Network Working Group
Internet Draft
Category: Standard Track
Expires: March 2011

Simon Delord, Telstra
Raymond Key, Telstra
Frederic Jounay, France Telecom
Yuji Kamite, NTT Communications
Zhihua Liu, China Telecom
Manuel Paul, Deutsche Telekom
Ruediger Kunze, Deutsche Telekom
Mach Chen, Huawei
Lizhong Jin, ZTE

September 28, 2010

Extension to LDP-VPLS for Ethernet Broadcast and Multicast
draft-delord-l2vpn-ldp-vpls-broadcast-exten-03.txt

Status of this Memo

Abstract

This document proposes a simple extension to LDP-VPLS to improve bandwidth efficiency for Ethernet broadcast/multicast traffic within a carrier’s network. It makes use of unidirectional point-to-multipoint PseudoWires to minimise payload frame duplication on physical links.

Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

Table of Contents

1. Introduction

This document proposes a simple extension to LDP-VPLS [RFC4762] to improve bandwidth efficiency for Ethernet broadcast/multicast traffic within a carrier's network. This bandwidth improvement is achieved by adding to the existing full-mesh of bidirectional point-to-point PseudoWires (P2P PWs) unidirectional point-to-multipoint PseudoWires (P2MP PWs) between selected PEs within a VPLS instance. With P2MP PWs, the ingress PE is not responsible for replicating the payload frame on each P2P PW towards the egress PE, instead the network elements along the physical path participate in the replication process. The replication is done by the underlying point-to-multi-point label switched path. This proposal allows for a large number of P2MP PWs to be carried through a single MPLS P2MP tunnel, thus, it is never necessary to maintain state in the network core for individual P2MP PWs.

2. Problem Statement & Motivation

2.1. Problem Statement

[RFC5501] provides an in-depth discussion on broadcast/multicast related requirements for VPLS. It highlights two specific issues:

- issue A: replication to non-member site.
- issue B: replication of PWs on shared physical path.

2.1.1. Issue A:

The current standard VPLS is a L2VPN service agnostic to customer’s Layer 3 traffic, hence does not maintain any information about IP multicast group membership. Although a Layer 3 IP multicast packet is encapsulated in a Layer 2 Ethernet multicast frame, the current standard VPLS treats Ethernet multicast frame in exactly the same way as Ethernet broadcast frame. There is therefore an issue that multicast traffic is sent to sites with no members. Since the upstream PE does not maintain downstream membership information, it simply floods frames to all downstream PEs, and the downstream PEs forward them to directly connected CEs; however, those CEs might not be the members of any multicast group.

There are therefore two elements to Issue A:

- the PE to CE section (e.g. the AC), where a CE will receive unintended traffic.

- the PE to PE section within a VPLS instance, where a PE will receive multicast traffic even when it has no CE being member of any multicast group.

To address the PE to CE part, a PE might have to maintain multicast group information for CEs that are not kept in the existing VPLS solutions.

To address the PE to PE part and limit the flooding scope across the backbone, a PE needs to discover multicast group information from other remote PEs.

Both elements will present scalability concerns about state resources (memory, CPU, etc.) and their maintenance complexity.

Finally, if Layer-3 information is checked for transport, the following [RFC4665] requirement "a L2VPN service SHOULD be agnostic to customer’s Layer3 traffic" can no longer be met.

2.1.2. Issue B:

Issue B on the other hand can still be improved without making use of any Layer-3-related information.

Issue B may still be considered acceptable when:

- Ethernet broadcast/multicast traffic volume is low; and
- The number of replications on each outgoing physical interface for a VPLS instance is small (e.g. not many PEs per VPLS instance).

However, with more broadcast/multicast applications (e.g. broadcast TV), Ethernet broadcast/multicast traffic volume may increase to a significant level. Assuming HDTV requires 10Mbps per channel, a bundle of 100 channels will require 1Gbps.

Furthermore, as MPLS networks expand from the core towards aggregation/access, more PEs may participate in a single VPLS instance. The number of replications on each outgoing physical interface for a VPLS instance is likely to increase.

2.2. Motivation

Based on the previous section, it may still be desirable for some carriers to look at improving issue B without having to look at Layer 3 information (Issue A).

One reason for this is that sometimes there is no L3 data to snoop. Another reason may be that some carriers may not be allowed to look above the L2 header, for example there may be regulatory issues with inspecting the customer payload. Also, some carriers may not want to do L3 snooping as Operations will naturally become more complicated if the number of managed objects (e.g. multicast groups) increases.

Another important point is that some carriers may want a manual and granular optimisation process that allows optimizations to certain services or areas but does not impact the rest of the network. For example, the bandwidth improvement process may only be required at specific locations in the network where bandwidth-intensive multicast broadcast Ethernet flows exist. It would also be beneficial if the optimisation process were incrementally deployable, so that the optimisation can still be leveraged even if there are portions of the network that are not able to support the features required by the optimisation process. A potential case would be a VPLS instance composed of both PEs supporting the proposed protocol extension and PEs not supporting it, the enhancement is then achieved between the compliant PEs only.

Finally, some carriers may also prefer a deterministic process to an entirely automated path selection algorithm that is network driven. [RFC5501] gives several reasons on why this may be the case:

- Accounting for various operator policies where the logical multicast topology within a carrier’s network does not change dynamically in conjunction with a customer’s multicast routing.

- Operations will naturally become more complicated if topology changes occur more frequently.

- Troubleshooting will tend to be difficult if a solution supports frequent dynamic membership changes with optimized transport within the carrier’s network.

[VPLS-Multicast-BGP] is a solution that looks at solving both Issue A and Issue B. However, [VPLS-Multicast-BGP] proposes that, even for carriers who currently use [RFC4762] without auto-discovery mechanisms, BGP be introduced (section 7). This may also present operational challenges and complexities for some carriers, or this feature may simply not be supported on some of the network elements deployed.

2.3. Scope of the proposed solution

This draft therefore explores whether there is a way to improve Layer 2 Ethernet broadcast/multicast bandwidth simply and predictably with:

- Minimal extension to [RFC4762] and without the need to add BGP (e.g. no auto-discovery)
- Minimal impact to existing [RFC4762] deployed networks
- Operator driven optimisation (i.e. the operator decides where and how the bandwidth improvement should occur) to minimise the number of states and the potential operational complexities associated with dynamic changes within a carrier’s core network.

3. Terminology

This document uses terminology described in [RFC4762] and [P2MP-PW-REQ].

4. Relevant IETF technologies for the proposed solution

The proposed solution relies on [RFC4762] existing mechanisms and complements them with extensions (P2MP LSPs and P2MP PWs) already standardised ([RFC4875]) or currently under development by the IETF ([mLDP] and [P2MP-PW-LDP]).

4.1. P2MP LSPs

Similarly to what is defined in [RFC4762] where P2P PWs are multiplexed onto P2P LSPs, before the operator can start deploying P2MP PWs, an appropriate underlying layer made of P2MP LPSs needs to be configured (section 3.2 of [P2MP-PW-REQ]).

P2MP LSPs are used to minimise packet replication on specific physical links and to allow P routers in an MPLS domain to be transparent to services (e.g. a P Router will join the P2MP PSN tunnel operation but will have no knowledge of the P2MP PWs, same as [RFC4762]).

The mapping of the P2MP LSP over the physical topology is a key component of the bandwidth enhancement exercise and the operator needs to carefully consider where and how these P2MP LSPs should be deployed (see Appendix A for an example of a possible deployment).

Once configured, it is then possible to aggregate P2MP PWs over a particular P2MP LSP (similar to [RFC4762]).

4.2. P2MP PWs

P2MP PWs can be configured statically (e.g. by the operator) or via LDP on top of the P2MP LSPs. This configuration is done on a per PE per VPLS instance basis.

In a P2MP PW, the operator decides to connect one Root PE to at least two Leaf PEs (section 3.1 of [P2MP-PW-REQ]).

The Root PE is the headend of the P2MP PW (where a big Ethernet multicast/broadcast talker is connected - see example in Appendix A).

The Leaf PEs are the endpoints of the P2MP PW (they constitute the receivers where the broadcast/multicast traffic needs to be distributed to).

A Root PE may map more than one P2MP PW to a specific VPLS instance. In this case, the Root PE MUST NOT associate a leaf PE to more than one P2MP PW for a specific VPLS instance (this is to avoid a Leaf PE to receive duplicate copies of the same Ethernet frame from different P2MP PWs).

P2MP PWs are defined in [P2MP-PW-REQ] and one solution using LDP as the signalling mechanism between PEs is defined in [P2MP-PW-LDP].

5. Proposed extension to [RFC4762]

This section updates [RFC4762] by describing the extra rules to be applied within a VPLS when unidirectional P2MP PWs are added to the existing full-mesh of P2P PWs.

5.1. VPLS Reference Model

Figure 1 shows a topological model (not the physical topology) of a VPLS between four PEs with an arbitrary set of ACs attached to each VSI.

```
           +---------+            +---------+
           |   PE1   |            |   PE2   |
           |  +---+  |            |  +---+  |
           |  |   |  |            |  |   +-------AC4---
           |  | V |  |            |  | V |  |
           |  |   |  |            |  |   |  |
----AC1----+--+   |  |  Ethernet  |  |   +--+----AC5---
           |  | S +--+-----PW-----+--+ S |  |
           |  |   |  |            |  |   |  |
----AC2----+--+   |  |            |  |   |  |
           |  | I |  |            |  | I |  |
           |  |   |  | Ethernet   |  |   |  |
----AC3----+--+   |  |   PW +-----+--+   |  |
           |  +-+-+  |    /       |  +-+-+  |
           |    |  \ |   /        |    |    |
           +----+---\+  /         +----+----+
                |    \ /               |
        Ethernet|     \/               |Ethernet
              PW|     /\               |PW
                |    /  \Ethernet      |
           +----+---/+   \PW      +----+----+
           |PE3 |  / |    \       |PE4 |    |
           |  +-+-+  |     \      |  +-+-+  |
----AC6----+--+   |  |      \     |  |   +--+----AC8----
           |  | V |  |       +----+--+ V |  |
           |  |   |  |            |  |   |  |
----AC7----+--+   |  |  Ethernet  |  |   +--+----AC9----
           |  | S +--+-----PW-----+--+ S |  |
           |  |   |  |            |  |   |  |
           |  |   |  |            |  |   |  |
           |  | I |  |            |  | I |  |
           |  |   |  |            |  |   |  |
           |  |   |  |            |  |   |  |
           |  +---+  |            |  +---+  |
           |         |            |         |
           +---------+            +---------+
```

Figure 1: Reference Diagram for VPLS

Figure 2 shows the proposed extensions to VPLS for Ethernet broadcast and multicast. On top of the topology presented in Figure 2, two P2MP PWs have been added to the existing set of P2P PWs.

```
                    +---------+            +---------+
                    |   PE1   |            |   PE2   |
                    |  +---+  |            |  +---+  |
                    |  |   |  |            |  |   +-------AC4---
                    |  | V |  |  Ethernet  |  | V |  |
                    |  |   |  |    P2MP    |  |   |  |
         ----AC1----+--+   |  |     PW1    |  |   +--+----AC5---
                    |  | S +--+->--+-->----+--+ S |  |
                    |  |   |  |    |       |  |   |  |
         ----AC2----+--+   |  |    |       |  |   |  |
                    |  | I |  |    |   +->-+--+ I |  |
                    |  |   |  |    |   |   |  |   |  |
         ----AC3----+--+   |  |    |   |   |  |   |  |
                    |  +---+  |    |   |   |  +---+  |
                    |         |    |   |   |         |
                    +---------+    |   |   +---------+
                                   |   |
                                   |   |
                                   |   |
                                   |   |
                    +---------+    |   |   +---------+
                    |PE3      |    |   |   |PE4      |
                    |  +---+  |    |   |   |  +---+  |
         ----AC6----+--+   |  |    +->-----+--+   +--+----AC8----
                    |  | V |  |        |   |  | V |  |
                    |  |   |  |        |   |  |   |  |
         ----AC7----+--+   |  |        |   |  |   +--+----AC9----
                    |  | S +--+->------+->-+--+ S |  |
                    |  |   |  |   Ethernet |  |   |  |
                    |  |   |  |     P2MP   |  |   |  |
                    |  | I |  |     PW2    |  | I |  |
                    |  |   |  |            |  |   |  |
                    |  |   |  |            |  |   |  |
                    |  +---+  |            |  +---+  |
                    |         |            |         |
                    +---------+            +---------+
```

Figure 2: Proposed Extensions to VPLS for Ethernet broadcast/multicast

P2MP PW1 is composed of PE1 as the root PE and PE2 and PE4 as leaf PEs.

P2MP PW2 is composed of PE3 as the root PE and PE2 and PE4 as leaf PEs.

Note that for sake of clarity, Figure 2 does not show the full-mesh of P2P PWs presented in Figure 1.

Also note that the solution does not require that P2MP PWs be used on all PEs in the VPLS, for example there is only a P2P PW between PE1 and PE3 and a P2P PW between PE2 and PE4.

5.2. Choosing PEs for a specific VPLS to be connected by a P2MP PW

This updates section 4.3 of [RFC4762].

VPLS is a full-mesh of P2P PWs and optionally a number of unidirectional P2MP PWs. At the difference of P2P PWs, not all PEs in a VPLS instance need to be connected via P2MP PWs.

For each P2MP PW on this VPLS instance:

- The operator selects one PE as the Root of the P2MP PW.

- The operator also selects two or more PEs belonging to the same VPLS instance to be Leafs of the P2MP PW

- Because there is already a full-mesh of bidirectional P2P PWs between all PEs, the P2MP PW is unidirectional only (e.g. from the Root PE to all the Leaf PEs connected to it).

- The operator also needs to make sure that there is an active P2MP LSP setup between the Root PE and the Leaf PEs:

- If there is already an active P2MP LSP setup between the Root PE and the Leaf PEs, then procedures described in 5.3 can be followed.

- If there is no P2MP LSP between the Root PE and the Leaf PEs then the operator needs to create first a P2MP LSP in order for procedures in 5.3 to be followed. Procedures to setup a P2MP LSP will vary based on the technology used and are described in [mLDP] and [RFC4875].

5.3. Create and associate the P2MP PW to a specific VPLS Instance

This updates section 4.3 of [RFC4762].

Once that the endpoints of the P2MP PW have been selected and that there is an active P2MP LSP between them, the operator can then create and associate the P2MP PW to a specific VPLS instance. This activity can be done statically or via LDP [P2MP-PW-LDP].

Because P2MP PWs are used to demultiplex encapsulated Ethernet frames from multiple VPLS instances that are aggregated over the same P2MP transport LSP, it is necessary that a Leaf PE can associate unambiguously a P2MP PW aggregated within a P2MP LSP to both a specific VPLS instance and a Root PE.

In the static case, the operator is responsible for configuring all the required information on all PEs belonging to the P2MP PW.

In the LDP case, the P2MP PW is initiated by the Root PE by sending a P2MP PW LDP Label Mapping Message to each of the Leaf PEs.

