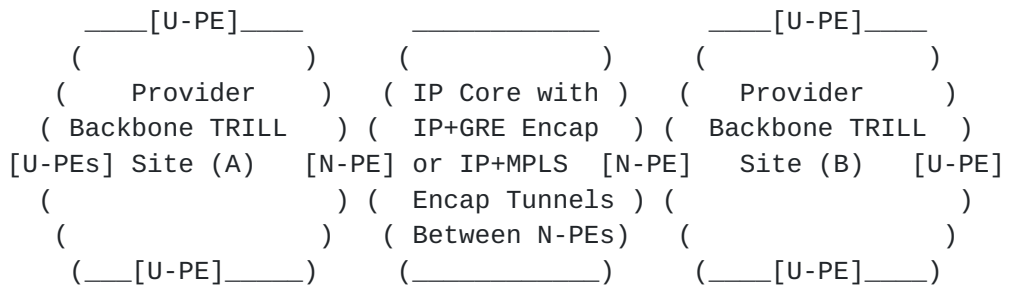


Figure 2.0 : TRILL based Provider backbone sites inter-connectivity

o Providing Layer 2 extension capabilities amongst different Provider Backbone Layer 2 clouds that need connectivity with each other.

o Recognizing MAC Moves across Provider Backbone Layer 2 Clouds and within a single site Layer 2 Cloud to enjoin disparate sites to look and feel as one big Layer 2 Cloud.

o Provide a solution agnostic to the technology used in the service provider network

o Provide a cost effective and simple solution to the above.

o Provide auto-configured tunnels instead of pre-configured ones in the transport network.

o Provide additional facilities as part of the transport network for eg., TE, QoS etc

o Routing and forwarding state is to be maintained at the network edges and not within the site or the core of the transport network. This requires minimization of the state explosion required to provide this solution.

o These clouds could be part of the same provider but be far away from each other. The customers of these clouds could demand connectivity to their sites through these TRILL clouds. These TRILL clouds could offer Provider Layer 2 VLAN transport for each of their customers. Hence Provide a seamless connectivity wherever these sites are placed.

o So connectivity for end-customers is through U-PE onto N-PE onto remote-N-PE and onto remote U-PE.

1.2.3 Campus TRILL network requirements

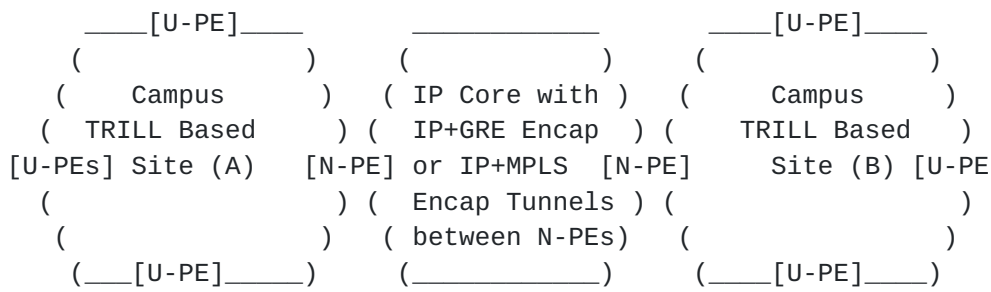


Figure 3.0 : TRILL based Campus inter-connectivity

o Providing Layer 2 extension capabilities amongst different disparate distantly located Campus Layer 2 clouds that need connectivity with each other.

o Recognizing MAC Moves across these Campus Layer 2 clouds and within a single site Campus cloud to enjoin disparate sites to look and feel as one Big Layer 2 Cloud.

o Provide a solution agnostic to the technology used in the service provider network.

o Provide a cost effective and simple solution to the above.

o Provide auto-configured tunnels instead of pre-configured ones in the transport network.

o Provide additional facilities as part of the transport network for eg., TE, QoS etc.

o Routing and Forwarding state optimizations as in 1.2.1 and 1.2.2.

o So connectivity for end-customers is through U-PE onto N-PE onto remote-N-PE and onto remote U-PE.

2. Architecture where the solution applies

2.1 Proposed Solution

The following section outlines (a) Campus TRILL topology or (b) TRILL Data Center topology or (c) Provider backbone Network topology for which solution is intended.

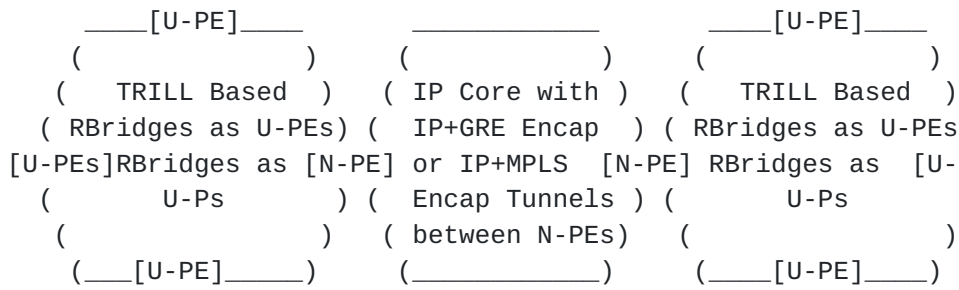


Figure 4.0 : Proposed Architecture

2.1.1 Control Plane

o Site network U-PEs still adopt learning function for source MACs bridged through their PE-CE links. For Campus TRILL networks (non-Provider-Backbone networks) the PE-CE links connect the regular hosts / servers. In the case of a data center the PE-CE links connect the servers in a rack to the U-PEs / Top of Rack Switches.

o End customer MACs are placed in BGP-MAC-VPN VRFs in the U-PE to customer PE-CE links. (at tier 2).

2.1.1.1 Nickname Collision Solution

o The near-end N-PE for a site has a forwarding adjacency for the Pseudo Level 2 area Pseudo-Interface to obtain trill nicknames of the next hop far-end N-PE's Level 2 Pseudo-Interface. This forwarding adjacency is built up during the course of BGP-MAC-VPN exchanges between the N-PEs. This forwarding adjacency is a kind of targeted IS-IS adjacency through the IP+GRE or IP+MPLS core. This forwarding adjacency exchange is accomplished through tweaking BGP to connect the near-end N-PE with the far-end N-PEs. Nickname election is done with N-PE Rbridge Pseudo-Interfaces participating in nickname election in Level 2 Area and their non-core facing interfaces which are Level 1 interfaces in the sites in the site considered to be a Level 1 area.

o The Nicknames of each site are made distinct within the site since the nickname election process PDUs for Level 1 area are NOT tunneled across the transport network to make sure that each U-P or U-PE or N-PE's Rbridge interface have knowledge of the nickname election process only in their respective sites / domains. If a new domain is connected as a site to an already existing network then the election process NEED NOT be repeated in the newly added site in order to make sure the nicknames are distinct as Multi-Level IS-IS takes care of forwarding from one site / domain to another. It is only the Pseudo-interface of the N-PE of the newly added site that will have to partake in an election to generate a new Pseudo Level 2 area Nickname

for itself.

2.1.1.2 U-PE BGP-MAC-VPN VRFs

o The Customer MACs are placed as routes in the MAC-VPN VRFs with Nexthops being the area number Nicknames of the U-PEs to which these customer MAC addresses are connected to. For MAC routes within the Level 1 area the Nicknames are those of the local U-PE itself while the MAC routes learnt from other sites have the area number of the site to which the remote U-PE belongs to. When the source learning happens the BGP-MAC-VPN-NLRI are communicated to the participating U-PEs in all the sites of the said customer. Refer to section A.1.1 in Appendix A.1 for more details on how forwarding takes place between the sites through the multi-level IS-IS mechanism orchestrated over the IP core network.

