TRILL Working Group
INTERNET-DRAFT
Intended Status: Proposed Standard
Expires: September 2012

Balaji Venkat Venkataswami
Bhargav Bhikkaji
Narayana Perumal Swamy
Dell-Force10

March 26, 2012

Interconnecting multiple TRILL sites deploying Traffic Engineering
draft-balaji-trill-te-multi-site-interconnect-00

Abstract

This document specifies the control plane procedures to support Traffic Engineering (TE) across TRILL sites where such sites are interconnected using [1] with the help of a Layer 3 core running IP+GRE or IP+MPLS. Traffic Engineering permits usage of a set of links that possess a certain characteristic like specified bandwidth, cost or even MTU. This draft aims at addressing how unicast frames travelling from one TRILL site to another across a Layer 3 core that supports IP+GRE and/or IP+MPLS can make use of the TE calculated paths in the sending site as well as the receiving site where such TE paths are pre-computed in both sites and need a mechanism to inter-link them together.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/1id-abstracts.html

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

Table of Contents

1 Introduction

This document specifies the control plane procedures to support Traffic Engineering (TE) across TRILL sites where such sites are interconnected using [1] with the help of a Layer 3 core running IP+GRE or IP+MPLS. Traffic Engineering permits usage of a set of links that possess a certain characteristic like specified bandwidth, cost or even MTU. This draft aims at addressing how unicast frames travelling from one TRILL site to another across a Layer 3 core that supports IP+GRE and/or IP+MPLS can make use of the TE calculated paths in the sending site as well as the receiving site where such TE paths are pre-computed in both sites and need a mechanism to inter-link them together.

This draft merely enables the above through a Area number aliasing mechanism. The mechanism to interconnect multiple TRILL sites and also which provides multi-tenancy in the sense that multiple customers of a Layer 3 core can make use of this proposal, is enabled by [1].

1.1 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].

2. Methodology

Assume the following topology consisting of two sites belonging to the same customer that are interconnected by a Layer 3 core network running IP+GRE and/or IP+MPLS as defined in [1]. The two sites are considered as IS-IS Level 1 areas having their own set of nicknames which may be non-unique between the two sites. That is the same nickname could be used in one site and the other or even a third or more if the need arises. The sites are connected using the N-PEs which are the border Rbridges that have one interface in the IS-IS Level 1 area and another Pseudo-interface in the Level 2 area which is actually Pseudo-Level-2 which in fact is the Layer 3 core interconnecting the two.

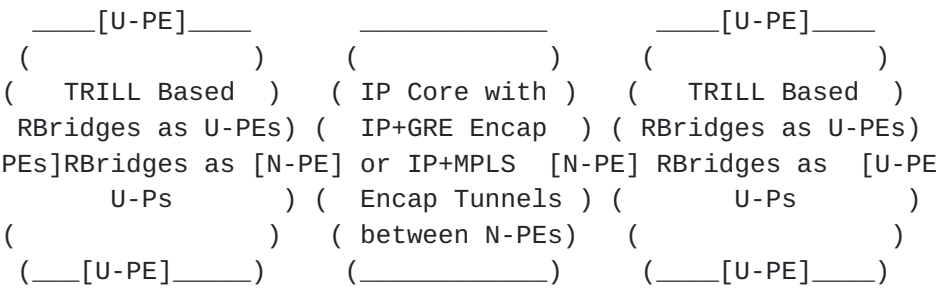


Figure 1.0 : Proposed Architecture

Legend :

U-PE : User-near PE device. 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 : 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.

As defined in [3] these separate sites are assigned unique area numbers so that the sites can be connected using multi-level IS-IS like configuration as specified in [1].

Here the MAC-routes and their corresponding Area number nicknames are placed in VRFs and using the BGP-MAC-VPN vrf methodology the sites are interconnected using BGP. BGP serves as the protocol that distributes the MAC-routes from one site to another. Specific Route Distinguishers (which are a capability of BGP based IP or MPLS VPNs) are used to assign uniqueness to the MAC-routes from amongst the sites. Route targets are used to export and import these routes in

and out of the N-PEs that interconnect these TRILL sites.

Here we advocate the use of a range Area number nicknames to be assigned to each site of a customer. Each Area number other than the default Area number which is used for non-TE based frames (both unicast and multicast), is assigned a significance for a specific pre-computed TE path within each site. We will call these Area numbers (other than the default Area number assigned for non-TE frames) as Site-TE-nicknames. These Site-TE-nicknames are distinct and unique for the set of customer sites interconnected by the Layer 3 core.j

Each of these Site-TE-nicknames represent a path computed on the basis of say a bandwidth, cost or MTU constraint. One could compute a path based on bandwidth that indicates that all the links in that TE-Path (represented by the Site-TE-nickname in that site) can carry traffic upto 10GB worth of traffic. Or the TE-Path computed may be based on cost indicating say that all the links in the TE-path have a cost over a threshold "X". Or the TE-path could be based on the fact that all the links in that path have a MTU over and above a threshold "M" or equal to "M". Combined constraints can also be used where bandwidth and MTU are to be considered.

So we assign specific Path Characteristics to a TE-Path and assign a Site-TE-nickname to it. Also the TE-nicknames for each TE-Path are assigned on each Rbridge and percolated / flooded within that site. The Site-TE-nicknames are then carried with Path Characteristic TLVs (to be defined for this purpose) through IS-IS in the site and advertised into BGP at the N-PE connecting the site to the Layer 3 core. The N-PE then uses a MP-BGP session to advertise these Site-TE-nicknames and the Path Characteristics in suitable format to other N-PEs across the Layer 3 core.