This label mapping contains, the VPLS instance the P2MP PW is associated to, the P2MP LSP used to transport the P2MP PW and the P2MP PW MPLS Label.

The P2MP PW MPLS Label is upstream assigned and allocated according to the rules in [RFC5331].

The root PE imposes the upstream-assigned label on the outbound packets sent over the P2MP-PW and using this label a Leaf PE can identify the inbound packets arriving over the P2MP PW.

Detailed LDP message formats and P2MP PW setup procedures are described in [P2MP-PW-LDP].

5.4. Mapping more than one P2MP PW to a specific VPLS Instance on a specific Root PE

The proposed solution allows for a Root PE to map more than one P2MP PW to a specific VPLS instance (see example in Appendix A).

However in this case, the Root PE MUST NOT associate a leaf PE to more than one P2MP PW for a specific VPLS instance (this is to avoid a Leaf PE to receive duplicate copies of the same Ethernet frame from different P2MP PWs).

5.5. Flooding and Forwarding

This section updates section 4.1. of [RFC4762].

A root PE MUST NOT flood frames simultaneously over P2MP PW and P2P PW toward the same leaf PE.

For the flooding of an Ethernet broadcast/multicast frame over PWs to remote PEs participating in the VPLS:

- If there is P2MP PW towards a remote PE, the P2P PW associated with this remote PE will not be used. One copy of the frame will be forwarded on the P2MP PW for all the remote PEs associated with it.
- If there is no P2MP PW towards a remote PE, the P2P PW associated with this remote PE is used.

It should be noted that local policy on the Root PE at the operator’s operational request can override any decision to flood and forward traffic over a P2MP PW for a VPLS instance. In that case, normal flooding procedures over P2P PWs described in 4.1 of [RFC4762] apply.

5.5.1. Flooding and Forwarding for Ethernet unknown unicast

In traditional Ethernet switched networks unknown unicast frames are handled the same way as broadcast and multicast Ethernet traffic (e.g. flooding). Similarly, current VPLS standards also handle unknown unicast traffic by flooding it across all P2P PWs.

The main purpose of this document is to address Ethernet broadcast and multicast traffic. For Ethernet unknown unicast frames there are two possibilities:

- forward the unknown unicast traffic on the P2MP PW, same as for Ethernet broadcast and multicast.
- keep the existing mechanism of [RFC4762] and flood over the mesh of P2P PWs.

Details on how Ethernet unknown unicast traffic should be handled will be added in a future revision of this document.

5.6. Address Learning

This section updates section 4.2. of [RFC4762].

A Leaf PE MUST support the ability to perform MAC address learning for packets received on a P2MP PW.

When a Leaf PE receives an Ethernet frame on a P2MP PW it:

- First determines the VSI associated to the P2MP PW
- Then determines the Root PE of the P2MP PW
- Then determines the P2P PW associated with that Root PE
- Finally, creates a forwarding state in the VPLS instance for the P2P PW associated with the Root PE with a destination MAC address being the same as the source MAC address being learned.

5.7. Loop Free Topology

This updates section 4.4. of [RFC4762]

Paragraph 2 "must not forward from one PW to another" is applicable to P2MP PW & P2P PW.

5.8. Hierarchical VPLS

H-VPLS considerations will be added in a later revision.

5.9 P2MP PW Status

In case of a P2MP PW status change to not operational as per [P2MP-PW-LDP], then this should be treated as if this P2MP PW does not exist.

6. Local PE Implementation

This section is OPTIONAL.

As described in section 2.1.1, a PE receiving an IP multicast frame, will forward it to all ACs, including those with no member of the specific IP multicast group attached.

Unnecessary traffic consumes bandwidth on the access link and may become a concern from the customer perspective. In some cases, it may also be a security concern as the multicast frame may be forwarded to an endpoint other than the intended destinations.

Consequently, the use of some L3 related supplementary information in order to improve bandwidth consumption on the AC may be considered. Enabling L3 snooping on an AC basis only has an impact on the PE where the AC belongs, it does not impact the number of P2MP PW/LSPs used within the carrier’s network and the state resources or the maintenance complexity associated with it.

Alternatives to L3 snooping such as static configuration of multicast Ethernet addresses & ports / interfaces for example are also possible.

7. Security Considerations

This section will be added in a future version.

8. IANA Considerations

There are no specific IANA considerations in this document.

9. Acknowledgments

This section will be added in a future version.

10. References

10.1. Normative References

[RFC2119]   Bradner, S., Key words for use in RFCs to Indicate Requirement Levels, BCP 14, RFC 2119, March 1997.

[RFC4762]   Lasserre & Kompella, Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling, January 2007

[P2MP-PW-LDP] L. Martini, F. Jounay, et. al, "Signaling Root-Initiated Point-to-Multipoint Pseudowires using LDP", draft-ietf-pwe3-p2mp-pw-00.txt, work in Progress, July 2010.

[mLDP] I. Minei, K. Kompella, I. Wijnands, B. Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", draft-ietf-mpls-ldp-p2mp-10, Work In Progress, July 2010.

[RFC4875] R. Aggarwal, Ed., D. Papadimitriou, Ed., S. Yasukawa, Ed., "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs).", rfc4875, May 2007.

[RFC5331] R. Aggarwal, Y. Rekhter, E. Rosen, "MPLS Upstream Label Assignment and Context Specific Label Space", RFC 5331, August 2008

10.2. Informative References

[RFC5501]   Kamite, et al., Requirements for Multicast Support in Virtual Private LAN Services, March 2009

[P2MP-PW-REQ] F. Jounay, et. al, "Requirements for Point to Multipoint Pseudowire", draft-ietf-pwe3-p2mp-pw-requirements-03.txt, Work in Progress, August 2010.

[VPLS-Multicast-BGP] Raggarwa, Kamite & Fang, "Multicast in VPLS", draft-ietf-l2vpn-vpls-mcast-07.txt, Work in Progress, September 2010

[RFC4665]   Augustyn, W. and Y. Serbest, "Service Requirements for Layer 2 Provider-Provisioned Virtual Private Networks", RFC 4665, September 2006

Author's Addresses

Simon Delord
Telstra
242 Exhibition Street
Melbourne, VIC, 3000, Australia
E-mail: simon.delord@gmail.com

Raymond Key
Telstra
242 Exhibition Street
Melbourne, VIC, 3000, Australia
E-mail: raymond.key@ieee.org

Frederic Jounay
France Telecom
2, avenue Pierre-Marzin
22307 Lannion Cedex
France
Email: frederic.jounay@orange-ftgroup.com

Yuji Kamite
NTT Communications Corporation
Granpark Tower
3-4-1 Shibaura, Minato-ku
Tokyo 108-8118
Japan
Email: y.kamite@ntt.com

Zhihua Liu
China Telecom
109 Zhongshan Ave., Guangzhou
510630, China
Email: zhliu@gsta.com

Manuel Paul
Deutsche Telekom
Goslarer Ufer 35
10589 Berlin, Germany
Email: manuel.paul@telekom.de

Ruediger Kunze
Deutsche Telekom
Goslarer Ufer 35
10589 Berlin, Germany
Email: ruediger.kunze@telekom.de

Mach(Guoyi) Chen
Huawei Technology Co., Ltd.
No. 9 Xinxi Road
Shangdi Information Industry Base
Hai-Dian District, Beijing 100085
China
EMail: mach@huawei.com

Lizhong Jin
ZTE Corporation
889, Bibo Road
Shanghai, 201203, China
Email: lizhong.jin@zte.com.cn

A. One example for broadcast video delivery

This section describes one deployment scenario in relation to broadcast video delivery and how the proposed solution would work.

One requirement of the model is that the application needs unicast data exchange (IP unicast transfer or control messages etc.) as a background environment. MAC-learning (and therefore VPLS) is effective to support it.

A.1. Broadcast Video Delivery Topology

Figure 3 presents the physical topology of one broadcast video deployment.

```
                              |     |     |
                              AC    AC    AC
                              |     |     |
                            +---+ +---+ +---+
                            |PE3| |PE4| |PE5|
                            +---+ +---+ +---+
                               \    |    /
                                \   |   /                 +----+
                                 \  |  /              +---|PE9 |--AC--
                                  \ | /              /    +----+
   Ethernet        +---+          +---+        +---+/     +----+
   Broadcast->-AC--|PE1|----------|P1 |--------|P3 |------|PE10|--AC--
   Source          +---+\        /+---+        +---+\     +----+
                         \      /   |            |   \    +----+
                          \    /    |            |    +---|PE11|--AC--
                           \  /     |            |        +----+
                            \/      |            |
                            /\      |            |
                           /  \     |            |        +----+
                          /    \    |            |    +---|PE12|--AC--
                         /      \   |            |   /    +----+
   Ethernet        +---+/        \+---+        +---+/     +----+
   Broadcast->-AC--|PE2|----------|P2 |--------|P4 |------|PE13|--AC--
   Source          +---+          +---+        +---+\     +----+
                                  / | \              \    +----+
                                 /  |  \              +---|PE14|--AC--
                                /   |   \                 +----+
                               /    |    \
                            +---+ +---+ +---+
                            |PE6| |PE7| |PE8|
                            +---+ +---+ +---+
                              |     |     |
                              AC    AC    AC
                              |     |     |
```

Figure 3 - Physical Topology for Broadcast video

Figure 3 is split in three logical components:

- The Core network composed of P1, P2, P3 & P4. These 4 network elements are P routers connected in a ring.

- The Data Centers. These are a few large PoPs with high resiliency that hold the video content. PE1 & PE2 are located in one Data Center and are dual-homed to the core network. An Ethernet broadcast source is connected to each PE in the Data Center.

- The Aggregation network. The Aggregation network is responsible for aggregating last mile technology towards endusers (direct fiber, GPON, DSL, etc.). PE3, PE4, until PE14 are VPLS PE routers in an aggregation PoP and single-homed to the Core network.

There are two different video distribution services organised as follows:

- PE1 is connected to PE3, PE4, ...PE14 via VPLS instance-1.
- One Ethernet broadcast source is connected to PE1 into VPLS instance-1.
- PE2 is connected to PE3, PE4 ...PE14 via VPLS instance-2.
- One Ethernet broadcast source is connected to PE2 into VPLS instance-2.

A.2. Impact of Physical Topologies on Ethernet Broadcast/multicast replication

Following the standard VPLS ingress replication mechanism, each time PE1 receives one broadcast frame from the ethernet broadcast source on VPLS-1, PE1 will replicate 12 times the incoming frame.

Similarly, each time PE2 receives one broadcast frame from the ethernet broadcast source on VPLS-2, PE2 will replicate 12 times the incoming frame.

A.3. Proposed enhancement of Ethernet broadcast/multicast

The proposed enhancements are done in three steps:

- create P2MP LSPs for the infrastructure. These P2MP LSPs are used to carry one or more P2MP PWs.
- create unidrectional P2MP PWs by selectively choosing PEs where the optimisation should occur.
- forward ethernet broadcast/multicast frames onto the P2MP PWs where these P2MP PWs have been created.

It is up to the network operator to decide how the distribution of the loading on physical link should occur.

Two different examples are presented below.

A.3.1. One possible enhancement scenario

A.3.1.1. Initial Deployment

In this scenario, the operator decides to create the following P2MP LSPs:

- PE1->PE3-5 via P1 called LSP1
- PE1->PE6-8 via P2 called LSP2
- PE1->PE9-11 via P3 called LSP3

- PE2->PE3-5 via P1 called LSP4
- PE2->PE6-8 via P2 called LSP5
- PE2->PE9-11 via P3 called LSP6

The operator then creates the following P2MP PWs:

- PE1->PE3-5 via P2MP PW1 over LSP1
- PE1->PE6-8 via P2MP PW2 over LSP2
- PE1->PE9-11 via P2MP PW3 over LSP3

- PE2->PE3-5 via P2MP PW4 over LSP4
- PE2->PE6-8 via P2MP PW5 over LSP5
- PE2->PE9-11 via P2MP PW6 over LSP6

There is no P2MP PWs between PE1 and PE12, PE13 and PE14.
There is no P2MP PWs between PE2 and PE12, PE13 and PE14.

There are several reasons why a P2MP PW may not be available on this part of the network (e.g. PE12, PE13 and PE14), for example:

- the hardware/software may not allow the support of the required features (P2MP LSPs and/or P2MP PWs).
- the operator does not need to improve multicast/broadcast services there (e.g. no specific bandwidth issue).
- the operator is currently under a migration phase where only part of the network is migrated at a time.

In this case, when PE1 receives one broadcast frame from the Ethernet broadcast source on VPLS-1:

- PE1 sends one copy of the broadcast frame onto P2MP PW1
- PE1 sends one copy of the broadcast frame onto P2MP PW2
- PE1 sends one copy of the broadcast frame onto P2MP PW3
- PE1 sends one copy onto the P2P PW towards PE12
- PE1 sends one copy onto the P2P PW towards PE13
- PE1 sends one copy onto the P2P PW towards PE14

PE1 only replicates 6 copies now (this is an improvement from 12 copies if only using P2P PWs).

A.3.1.2 Multiple P2MP PWs

Let’s assume now that a new broadcast service is targeted at covering endusers geographically connected to PE9, PE10 and PE11.

For example, this could be a wholesale service, where another carrier with limited footprint for the region covered by PE9, PE10 and PE11 is seeking access for deploying its own broadcast application.

Based on the proposal in this document, and assuming that the application also needs unicast data exchange, if the new broadcast source is connected to PE1, it is then possible to:

- Create a new VPLS instance on PE1, PE9, PE10 and PE11 and a full-mesh of P2P PWs between all 4 PEs.

- Build a new P2MP PW, called P2MP PW7 between PE1, PE9, PE10 & PE11 that uses the existing P2MP LSP - LSP3.

This proposal allows for both P2MP PW3 and P2MP PW7 to be carried through a single MPLS P2MP tunnel, thus, removing the need to maintain state in the network core for individual P2MP PWs. The P routers in the core only need to be aware of the P2MP LSPs.

A.3.2. Another possible enhancement scenario

In this scenario, the operator decides to create the following two P2MP LSPs:

- PE1-> PE3-14 via LSP1:
    - P1 as a branch towards PE3, PE4, PE5, P2 and P3
    - P2 as a branch towards PE6, PE7, PE8 and P4
    - P3 as a branch towards PE9, PE10 and PE11
    - P4 as a branch towards PE12, PE13 and PE14

- PE2-> PE3-14 via LSP2:
    - P2 as a branch towards PE6, PE7, PE8, P1 and P4
    - P1 as a branch towards PE3, PE4, PE5 and P3
    - P3 as a branch towards PE9, PE10 and PE11
    - P4 as a branch towards PE12, PE13 and PE14

The operator then creates the following P2MP PWs:

- PE1-> PE3-14 via P2MP PW1 over LSP1
- PE2-> PE3-14 via P2MP PW2 over LSP2

This case improves the P2P PW scenario as PE1 only replicates a single copy of the broadcast frame received from the ethernet broadcast source.