Format of the BGP-MAC-VPN VRF on a U-PE / ARB

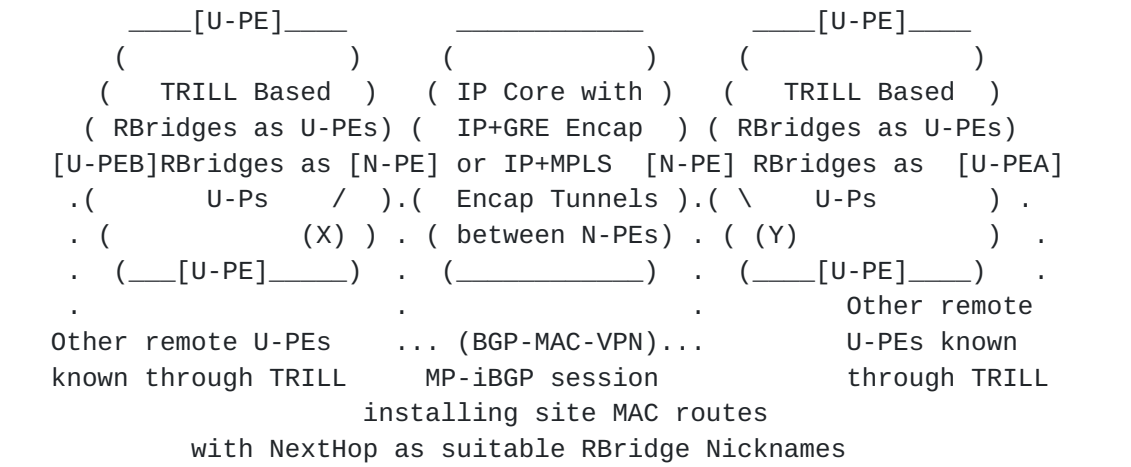
MAC address	U-PE Nickname
00:be:ab:ce:fg:9f (local)	<16-bit U-PE Nickname>
00:ce:cb:fe:fc:0f (Non-local)	<16-bit U-PE Area Num>
....	
....	

o A VRF is allocated for each customer who in turn may have multiple VLANs in their end customer sites. So in theory a total of 4K VLANs can be supported per customer. The P-VLAN or the provider VLAN in the case of a Provider Backbone category can also be 4K VLANs. So in effect in this scheme upto 4K customers could be supported if P-VLAN encapsulation is to be used to differentiate between multiple customers.

o ISIS for Layer 2 is run atop the Rbridges in the site / Tier-2 network

o ISIS for Layer 2 disseminates MACs reachable via the TRILL nexthop nicknames of site / Tier-2 network Rbridges amongst the Rbridges in the network site.

o N-PEs have VRFs for each tier-2 access network that gain connectivity through the IP+GRE or IP+MPLS core.



Legend :
(X) - Customer A Site 1 MAC-VPN-VRF
(Y) - Customer A Site 2 MAC-VPN-VRF

U-PEs are edge devices a.k.a Access Rbridges (ARBs)
U-Ps a.k.a Core Rbridges (CRBs) are core devices that interconnect U-PEs.

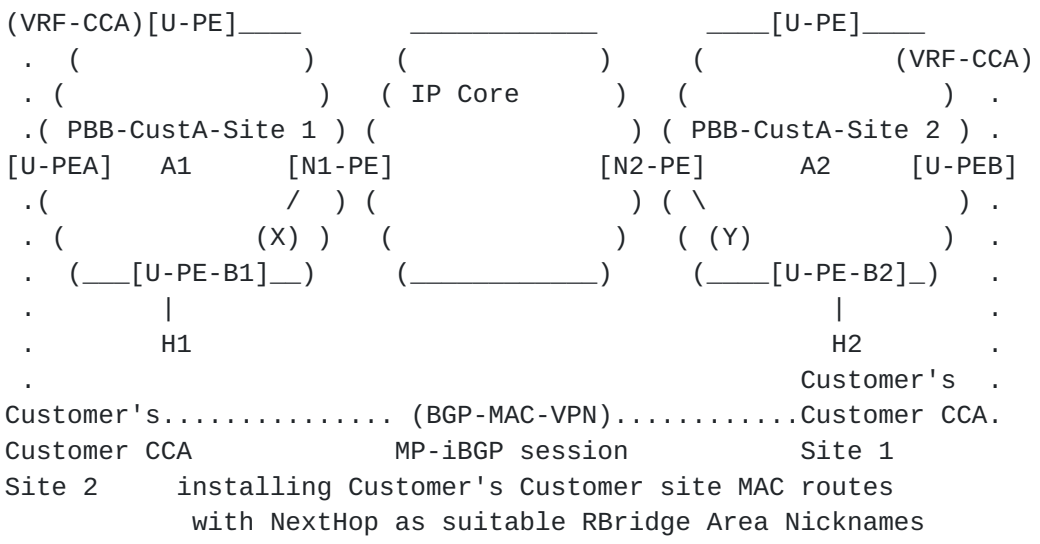
Figure 5.0 : BGP-MAC-VPN VRFs amongst N-PEs

o N-PEs re-distribute the MAC routes in their respective VRFs into the IS-IS Level 1 area after export / import amongst the N-PEs is done. The reverse re-distribution from IS-IS to BGP is also done at each N-PE for its tier-2 customer site.

o N-PEs exchange BGP information through route-targets for various customer sites with other N-PEs. The MAC routes for the various customer sites are placed in the BGP-MAC-VPN VRF of each N-PE for each customer site it connects to on the same lines as U-PE MAC-VPN-VRFs. The MAC routes placed in the VRFs of the N-PEs indicate the MAC addresses for the various Rbridges of the remote tier-2 customer sites with the respective next-hops being the Nicknames of the Level 2 pseudo-interface of the far-end N-PE through which these MAC routes are reachable.

o U-PE and U-P Rbridges MACs and TRILL nicknames are placed in BGP-MAC-VPN vrf on the N-PEs.

o Routes to various end customer MACs within a tier-2 customer's sites are exchanged through BGP MAC-VPN sessions between U-PEs. IP connectivity is provided through IP addresses on same subnet for participating U-PEs.



Legend :
A1, A2 - Area Nicknames of the customer sites in TRILL
N1, N2 - These are the N-PEs connecting A1 and A2 running BGP sessions
B1, B2 - U-PEs in A1 and A2 respectively running BGP sessions
H1, H2 - Hosts connected to B1 and B2 U-PEs.
Figure 6.0 : BGP-MAC-VPN VRFs between U-PE amongst various sites

2.1.1.3 Control Plane explained in detail.

1) B1 and B2 exchange that MACs of H1 and H2 are reachable via BGP. Example., H2-MAC is reachable via B2-MAC through area Nickname A2.

2) N1 and N2 exchange that A1 and A2 are reachable through N1 Nickname and N2 Nickname respectively via BGP.

3) N1 and N2 also exchange the MACs of U-PEs B1 and B2.

4) The routes in the N1 and N2 are re-distributed into IS-IS to end up with the following correlated routing state.

Now the correlated route in B1 is that H2 -> reachable via -> B2 -> reachable via A2 -> reachable via N1 Nickname.

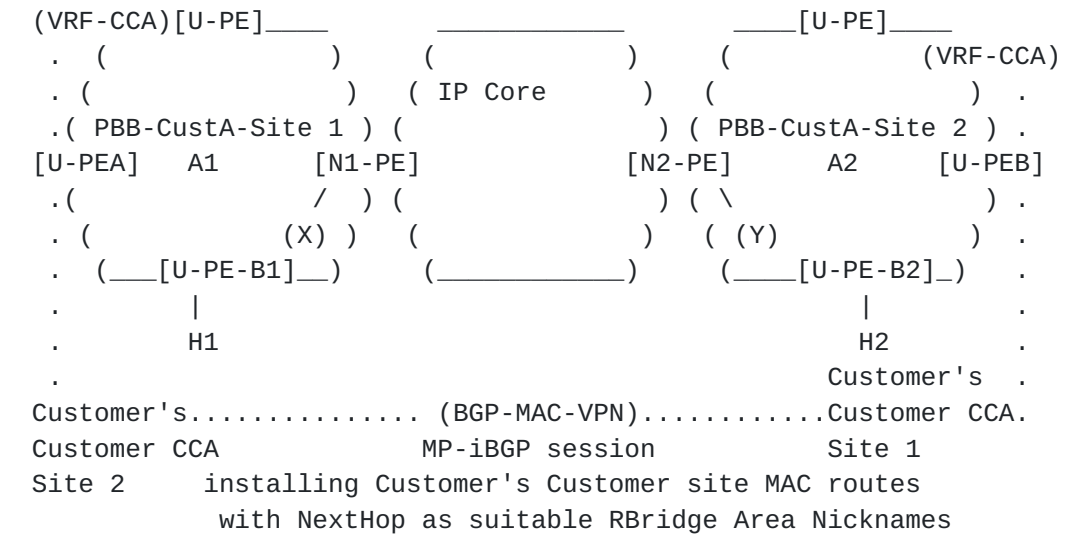
And the correlated route in B2 is that H1 -> reachable via -> B1 -> reachable via A1 -> reachable via N2 Nickname.

And the correlated route in N1 is that B2 -> reachable via -> A2 -> reachable via Nickname N2

And the correlated route in N2 is that B1 -> reachable via -> A1 ->

reachable via Nickname N1

2.1.2 Corresponding Data plane for the above control plane example.



Legend :
A1, A2 - Area Nicknames of the customer sites in TRILL
N1, N2 - These are the N-PEs connecting A1 and A2 running BGP sessions
B1, B2 - U-PEs in A1 and A2 respectively running BGP sessions
H1, H2 - Hosts connected to B1 and B2 U-PEs.
Figure 6.0 : BGP-MAC-VPN VRFs between U-PE amongst various sites

1) H1 sends a packet to B1 with SourceMac as H1-MAC and DestMac as H2-MAC and C-VLAN as C1. This frame is named F1.

2) B1 encapsulates this packet in a P-VLAN (Provider VLAN) packet with outer SourceMac as B1-MAC and DestMac as B2-MAC with P-VLAN PV1. This frame is named F2.

3) B1 being and Rbridge encapsulates a TRILL header on top of F2, with Ingress Rbridge as B1 and Egress Rbridge as A2.