On the receiving N-PE the information is re-distributed into IS-IS TLVs and the Site-TE-nicknames reach all the Rbridges within the receiving site.

The reachability information in a Rbridge within the TE-Path based unicast frame sending site would be that the destination exists in another site whose Site-TE-nickname is reachable through the near end N-PE. A TE-Path within the local / sending site would have been computed based on bandwidth , cost or MTU or combination of these would have been constructed to the N-PE if such links possessing the characteristics exist.

Now an Ingress Rbridge can use its local Site-TE-nickname (for that TE-Path to the N-PE) if such a TE-Path the meets the constraints is available, as a Egress Nickname in the TRILL header to get a frame to

flow through its site through the locally available TE-Path and reach the connected N-PE within the site. At the N-PE the TRILL header is decapsulated and the Egress Nickname of the incoming frame looked up for equivalent path characteristics in the set of Site-TE-nicknames of the site where the MAC-route says the target host exists. If a match occurs then the local N-PE puts in the suitable Egress Nickname as that Site-TE-nickname and sends the packet with the TRILL header across to the remote site.

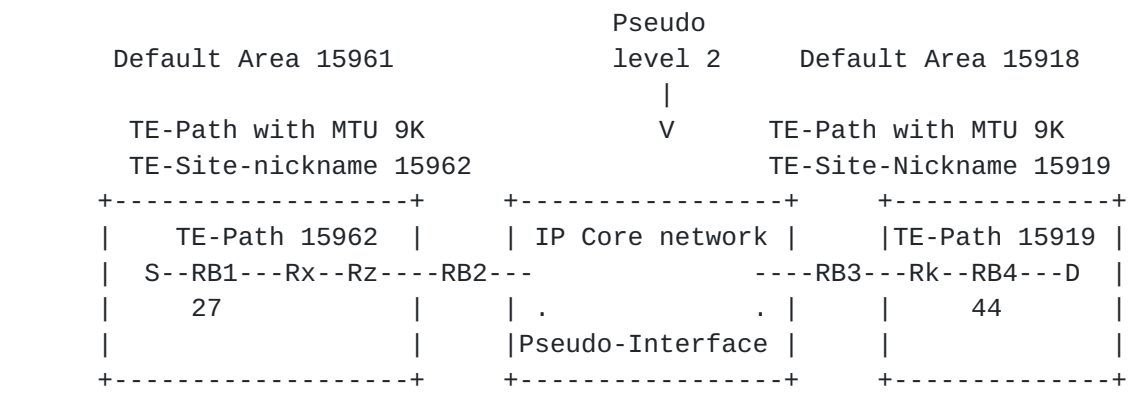
At the receiving site the Egress Rbridge Nickname in the TRILL header is inspected and the appropriate TE-Path from the receiving N-PE (remote site N-PE) to the Egress Rbridge terminating the TE-Path represented by the Site-TE-nickname is used to carry the unicast frame to its destination.

If no match exists for the path characteristics then the decapsulation still takes place and the normal discarding of the TRILL header over the L3 core as specified in [1] is done and the frame sent across to the other side. At the receiving end one of the existing normal paths (other than the TE-Paths) is used to get the packet to the target host.

If the sending site does not have a TE-Path to its local N-PE that meets the constraints then it would choose to send the unicast frame through normal means as in [1].

In the figure below we demonstrate how the data path is taken from sender to receiver assuming the sender is in one site and receiver in other.

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
   and through remote TE-Site-Nickname 15919 with MTU 9K and
   through local N-PE RB2. and through local TE-Site-Nickname
   15962 with MTU 9K through RB2.
2) RB3 has learned that D is attached to nickname 44.
   and through TE-Site-Nickname 15919 with MTU 9K

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 = 15962.

- RB2 has announced in the Level 1 IS-IS instance in area 15961, that it is attached to all the area nicknames, including 15918 and 15919 which is just an TE-Alias for 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.)

- RB2, when transitioning the frame from Level 1 to Level 2,

replaces the ingress RBridge nickname with the area nickname, so replaces 27 with 15962. Within Level 2, the ingress RBridge field in the TRILL header will therefore be 15962, and the egress RBridge field will be 15919 after the path characteristics matching and the choice of 15919 as the Egress Rbridge satisfying the said constraints for the TE-Path in 15919. Also RB2 learns that S is attached to nickname 27 in area 15962 to accommodate return traffic. This is thus a bi-directional TE-Path that satisfies the constraints chosen by the Ingress Rbridge RB1.

- The frame is forwarded through Level 2, to RB3, which has advertised, in Level 2, reachability to the nickname 15918 and 15919.

- RB3, when forwarding into area 15918, keeps the egress nickname in the TRILL header as 15919 nickname which is the TE-Path to RB4 whose actual nickname is 44. So, within the destination area, the ingress nickname will be 15962 and the egress nickname will be 15919.

- RB4, when decapsulating, learns that S is attached to nickname 15962, which is the TE-Path Site-TE-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 default 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 which is the MVPN PIM-bidir tree as in [1].

2.1 Extensions in IS-IS and BGP

The TLVs in IS-IS to be added and the BGP extensions will be dealt with in more detail in later versions of this draft. The other TLVs that support creation of TE-Paths as in [7] will remain as is.

3 Security Considerations

TBD.

4 IANA Considerations

Suitable IANA requests will be detailed in upcoming versions of the draft.

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-balaji-trill-over-ip-multi-level-04.txt, Bhargav Bhikkaji et.al, March 2012, Work in Progress

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

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

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

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

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

[7] draft-hu-trill-traffic-engineering-00.txt Fanwei Hu et.al, January 11 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

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

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

Email: Bhargav_Bhikkaji@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