L2VPN Working Group
Internet Draft
Intended status: Standard
Expires: January 21, 2011

Pranjal Kumar Dutta
Florin Balus
Alcatel-Lucent

Olen Stokes
Extreme Networks

Geraldine Calvignac
France Telecom

October 25, 2010

LDP Extensions for Optimized MAC Address Withdrawal in H-VPLS
draft-ietf-l2vpn-vpls-ldp-mac-opt-03.txt

Status of this Memo

Copyright Notice

Abstract

[RFC4762] describes a mechanism to remove or unlearn MAC addresses that have been dynamically learned in a VPLS Instance for faster convergence on topology change. The procedure also removes MAC addresses in the VPLS that do not require relearning due to such topology change.

This document defines an enhancement to the MAC Address Withdrawal procedure with empty MAC List [RFC4762], which enables a Provider Edge(PE) device to remove only the MAC addresses that need to be relearned.

Additional extensions to [RFC4762] MAC Withdrawal procedures are specified to provide optimized MAC flushing for the PBB-VPLS specified in [PBB-VPLS Model].

Table of Contents

1.1. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC-2119.

This document uses the terminology defined in [PBB-VPLS Model], [RFC5036], [RFC4447] and [RFC4762]. Throughout this document VPLS means the emulated bridged LAN service offered to a customer. H-VPLS means the hierarchical connectivity or layout of MTU-s and PE devices offering the VPLS [RFC4762]. The terms spoke node and MTU-s in H-VPLS are used interchangeably.

2. Introduction

A method of Virtual Private LAN Service (VPLS), also known as Transparent LAN Service (TLS) is described in [RFC4762]. A VPLS is created using a collection of one or more point-to-point pseudowires (PWs) [RFC4664] configured in a flat, full-mesh topology. The mesh topology provides a LAN segment or broadcast domain that is fully capable of learning and forwarding on Ethernet MAC addresses at the PE devices.

This VPLS full mesh core configuration can be augmented with additional non-meshed spoke nodes to provide a Hierarchical VPLS (H-VPLS) service [RFC4762].

[PBB-VPLS Model] describes how Provider Backbone Bridging (PBB) can be integrated with VPLS to allow for useful PBB capabilities while continuing to avoid the use of MSTP in the backbone. The combined solution referred to as PBB-VPLS results in better scalability in terms of number of service instances, PWs and C-MACs that need to be handled in the VPLS PEs.

A MAC Address Withdrawal mechanism for VPLS is described in [RFC4762] to remove or unlearn MAC addresses for faster convergence on topology change in resilient H-VPLS topologies.

An example of usage of the MAC Flush mechanism is the dual-homed H-VPLS where an edge device termed as MTU-s is connected to two PE devices via primary spoke PW and backup spoke PW respectively. Such redundancy is designed to protect against the failure of primary spoke PW or primary PE device. When the MTU-s switches over to the backup PW, it is required to flush the MAC addresses learned in the corresponding VSI in peer PE devices participating in full mesh, to avoid black holing of frames to those addresses. Note that forced switchover to backup PW can be also performed at MTU-s administratively due to maintenance activities on the primary spoke PW. When the backup PW is made active by the MTU-s, it triggers LDP Address Withdraw Message with a list of MAC addresses to be flushed. The message is forwarded over the LDP session(s) associated with the newly activated PW. In order to minimize the impact on LDP convergence time and scalability when a MAC List TLV contains a large number of MAC addresses, many implementations use a LDP Address Withdraw Message with an empty MAC List. Throughout this document the term MAC Flush Message is used to specify LDP Address Withdraw Message with empty MAC List described in [RFC4762] unless specified otherwise.

As per the MAC Address Withdrawal processing rules in [RFC4762] a PE device on receiving a MAC flush message removes all MAC addresses associated with the specified VPLS instance (as indicated in the FEC TLV) except the MAC addresses learned over the newly activated PW. The PE device further triggers a MAC flush message to each remote PE device connected to it in the VPLS full mesh.

This method of MAC flushing is modeled after Topology Change Notification (TCN) in Rapid Spanning Tree Protocol (RSTP)[802.1w]. When a bridge switches from a failed link to the backup link, the bridge sends out a TCN message over the newly activated link. The upstream bridge upon receiving this message flushes its entire MAC addresses except the ones received over this link and sends the TCN message out of its other ports in that spanning tree instance. The message is further relayed along the spanning tree by the other bridges. When a PE device in the full-mesh of H-VPLS receives a MAC flush message it also flushes MAC addresses which are not affected due to topology change, thus leading to unnecessary flooding and relearning. This document describes the problem and a solution to optimize the MAC flush procedure in [RFC4762] so it flushes only the set of MAC addresses that require relearning when topology changes in H-VPLS. The solution proposed in this document is generic and is applicable when MS-PWs are used in interconnecting PE devices in H-VPLS.

[PBB-VPLS Model] describes how PBB can be integrated with VPLS to allow for useful PBB capabilities while continuing to avoid the use of MSTP in the backbone. The combined solution referred as PBB-VPLS results in better scalability in terms of number of service instances, PWs and C-MACs that need to be handled in the VPLS PEs.

This document describes also extensions to LDP MAC Flush procedures described in [RFC4762] required to build desirable capabilities to PBB-VPLS solution.

Section 3 covers the problem space. Section 4 describes the solution and the required TLV extensions.

3. Problem Description

3.1. MAC Flush in regular H-VPLS

Figure 1 describes a dual-homed H-VPLS scenario for a VPLS instance where the problem with the existing MAC flush method in [RFC4762] is explained.

```
                                 PE-1                         PE-3
                               +--------+                  +--------+
                               |        |                  |        |
                               |   --   |                  |   --   |
   Customer Site 1             |  /  \  |------------------|  /  \  |->
     CE-1               /------|  \ s/  |                  |  \S /  |
       \     primary spoke PW  |   --   |           /------|   --   |
        \             /        +--------+          /       +--------+
         \    (MTU-s)/              |    \        /             |
          +--------+/               |     \      /              |
          |        |                |      \    /               |
          |   --   |                |       \  /                |
          |  /  \  |                |      H-VPLS Full Mesh Core|
          |  \S /  |                |       / \                 |
          |   --   |                |      /   \                |
         /+--------+\               |     /     \               |
        /     backup spoke PW       |    /       \              |
       /              \        +--------+         \--------+--------+
      CE-2             \       |        |                  |        |
   Customer Site 2      \------|  --    |                  |  --    |
```

```
                              | /  \   |------------------| /  \   |->
                              | \s /   |                  | \S /   |
                              |  --    |                  |  --    |
                              +--------+                  +--------+
                                PE-2                         PE-4
```

Figure 1: Dual homed MTU-s in two tier hierarchy H-VPLS

In Figure 1, the MTU-s is dual-homed to PE-1 and PE-2. Only the primary spoke PW is active at MTU-s, thus PE-1 is acting as the active device to reach the full mesh in the VPLS instance. The MAC addresses of nodes located at access sites (behind CE1 and CE2) are learned at PE-1 over the primary spoke PW. PE-2, PE-3 and PE-4 learn those MAC addresses on their respective mesh PWs terminating to PE-1. When MTU-s switches to the backup spoke PW and activates it, PE-2 becomes the active device to reach the full mesh core. Traffic entering the H-VPLS from CE-1 and CE-2 is diverted by the MTU-s to the backup spoke PW. For faster convergence MTU-s may desire to unlearn or remove the MAC addresses that have been learned in the upstream VPLS full-mesh through PE-1. MTU-s may send a MAC flush message to PE-2 once the backup PW has been made active. As per the processing rules defined in [RFC4762], PE-2 flushes the MAC addresses learned in the VPLS from the PWs terminating at PE-1, PE-3 and PE-4.

In the H-VPLS core, PE devices are connected in full mesh unlike the spanning tree connectivity in bridges. So the MAC addresses that require flushing and relearning at PE-2 are only the MAC addresses those have been learned on the PW connected to PE-1.

PE-2 further relays MAC flush messages to all other PE devices in the full mesh. Same processing rule applies at all those PE devices. For example, at PE-3 all of the MAC addresses learned from the PWs connected to PE-1 and PE-4 are flushed and relearned subsequently. As the number of PE devices in the full-mesh increases, the number of unaffected MAC addresses flushed in a VPLS instance also increases, thus leading to unnecessary flooding and relearning. With large number of VPLS instances provisioned in the H-VPLS network topology the amount of unnecessary flooding and relearning increases.

3.2. Black holing issue in PBB-VPLS

In PBB-VPLS solution a B-component VPLS (B-VPLS) may be used as infrastructure for one or more I-component instances. B-VPLS control plane (LDP Signaling) replaces I-component control plane throughout the MPLS core. This is raising an additional challenge related to black hole avoidance in the I-component domain as described in this section. Figure 2 describes the case of a CE device (node A) dual-homed to two I-component instances located on two PBB-VPLS PEs (PE1 and PE2).

```
                          IP/MPLS Core
                        +--------------+
                        |PE2           |
                       +-----+         |
                       |PBB |   +-+    |
                   _   |VPLS|---|P|    |
                     S/+-----+  /+-+\   |PE3
                     / +-----+ /     \+-----+
               +---+/  |PBB |/  +-+  |PBB |   +---+
      CMAC X--|CE |---|VPLS|---|P|--|VPLS|---|CE |--CMAC Y
               +---+ A +-----+   +-+  +-----+   +---+
                 A      |PE1           |        B
                        |              |
                        +--------------+
```

Figure 2: PBB Black holing Issue - CE Dual-Homing use case

The link between PE1 and CE A is active (marked with A) while the link between CE A and PE2 is in Standby/Blocked status. In the network diagram CMAC X is one of the MAC addresses located behind CE A in the customer domain, CMAC Y is behind CE B and the BVPLS instances on PE1 are associated with backbone MAC (BMAC) B1 and PE2 with BMAC B2.

As the packets flow from CMAC X to CMAC Y through PE1 of BMAC B1, the remote PEs participating in the IVPLS (for example, PE3) will learn the CMAC X associated with BMAC B1 on PE1. Under failure of the link between CE A and PE1 and activation of link to PE2, the remote PEs (for example, PE3) will black-hole the traffic destined for customer MAC X to BMAC B1 until the aging timer expires or a packet flows from X to Y through the PE B2. This may take a long time (default aging timer is 5 minutes) and may affect a large number of flows across multiple I-components.

A possible solution to this issue is to use the existing LDP MAC Flush as specified in [RFC4762] to flush in the BVPLS domain the BMAC associated with the PE where the failure occurred. This will automatically flush the CMAC to BMAC association in the remote PEs. This solution though has the disadvantage of producing a lot of unnecessary MAC flush in the B-VPLS domain as there was no failure or topology change affecting the Backbone domain.

A better solution is required to propagate the I-component events through the backbone infrastructure (B-VPLS) in order to flush only the customer MAC to BMAC entries in the remote PBB-VPLS PEs. As there are no IVPLS control plane exchanges across the PBB backbone, extensions to B-VPLS control plane are required to propagate the I-component MAC Flush events across the B-VPLS.

4. Solution description

4.1. MAC Flush Optimization for regular H-VPLS

The basic principle of the optimized MAC flush mechanism is explained with reference to Figure 1. On switching over to the backup spoke PW when MTU-s triggers MAC flush message to PE-2, it also communicates the unique PW endpoint identifier (PE-ID) in PE-1, the formerly active PE device. In VPLS a PW terminates on a Virtual Switching Instance (VSI) in a PE device. The PE-ID is relayed in all the subsequent MAC flush messages triggered by PE-2 to its peer PE devices in the full mesh. Each PE device that receives the message identifies the VPLS (From FEC TLV) and its respective PW that terminates in PE-1 (from PE-ID). Thus the PE device flushes only the MAC addresses learned from that PW connected to PE-1.

This section defines a PW Endpoint Identifier (PE-ID) TLV for LDP [RFC5036]. The PE-ID TLV carries the unique identifier of a generic PW endpoint.

4.1.1. PE-ID TLV Format

The encoding of PE-ID TLV follows standard LDP TLV encoding in [RFC5036]. A PE-ID TLV contains a list of one or more PE-ID Elements. Its encoding is:

```
      0                   1                   2                   3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |1|0|  PE-ID  TLV (0x0405)      |             Length            |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |                        PE-ID Element 1                        |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |                                                               |
     ~                                                               ~
     |                                                               |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |                        PE-ID Element n                        |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

U (Unknown) bit of thus LDP TLV MUST be set to 1. If the PE-ID TLV is not understood then it is ignored the receiving device.

F (Forward) MUST be set to 0. Since the LDP mechanism used here is targeted, the TLV is not forwarded if it is not understood by the receiving device.

The Type field MUST be set to 0x405 (subject to IANA approval). This identifies the TLV type as PE-ID TLV.

Length field specifies the total length in octets of the Value in PE-ID TLV.

PE-ID Element 1 to PE-ID Element n: there are several types of PE-ID Elements. The PE-ID Element Encoding depends on the type of the PE-ID Element. A PE-ID Element uniquely identifies a PW Endpoint.

A PE-ID Element value is encoded as 1 octet field that specifies the element type, 1 octet field that identifies the length in octets of the element value, and a variable length field that is type dependent element value.

The PE-ID Element value encoding is:

PE-ID name	Type	Length	Value
FEC-128 specific	0x01	12 octets	See below.
FEC-129 specific	0x02	Variable	See below.

The type of PE-ID Element depends on the type of FEC Element used to provision the respective PW. [RFC4447] defines two types of FEC elements that may be used for provisioning PWs - Pwid FEC (type 128) and the Generalized ID (GID) FEC (type 129). The Pwid FEC element includes a fixed-length 32 bit value called the PWid. The same PWid value must be configured on the local and remote PE prior to PW setup. The GID FEC element includes TLV fields for attachment individual identifiers (AII) that, in conjunction with an attachment group identifier (AGI), serve as PW endpoint identifiers. The endpoint identifier on the local PE (denoted as <AGI, source AII or SAII>) is called the source attachment identifier (SAI) and the endpoint identifier on the remote PE (denoted as <AGI, target AII or TAII>) is called the target attachment identifier (TAI). The SAI and TAI can be distinct values. This is useful for provisioning models where the local PE (with a particular SAI) does not know and must somehow learn (e.g. via MP-BGP auto-discovery) of remote TAI values prior to launching PW setup messages towards the remote PE.

FEC-128 specific PE-ID Element

This sub-type is to be used to identify a PW endpoint only if Pwid FEC Element is used for signaling the PW. The encoding of this PE-ID element is as follows:

```
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     0x01      |    Length     |           PW type             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                             PW ID                             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       Endpoint Address                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

PW type: The PW Type value from PWid FEC element.

PW ID: The PW ID value from the Pwid FEC element.

Endpoint Address: 32-bit LSR-ID from the LDP-ID used in LDP signaling Session by a PW endpoint.