4) This reaches N1 where N1 decapsulates the TRILL header and sends frame F2 inside a IP+GRE header with GRE key as Cust-A's VRF id.

5) Packet reaches N2 where N2 looks up the GRE key to identify which customer / VRF to be looked into.

6) In that VRF table N2 looks up B2 and encapsulates F2 with TRILL header with Ingress Rbridge as A1 and Egress Rbridge being B2.

7) Finally the packet reaches B2 and is decapsulated and sends F1 to

the host.

2.1.2.1 Control plane for regular Campus and Data center sites

For non-PBB like environments one could choose the same capabilities as a PBB like environment with all TORs for e.g in a data center having BGP sessions through BGP Route Reflectors with other TORs. By manipulating the Route Targets specific TORs could be tied in together in the topology within a site or even across sites. The easier way to go about the initial phase of deployment would be to restrict the MP-BGP sessions between N-PEs alone within Campus networks and Data centers and let IS-IS do the job of re-distributing into BGP. Flexibility however can be achieved by letting the U-PEs in the Campus or data center networks too to have MP-BGP sessions. Different logical topologies could be achieved as the result of the U-PE BGP sessions.

2.1.2.1.1 First phase of deployment for Campus and Data Center sites

For the first phase of deployment it is recommended that MP-BGP sessions be constructed between N-PEs alone in case of Data Center and Campus sites. This is necessary as PBB tunnels are not involved. The exchanges remain between the N-PEs about the concerned sites alone and other peering sessions of BGP are not needed since connectivity is the key. When TOR silo based topologies need to be executed then MP-BGP sessions between TORs on the near site and the remote sites can be considered. This will be explored in other documents in the future.

2.1.2.1.2 Control Plane for Data Centers and Campus

1) N1 and N2 exchange that A1 and A2 are reachable through N1 Nickname and N2 Nickname respectively via BGP.

2) N1 and N2 also exchange that B1 and B2 are within A1 and A2 and that H1 and H2 are attached to B1 and B2 respectively.

3) N1 and N2 also exchange the MACs of ARBs B1 and B2.

4) The routes in the N1 and N2 are re-distributed into IS-IS to end up with the following correlated routing state.

5) The corresponding ESADI protocol routes for end stations will also be exchanged between N-PEs using BGP. The Nickname of the nexthop will be the Area number from which the route originated.

Now the correlated route in B1 is that H2 -> reachable via -> B2 -> reachable via A2 -> reachable via N1 Nickname.

And the correlated route in B2 is that H1 -> reachable via -> B1 -> reachable via A1 -> reachable via N2 Nickname.

And the correlated route in N1 is that B2 -> reachable via -> A2 -> reachable via Nickname N2

And the correlated route in N2 is that B1 -> reachable via -> A1 -> reachable via Nickname N1

2.1.2.1.3 Data Plane for Data Centers and Campus

1) H1 sends a packet to B1 with SourceMac as H1-MAC and DestMac as H2-MAC and C-VLAN as C1. This frame is named F1.

2) B1 encapsulates this packet with outer SourceMac as B1-MAC and DestMac as B2-MAC. This frame is named F2.

3) B1 being and Rbridge encapsulates a TRILL header on top of F2, with Ingress Rbridge as B1 and Egress Rbridge as A2.

4) This reaches N1 where N1 decapsulates the TRILL header and sends frame F2 inside a IP+GRE header with GRE key as Cust-A's VRF id.

5) Packet reaches N2 where N2 looks up the GRE key to identify which customer / VRF to be looked into.

6) In that VRF table N2 looks up B2 and encapsulates F2 with TRILL header with Ingress Rbridge as A1 and Egress Rbridge being B2.

7) Finally the packet reaches B2 and is decapsulated and sends F1 to the host.

2.1.2.1.4 Control Plane for Data Centers and Campus networks with more optimizations

In order to avoid double encapsulations at the U-PE / Access Rbridge level it can also be proposed that the U-PE/ARB contain only the Customer MAC addresses and include the N-PE in the default / unknown flood tree. This way the unknown MACs are sent across all participating sites in a VPN. The response will point to the nearest N-PE. The Customer MACs will be learnt by the N-PE for over the core conversations. This way we get rid of the double encapsulations at the ARB level.

1) N1 and N2 exchange that A1 and A2 are reachable through N1 Nickname and N2 Nickname respectively via BGP.

2) N1 and N2 also exchange that B1 and B2 are within A1 and A2 and

that H1 and H2 are attached to B1 and B2 respectively.

4) The routes in the N1 and N2 are re-distributed into IS-IS to end up with the following correlated routing state.

5) The corresponding ESADI protocol routes for end stations will also be exchanged between N-PEs using BGP. The Nickname of the nexthop will be the Area number from which the route originated.

Now the correlated route in B1 is that H2 -> -> reachable via A2 -> reachable via N1 Nickname.

And the correlated route in B2 is that H1 -> -> reachable via A1 -> reachable via N2 Nickname.

2.1.2.1.5 Data Plane Optimizations for Data Centers and Campus

1) H1 sends a packet to B1 with SourceMac as H1-MAC and DestMac as H2-MAC and C-VLAN as C1. This frame is named F1.

2) B1 being and Rbridge encapsulates a TRILL header on top of F2, with Ingress Rbridge as B1 and Egress Rbridge as A2.

3) This reaches N1 where N1 decapsulates the TRILL header and sends frame F2 inside a IP+GRE header with GRE key as Cust-A's VRF id.

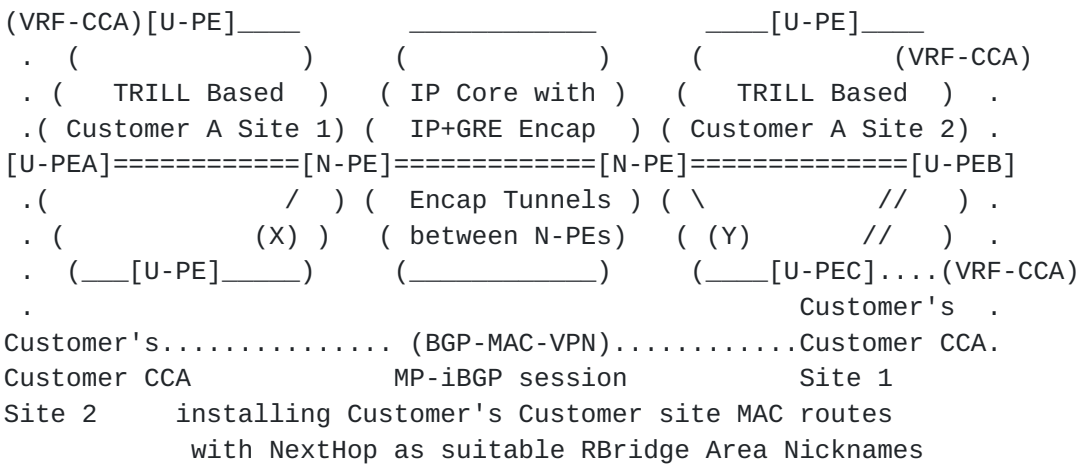
5) Packet reaches N2 where N2 looks up the GRE key to identify which customer / VRF to be looked into.

6) In that VRF table N2 looks up H2-MAC and encapsulates F1 with TRILL header with Ingress Rbridge as A1 and Egress Rbridge being B2.

7) Finally the packet reaches B2 and is decapsulated and sends F1 to the host.

2.1.2.2 Other Data plane particulars.

Default Dtree which is spanning all sites is setup for P-VLAN for Customer's Customer CCA supported on all Tier-2 sites. Denoted by ===, //.



Legend :
(X) - Customer A Site 1 MAC-VPN-VRF
(Y) - Customer A Site 2 MAC-VPN-VRF
(VRF-CCA) - MAC-VPN-VRF for Customer's Customer A (CCA) Site 1
(VRF-CCA) - MAC-VPN-VRF for Customer's Customer A (CCA) Site 2
(VRF-CCA) - MAC-VPN-VRF for Customer's Customer A (CCA) Site 3

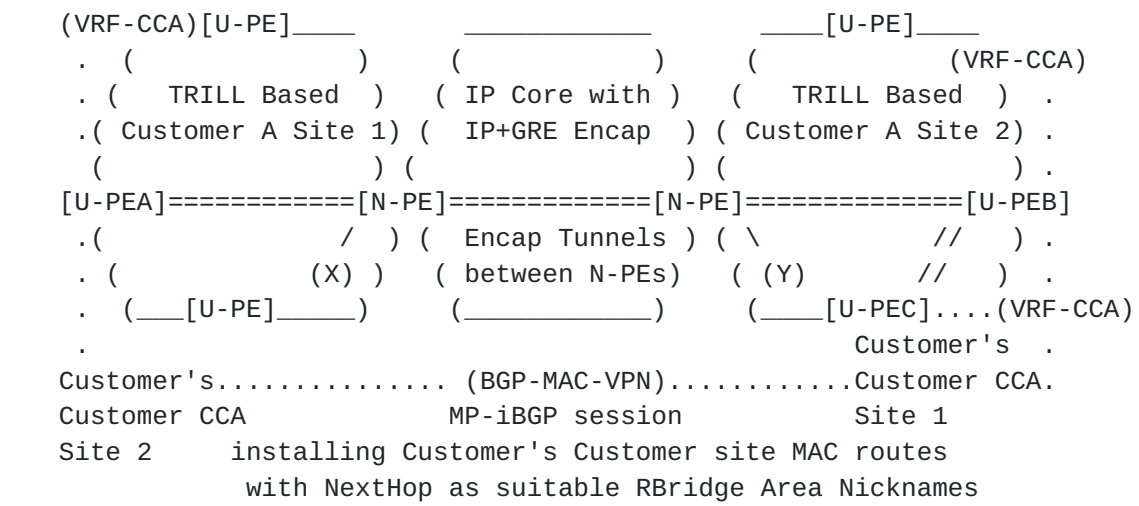
Figure 8.0 : Dtree spanning all U-PEs for unknown floods.