FEC-129 specific PE-ID element

This sub-type is to be used to indentify a PW endpoint only if GID FEC Element is used for signaling the PW. The encoding of this PE-ID element is as follows:

```
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     0x02      |    Length     |           PW type             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            AGI TLV                            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            AII TLV                            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

PW type: The PW Type value from GID FEC element.

PW ID: The PW ID value from the GID FEC element.

AGI TLV: The AGI from the corresponding GID Element

AII TLV: The AII associated with the PW endpoint.

4.1.2. Application of PE-ID TLV in Optimized MAC Flush

For optimized MAC flush, the PE-ID TLV MAY be sent as an OPTIONAL parameter in existing LDP Address Withdraw Message with empty MAC List. The PE-ID TLV carries the unique PW endpoint identifier in a VPLS as described in section 4.

It is to note that for optimized MAC flush the PE-ID TLV carries sufficient information for identifying the VPLS instance and the unique VSI Identifier. For backward compatibility with MAC flush procedures in [RFC4762] both FEC TLV and PE-ID TLV should be sent in the MAC flush message. However the inclusion of the FEC-TLV should be based on what would be the desired effect should the PE-ID not be understood by the receiver. In cases where the desired action when the PE-ID is not understood would be to behave as described in [RFC4762], then the FEC TLV SHOULD be always included. In cases where the desired action when the PE-ID is not understood is no mac flushing, then the FEC TLV SHOULD NOT be included. The PE-ID TLV SHOULD carry the unique VSI identifier in the VPLS instance (specified in the FEC TLV). The PE-ID TLV SHOULD be placed after the existing TLVs in MAC Flush message in [RFC4762].

4.1.3. PE-ID TLV Processing Rules

This section describes the processing rules of PE-ID TLV that SHOULD be followed in the context of MAC flush procedures in an H-VPLS.

When an MTU-s triggers MAC flush after activation of backup spoke PW, it MAY send the PE-ID TLV that identifies VSI in the formerly active

PE device. There may be cases where a PE device in full mesh initiates MAC flush towards the core when it detects a spoke PW failure. In such a case the PE-ID TLV in MAC flush message MAY identify its own VSI. Irrespective of whether it is the MTU-s or PE device that initiates the MAC flush, a PE device receiving the PE-ID TLV SHOULD follow the same processing rules as described in this section.

Note that if MS-PW is used in VPLS then a MAC flush message is processed only at the T-PE nodes since S-PE(s) traversed by the MS-PW propagate MAC flush messages without any action. In this section, a PE device signifies only T-PE in MS-PW case unless specified otherwise.

When a PE device receives a MAC flush with PE-ID TLV, it SHOULD flush all the MAC addresses learned from the PW that terminates in the remote VSI identified by the PE-ID element.

If a PE-ID element received in the MAC flush message identifies the local VSI, it SHOULD flush the MAC addresses learned from its local spoke PW(s) in the VPLS instance.

If a PE device receives a MAC flush with the PE-ID TLV option and a valid MAC address list, it SHOULD ignore the option and deal with MAC addresses explicitly as per [RFC4762].

If a PE device that doesn’t support PE-ID TLV receives a MAC flush message with this option, it MUST ignore the option and follow the processing rules as per [RFC4762].

4.1.4. Optimized MAC Flush Procedures

This section explains the optimized MAC flush procedure in the scenario in Figure 1. When the backup PW is activated by MTU-s, it may send MAC flush message to PE-2 with the FEC TLV and the optional PE-ID TLV. The PE-ID element carries the VSI identifier in PE-1 for the VPLS. Upon receipt of the MAC flush message, PE-2 identifies the VPLS instance that requires MAC flush from the FEC element in the FEC TLV. From the PE-ID TLV, PE-2 identifies the PW in the VPLS that terminates in PE-1. PE-2 removes all MAC addresses learned from that PW. PE-2 relays MAC flush messages with the received PE-ID to all its peer PE devices. When the message is received at PE-3, it identifies the PW that terminates in the remote VSI in PE-1. PE-3 removes all MAC addresses learned on the PW that terminated in PE1. There may be redundancy scenerios where a PE device in the full mesh may be required to initiate optimized MAC Address Withdrawal. Figure 3 shows a redundant H-VPLS topology to protect against failure of MTU-s device. Provider RSTP may be used as selection algorithm for active and backup PWs in order to maintain the connectivity between MTU devices and PE devices at the edge. It is assumed that PE devices can detect failure on PWs in either direction through OAM mechanisms such as VCCV procedures for instance.

```
            MTU-1================PE-1===============PE-3
              ||                  || \             /||
              ||  Redundancy      ||  \           / ||
              ||  Provider RSTP   ||   Full-Mesh .  ||
              ||                  ||  /           \ ||
              ||                  || /             \||
            MTU-2----------------PE-2===============PE-4
                   Backup PW
```

Figure 3: Redundancy with Provider RSTP

MTU-1, MTU-2, PE-1 and PE-2 participate in provider RSTP. By configuration in RSTP it is ensured that the PW between MTU-1 and PE-1 is active and the PW between MTU-2 and PE-2 is blocked (made backup) at MTU-2 end. When the active PW failure is detected by RSTP, it activates the PW between MTU-2 and PE-2. When PE-1 detects the failing PW to MTU-1, it may trigger MAC flush into the full mesh with PE-ID TLV that carries its own VSI identifier in the VPLS. Other PE devices in the full mesh that receive the MAC flush message identify their respective PWs terminating on PE-1 and flush all the MAC addresses learned from it.

By default, MTU-2 should still trigger MAC flush as currently defined in [RFC4762] after the backup PW is made active by RSTP. Mechanisms to prevent two copies of MAC withdraws to be sent in such scenarios is out of scope of this document.

[RFC4762] describes multi-domain VPLS service where fully meshed VPLS networks (domains) are connected together by a single spoke PW per VPLS service between the VPLS "border" PE devices. To provide redundancy against failure of the inter-domain spoke, full mesh of inter-domain spokes can be setup between border PE devices and provider RSTP may be used for selection of the active inter-domain spoke. In case of inter-domain spoke PW failure, PE initiated MAC withdrawal may be used for optimized MAC flushing within individual domains.

4.2. LDP MAC Withdraw Extensions for PBB-VPLS

The use of Address Withdraw message with MAC List TLV is proposed in [RFC4762] as a way to expedite removal of MAC addresses as the result of a topology change (e.g. failure of a primary link of a VPLS PE and implicitly the activation of an alternate link in a dual-homing use case). These existing procedures apply individually to B-VPLS and I-component domains.

When it comes to reflecting topology changes in access networks connected to I-component across the B-VPLS domain certain additions should be considered as described below.

MAC Switching in PBB is based on the mapping of Customer MACs (CMACs) to Backbone MAC(s) (BMACs). A topology change in the access (I-domain) should just invoke the flushing of CMAC entries in PBB PEs' FIB(s) associated with the I-component(s) impacted by the failure. There is a need to indicate the PBB PE (BMAC source) that originated the MAC Flush message to selectively flush only the MACs that are affected.

These goals can be achieved by adding a new MAC Flush Parameters TLV in the LDP Address Withdraw message to indicate the particular domain(s) requiring MAC flush. On the other end, the receiving PEs may use the information from the new TLV to flush only the related FIB entry/entries in the I-component instance(s).

4.2.1. MAC Flush Parameters TLV format

The MAC Flush Parameters TLV is described as below:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|1|1| MAC Flush Params TLV(TBD) |          Length               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Flags     | Sub-TLV Type  |        Sub-TLV Length         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                Sub-TLV Variable Length Value                  |
|                              "                                |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

The U and F bits are set to forward if unknown so that potential intermediate VPLS PEs unaware of the new TLV can just propagate it transparently. The MAC Flush Parameters TLV type is to be assigned by IANA. The encoding of the TLV follows the standard LDP TLV encoding in [RFC5036].

The TLV value field contains an one byte Flag field used as described below. Further the TLV value may carry one or more sub-TLVs. Any sub-TLV definition to the above TLV MUST address the actions in combination with other existing sub-TLVs.

The detailed format for the Flags bit vector is described below:

```
 0 1 2 3 4 5 6 7

+-+-+-+-+-+-+-+-+

|C|N|    MBZ    | (MBZ = MUST Be Zero)

+-+-+-+-+-+-+-+-+
```

1 Byte Flag field is mandatory. The following flags are defined :

C flag, used to indicate the context of the PBB-VPLS component in which MAC flush is required. For PBB-VPLS there are two contexts of MAC flushing - The Backbone VPLS (B-component VPLS) and Customer VPLS (I-component VPLS). C flag MUST be ZERO (C=0) when a MAC Flush for the B-VPLS is required. C flag MUST be set (C=1) when the MAC Flush for I-VPLS is required.

N flag, used to indicate whether a positive (N=0, Flush-all-but-mine) or negative (N=1 Flush-all-from-me) MAC Flush is required. The source (mine/me) is defined either as the PW associated with the LDP session on which the LDP MAC Withdraw was received or with the BMAC(s) listed in the BMAC Sub-TLV.

MBZ flags, the rest of the flags should be set to zero on transmission and ignored on reception.

The following sub-TLVs MUST be included in the MAC Flush Parameters TLV if the C-flag is set to 1:

- PBB BMAC List sub-TLV:

Type: 0x01

Length: value length in octets. At least one BMAC address must be present in the list.

Value: one or a list of 48 bits BMAC addresses. These are the source BMAC addresses associated with the B-VPLS instance that originated the MAC Withdraw message. It will be used to identify the CMAC(s) mapped to the BMAC(s) listed in the sub-TLV.

- PBB ISID List sub-TLV:

Type: 0x02,

Length: value length in octets. Zero indicates an empty ISID list. An empty ISID list means that the flush applies to all the ISIDs mapped to the B-VPLS indicated by the FEC TLV.

Value: one or a list of 24 bits ISIDs that represent the I-component FIB(s) where the MAC Flush needs to take place.

4.2.2. MAC Flush Parameters TLV Processing Rules

The following steps describe the details of the processing for the related LDP Address Withdraw message:

. The LDP MAC Withdraw Message, including the MAC Flush Parameters TLV is initiated by the PBB PE(s) experiencing a Topology Change event in one or multiple customer I-component(s).

  o The flags are set accordingly to indicate the type of MAC Flush required for this event: N=0 (Flush-all-but-mine), C=1 (Flush only CMAC FIBs).

  o The PBB Sub-TLVs (BMAC and ISID Lists) are included according to the context of topology change.

. On reception of the LDP Address Withdrawal message, the B-VPLS instances corresponding to the FEC TLV in the message must interpret the content of MAC Flush Parameters TLV. If the C-bit is set to 1 then Backbone Core Bridges (BCB) in the PBB-VPLS SHOULD NOT flush their BMAC FIBs. The B-VPLS control plane SHOULD propagate the MAC Flush following the split-horizon grouping and the established B-VPLS topology.

. The usage and processing rules of MAC Flush Parameters TLV in the context of Backbone Edge Bridges (BEB) is as follows:

.

o The PBB ISID List is used to determine the particular ISID FIBs (I-VPLS) that need to be flushed. If the ISID List is empty then all the ISID FIBs associated with the receiving B-VPLS SHOULD be flushed.

o The PBB BMAC List is used to identify from the ISID FIBs in the previous step to selectively flush BMAC to CMAC associations depending on the N flag specified below.

. Next, depending on the N flag value the following actions apply:

o N=0, all the CMACs in the selected ISID FIBs SHOULD be flushed with the exception of the resulted CMAC list from the BMAC List mentioned in the message. ("Flush all but the CMACs associated with the BMAC(s) in the BMAC List Sub-TLV from the FIBs associated with the ISID list").

o N=1, the resulted CMAC list SHOULD be flushed ("Flush all the CMACs associated with the BMAC(s) in the BMAC List Sub-TLV from the FIBs associated with the ISID list").

4.2.3 Applicability of MAC Flush Parameters TLV

If MAC Flush Parameters TLV is received by a BEB in a PBB-VPLS that does not understand the TLV then it may result in undesirable MAC flushing action. It is RECOMMENDED that all PE devices participating in PBB-VPLS support MAC Flush Parameters TLV.

The MAC Flush Parameters TLV is also applicable to regular VPLS context as well. To achieve negative MAC Flush (flush-all-from-me) in regular VPLS context, the MAC Flush Parameters TLV SHOULD be encoded with C=0 and N = 1 without inclusion of any Sub-TLVs. Negative MAC flush is highly desirable in scenarios when VPLS access redundancy is provided by Ethernet Ring Protection as specified in ITU-T G.8032 specification etc.

5. Security Considerations

Control plane aspects:

- LDP security (authentication) methods as described in [RFC5036] is applicable here. Further this document implements security considerations as in [RFC4447] and [RFC4762].

Data plane aspects:

- This specification does not have any impact on the VPLS forwarding plane.

6. IANA Considerations

The Type field in PE-ID TLV is defined as 0x405 and is subject to IANA approval.

The Type field in MAC Flush Parameters TLV is defined as 0x406 and is subject to IANA approval.

7. Acknowledgments

The authors would like to thank the following people who have provided valuable comments and feedback on the topics discussed in this document: Marc Lasserre, Dimitri Papadimitriou, Jorge Rabadan, Prashanth Ishwar, Vipin Jain, John Rigby, Ali Sajassi, Wim Henderickx, Jorge Rabadan and Maarten Vissers.

8. References

8.1. Normative References

[RFC4762] Lasserre, M. and Kompella, V. (Editors), "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.

[RFC5036] Andersson, L., et al. "LDP Specification", RFC5036, October 2007.

[RFC4447] Martini. and et al., "Pseudowire Setup and Maintenance Using Label Distribution Protocol (LDP)", RFC 4447, April 2006.

8.2. Informative References

[PBB-VPLS Model] F. Balus, et Al. "Extensions to VPLS PE model for Provider Backbone Bridging", draft-ietf-l2vpn-pbb-vpls-pe-model-00.txt, May 2009 (work in progress)

[RFC4664] Andersson, L., et al. "Framework for Layer 2 Virtual Private Networks (L2VPNs)", RFC 4664, September 2006.

[802.1w] "IEEE Standard for Local and metropolitan area networks. Common specifications Part 3: Media Access Control (MAC) Bridges. Amendment 2: Rapid Reconfiguration", IEEE Std 802.1w-2001.

Author's Addresses

Pranjal Kumar Dutta
Alcatel-Lucent
701 E Middlefield Road,
Mountain View, CA 94043
USA
Email: pranjal.dutta@alcatel-lucent.com

Florin Balus
Alcatel-Lucent
701 E. Middlefield Road
Mountain View, CA, USA 94043
Email: florin.balus@alcatel-lucent.com

Geraldine Calvignac
France Telecom
2, avenue Pierre-Marzin
22307 Lannion Cedex
France
Email: geraldine.calvignac@orange-ftgroup.com

Olen Stokes
Extreme Networks
PO Box 14129
RTP, NC 27709
USA
Email: ostokes@extremenetworks.com

Internet Working Group
Internet Draft
Intended status: Standards Track
Expires: April 2011

Y. Jiang
L. Yong
Huawei
M. Paul
Deutsche Telekom
F. Jounay
France Telecom Orange

October 25, 2010

VPLS PE Model for E-Tree Support
draft-jiang-l2vpn-vpls-pe-etree-02.txt

Abstract

A generic VPLS solution for E-Tree services is proposed which uses VLANs to indicate root/leaf traffic. A VPLS Provider Edge (PE) model is illustrated as an example for the solution. In the solution, E-Tree VPLS PEs are interconnected by full mesh tagged PWs, the MAC address based Ethernet forwarding engine and the PW works in the same way as before. A signaling mechanism for E-Tree capability and VLAN mapping notification is further described.

Table of Contents

1. Introduction

E-Tree service is defined in Metro Ethernet Forum (MEF) as rooted multi-point EVC service, where traffic from a root can reach any root or leaf, and traffic from a leaf can reach any root, but should never reach a leaf. Although VPMS or P2MP multicast is a somewhat simplified version of this service, in fact there is no exact corresponding terminology in IETF.

[Etree-req] gives the requirements to provide E-Tree solutions in the VPLS and the need to filter leaf to leaf traffic in the VPLS.

[vpls-etree] describes a PW control word based E-Tree solution, where a bit in the PW control word is used to indicate the root/leaf attribute for a packet. The Ethernet forwarder in the VPLS is also extended to filter the leaf-leaf traffic based on the <ingress port, egress port, CW L-bit> tuple.

[Etree-2PW] proposes another E-Tree solution where root and leaf traffic are classified and forwarded in the same VSI but with two separate PWs.

Both solutions are only applicable to "VPLS only" networks.

In fact, VPLS PE usually consists of a bridge module itself [RFC4664], moreover E-Tree services may cross both Ethernet and VPLS domains. Therefore, the support of interconnection between Ethernet and VPLS for an E-Tree service is indispensable.

IEEE 802.1 has incorporated the generic E-Tree solution in the latest version of 802.1Q [802.1aq], which is just an improvement on the traditional asymmetric VLAN mechanism. In the solution, VLANs are used to indicate root/leaf attribute of a packet: one VLAN is used to carry traffic originated from the roots and another VLAN is used to carry traffic originated from the leaves. The bridge can then filter on each leaf port all the traffic received on the VLANs associated with the leaves. Thus it is better to use the same mechanism in VPLS rather than develop a new mechanism which may not interwork with Ethernet.

This document introduces how the Ethernet VLAN solution can be used to support generic E-Tree services in the VPLS. This solution is fully compatible with the IEEE bridge architecture and the IETF PWE3 technology, and VPLS scalability and simplicity is also well kept. With this mechanism it is also possible to deploy a converged E-Tree service across both Ethernet and MPLS networks.

As an example, a typical VPLS PE model is firstly introduced and extended which consists of a Tree VSI connected to an S-VLAN bridge with a dual-VLAN interface. However, this model is applicable to a PE with C-VLAN or B-VLAN as its service demarcation’s encapsulation.

This document then discusses the PW encapsulation and PW processing such as VLAN mapping options for transporting E-Tree services in a VPLS.

Finally, the extensions needed to support the signaling of E-Tree capability and VLAN mapping are also discussed.

2. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

3. Terminology

Most of the terminology used here is from [IEEE802.1Q], [IEEE802.1ad], [RFC4664] and [RFC4762]. Terminology specific to this document is introduced as needed in later sections.

4. PE Model with E-Tree Support

"VPLS only" PE architecture as outlined in Fig. 1 of [Etree-req] is a simplification of the PWE3 architecture, the more common VPLS PE architectures are discussed in more details in [RFC 4664] and [vpls-interop].

Therefore, VLAN based E-Tree solution are demonstrated with the help of a typical VPLS PE model. Other PE models are further discussed in Appendix A.

4.1. Existing PE Models

According to [RFC4664], there are at least three models possible for a VPLS PE, including:

o A single bridge module, a single VSI;

o A single bridge module, multiple VSIs;

o Multiple bridge modules, each attaches to a VSI.

The second PE model as depicted in Fig. 1 and Fig. 2 is a typical one for VPLS [vpls-interop], where the S-VLAN bridge module is connected to multiple VSIs each with a single VLAN interface.

```
                      +--------------------------------+
                      |  802.1ad Bridge Module Model   |
                      |                                |
           +---+      |  +------+      +-----------+   |
           |CE |---------|C-VLAN|------|           |   |
           +---+      |  |bridge|------|           |   |
                      |  +------+      |           |   |
                      |     o          |   S-VLAN  |   |
                      |     o          |           |   |
                      |     o          |   Bridge  |   |
           +---+      |  +------+      |           |   |
           |CE |---------|C-VLAN|------|           |   |
           +---+      |  |bridge|------|           |   |
                      |  +------+      +-----------+   |
                      +--------------------------------+
```

Figure 1 The Model of 802.1ad Bridge Module

```
           +-----------------------------------------+
           |           VPLS-capable PE model         |
           |   +----------------+         +-------+  |
           |   |                |         |VSI-1  |------------
           |   |                |==========|      |------------ PWs
           |   |      Bridge    ------------      |------------
           |   |                | S-VLAN-1 +-------+  |
           |   |      Module    |             o       |
           |   |                |             o       |
           |   |    (802.1ad    |             o       |
           |   |     bridge)    |             o       |
           |   |                |             o       |
           |   |                | S-VLAN-n +-------+  |
           |   |                ------------VSI-n |-------------
           |   |                |==========|      |------------- PWs
           |   |                |     ^    |      |-------------
           |   +----------------+     |    +-------+  |
           |                          |               |
           +--------------------------|---------------+
                                LAN emulation Interface
```

Figure 2 VPLS-capable PE Model

In the PE model above, Ethernet service from the CEs will cross multiple stages of bridge modules (i.e., C-VLAN and S-VLAN bridge) in a PE to access the egress PWs. Therefore, the association of an AC port and a PW in a single forwarding engine as required in [vpls-etree] or [Etree-2PW] is difficult, sometimes even impossible.

This model could be further enhanced by the introduction of a trunk VLAN and a branch VLAN as Ethernet frames enter the PE. To be more precise, they are called root and leaf VLAN respectively in this document. All the traffics from the root VLAN are received both on the roots and the leaves, while traffics from the branch VLAN are received on the roots and dropped on the leaves. It was demonstrated in [802.1aq] that E-Tree on Ethernet could be well supported with this mechanism.

Assume this mechanism is implemented in the bridge module, then it is quite straightforward to infer a VPLS PE model with two VSIs (as shown in Fig. 3) to support the E-Tree. But this model will require two VSIs per PE and two sets of full meshed PWs per E-Tree service, which is poorly scalable in a large MPLS/VPLS network.

```
+----------------------------------------+
|           VPLS-capable PE model        |
|   +---------------+         +------+  |
|   |               |         |VSI-1 |------------
|   |               |==========|      |------------ PWs
|   |     Bridge    ------------      |------------
|   |               | Root     +------+  |
|   |     Module    | S-VLAN      o      |
|   |               |             o      |
|   |   (802.1ad    |             o      |
|   |    bridge)    |             o      |
|   |               | Leaf        o      |
|   |               | S-VLAN   +------+  |
|   |               ------------VSI-2 |-------------
|   |               |==========|      |------------- PWs
|   |               |     ^    |      |-------------
|   +---------------+     |    +------+  |
|                         |              |
+-------------------------|--------------+
                     LAN emulation Interface
```

Figure 3 VPLS PE Model with E-Tree Support

4.2. A New PE Model with E-Tree Support

To provide for the E-Tree support in a more scalable way, a new VPLS PE model is proposed and depicted in Fig. 4, where the S-VLAN bridge module is connected to the Tree VSI (T-VSI, a VSI with E-Tree support) with a dual-VLAN virtual interface. That is, both the root S-VLAN and the leaf S-VLAN are connected to the Tree VSI (T-VSI). In this way, only one VPLS instance and one set of PWs is needed per E-Tree service. With this model, multiple E-Trees can also be provided by the same T-VSI if needed, and further increase the scalability of VPLS.

```
        +-----------------------------------------+
        |           VPLS-capable PE model         |
        |   +---------------+          +------+   |
        |   |               |==========|TVSI-1|------------
        |   |               ------------      |------------ PWs
        |   |     Bridge    ------------      |------------
        |   |               | Root &   +------+   |
        |   |     Module    | Leaf VLAN   o       |
        |   |               |             o       |
        |   |    (802.1ad   |             o       |
        |   |     bridge)   |             o       |
        |   |               |             o       |
        |   |               | S-VLAN-n +------+   |
        |   |               ------------VSI-n |-------------
        |   |               |==========|      |------------- PWs
        |   |               |     ^    |      |-------------
        |   +---------------+     |    +------+   |
        |                         |               |
        +-------------------------|---------------+
                          LAN emulation Interface
```

Figure 4 E-Tree VPLS-capable PE Model

Both VLANs should share the same FIB and work in shared VLAN learning. The traffic from the root UNIs are firstly tagged with root C-VLAN by the C-VLAN bridge module, and then tagged with root S-VLAN by the S-VLAN bridge module, thus can only be transported on the root S-VLAN. Similarly, the traffic from the leaves can only be transported on the leaf S-VLAN.

In fact, this model can also be applied to a PE with C-VLAN (customer sites attached to the PEs with untagged ports), or B-VLAN (with a PBB bridge module embedded in the PE) as a provider’s tag encapsulation. Therefore, the document will use the VLAN tag as a generalized form in the latter sections.

5. PW for E-Tree Support

A pair of T-VSIs in a VPLS is interconnected with a bidirectional PW. The VLAN indicating root/leaf attribute of the packet is carried in the PW, and the peer PE must drop the packet with a leaf VLAN on the egress AC of leaf UNI.

There are three ways of manipulating VLANs for an E-Tree:

- o Provisioning two global VLANs across both the Ethernet and the VPLS instance domain;

- o Provisioning two local VLANs in the VLAN space for each Ethernet domain and two global VLANs in the VPLS network domain, the VLAN mapping is done completely in the Ethernet domains (e.g., in the bridge module of the PE).

- o Provisioning two local VLANs independently for each Ethernet domain and two local VLANs on each PE for better scalability. That is, the assignment of VLANs in the PE may be local to improve the scalability.

The first method is called global VLAN based and no VLAN mapping is needed, but two unique VLANs must be allocated in the VPLS for them. The second method is called partial global VLAN based, which needs a VLAN mapping in the bridge module or in the Ethernet device attached to the PE. The last method is called local VLAN based and more scalable, but needs a VLAN mechanism in the PW. VLAN mapping is elaborated in the following section.

5.1. VLAN Mapping

In order to carry both VLANs (root and leaf VLAN) in a single PW and map those into the remote peer’s VLANs, cares must be taken on both the PEs associated with the PW.

Two options of VLAN mapping are possible:

- o Local mapping, that is, the remote PE is responsible for mapping VLANs into its local VLANs. For the local VLAN based method, VLAN mapping is done when a frame exits the PW; for the partial global VLAN based method, VLAN mapping is done when a frame exits the bridge module.

o Remote mapping, that is, the local PE is responsible for mapping VLANs into the remote PE’s VLANs. For the local VLAN based method, VLAN mapping is done when a frame enters the PW; for the partial global VLAN based method, VLAN mapping is done when a frame enters the bridge module.

Normally, each PE does its own local mapping. But when a PE is not capable of VLAN mapping, remote mapping can be done on its peer.

If no PE is capable of VLAN mapping, global VLAN based method can be used instead.

5.2. Tagged Mode PW Encapsulation

For a VPLS instance to support an E-Tree as described above, the Ethernet PW should work in the tagged mode (PW type 0x0004) as described in [RFC4448], and a C-VLAN, S-VLAN, or B-VLAN tag must be carried in each frame in the PW to indicate the E-Tree root/leaf attribute.

For global VLAN based method, it is the global VLAN tag to be carried and no VLAN mapping needed in the VPLS.

For the local VLAN or partial global VLAN based method, either the local or the remote VLAN tag could be carried depending on the mapping option. In the local mapping mode, the remote VLANs are carried with no change, while in the remote mapping mode, the local VLANs are carried instead.

The mapping between the local VLAN and the remote VLAN (local root VLAN <-> remote root VLAN; local leaf VLAN <-> remote leaf VLAN) should be provisioned by management or signaled by a control protocol such as LDP. The signaling extensions for E-Tree support are provided in Section 6 and 7.

5.3. PW Processing

5.3.1. PW Processing in the Normal Mode

In the normal mode, two VPLS PEs with a T-VSI in each of them are inter-connected and both sides are miscellaneously attached with roots and leaves, as shown in the scenario of Fig. 5. At the PE where a frame exits the PW, if a frame with the remote leaf VLAN is received, then it is mapped to the local leaf VLAN, otherwise, if a frame with the remote root VLAN is received, then it is mapped to the local root VLAN. Packets over both VLANs are processed in the same I-VSI and are further forwarded or dropped in the exit bridge module using the mechanism as described in 802.1Q.

```
          +---------------------------------+
          |  VPLS PE with T-VSI             |
          |                                 |
+----+    | +------+   +-------+   +-----+ |  PW
|Root|------|C-VLAN|---|S-VLAN |---|T-VSI|----------
+----+    | | BRG  |   | BRG   |   |     |----------
+----+    | |      |---|       |---|     |----------
|Leaf|------|      |   |       |   |     |---------+
+----+    | +------+   +-------|   +-----+ |        |
          |                                 |        |
          +---------------------------------+        |
                                                     |
          +---------------------------------+        |
          |  VPLS PE with T-VSI             |        |
          |                                 |        |
+----+    | +------+   +-------+   +-----+ |  PW    |
|Root|------|C-VLAN|---|S-VLAN |---|T-VSI|---------+
+----+    | | BRG  |   | BRG   |   |     |----------
+----+    | |      |---|       |---|     |----------
|Leaf|------|      |   |       |   |     |----------
+----+    | +------+   +-------|   +-----+ |
          |                                 |
          +---------------------------------+
```

Figure 5 T-VSI Interconnected in the Normal Mode

5.3.2.PW Processing in the Compatibility Mode

The new VPLS PE model can work in a traditional VPLS network seamlessly in the compatibility mode. As shown in Fig. 5, the VPLS PE with T-VSI can access both root and leaf node, while the VPLS PE with a traditional VSI can only access the root node.

```
                +--------------------------------+
                |  VPLS PE with T-VSI            |
                |                                |
      +-----+   | +-------+   +--------+   +-----+ |  PW
      |Root|------|C-VLAN|---|S-VLAN |---|T-VSI|----------
      +-----+   | | BRG  |   | BRG    |   |     |----------
      +-----+   | |      |---|        |---|     |----------
      |Leaf|------|      |   |        |   |     |----------+
      +-----+   | +-------+   +--------|   +-----+ |        |
                |                                |        |
                +--------------------------------+        |
                                                          |
                +--------------------------------+        |
                |  VPLS PE with VSI              |        |
                |                                |        |
      +-----+   | +-------+   +--------+   +-----+ |  PW    |
      |Root|------|C-VLAN|---|S-VLAN |---|VSI  |----------+
      +-----+   | | BRG  |   | BRG    |   |     |----------
      +-----+   | |      |---|        |   |     |----------
      |Root|------|      |   |        |   |     |----------
      +-----+   | +-------+   +--------|   +-----+ |
                |                                |
                +--------------------------------+
```

Figure 6 T-VSI interconnected with Traditional VSI

In this case, the PE with a T-VSI in it must work in the compatibility mode, that is, the egress PW of the T-VSI must translate frames received over both local root and leaf VLAN into a PW with a single VLAN (i.e., local root VLAN if the peer is capable of rewriting the VLAN, or the remote peer’s VLAN otherwise), while the ingress PW only translates the frames received over the PW into the local root VLAN.

5.3.3.PW Processing in the Optimization Mode

When two VPLS PE with T-VSI are inter-connected and one side is attached with pure leaves, as shown in the scenario of Fig. 6, the egress PW of the miscellaneous attached PE then should work in the optimization mode, that is, the PE can drop all the frames received over the local leaf VLAN rather than transport them over the PW and be discarded on the remote PE. Thus bandwidth efficiency of the VPLS can be improved.