(1) When a packet comes into a U-PE from the near-end the source MAC is learned and placed in the near-end U-PE BGP-MAC-VPN VRF. This is done in a sub-table depending on which VLAN they belong to in the end-customer's VLANs. The destination MAC if unknown is flooded through a default Spanning tree (could be a dtree) constructed for that provider VLAN which is mapped to carry traffic for the end-customer VLAN in the customer's network sites involved.

Default Dtree which is spanning all sites is setup for P-VLAN for Customer's Customer CCA supported on all Tier-2 sites.

Denoted by ===, //.

Forwarding for unknown frames using the default Dtree spanning all customer sites and their respective U-PEs and onto their customers.



Legend :
(X) - Customer A Site 1 MAC-VPN-VRF
(Y) - Customer A Site 2 MAC-VPN-VRF
(VRF-CCA) - MAC-VPN-VRF for Customer's Customer A (CCA) Site 1
(VRF-CCA) - MAC-VPN-VRF for Customer's Customer A (CCA) Site 2
(VRF-CCA) - MAC-VPN-VRF for Customer's Customer A (CCA) Site 3

Figure 9.0 : Unknown floods through Dtree spanning for that P-VLAN

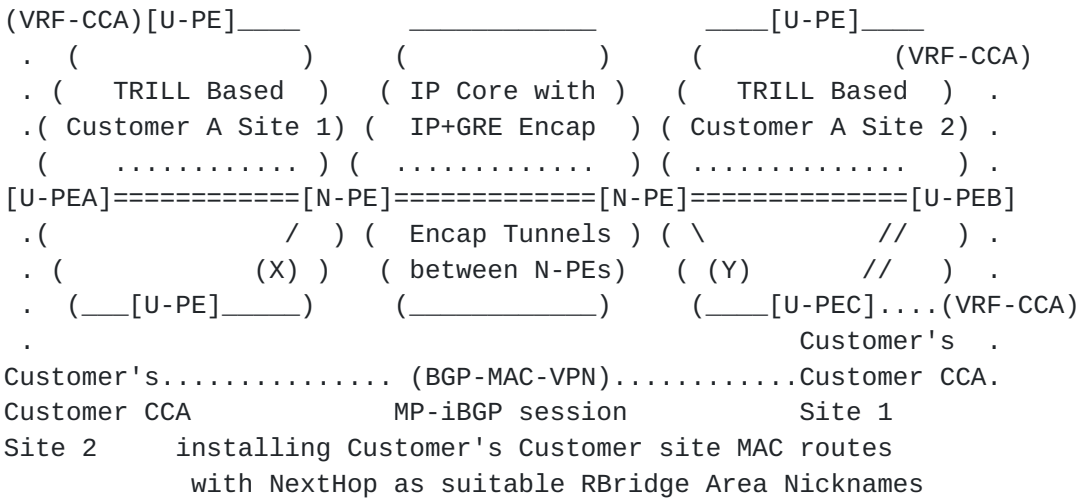
(2) The Spanning tree (which could be a dtree for that VLAN) carries that packet through site network switches all the way to N-PEs bordering that network site. U-PEs can drop the packet if there exist no ports for that customer VLAN on that U-PE. The Spanning tree includes auto-configured IP-GRE tunnels or MPLS LSPs across the IP+GRE and/or IP+MPLS cloud which are constituent parts of that tree and hence the unknown flood is carried over to the remote N-PEs participating in the said Dtree. The packet then heads to that remote-end (leaf) U-PEs and on to the end customer sites. For purposes of connecting multiple N-PE devices for a Dtree that is being used for unknown floods, a mechanism such as PIM-Bidir overlay using the MVPN mechanism in the core of the IP network can be used. This PIM-Bidir tree would stitch together all the N-PEs of a specific customer.

(3) BGP-MAC-VPN VRF exchanges between N-PEs carry the routes for MACs

of the near-end Rbridges in the near-end site network to the remote-end site network. At the remote end U-PE a correlation between near-end U-PE and the customer MAC is made after BGP-MAC-VPN VRF exchanges between near-end and far-end U-PEs. The MPLS inner label or the GRE key indicates which VRF to consult for an incoming encapsulated packet at an ingress N-PE and at the outgoing N-PE in the IP core.

(4) From thereon the source MAC so learnt at the far end is reachable just like a Hierarchical VPN case in MPLS Carrier Supporting Carrier. The only difference is that the nicknames of the far-end U-PEs/U-Ps may be the same as the nicknames of the near-end U-PEs/U-Ps. In order to overcome this, the MAC-routes exchanged between the U-PEs have the next-hops as Area nicknames of the far-end U-PE and then the Area number nickname is resolved to the near-end N-PE/N-PEs in the local site that provide connectivity to the far-end U-PE in question.

<srcMac, DstMac> srcMac is known at U-PEA, so advertize to other U-PEs through BGP in the other customer sites for Customer A that srcMAC is reachable via U-PEA. This is received at the BGP-MAC-VPN VRFs in U-PEB and U-PEC.

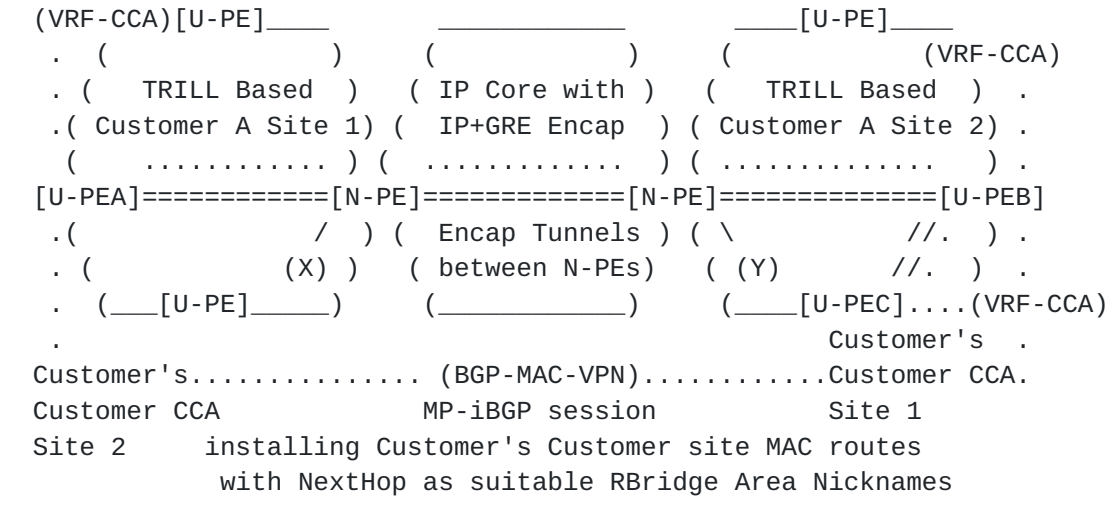


Legend :
(X) - Customer A Site 1 MAC-VPN-VRF
(Y) - Customer A Site 2 MAC-VPN-VRF
(VRF-CCA) - MAC-VPN-VRF for Customer's Customer A (CCA) Site 1
(VRF-CCA) - MAC-VPN-VRF for Customer's Customer A (CCA) Site 2
(VRF-CCA) - MAC-VPN-VRF for Customer's Customer A (CCA) Site 3

Figure 10.0 : Distributing MAC routes through BGP-MAC-VPN

<srcMac, DstMac>

Flooding when DstMAC is unknown. The flooding reaches all U-PEs and is forwarded to the customer devices (Customer's customer devices).