```
                +---------------------------------+
                |  VPLS PE with T-VSI             |
                |                                 |
      +-----+   | +-------+   +--------+   +-----+ |  PW
      |Root|------|C-VLAN|---|S-VLAN |---|T-VSI|----------
      +-----+   | | BRG  |   | BRG    |   |     |----------
      +-----+   | |      |---|        |---|     |----------
      |Leaf|------|      |   |        |   |     |---------+
      +-----+   | +-------+   +--------|   +-----+ |        |
                |                                 |        |
                +---------------------------------+        |
                                                           |
                +---------------------------------+        |
                |  VPLS PE with T-VSI             |        |
                |                                 |        |
      +-----+   | +-------+   +--------+   +-----+ |  PW    |
      |Leaf|------|C-VLAN|---|S-VLAN |---|T-VSI|---------+
      +-----+   | | BRG  |   | BRG    |   |     |----------
      +-----+   | |      |---|        |---|     |----------
      |Leaf|------|      |   |        |   |     |----------
      +-----+   | +-------+   +--------|   +-----+ |
                |                                 |
                +---------------------------------+
```

Figure 7 T-VSI interconnected with 1-side of pure Leaves

6. LDP Extensions for E-Tree Support

To dynamically provision the E-Tree service using the signaling procedures specified in [RFC4447], an E-Tree specific interface parameter sub-TLV is proposed as follows:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  E-Tree       |  Length=8     |           Reserved        |P|R|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|          Root VLAN ID         |          Leaf VLAN ID         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 8 E-Tree Sub-TLV

Where:

- o E-Tree is the sub-TLV identifier to be assigned by IANA.

- o Length is the length of the sub TLV in octets.

- o Reserved bits MUST be set to zero on transmit and be ignored on receive.

- o P is a Pure Leaf bit, it is set to 1 to indicate that the PE is attached with all leaves, and set to 0 otherwise.

- o R is a request bit of Remote VLAN Translation. If a PE is capable of translating VLANs, then set R to 0, otherwise set R to 1. If a PE receives R=1 from its peer, then it must do VLAN translation for this peer, otherwise local mapping rule applies.

- o Root VLAN ID is the value of the local root VLAN.

- o Leaf VLAN ID is the value of the local leaf VLAN.

When the VPLS supporting an E-Tree service is setting up the PW, the PW endpoints negotiate the E-Tree support using the above E-Tree sub-TLV. Note PW type of 0x0004 should be used during the PW negotiation.

A PE that wishes to support E-Tree service includes an E-Tree Sub-TLV in its PW label mapping message, together with its local root VLAN and leaf VLAN carried in the Root VLAN ID and Leaf VLAN ID field respectively. A PE that has E-Tree capability and willing to support it MUST include an E-Tree Sub-TLV with its own local root VLAN and leaf VLAN. A PE that is incapable of translating VLANs MUST set the R bit to 1, while a PE that is capable of translating VLANs MAY set the

R bit to 1 to indicate remote mapping is preferred. And a PE is attached with pure leaves SHOULD set the P bit to 1.

If a PE incapable of VLAN mapping has received an E-Tree Sub-TLV with the bit "R" set, and either the root VLAN ID or the leaf VLAN ID in the message does not match the local root VLAN or the local leaf VLAN, then the PW should not be set up and a label release message with the error code "E-Tree VLAN mapping not supported" must be sent.

If a PE has sent an E-Tree Sub-TLV and has received an E-Tree Sub-TLV, then it must work as described in Section 5.3.1. If the bit "L" is set, then it should work as described in Section 5.3.3.

If a PE has sent an E-Tree Sub-TLV and does not receive an E-Tree Sub-TLV, then it must work in the mode of compatibility as described in Section 5.3.2.

7. BGP Extensions for E-Tree Support

BGP may also be used to distribute the E-Tree and VLAN mapping information. It is to be specified in the next version.

8. Applicability

The solution is applicable to LDP VPLS [RFC4762] and may also be applicable to BGP VPLS [RFC 4761].

The solution is applicable to both "VPLS Only" network and VPLS with Ethernet aggregation network.

9. Security Considerations

To be added in the next version.

10. IANA Considerations

IANA is requested to allocate a value for E-Tree in the Pseudowire Interface Parameters Sub-TLV type registry.

Parameter ID	Length	Description
TBD	8	E-Tree

IANA is requested to allocate a new LDP status code from the registry of name "STATUS CODE NAME SPACE". The following value is suggested:

Range/Value	E	Description
TBD	0	E-Tree VLAN mapping not supported

11. References

11.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC4447] Martini, L., and et al, "Pseudowire Setup and Maintenance Using Label Distribution Protocol (LDP)", RFC 4447, April 2006.

[RFC4448] Martini, L., and et al, "Encapsulation Methods for Transport of Ethernet over MPLS Networks", RFC 4448, April 2006.

[RFC4762] Lasserre, M. and Kompella, V., "Virtual Private LAN Services using LDP", RFC 4762, January 2007.

11.2. Informative References

[RFC3985] Bryant, S., and Pate, P., "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, March 2005.

[RFC4664] Andersson, L., and Rosen, E., "Framework for Layer 2 Virtual Private Networks (L2VPNs)", RFC 4664, September 2006.

[vpls-interop] Sajassi, A., and et al, "VPLS Interoperability with CE Bridges", draft-ietf-l2vpn-vpls-bridge-interop-05, March 2010

[ETree-req] Key, R., et al, "Requirements for MEF E-Tree Support in VPLS", draft-key-l2vpn-vpls-etree-reqt-02, October 2010

[vpls-etree] Delord, S., and et al, "Extension to VPLS for E-Tree", draft-key-l2vpn-vpls-etree-02, January 2010

[802.1aq] IEEE 802.1aq D3.0, Virtual Bridged Local Area Networks - Amendment 9: Shortest Path Bridging, June 2010

[Etree-2PW] Ram, R., and et al., Extension to LDP-VPLS for E-Tree Using Two PW, draft-ram-l2vpn-ldp-vpls-etree-2pw-00.txt, October 2010

12. Acknowledgments

The authors would like to thank Adrian Farrel and Susan Hares for their valuable comments and advices.

Appendix A. Other PE Models for E-Tree

A.1. PE Model With a VSI and No bridge

If there is no bridge module in a PE, the PE may consist of Native Service Processors (NSPs) as shown in Figure A.1 (adapted from Fig. 5 of [RFC3985]) which may apply any transformation operation for VLANs (e.g., VLAN insertion/removal or VLAN mapping). Thus a root VLAN or leaf VLAN is added by the NSP depending on the UNI type of the AC over which the packet arrives.

Further, when a packet with a leaf VLAN exits a forwarder and arrives at the NSP, the NSP must drop the packet if the egress AC is a leaf UNI.

Tagged PW and VLAN mapping work in the same way as in the typical PE model.

```
        +-----------------------------------------+
        |                PE Device                |
Multiple+-----------------------------------------+
AC      |      |          |       Single          | PW Instance
<------>o  NSP #          +      PW Instance     X<----------->
        |      |          |                       |
        |------|  VSI     |-----------------------|
        |      |          |       Single          | PW Instance
<------>o  NSP #Forwarder +      PW Instance     X<----------->
        |      |          |                       |
        |------|          |-----------------------|
        |      |          |       Single          | PW Instance
<------>o  NSP #          +      PW Instance     X<----------->
        |      |          |                       |
        +-----------------------------------------+
```

Figure A.1 PE model with a VSI and no bridge module

Authors' Addresses

Yuanlong Jiang
Huawei Technologies Co., Ltd.
Bantian industry base, Longgang district
Shenzhen, China
Email: yljiang@huawei.com

Lucy Yong
Huawei USA
1700 Alma Dr. Suite 500
Plano, TX 75075, USA
Email: lucyyong@huawei.com

Manuel Paul
Deutsche Telekom
Goslarer Ufer 35
10589 Berlin, Germany
Email: manuel.paul@telekom.de

Frederic Jounay
France Telecom Orange
2, avenue Pierre-Marzin
22307 Lannion Cedex, France
Email: frederic.jounay@orange-ftgroup.com

Network Working Group
Internet Draft
Category: Informational
Expires: April 2011

Raymond Key, Telstra
Simon Delord, Telstra
Frederic Jounay, France Telecom
Lu Huang, China Mobile
Zhihua Liu, China Telecom
Manuel Paul, Deutsche Telekom
Ruediger Kunze, Deutsche Telekom
Nick Del Regno, Verizon
Joshua Rogers, Time Warner Cable

October 7, 2010

Requirements for MEF E-Tree Support in VPLS
draft-key-l2vpn-vpls-etree-reqt-02.txt

Status of this Memo

Abstract

This document provides functional requirements for Metro Ethernet Forum (MEF) Ethernet Tree (E-Tree) support in Virtual Private LAN Service (VPLS). It is intended that potential solutions will use these requirements as guidelines.

Table of Contents

Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

1. Introduction

This document provides functional requirements for Metro Ethernet Forum (MEF) Ethernet Tree (E-Tree) support in Virtual Private LAN Service (VPLS). It is intended that potential solutions will use these requirements as guidelines.

Considerable number of service providers have adopted VPLS to provide MEF Ethernet LAN (E-LAN) services to customers. Service Providers currently need a simple and effective solution to emulate E-Tree services in addition to E-LAN services on their MPLS networks.

2. Virtual Private LAN Service

VPLS is a L2VPN service that provides multipoint-to-multipoint connectivity for Ethernet across an IP or MPLS-enabled IP Packet Switched Network. VPLS emulates the Ethernet VLAN functionality of traditional Ethernet network.

VPLS is a current IETF standard, please refer to [RFC4761] [RFC4762].

Data frame is Ethernet frame.

Data forwarding is MAC-based forwarding, which includes MAC address learning and aging.

3. MEF Multipoint Ethernet Services

MEF has defined two multipoint Ethernet Service types:
- E-LAN (Ethernet LAN), multipoint-to-multipoint service
- E-Tree (Ethernet Tree), rooted-multipoint service

For full specification, please refer to [MEF6.1] [MEF10.2].

3.1. Similarity between E-LAN and E-Tree

Data frame is Ethernet frame.

Data forwarding can be MAC-based forwarding or something else, to be specified by service provider as service frame delivery attributes in the particular service definition.

A generic E-LAN/E-Tree service is always bidirectional in the sense that ingress frames can originate at any endpoint in the service.

3.2. Difference between E-LAN and E-Tree

Within the context of a multipoint Ethernet service, each endpoint is designated as either a Root or a Leaf. A Root can communicate with all other endpoints in the same multipoint Ethernet service, however a Leaf can only communicate with Roots but not Leafs.

The only difference between E-LAN and E-Tree is:

- E-LAN has Root endpoints only, which implies there is no communication restriction between endpoints
- E-Tree has both Root and Leaf endpoints, which implies there is a need to enforce communication restriction between Leaf endpoints

3.3. E-Tree Use Cases

Table 1 below presents some major E-Tree use cases.

	Use Case	Root	Leaf
1	Hub & Spoke VPN	Hub Site	Spoke Site
2	Wholesale Access	Customer's Interconnect	Customer's Subscriber
3	Mobile Backhaul	RAN NC	RAN BS
4	IEEE 1588 PTPv2 Clock Synchronisation	PTP Server	PTP Client
5	Internet Access	BNG Router	Subscriber
6	Broadcast Video (unidirectional only)	Video Source	Subscriber
7	Broadcast/Multicast Video plus Control Channel	Video Source	Subscriber
8	Device Management	Management System	Managed Device

Table 1: E-Tree Use Cases

Common to all use cases, direct layer 2 Leaf-to-Leaf communication is not required. For Mobile backhaul, this may not be valid for LTE X2 interfaces in the future.

If direct layer 2 Leaf-to-Leaf communication is not allowed due to security concern, then E-Tree should be used to prohibit communication between Leaf endpoints, otherwise E-LAN is also a feasible option.

3.4. Generic E-Tree Service

A generic E-Tree service supports multiple Root endpoints. The need for multiple Root endpoints is usually driven by redundancy requirement. Whether a particular E-Tree service needs to support single or multiple Roots depends on the target application.

A generic E-Tree service supports all the following traffic flows:
- Ethernet Unicast from Root to Leaf
- Ethernet Unicast from Leaf to Root
- Ethernet Unicast from Root to Root
- Ethernet Broadcast/Multicast from Root to Roots & Leafs
- Ethernet Broadcast/Multicast from Leaf to Roots

A particular E-Tree service may need to support all the above or only a subset depending on the target application.

4. Problem Statement

4.1. Motivation

VPLS can be used to emulate MEF E-LAN service over MPLS network provided that the E-LAN service uses MAC-based forwarding as service frame delivery attributes.

Considerable number of service providers have adopted VPLS to provide MEF E-LAN services to customers. Service Providers currently need a simple and effective solution to emulate E-Tree services in addition to E-LAN services on their MPLS networks.

4.2. Leaf-to-Leaf Communication Restriction

Current standard VPLS treats all ACs equal (i.e. not classified into Root or Leaf) and provides any-to-any connectivity among all ACs. The current standard VPLS does not include any mechanism of communication restriction between specific ACs, therefore is insufficient for emulating generic E-Tree service over MPLS network.

A problem occurs when there are two or more PEs with both Root AC and Leaf AC.

Let’s look at the scenario illustrated in Figure 1 below. VPLS is used to emulate an E-Tree service over a MPLS network.