Legend :
(X) - Customer A Site 1 MAC-VPN-VRF
(Y) - Customer A Site 2 MAC-VPN-VRF
(VRF-CCA) - MAC-VPN-VRF for Customer's Customer A (CCA) Site 1
(VRF-CCA) - MAC-VPN-VRF for Customer's Customer A (CCA) Site 2
(VRF-CCA) - MAC-VPN-VRF for Customer's Customer A (CCA) Site 3

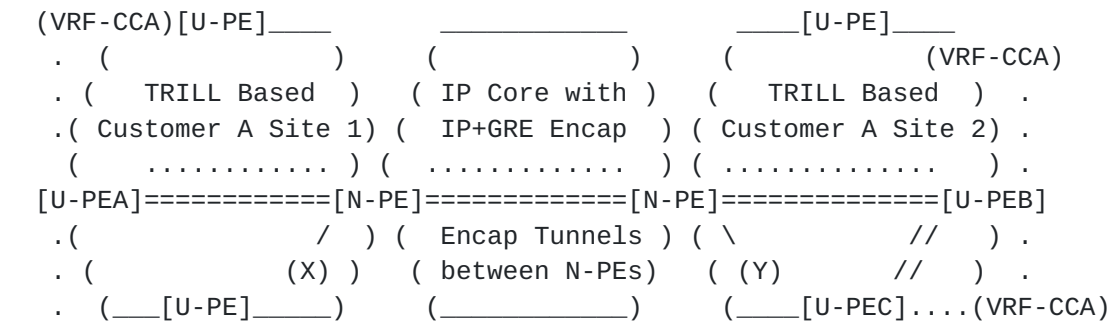
Figure 11.0 : Forwarding when DstMAC is unknown.

<srcMac, DstMac>
When DstMAC is known. Payload is carried in the following fashion in the IP core.

(<Outer Ethernet Header, IP+GRE,VRF in GRE key>,
In PBB like environments / sites interconnected, the payload is P-VLAN headers encapsulating actual payload.

<Outer Ethernet header, P-VLAN header>
<Payload = Ethernet header, Inner VLAN header>, <Actual Payload>)

In Campus and Data Center environments only the latter is carried. There is no P-VLAN header required.



Legend :
(X) - Customer A Site 1 MAC-VPN-VRF
(Y) - Customer A Site 2 MAC-VPN-VRF
(VRF-CCA) - MAC-VPN-VRF for Customer's Customer A (CCA) Site 1
(VRF-CCA) - MAC-VPN-VRF for Customer's Customer A (CCA) Site 2
(VRF-CCA) - MAC-VPN-VRF for Customer's Customer A (CCA) Site 3

Figure 12.0 : Forwarding when the DstMAC is known.

(5) The reverse path would do the same for reachability of the near-end from the far-end.

(6) Connectivity is thus established between end customer-sites through site networks and through the IP+GRE and/or IP+MPLS core.

(7) End customer packets are carried IP+GRE tunnels or IP+MPLS LSPs through access network site to near-end N-PE in the near-end. N-PE encapsulates this in auto-configured MPLS LSPs or IP+GRE tunnels to

far-end N-PEs through the IP+GRE and/or IP+MPLS core. The label is stripped at the far-end N-PE and the inner frame continues to far-end U-PE and onto the customer.

2.1.3 Encapsulations

2.1.3.1 IP + GRE

(<Outer Ethernet Header, IP+GRE, VRF in GRE key>,

In PBB like environments...

<Outer Ethernet header, P-VLAN header>,
<Payload = Ethernet header, Inner VLAN header>, <Actual Payload>)

In non-PBB like environments such as Campus and Data Center the Ethernet header with P-VLAN header is not required.

2.1.3.2 IP + MPLS

(<Outer Ethernet Header, MPLS header, VRF in Inner MPLS label>,

In PBB like environments...

<Outer Ethernet header, P-VLAN header>,
<Payload = Ethernet header, Inner VLAN header>, <Actual Payload>)

In non-PBB like environments such as Campus and Data Center the Ethernet header with P-VLAN header is not required.

2.2 Other use cases

o Campus to Campus connectivity can also be achieved using this solution. Multi-homing where multiple U-Pes connect to the same customer site can also facilitate load-balancing if a site-id (can use ESI for MAC-VPN-NLRI) is incorporated in the BGP-MAC-VPN NLRI. Mac Moves can be detected if the site-id of the advertised MAC from U-Pes is different from the older ones available.

2.3 Novelty

o TRILL MAC routes and their associated nexthops which are TRILL nicknames Are re-distributed into BGP from IS-IS

o Thus BGP-MAC-VPNs on N-Pes in the transport network contain MAC routes with nexthops as TRILL Area nicknames.

o The customer edge Rbridges / Provider bridges too contain MAC routes with associated nexthops as TRILL nicknames. This proposal is an extension of BGP-MAC-VPN I-D to include MAC routes with TRILL Area nicknames as Nexthops.

2.4 Uniqueness and advantages

o Uses existing protocols such as IS-IS for Layer 2 and BGP to achieve this. No changes to IS-IS except for redistribution into BGP at the transport core edge and vice-versa.

o Uses BGP-MAC-VPNs for transporting MAC-updates of customer devices between edge devices only.

o Employs a hierarchical MAC-route hiding from the core Rbridges of the site. Employs a hierarchical VPN like solution to avoid routing state of sites within the transport core.

o Multi-tenancy through the IP+GRE or IP+MPLS core is possible when N-PEs at the edge of the L3 core place various customer sites using the VPN VRF mechanism. This is otherwise not possible in traditional networks and using other mechanisms suggested in recent drafts.

o The VPN mechanism also provides ability to use overlapping MAC address spaces within distinct customer sites interconnected using this proposal.

o Multi-tenancy within each data center site is possible by using VLAN separation within the VRF.

o Mac Moves can be detected if source learning / Grauitous ARP combined with the BGP-MAC-VPN update triggers a change in the concerned VRF tables.

o PBB like functionality supported where P-VLAN and Customer VLAN are different spaces.

o Uses regular BGP supporting MAC-VPN features, between transport core edge devices and the Tier-2 customer edge devices.

o When new TRILL sites are added then no re-election in the Level 1 area is needed. Only the Pseudo-interface of the N-PE has to be added to the mix with the transport of the election PDUs being done across the transport network core.

2.4.1 Multi-level IS-IS

Akin to TRILL IS-IS multi-level draft where each N-PE can be considered as a ABR having one nickname in a customer site which in turn is a level-1 area and a Pseudo Interface facing the core of the transport network which belongs to a Level 2 Area, the Pseudo Interface would do the TRILL header decapsulation for the incoming packet from the Level 1 Area and throw away the TRILL header within the Pseudo Level 2 Area and transport the packets across the Layer 3 core (IP+GRE and/or IP+MPLS) after an encapsulation in IP+GRE or IP+MPLS. Thus we should have to follow a scheme with the NP-E core facing Pseudo-interface in the Level 2 Pseudo-Area doing the TRILL encapsulation and decapsulation for outgoing and incoming packets respectively from and to the transport core. The incoming packets from the Level 1 area are subject to encapsulation in IP+GRE or IP+MPLS by the sending N-PE's Pseudo-Interface and the outgoing packets from the transport core are subject to decapsulation from their IP+GRE or IP+MPLS headers by the Pseudo-Interface on the receiving N-PE.

2.4.2 Benefits of the VPN mechanism

Using the VPN mechanism it is possible that MAC-routes are placed in distinct VRFs in the N-PEs thus providing separation between customers. Assume customer A and customer B have several sites that need to be interconnected. By isolating the routes within specific VRFs multi-tenancy across the L3 core can be achieved. Customer A's sites talk to customer A's sites alone and the same is applicable with Customer B.

The same mechanism also provides for overlapping MAC addresses amongst the various customers. Customer A could use the same MAC-addresses as Customer B. This is otherwise not possible with other mechanisms that have been recently proposed.

2.4.3 Inter-working with other VXLAN, NVGRE sites

Without TRILL header it is possible to inter-work with STP sites, VXLAN sites, NVGRE sites and with other TRILL sites.

For this purpose if for example TRILL site has to inter-operate with VXLAN sites then the VXLAN site has to have a VXLAN gateway that translated plain Ethernet packets coming in from the WAN core into VXLAN packets with the VRF signifying the VXLAN-ID or the VNI.

2.4.4 Benefits of using Multi-level

The benefits of using Multi-level are choosing appropriate Multicast

Trees in other sites through the inter-area multicast method as proposed by Radia Perlman et.al.

2.5 Comparison with OTV and VPN4DC and other schemes

o OTV requires a few proprietary changes to IS-IS. There are less proprietary changes required for this scheme with regard to IS-IS compared to OTV.

o VPN4DC is a problem statement and is not yet as comprehensive as the scheme proposed in this document.

o [4] deals with Pseudo-wires being setup across the transport core. The control plane protocols for TRILL seem to be tunneled through the transport core. The scheme in the proposal we make do NOT require anything more than Pseudo Level 2 area number exchanges and those for the Pseudo-interfaces. BGP takes care of the rest of the routing. Also [4] does not take care of nick-name collision detection since the control plane TRILL is also tunneled and as a result when a new site is sought to be brought up into the inter-connection amongst existing TRILL sites, nick-name re-election may be required.

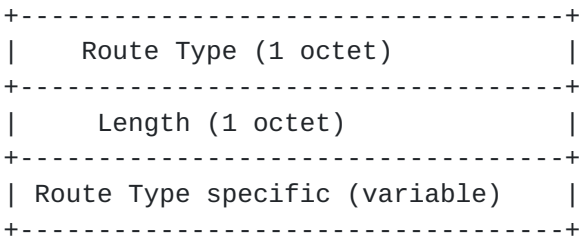
o [5] does not have a case for TRILL. It was intended for other types of networks which exclude TRILL since [5] has not yet proposed TRILL Nicknames as nexthops for MAC addresses.

2.6 Multi-pathing

By using different RDs to export the BGP-MAC routes with their appropriate Nickname next-hops from more than one N-PE we could achieve multi-pathing over the transport IP+GRE and/or IP+MPLS core.

2.7 TRILL extensions for BGP

2.7.1 Format of the MAC-VPN NLRI



The Route Type field defines encoding of the rest of MAC-VPN NLRI (Route Type specific MAC-VPN NLRI).

The Length field indicates the length in octets of the Route Type specific field of MAC-VPN NLRI.

This document defines the following Route Types:

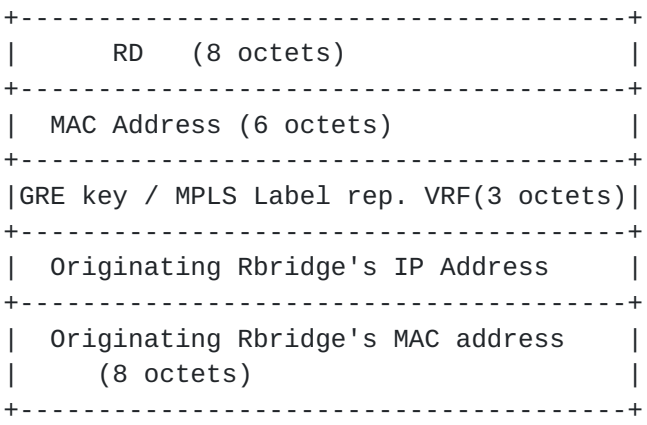
+ 1 - Ethernet Tag Auto-Discovery (A-D) route
+ 2 - MAC advertisement route
+ 3 - Inclusive Multicast Ethernet Tag Route
+ 4 - Ethernet Segment Route
+ 5 - Selective Multicast Auto-Discovery (A-D) Route
+ 6 - Leaf Auto-Discovery (A-D) Route
+ 7 - MAC Advertisement Route with Nexthop as TRILL Nickname

Here type 7 is used in this proposal.

2.7.2. BGP MAC-VPN MAC Address Advertisement

BGP is extended to advertise these MAC addresses using the MAC advertisement route type in the MAC-VPN-NLRI.

A MAC advertisement route type specific MAC-VPN NLRI consists of the following:



The RD MUST be the RD of the MAC-VPN instance that is advertising the NLRI. The procedures for setting the RD for a given MAC VPN are described in section 8 in [3].

The encoding of a MAC address is the 6-octet MAC address specified by IEEE 802 documents [802.1D-ORIG] [802.1D-REV].

If using the IP+GRE and/or IP+MPLS core networks the GRE key or MPLS label MUST be the downstream assigned MAC-VPN GRE key or MPLS label that is used by the N-PE to forward IP+GRE or IP+MPLS encapsulated ethernet packets received from remote N-PEs, where the destination MAC address in the ethernet packet is the MAC address advertised in

the above NLRI. The forwarding procedures are specified in previous sections of this document. A N-PE may advertise the same MAC-VPN label for all MAC addresses in a given MAC-VPN instance. Or a N-PE may advertise a unique MAC-VPN label per MAC address. All of these methodologies have their tradeoffs.

Per MAC-VPN instance label assignment requires the least number of MAC-VPN labels, but requires a MAC lookup in addition to a GRE key or MPLS lookup on an egress N-PE for forwarding. On the other hand a unique label per MAC allows an egress N-PE to forward a packet that it receives from another N-PE, to the connected CE, after looking up only the GRE key or MPLS labels and not having to do a MAC lookup.

The Originating Rbridge's IP address MUST be set to an IP address of the PE (U-PE or N-PE). This address SHOULD be common for all the MAC-VPN instances on the PE (e.,g., this address may be PE's loopback address).

2.7.2.1 Next hop field in MP_REACH_NLRI

The Next Hop field of the MP_REACH_NLRI attribute of the route MUST be set to the Nickname of the N-PE or in the case of the U-PE the Area Nickname of the Rbridge one whose MAC address is carried in the Originating Rbridge's MAC Address field.

The BGP advertisement that advertises the MAC advertisement route MUST also carry one or more Route Target (RT) attributes.

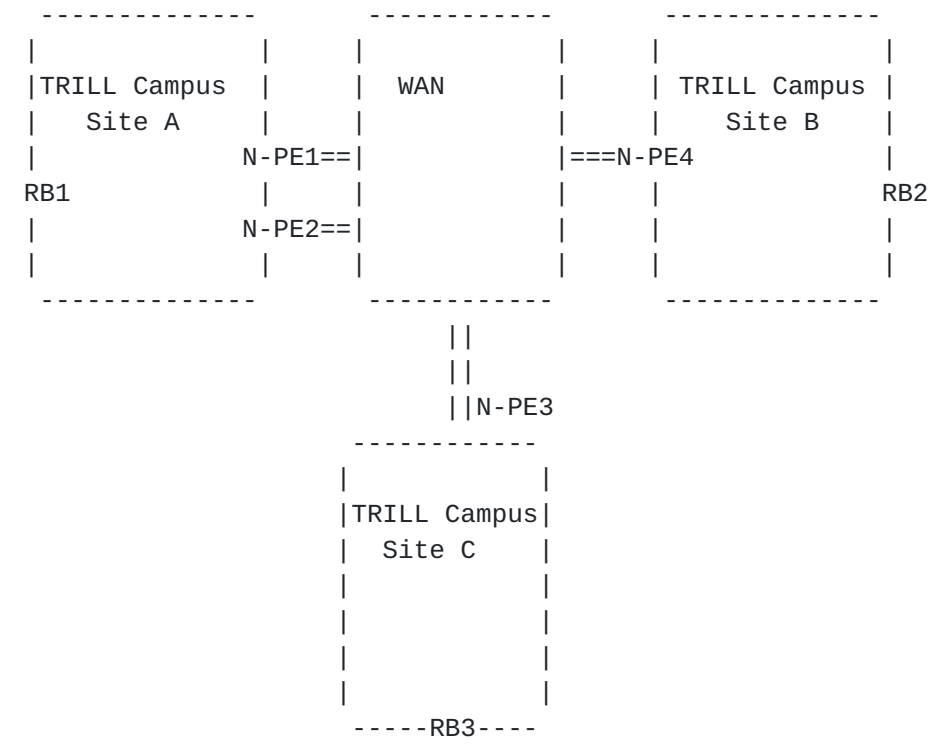
It is to be noted that document [3] does not require N-PEs/U-PEs to create forwarding state for remote MACs when they are learned in the control plane. When this forwarding state is actually created is a local implementation matter. However the proposal in this document requires that forwarding state be established when these MAC routes are learned in the control plane.

2.7.2.2 Route Reflectors for scaling

It is recommended that Route Reflectors SHOULD be deployed to mesh the U-PEs in the sites with other U-PEs at other sites (belonging to the same customer) and the transport network also have RRs to mesh the N-PEs. This takes care of the scaling issues that may arise if full mesh is deployed amongst U-PEs or the N-PEs.

2.7.3 Multicast Operations in Interconnecting TRILL sites

For the purpose of multicast it is possible that the IP core can have a Multicast-VPN based PIM-bidir tree (akin to Rosen or NGEN-MVPN) for each customer that will connect all the N-PEs related to a customer and carry the multicast traffic over the transport core thus connecting site to site multicast trees. Each site that is connected to the N-PE would have the N-PE as the member of the MVPN PIM-Bidir Tree connecting that site to the other sites' chosen N-PE. Thus only one N-PE from each site is part of the MVPN PIM-Bidir tree so constructed. If there exists more than one N-PE per site then that other N-PE is part of a different MVPN PIM-Bidir tree. Consider the following diagram that represents three sites that have connectivity to each other over a WAN. The Site A has 2 N-PEs connected from the WAN to itself and the others B and C have one each. It is to be noted that two MVPN Bidir-Trees are constructed one with Site A's N-PE1 and Site B and C's N-PE respectively while the other MVPN Bidir-tree is constructed with Site A's N-PE2 and site B and C's respective N-PEs. It is possible to load-balancing of multicast groups among the sites. The method of interconnecting trees from the respective Level 1 areas (that is the sites) to each other is akin to stitching the Dtrees that have the N-PEs as their stitch end-points in the Pseudo-Level 2 area with the MVPN Bidir tree acting as the conduit for such stitching. The tree-ids in each site are non-unique and need not be distince across sites. It is only that the N-PEs which have their one foot in the Level 1 area are stitched together using the MVPN Bidir overlay in the Layer 3 core.



Here N-PE1, N-PE3 and N-PE4 form a MVPN Bidir-tree amongst themselves

to link up the multilevel trees in the 3 sites. While N-PE2, N-PE3 and N-PE4 form a MVPN Bidir-tree amongst themselves to up the multilevel trees in the 3 sites.

There exist 2 PIM-Bidir overlay trees that can be used to load-balance say Group G1 on the first and G2 on the second. Lets say the source of the Group G1 lies within Site A and the first overlay tree is chosen for multicasting the stream. When the packet hits the WAN link on N-PE1 the packet is replicated to N-PE3 and N-PE4. It is important to understand that a concept like Group Designated Border Rbridge (GDBR) is applied in this case where group assignments are made to specific N-PEs such that only one of them is active for a particular group and the other does not send it across the WAN using the respective MVPN PIM-Bidir tree. Now Group G2 could then use the MVPN PIM-bidir based tree for its transport. The procedures for election of Group Designated Border Rbridge within a site will be further discussed in detail in future versions of this draft or may be taken to a separate document. VLAN based load-balancing of multicast groups is also possible and feasible in this scenario. It also can be VLAN, Multicast MAC-DA based. The GDBR scheme is applicable only for packets that N-PEs receive as TRILL decapsulated MVPN PIM-Bidir tree frames from the Layer 3 core. If a TRILL encapsulated multicast frame arrives at a N-PE only the GDBR for that group can decapsulate the TRILL header and send it across the Layer 3 core. The other N-PEs can however forward these multi-destination frames coming from N-PEs across the core belonging to a different site.

When the packet originates from the source host the Egress Nickname of the multicast packet is set to the Dtree root at the Level 1 area where the source is originating the stream from. The packet flows along the multicast distribution tree to all Rbridges which are part of the Dtree. Now the N-PE that provides connectivity to the Pseudo-Level 2 area and to other sites beyond it, also recieves the packet. The MVPN PIM-bidir tree is used by the near end N-PE to send the packet to all the other member N-PEs of the customer sites and appropriate TRILL encapsulation is done at the ingress N-PE for this multicast stream with the TRILL header containing a local Dtree root on the receiving site and packet streamed to the said receivers in that site. Source suppression such that the packet is not put back on the core, is done by looking at the Group Designated Border Rbridge information at the receiving site. If then other N-PEs which connect the site to the Layer 3 core receive the multicast packet sent into the site by the GDBR for that group then the other N-PEs check if they are indeed the GDBR for the said group and if not they do not forward the traffic back into the core.

It is to be noted that the Group Address TLV is transported by BGP

from across the other sites into a site and it is the GDBR for that group from the remote side that enables this transport. This way the MVPN PIM-bidir tree is pointed to from within each site through the configured GDBR N-PEs for a said group. The GDBR thus lies as one of the receivers in the Dtree for a said group within the site where the multicast stream originates.

2.7.4 Comparison with PBB-EVPN

With respect to PBB-EVPN scheme outlined in [PBB-EVPN], the scheme explained in this document has the following advantages over and above the PBB-EVPN scheme.

2.7.4.1 No nickname integration issues in our scheme

Existing TRILL based sites can be brought into the interconnect without any re-election / re-assignment of nicknames. The one benefit it seems to have vs PBB-EVPN is that adding a new site to a VPN, or merging 2 VPNs, doesn't cause issues with nickname clashes. This is a major advantage since the new TRILL site can hit the ground running without any interruptions to the existing sites in the interconnect.

2.7.4.2 Hierarchical Nicknames and their disadvantages in the PBB-EVPN scheme

The PBB-EVPN scheme advocates the use of Hierarchical Nicknames where the nickname is split into the Site-ID and the Rbridge-ID. The use of the nicknames has the following corollary disadvantages.

(a) The nickname is a 16 bit entity. With a interconnect where there are for eg., 18 sites the PBB-EVPN scheme has to use 5 bits in the nickname bitspace for Site-ID. It wastes (32 - 18) = 14 Site-IDs. The number of sites is also limited to say at best 255 sites.

(b) The nickname is a 16 bit entity. With a interconnect where there are at least 4K Rbdriges in each site, the nickname space has to set aside 12 bits at the least in the nickname space for the Rbridge-ID. This means that the Sites cannot be more than 2^4 = 16.

Thus the use of the hierarchical scheme limits the Site-IDs and also the number of Rbridges within the site. If we want to have more Sites we set aside more bits for the Site-ID thus sacrificing maximum number of Rbridge-IDs within the site. If there are more RBridges within each site, then allocating more bits for the RBridge-ID would sacrifice the maximum number of Site-IDs possible.

For eg., in a branch office scenario if there are 32 sites and more than 255 Rbridges in each of the branch offices it would be difficult

to accomodate the set of sites along with the number of Rbridges using the hierarchical nickname scheme.

In the scheme outlined in this document, it is possible to set aside 1000 nicknames or 2000 nicknames or even 200 nicknames depending on the number of sites (since this is a range of nicknames without hierarchy in the nickname space), without compromising on the maximum number of Rbridges within each site. If M were the number of sites to be supported then the number of Rbridges would be 2^16 - M = X. This X number would be available to all sites since the nickname is site-local and not globally unique.

It would be possible to set aside a sizeable number within the nickname space for future expansion of sites without compromising on the number of Rbrdiges within the site.

2.7.4.3 Load-Balancing issues with respect to PBB-EVPN

While PBB-EVPN allows for active/active load-balancing the actual method of distributing the load leads to pinning the flow onto one of the multi-homed N-PEs for a specific site rather than the multi-path hashing based scheme that is possible with our scheme.

2.7.4.4 Technology Agnostic for interworking between TRILL and Non-TRILL sites

Our scheme provides a Technology agnostic method for inter-working between TRILL and non-TRILL sites such as STP-based sites, and other NVO3 schemes for example. This is because the TRILL header is not carried over the L3 core. This is provided as an option in the initial capability exchange between N-PEs when a said pair of N-PEs handshake for BGP. The PBB-EVPN scheme doesnt offer this capability.

2.7.5 Conversational C-MACs only in the N-PE VRF MAC table

It is possible to maintain only conversational MACs on the N-PE table in the case of Campus and Data Center networks by installing the C-MACs in the hardware learned through the site interface or through the Core facing interface only if there arises evidence of across-core conversations. Thus those C-MAC addresses that have been learnt as a result of conversations between Rbridges across sites connecting to hosts that actively communicate with each other are installed in the hardware. Locally switched conversations are not learnt. This is an optimization that will reduce the disadvantage of learning all possible C-MACs located in all the various sites of a VPN on the N-PE. If a one-sided C-MAC is evidenced in the data plane, the learnt

C-MAC is not installed in the hardware unless a reverse path conversation is heard across sites. This C-MAC initially is placed in the software table and the wait begins to hear a reverse path conversation flow. If the wait results in learning that a two-way conversation exists across sites then the software learns are actually programmed in the hardware.

2.7.5.1 VLAN filtering at U-PEs.

It is further possible for the U-PE to filter based on VLANs that it possesses and thus exclude those MAC addresses for VLANs that it does not converse with for the hosts attached to it. This optimizes on the table-size to a large degree since the U-PEs need to know only what they need and not hold all the C-MACs that are in vogue in that site for those conversing VLANs for that Rbridge.

2.7.6 Table sizes in hardware will increase

There may be a concern that table sizes in hardware may be a problem with respect to the C-MAC scaling. With the possibility of having more table sizes in merchant silicon this may no longer be a issue. Also with enhanced lookup tables which may be external to the merchant silicon this problem may no longer be a downside to the scheme proposed in this document.

2.7.7 The N-PE and its implementation

It is possible that the N-PE placed as the border Rbridge and router-PE device respectively on either side of the L3 core, the actual implementation would be in the form of two devices one acting as the border Rbridge and the other as the plain Provider Edge router. The link between the two would be an attachment circuit.

2.7.8 Hierarchical MAC addresses that shrink table sizes

In this section, we discuss how hierarchical MAC addresses can be doled out to the end stations thus reducing the size of the BGP-MAC-VPN VRFs in the IP+GRE or IP+MPLS edge devices. We also discuss how the MAC-Moves which involve changing the IP to MAC address associations where the IP addresses remain constant when VMs ot physical servers (without VMs) are removed from one part of the network and moved to another even between Trill Data Center sites.

Consider the case where the end stations with either Virtual Machines managed by hypervisors or physical servers exist behind the U-PEs at each Data Center site. The Hypervisor or the physical server when they are booted up and join the cloud behind the U-PE send Active Directory Requests to an AD-Service. These AD requests are

intercepted by a smart endnode proximal to the U-PE (ARB). The smart endnode (as in a cloudlet specified in [RadiaCloudlet]) requests the AD-Service with information on the U-PEs (ARBs) Rbridge Nickname in that site. This AD-Service is available for each site and has discontiguous sets of Hierarchical MAC address prefixes of length 20 bits to dole out for each U-PE (behind which end stations exist) within a site. These discontiguous sets are unique for all U-PEs in all sites belonging to a particular Trill Data Center inter-connection. The AD-Service replies with hierarchical prefix and the nodes are assigned their addresses based on arbitration from the smart endnode. It is also possible that the AD-service will return the complete MAC address with the hierarchical prefix of 20 bits at the beginning of the address. The smart endnode returns this AD-service request to the end station requesting it. The VM or the physical server whichever the case may be absorbs this address and uses this address to reach out to the other end stations within the site or across sites.

The N-PE begins to learn MAC prefixes alone of the MAC addresses passing through it and the ingress Rbridge nickname to which this prefix was reported from. For MAC prefixes belonging to other sites of the Data Center VPN the Area nickname of the other site from which it came from is learnt from. This is also conveyed to other N-PEs belonging or having BGP-MAC-VPN VRFs of the said VPN with participating DC sites. Thus the table sizes of the BGP-MAC-VPN VRFs is reduced to having only prefixes rather than having complete MAC addresses.

2.7.8.1 MAC-Moves with hierarchical MAC addresses

When the VMs or Physical servers are moved from one site to another VPN site then appropriate Gratuitous ARP requests are sent from the moved VM or end station which then helps the communicating end stations or VMs to that moving end station or VM to re-assign their IP to MAC address mapping. This is because the move would have changed the hierarchical prefix of the moving stations MAC address based on the U-PE to which the end station attaches to after the move. Appropriate mechanisms are already present to make this change.

3 Security Considerations

TBD.

4 IANA Considerations

A few IANA considerations need to be considered at this point. A proper AFI-SAFI indicator would have to be provided to carry MAC addresses as NLRI with Next-hops as Rbridbge Nicknames. This one AFI-SAFI indicator could be used for both U-PE MP-iBGP sessions and N-PE MP-iBGP sessions. For transporting the Group Address TLV suitable extensions to BGP must be done and appropriate type codes assigned for the tranport of such TLVs in the BGP-MAC-VPN VRF framework.

5 References

5.1 Normative References

[KEYWORDS] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC1776] Crocker, S., "The Address is the Message", RFC 1776, April 1 1995.

[TRUTHS] Callon, R., "The Twelve Networking Truths", RFC 1925, April 1 1996.

5.2 Informative References

[1] draft-xl-trill-over-wan-00.txt, XiaoLan. Wan et.al December 11th ,2011 Work in Progress

[2] draft-perlman-trill-rbridge-multilevel-03.txt, Radia Perlman et.al October 31, 2011 Work in Progress

[3] draft-raggarwa-mac-vpn-01.txt, Rahul Aggarwal et.al, June 2010, Work in Progress.

[4] draft-yong-trill-trill-o-mpls, Yong et.al, October 2011, Work in Progress.

[5] draft-raggarwa-sajassi-l2vpn-evpn Rahul Aggarwal et.al, September 2011, Work in Progress.

[RadiaCloudlet] draft-perlman-trill-cloudlet-00, Radia Perlman et.al, July 30 2012, Work in Progress.

[EVILBIT] Bellovin, S., "The Security Flag in the IPv4 Header", RFC 3514, April 1 2003.

[RFC5513] Farrel, A., "IANA Considerations for Three Letter Acronyms", RFC 5513, April 1 2009.

[RFC5514] Vyncke, E., "IPv6 over Social Networks", RFC 5514, April 1 2009.

Authors' Addresses

Bhargav Bhikkaji,
Dell-Force10,
350 Holger Way,
San Jose, CA
U.S.A

Email: Bhargav_Bhikkaji@dell.com

Balaji Venkat Venkataswami,
Dell-Force10,
Olympia Technology Park,
Fortius block, 7th & 8th Floor,
Plot No. 1, SIDCO Industrial Estate,
Guindy, Chennai - 600032.
TamilNadu, India.
Tel: +91 (0) 44 4220 8400
Fax: +91 (0) 44 2836 2446

EMail: BALAJI_VENKAT_VENKAT@dell.com

Ramasubramani Mahadevan,
Dell-Force10,
Olympia Technology Park,
Fortius block, 7th & 8th Floor,
Plot No. 1, SIDCO Industrial Estate,
Guindy, Chennai - 600032.
TamilNadu, India.

Tel: +91 (0) 44 4220 8400
Fax: +91 (0) 44 2836 2446

EMail: Ramasubramani_Mahade@dell.com

Shivakumar Sundaram,
Dell-Force10,
Olympia Technology Park,
Fortius block, 7th & 8th Floor,
Plot No. 1, SIDCO Industrial Estate,
Guindy, Chennai - 600032.
TamilNadu, India.
Tel: +91 (0) 44 4220 8400
Fax: +91 (0) 44 2836 2446

EMail: Shivakumar_sundaram@dell.com

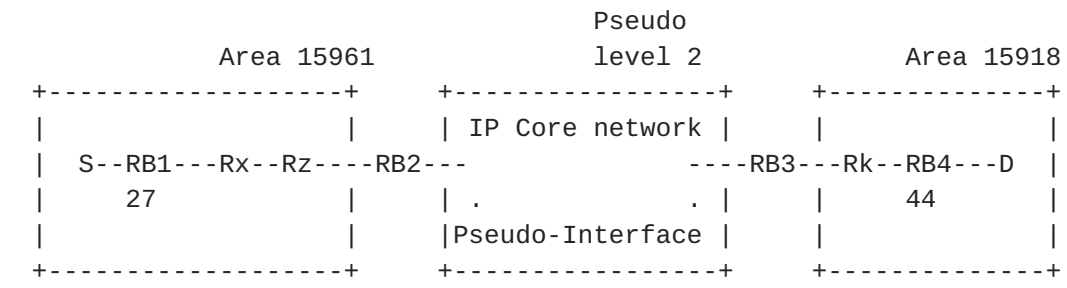
Narayana Perumal Swamy,
Dell-Force10,
Olympia Technology Park,
Fortius block, 7th & 8th Floor,
Plot No. 1, SIDCO Industrial Estate,
Guindy, Chennai - 600032.
TamilNadu, India.
Tel: +91 (0) 44 4220 8400
Fax: +91 (0) 44 2836 2446

Email: Narayana_Perumal@dell.com

A.1 Appendix I

A.1.1 Extract from Multi-level IS-IS draft made applicable to scheme

In the following picture, RB2 and RB3 are area border RBridges. A source S is attached to RB1. The two areas have nicknames 15961 and 15918, respectively. RB1 has a nickname, say 27, and RB4 has a nickname, say 44 (and in fact, they could even have the same nickname, since the RBridge nickname will not be visible outside the area).



Here RB2 and RB3 are N-PEs. RB4 and RB1 are U-PEs.

This sample topology could apply to Campus and data-center topologies. For Provider Backbone topologies S would fall outside the Area 15961 and RB1 would be the U-PE carrying the C-VLANs inside a P-VLAN for a specific customer.

Let's say that S transmits a frame to destination D, which is connected to RB4, and let's say that D's location is learned by the relevant RBridges already. The relevant RBridges have learned the following:

1) RB1 has learned that D is connected to nickname 15918
2) RB3 has learned that D is attached to nickname 44.

The following sequence of events will occur:

- S transmits an Ethernet frame with source MAC = S and destination MAC = D.

- RB1 encapsulates with a TRILL header with ingress RBridge = 27, and egress = 15918.

- RB2 has announced in the Level 1 IS-IS instance in area 15961, that it is attached to all the area nicknames, including 15918. Therefore, IS-IS routes the frame to RB2. (Alternatively, if a distinguished range of nicknames is used for Level 2, Level 1 RBridges seeing such an egress nickname will know to route to the nearest border router, which can be indicated by the IS-IS attached bit.)

In the original draft on multi-level IS-IS the following happens and

QUOTE...

- RB2, when transitioning the frame from Level 1 to Level 2, replaces the ingress RBridge nickname with the area nickname, so replaces 27 with 15961. Within Level 2, the ingress RBridge field in the TRILL header will therefore be 15961, and the egress RBridge

field will be 15918. Also RB2 learns that S is attached to nickname 27 in area 15961 to accommodate return traffic.

- The frame is forwarded through Level 2, to RB3, which has advertised, in Level 2, reachability to the nickname 15918.

- RB3, when forwarding into area 15918, replaces the egress nickname in the TRILL header with RB4's nickname (44). So, within the destination area, the ingress nickname will be 15961 and the egress nickname will be 44.

- RB4, when decapsulating, learns that S is attached to nickname 15961, which is the area nickname of the ingress.

Now suppose that D's location has not been learned by RB1 and/or RB3. What will happen, as it would in TRILL today, is that RB1 will forward the frame as a multi-destination frame, choosing a tree. As the multi-destination frame transitions into Level 2, RB2 replaces the ingress nickname with the area nickname. If RB1 does not know the location of D, the frame must be flooded, subject to possible pruning, in Level 2 and, subject to possible pruning, from Level 2 into every Level 1 area that it reaches on the Level 2 distribution tree.

UNQUOTE...

In the current proposal that we outline in this document, the TRILL header is done away with completely in the IP+GRE or IP+MPLS core. A re-look into the inner headers after decapsulation gives the appropriate information to carry the frame from the N-PE towards the destination U-PE.