```
                    <------------E-Tree------------>
                   +---------+            +---------+
                   |   PE1   |            |   PE2   |
+---+              |  +---+  |            |  +---+  |          +---+
|CE1+-----AC1----+--+     |  |            |  |    +--+----AC3-----+CE3|
+---+  (Root AC) |  | V |  |  Ethernet  |  | V |  | (Root AC)  +---+
                   |  | S +--+-----PW-----+--+ S |  |
+---+              |  | I |  |            |  | I |  |          +---+
|CE2+-----AC2----+--+     |  |            |  |    +--+----AC4-----+CE4|
+---+  (Leaf AC) |  +---+  |            |  +---+  | (Leaf AC)  +---+
                   +---------+            +---------+
```

Figure 1: Problem Scenario for Leaf-to-Leaf Communication Restriction

When PE2 receives a frame from PE1 via the Ethernet PW,
- PE2 does not know which AC on PE1 is the ingress AC
- PE2 does not know whether the ingress AC is a Leaf AC or not
- PE2 does not have sufficient information to enforce the Leaf-to-Leaf communication restriction

Examples:
- CE2 sends a Broadcast/Multicast frame to PE1 via AC2
- CE2 sends a Unicast frame to PE1 via AC2, destination address in Ethernet header equal to CE4's MAC address

Note: Figure 1 is a hypothetical case solely for explaining the problem, and not meant to represent a typical E-Tree service.

There are some possible ways to get around this problem that do not require extension to the current standard VPLS but they all come with significant design complexity or deployment constraints, please refer to [Draft ETree Frwk] Appendix A.

5. Requirements

5.1. Functional Requirements

A solution MUST prohibit communication between any two Leaf ACs in a VPLS instance.

A solution MUST allow multiple Root ACs in a VPLS instance.

A solution MUST allow Root AC and Leaf AC of a VPLS instance co-exist on any PE.

5.2. Applicability

There are two distinct VPLS standards, performing similar functions in different manners.

- [RFC4761], commonly known as BGP-VPLS

- [RFC4762], commonly known as LDP-VPLS

A solution MUST identify which VPLS standards the solution is applicable to, [RFC4761] or [RFC4762] or both.

Service providers may use single or multiple technologies to deliver an end-to-end E-Tree service.

- Case 1: Single technology "VPLS Only"

- Case 2: Multiple technologies "VPLS + Others"
    - e.g. VPLS + Ethernet network, VPLS + OTN

- Case 3: Single/multiple technologies "No VPLS"
    - e.g. Ethernet network, Ethernet network + OTN
    - out of scope for this document

A solution MUST identify which of the above cases the solution is applicable to. For Case 2, further details may be required to specify the applicable deployment scenarios.

5.3. Backward Compatibility

A solution SHOULD minimise the impact on existing VPLS solution, especially for the MEF E-LAN services already in operation.

A solution SHOULD be backward compatible with the existing VPLS solution. It SHOULD allow a case where a common VPLS instance is composed of both PEs supporting the solution and PEs not supporting it, and the Leaf-to-Leaf communication restriction is enforced within the scope of the compliant PEs.

6. Security Considerations

This will be added in later version of this document.

7. IANA Considerations

This will be added in later version of this document.

8. Acknowledgements

This will be added in later version of this document.

9. References

9.1. Normative References

[MEF6.1] Metro Ethernet Forum, Ethernet Services Definitions - Phase 2, April 2008

[MEF10.2] Metro Ethernet Forum, Ethernet Services Attributes Phase 2, October 2009

[RFC2119] Bradner, S., Key words for use in RFCs to Indicate Requirement Levels, BCP 14, RFC 2119, March 1997

[RFC4761] Kompella & Rekhter, Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling, January 2007

[RFC4762] Lasserre & Kompella, Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling, January 2007

9.2. Informative References

[Draft ETree Frwk] Key, et al., A Framework for E-Tree Service over MPLS Network, draft-key-l2vpn-etree-frwk-04.txt, October 2010

[Draft VPMS Frmwk] Kamite, et al., Framework and Requirements for Virtual Private Multicast Service (VPMS), draft-ietf-l2vpn-vpms-frmwk-requirements-03.txt, July 2010

Appendix A. Frequently Asked Questions

A.1. Are E-Tree requirements addressed in the VPMS requirement draft?

VPMS is Virtual Private Multicast Service. VPMS requirement draft refers to [Draft VPMS Frmwk].

The focus of VPMS is to provide point-to-multipoint connectivity.

VPMS provides single coverage of receiver membership (i.e. there is no distinct differentiation for multiple multicast groups). A VPMS service supports single Root AC. All traffic from the Root AC will be forwarded to all Leaf ACs (i.e. P2MP, from Root to all Leafs). Destination address in Ethernet frame is not used in data forwarding. As an optional capability, a VPMS service may support reverse traffic from a Leaf AC to the Root AC (i.e. P2P, from Leaf to Root).

In contrast, the focus of MEF E-Tree is that a Leaf can only communicate with Roots but not Leafs.

A generic MEF E-Tree service supports multiple Root endpoints. Whether a particular E-Tree service needs to support single or multiple Root endpoints depends on the target application.

A generic MEF E-Tree service supports all the following traffic flows:

- Ethernet Unicast bidirectional Root to/from Root
- Ethernet Unicast bidirectional Root to/from Leaf
- Ethernet Broadcast/Multicast unidirectional Root to all Roots & Leafs
- Ethernet Broadcast/Multicast unidirectional Leaf to all Roots.

A particular E-Tree service may need to support all the above or only a subset depending on the target application.

IETF’s VPMS definition and MEF’s E-Tree definition are significantly different.

Only for special case E-Tree service where

- Single Root only
- No Unicast traffic from Root destined for a specific Leaf (or there is no concern if such Unicast traffic are forwarded to all Leafs)

VPMS will be able to meet the requirement. An example is single-root E-Tree service for content delivery application.

For generic E-Tree service, VPMS will not be able to meet the requirements.

A.2. Are there any potential deployment scenarios for a "VPLS Only" solution?

This refers to Section 5.2. Applicability, Case 1: Single technology "VPLS Only".

Yes, there are potential deployment scenarios for a "VPLS Only" solution, some examples below.

Example 1 -

```
                                                   Enhanced VPLS with
              <-----Physical P2P Service------><-----E-Tree Support-----
                                               +---------+
              +---+                            |   PE1   |
   +---+      |NTU|                            |  +---+  |
   |CE1+------+---+--V1-----------------AC1----+--+   |  |
   +---+ Root +---+                  (Root AC) |  | V |  |
                                               |  |   |  | Ethernet
              +---+                            |  | S +--+----PW--->PE2
   +---+      |NTU|                            |  |   |  |
   |CE2+------+---+--V2-----------------AC2----+--+ I |  |
   +---+ Root +---+                  (Root AC) |  |   |  |
                                               |  |   |  |
              +---+                            |  |   |  |
   +---+      |NTU|                            |  |   |  |
   |CE3+------+---+--V3-----------------AC3----+--+   |  |
   +---+ Root +---+                  (Root AC) |  |   |  |
                                               |  |   |  |
              +---+                            |  |   |  |
   +---+      |NTU|                            |  |   |  |
   |CE4+------+---+--V4-----------------AC4----+--+   |  |
   +---+ Leaf +---+                  (Leaf AC) |  |   |  |
                                               |  |   |  |
              +---+                            |  |   |  |
   +---+      |NTU|                            |  |   |  |
   |CE5+------+---+--V5-----------------AC5----+--+   |  |
   +---+ Leaf +---+                  (Leaf AC) |  +---+  |
                                               +---------+
```

Example 2 -

```
                  Logical P2P Service          Enhanced VPLS with
           <-------via Access Switch-------><-----E-Tree Support-----
                      +---------+          +---------+
                      | Access  |          |   PE1   |
           +---+      | Switch  |          |         |
 +---+     |NTU|      |         |          |  +---+  |
 |CE1+------+---+--V1--+--VLAN1--+--V1--AC1--+--+   |  |
 +---+ Root +---+      |         | (Root AC) |  | V |  |
                      +---------+          |  |   |  | Ethernet
                                           |  | S +--+----PW--->PE2
                      +---------+          |  |   |  |
                      | Access  |          |  | I |  |
           +---+      | Switch  |          |  |   |  |
 +---+     |NTU|      |         |          |  |   |  |
 |CE2+------+---+--V2--+--VLAN2--+--V2--AC2--+--+   |  |
 +---+ Root +---+      |         | (Root AC) |  |   |  |
                      |         |          |  |   |  |
           +---+      |         |          |  |   |  |
 +---+     |NTU|      |         |          |  |   |  |
 |CE3+------+---+--V3--+--VLAN3--+--V3--AC3--+--+   |  |
 +---+ Root +---+      |         | (Root AC) |  |   |  |
                      |         |          |  |   |  |
           +---+      |         |          |  |   |  |
 +---+     |NTU|      |         |          |  |   |  |
 |CE4+------+---+--V4--+--VLAN4--+--V4--AC4--+--+   |  |
 +---+ Leaf +---+      |         | (Leaf AC) |  |   |  |
                      |         |          |  |   |  |
           +---+      |         |          |  |   |  |
 +---+     |NTU|      |         |          |  |   |  |
 |CE5+------+---+--V5--+--VLAN5--+--V5--AC5--+--+   |  |
 +---+ Leaf +---+      |         | (Leaf AC) |  +---+  |
                      +---------+          +---------+
```

Example 3 -

```
                     Ethernet Switching            Enhanced VPLS with
              <------with Split Horizon-------><-----E-Tree Support-----
                         +---------+           +---------+
                         | Access  |           |   PE1   |
                         | Switch  |           |         |
                         |         |           |         |
                         |  +---+  |           |  +---+  |
              +---+      |  | V |  |           |  |   |  |
   +---+      |NTU|      |  | L |  |           |  | V |  |
   |CE1+------+---+--V1--+--+ A +--+--V1--AC1--+--+   |  |
   +---+ Root +---+      |  | N |  | (Root AC) |  | S |  |
                         |  | 1 |  |           |  |   |  | Ethernet
                         |  +---+  |           |  | I +--+----PW--->PE2
                         +---------+           |  |   |  |
                                               |  |   |  |
                         +---------+           |  |   |  |
                         | Access  |           |  |   |  |
                         | Switch  |           |  |   |  |
              +---+      |         |           |  |   |  |
   +---+      |NTU|      |  +---+  |           |  |   |  |
   |CE2+------+---+--V2--+--+ V |  |           |  |   |  |
   +---+ Root +---+      |  | L |  |           |  |   |  |
                         |  | A +--+--V2--AC2--+--+   |  |
              +---+      |  | N |  | (Root AC) |  |   |  |
   +---+      |NTU|      |  | 2 |  |           |  |   |  |
   |CE3+------+---+--V2--+--+   |  |           |  |   |  |
   +---+ Root +---+      |  +---+  |           |  |   |  |
                         |         |           |  |   |  |
              +---+      |         |           |  |   |  |
   +---+      |NTU|      |  +---+  |           |  |   |  |
   |CE4+------+---+--V4--+SH+ V |  |           |  |   |  |
   +---+ Leaf +---+      |  | L |  |           |  |   |  |
                         |  | A +--+--V4--AC4--+--+   |  |
              +---+      |  | N |  | (Leaf AC) |  |   |  |
   +---+      |NTU|      |  | 4 |  |           |  |   |  |
   |CE5+------+---+--V4--+SH+   |  |           |  |   |  |
   +---+ Leaf +---+      |  +---+  |           |  +---+  |
                         +---------+           +---------+
```

Note:
- Group Roots and Leafs into two separate VLANs on Access Switch
- SH means member of split horizon group on Access Switch

Authors' Addresses

Raymond Key
Telstra
242 Exhibition Street, Melbourne
VIC 3000, Australia
Email: raymond.key@team.telstra.com

Simon Delord
Telstra
242 Exhibition Street, Melbourne
VIC 3000, Australia
Email: simon.a.delord@team.telstra.com

Frederic Jounay
France Telecom
2, avenue Pierre-Marzin
22307 Lannion Cedex, France
Email: frederic.jounay@orange-ftgroup.com

Lu Huang
China Mobile
Unit 2, 28 Xuanwumenxi Ave, Xuanwu District
Beijing 100053, China
Email: huanglu@chinamobile.com

Zhihua Liu
China Telecom
109 Zhongshan Ave., Guangzhou
510630, China
Email: zhliu@gsta.com

Manuel Paul
Deutsche Telekom
Goslarer Ufer 35
10589 Berlin, Germany
Email: manuel.paul@telekom.de

Ruediger Kunze
Deutsche Telekom
Goslarer Ufer 35
10589 Berlin, Germany
Email: ruediger.kunze@telekom.de

Nick Del Regno
Verizon
400 International Pkwy
Richardson, TX 75081, USA
Email: nick.delregno@verizon.com

Joshua Rogers
Time Warner Cable
11921 N. MoPac Expwy
Austin, TX 78759, USA
Email: josh.rogers@twcable.com

Copyright Notice

Network Working Group R. Aggarwal
Internet Draft Juniper Networks
Category: Standards Track
Expiration Date: April 2011 A. Sajassi
Cisco

W. Henderickx
Alcatel-Lucent

A. Isaac
Bloomberg

J. Uttaro
AT&T

N. Bitar
Verizon

R. Shekhar
Juniper Networks

F. Balus
Alcatel-Lucent

K. Patel
S. Boutros
Cisco

October 18, 2010

BGP MPLS Based Ethernet VPN

draft-raggarwa-sajassi-l2vpn-evpn-00.txt

Status of this Memo

Abstract

This document describes procedures for BGP MPLS based MAC VPNs (E-VPN).

Table of Contents

1. Specification of requirements

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 [RFC2119].

2. Contributors

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

Quaizar Vohra
Kireeti Kompella
Apurva Mehta
Juniper Networks

3. Introduction

This document describes procedures for BGP MPLS based Ethernet VPNs (E-VPN). The procedures described here are intended to meet the requirements in [E-VPN-REQ]. Please refer to [E-VPN-REQ] for the detailed requirements and motivation.

This document proposes a MPLS based technology, referred to as MPLS-based E-VPN (E-VPN). E-VPN requires extensions to existing IP/MPLS protocols as described in section 5. In addition to these extensions E-VPN uses several building blocks from existing MPLS technologies.

4. Terminology

CE: Customer Edge device e.g., host or router or switch
MES: MPLS Edge Switch
EVI: E-VPN Instance
ESI: Ethernet segment identifier
LACP: Link Aggregation Control Protocol
MP2MP: Multipoint to Multipoint
P2MP: Point to Multipoint
P2P: Point to Point

5. BGP MPLS Based E-VPN Overview

This section provides an overview of E-VPN.

An E-VPN comprises CEs that are connected to PEs or MPLS Edge Switches (MES) that comprise the edge of the MPLS infrastructure. A CE may be a host, a router or a switch. The MPLS Edge Switches provide layer 2 virtual bridge connectivity between the CEs. There may be multiple E-VPNs in the provider’s network. This document uses the terms E-VPN and E-VPN inter-changeably. A E-VPN routing and forwarding instance on a MES is referred to as a E-VPN Instance (MVI).

The MESes are connected by a MPLS LSP infrastructure which provides the benefits of MPLS such as fast-reroute, resiliency etc.

In a E-VPN, learning between MESes occurs not in the data plane (as happens with traditional bridging) but in the control plane. Control plane learning offers much greater control over the learning process, such as restricting who learns what, and the ability to apply policies. Furthermore, the control plane chosen for this is BGP (very similar to IP VPNs (RFC 4364)), providing much greater scale, and the ability to "virtualize" or isolate groups of interacting agents (hosts, servers, Virtual Machines) from each other. In E-VPNs MESes advertise the MAC addresses learned from the CEs that are connected to them, along with a MPLS label, to other MESes in the control plane. Control plane learning enables load balancing and allows CEs to connect to multiple active points of attachment. It also improves convergence times in the event of certain network failures.

However, learning between MESes and CEs is done by the method best suited to the CE: data plane learning, IEEE 802.1x, LLDP, 802.1aq or other protocols.

It is a local decision as to whether the Layer 2 forwarding table on a MES contains all the MAC destinations known to the control plane or implements a cache based scheme. For instance the forwarding table may be populated only with the MAC destinations of the active flows transiting a specific MES.

The policy attributes of a E-VPN are very similar to an IP VPN. A E-VPN instance requires a Route-Distinguisher (RD) and a E-VPN requires one or more Route-Targets (RTs). A CE attaches to a E-VPN on a MES in a particular MVI on a VLAN or simply an ethernet interface. When the point of attachment is a VLAN there may be one or more VLANs in a particular E-VPN. Some deployment scenarios guarantee uniqueness of VLANs across E-VPNs: all points of attachment of a given E-VPN use the same VLAN, and no other E-VPN uses this VLAN. This document refers to this case as a "Default Single VLAN E-VPN" and describes simplified procedures to optimize for it.

6. Ethernet Segment Identifier

If a CE is multi-homed to two or more MESes, the set of attachment circuits constitutes an "Ethernet segment". An Ethernet segment may appear to the CE as a Link Aggregation Group (LAG). Ethernet segments have an identifier, called the "Ethernet Segment Identifier" (ESI). A single-homed CE is considered to be attached to a Ethernet segment with ESI 0. Otherwise, an Ethernet segment MUST have a unique non-zero ESI. The ESI can be assigned using various mechanisms:

1. The ESI may be configured. For instance when E-VPNs are used to provide a VPLS service the ESI is fairly analagous to the Multi-homing site ID in [BGP-VPLS-MH].

2. If LACP is used, between the MESes and CEs, then the ESI is determined by LACP. This is the LAG system ID (48 bit MAC address) and the CE’s LAG Aggregator Key. This is the 48 bit virtual MAC address of the CE for the LACP link bundle and the CE’s LAG Aggregator Key. As far as the CE is concerned it would treat the multiple MESes that it is homed to as the same switch. This allows the host to aggregate links to different MESes in the same bundle.

3. If LLDP is used, between the MESes and CEs that are hosts, then the ESI is determined by LLDP. The ESI will be specified in a following version.

4. In the case of indirectly connected hosts and a bridged LAN between the hosts and the MESes, the ESI is determined based on the Layer 2 bridge protocol as follows: If STP is used then the value of the ESI is derived by listening to BPDUs on the ethernet segment. The MES does not run STP. However it does learn the Switch ID, MSTP ID and Root Bridge ID by listening to BPDUs. The ESI is as follows:

{Switch ID (6 bits), MSTP ID (6 bits), Root Bridge ID (48 bits)}

7. BGP E-VPN NLRI

This document defines a new BGP NLRI, called the E-VPN NLRI.

Following is the format of the E-VPN NLRI:

```
+-----------------------------------+
|    Route Type (1 octet)           |
+-----------------------------------+
|     Length (1 octet)              |
+-----------------------------------+
| Route Type specific (variable)    |
+-----------------------------------+
```

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

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

This document defines the following Route Types:

+ 1 - Ethernet Tag Auto-Discovery (A-D) route
+ 2 - MAC advertisement route
+ 3 - Inclusive Multicast Route
+ 4 - Ethernet Segment Route
+ 5 - Selective Multicast Auto-Discovery (A-D) Route
+ 6 - Leaf Auto-Discovery (A-D) Route

The detailed encoding and procedures for these route types are described in subsequent sections.

The E-VPN NLRI is carried in BGP [RFC4271] using BGP Multiprotocol Extensions [RFC4760] with an AFI of TBD and an SAFI of E-VPN (To be assigned by IANA). The NLRI field in the MP_REACH_NLRI/MP_UNREACH_NLRI attribute contains the E-VPN NLRI (encoded as specified above).

In order for two BGP speakers to exchange labeled E-VPN NLRI, they must use BGP Capabilities Advertisement to ensure that they both are capable of properly processing such NLRI. This is done as specified in [RFC4760], by using capability code 1 (multiprotocol BGP) with an AFI of TBD and an SAFI of E-VPN.

7.1. Ethernet Tag Auto-Discovery Route

A Ethernet Tag A-D route type specific E-VPN NLRI consists of the following:

```
+---------------------------------------+
|      RD   (8 octets)                  |
+---------------------------------------+
| Ethernet Segment Identifier (8 octets)|
+---------------------------------------+
|  Ethernet Tag ID (4 octets)           |
+---------------------------------------+
|  MPLS Label (3 octets)                |
+---------------------------------------+
|   Originating Router's IP Addr        |
+---------------------------------------+
```

For procedures and usage of this route please see the sections on "Auto-Discovery of Ethernet Tags on Ethernet Segments", "Designated Forwarder Election" and "Load Balancing".

7.2. MAC Advertisement Route

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

```
+---------------------------------------+
|      RD   (8 octets)                  |
+---------------------------------------+
| Ethernet Segment Identifier (8 octets)|
+---------------------------------------+
|  Ethernet Tag ID (4 octets)           |
+---------------------------------------+
|  MAC Address Length (1 octet)         |
+---------------------------------------+
|  MAC Address (6 octets)               |
+---------------------------------------+
|  MPLS Label (n * 3 octets)            |
+---------------------------------------+
|  Originating Router's IP Addr         |
+---------------------------------------+
```

For procedures and usage of this route please see the sections on ""Determining Reachability to Unicast MAC Addresses" and "Load Balancing of Unicast Packets".

7.3. Inclusive Multicast Ethernet Tag Route

An Inclusive Multicast Ethernet Tag route type specific E-VPN NLRI consists of the following:

```
+---------------------------------------+
|      RD   (8 octets)                  |
+---------------------------------------+
| Ethernet Segment Identifier (8 octets)|
+---------------------------------------+
|  Ethernet Tag ID (4 octets)           |
+---------------------------------------+
|   Originating Router's IP Addr        |
+---------------------------------------+
```

For procedures and usage of this route please see the sections on "Handling of Multi-Destination Traffic", "Unknown Unicast Traffic" and "Multicast".

7.4. Ethernet Segment Route

An Ethernet Segment route type specific E-VPN NLRI consists of the following:

For procedures and usage of this route please see the sections on "Multi-Homed Ethernet Segment Auto-Discovery", "Designated Forwarder Election" and "Split Horizon".

```
+---------------------------------------+
|      RD   (8 octets)                  |
+---------------------------------------+
| Ethernet Segment Identifier (8 octets)|
+---------------------------------------+
|  MPLS Label (3 octets)                |
+---------------------------------------+
|   Originating Router's IP Addr        |
+---------------------------------------+
```

8. ES-Import Extended Community

This extended community is a new transitive extended community and it includes all the MESes connected to the same multi-homed site. It is used to distribute Ethernet Segment routes. The value is derived automatically from the ESI by encoding the 6-byte system MAC address of the ESI in this RT. In order to derive this RT automatically, it is assumed that the system MAC address of the CE is unique in the network.

Each ES-Import extended community is encoded as a 8-octet value as follows:

```
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| 0x44          |   Sub-Type    |          ES-Import            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     ES-Import Cont'd                          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

9. Auto-Discovery

EVPN requires the following types of auto-discovery procedures:

+ EVPN Auto-Discovery, which allows an MES to discover the other MESes in the EVPN. Each MES advertises one or more "Inclusive Multicast Tag Routes". The procedures for advertising these routes are described in the section on "Handling of Multi-Destination Traffic".

+ Auto-Discovery of Ethernet Tags on Ethernet Segments, in a particular EVPN. The procedures are described in section "Auto-Discovery of Ethernet Tags on Ethernet Segments".

+ Ethernet Segment Auto-Discovery used for auto-discovery of MESes that are multi-homed to the same ethernet segment. The procedures are described in section XXX and XXX.

10. Auto-Discovery of Ethernet Tags on Ethernet Segments

If a CE is multi-homed to two or more MESes on a particular ethernet segment, each MES MUST advertise to other MSEs in the E-VPN, the information about one or more Ethernet Tags (e.g., VLANs) on that ethernet segment. If a CE is not multi-homed, then the MES that it is attached to MAY advertise the information about Ethernet Tags (e.g., VLANs) on the ethernet segment connected to the CE.

The information about an Ethernet Tag on a particular ethernet segment is advertised using a "Ethernet Tag Auto-Discovery route (Ethernet Tag A-D route)". This route is advertised using the E-VPN NLRI.

The Ethernet Tag Auto-discovery information is used for Designated Forwarder (DF) election as described in section "Desiginated Forwarder Election". It is also used to enable equal cost multi-path as described in section "Load Balancing of Unicast Packets". Further, it can be used to optimize withdrawl of MAC addresses as described in section "Convergence".

This section describes procedures for advertising one or more Ethernet Tag A-D routes per E-VPN. We will call this as "Ethernet Tag A-D route per E-VPN". This section also describes procedures to advertise and withdraw a single Ethernet Tag A-D route per Ethernet Segment. We will call this as "Ethernet Tag A-D route per Segment".

10.1. Constructing the Ethernet Tag A-D Route

The format of the Ethernet Tag A-D NLRI is specified in section "BGP E-VPN NLRI".

10.1.1. Ethernet Tag A-D Route per E-VPN

This section describes procedures to construct the Ethernet Tag A-D route when one or more such routes are advertised by a MES for a given E-VPN instance.

Route-Distinguisher (RD) MUST be set to the RD of the E-VPN instance that is advertising the NLRI. A RD MUST be assigned for a given E-VPN instance on a MES. This RD MUST be unique across all E-VPN instances on a MES. This can be accomplished by using a Type 1 RD [RFC4364]. The value field comprises an IP address of the MES (typically, the loopback address) followed by a number unique to the MES. This number may be generated by the MES, or, in the Default Single VLAN E-VPN case, may be the 12 bit VLAN ID, with the remaining 4 bits set to

0.

Ethernet Segment Identifier MUST be an 8 octet entity as described in section "Ethernet Segment Identifier". This MAY be set to 0.

The Ethernet Tag ID is the identifier of an Ethernet Tag on the ethernet segment. This value may be a two octet VLAN ID or it may be another Ethernet Tag used by the E-VPN. It MAY be set to the default Ethernet Tag on the ethernet segment or 0.

Note that the above allows the Ethernet Tag A-D route to be advertised with one of the following granularities:

+ One Ethernet Tag A-D route for a given <ESI, Ethernet Tag ID> tuple per E-VPN

+ One Ethernet Tag A-D route for a given <ESI> in a given E-VPN where the Ethernet Tag ID is set to 0.

+ One Ethernet Tag A-D route for a given <Ethernet Tag ID> in a given E-VPN where the ESI is set to 0.

+ One Ethernet Tag A-D route for the E-VPN where both ESI and Ethernet Tag ID are set to 0.

E-VPNs support both the non-qualified and qualified learning model. When non-qualified learning is used the Ethernet Tag Identifier specified in this section and in other places in this document MUST be set to a default value. When qualified learning is used and the Ethernet Tags been MESes and CEs in the E-VPN are consistantly assigned for a given broadcast domain, the Ethernet Tag Identifier MUST be set to the Ethernet Tag for the concerned broadcast domain between the advertising MES and the CE. When qualified learning is used and the Ethernet Tags been MESes and CEs in the E-VPN are not consistantly assigned for a given broadcast domain, the Ethernet Tag Identifier MUST be set to an E-VPN provider assigned tag that maps locally on the advertising MES to an ethernet broadcast domain identifier such as a VLAN ID.

The usage of the MPLS label is described in section on "Load Balancing of Unicast Packets".

The Originating Router’s IP address MUST be set to an IP address of the PE. This address SHOULD be common for all the MVIs on the PE (e.,g., this address may be PE’s loopback address).

The Next Hop field of the MP_REACH_NLRI attribute of the route MUST be set to the same IP address as the one carried in the Originating Router's IP Address field.

10.1.1.1. Ethernet Tag A-D Route Targets

The Ethernet Tag A-D route MUST carry one or more Route Target (RT) attributes. RTs may be configured (as in IP VPNs), or may be derived automatically.

If a MES uses Route Target Constrain [RT-CONSTRAIN], the MES SHOULD advertise all such RTs using Route Target Constrains This allows each Ethernet Tag A-D route to reach only the relevant MESes.

10.1.1.1.1. Auto-Derivation from the Ethernet Tag ID

The following is the procedure for deriving the RT attribute automatically from the Ethernet Tag ID associated with the advertisement:

+ The Global Administrator field of the RT MUST be set to the Autonomous System (AS) number that the MES belongs to.

+ The Local Administrator field of the RT contains a 4 octets long number that encodes the Ethernet Tag-ID.

The above auto-configuration of the RT implies that a different RT is used for every Ethernet Tag in a E-VPN, if the E-VPN contains multiple Ethernet Tags. For the "Default Single VLAN E-VPN" this results in auto-deriving the RT from the Ethernet Tag for that E-VPN.

10.1.2. Ethernet Tag A-D Route per Ethernet Segment

This section describes procedures to construct the Ethernet Tag A-D route when a single such route is advertised by a MES for a given Ethernet Segment.

Route-Distinguisher (RD) MUST be a Type 1 RD [RFC4364]. The value field comprises an IP address of the MES (typically, the loopback address) followed 0.

Ethernet Segment Identifier MUST be an 8 octet entity as described in section "Ethernet Segment Identifier".

The Ethernet Tag ID MUST be set to 0.

10.1.2.1. Ethernet Tag A-D Route Targets

The Ethernet Tag A-D route MUST carry one or more Route Target (RT) attributes. These RTs MUST be the set of RTs associated with all the E-VPN instances to which the Ethernet Segment, corresponding to the Ethernet Tag A-D route, belongs.

10.2. Motivations for Ethernet Tag A-D Route per Ethernet Segment

This section describes various scenarios in which the Ethernet Tag A-D route should be advertised per Ethernet Segment.

10.2.1. Optimizing Control Plane Convergence

Ethernet Tag A-D route per Ethernet Segment should be advertised when it is desired to optimize the control plane convergence of the withdrawl of the Ethernet Segment A-D routes. If this is done then when an ethernet segment fails, the single Ethernet Tag A-D route corresponding to the segment can be withdrawn first. This allows all MESes that receive this withdrawl to invalidate the MAC routes learned from the ethernet segment.

Note that the Ethernet Tag A-D route per Ethernet Segment, when used to optimize control plane convergence, is advertised in addition to the Ethernet Tag A-D routes per EVPN.

10.2.2. Reducing number of Ethernet Tag A-D Routes

In certain scenarios advertising Ethernet Tag A-D routes per ethernet segment, instead of per E-VPN, may reduce the number of Ethernet Tag A-D routes in the network. In these scenarios Ethernet Tag A-D routes may be advertised per ethernet segment instead of per E-VPN.

11. Multi-Homed Ethernet Segment Auto-Discovery

Each MES advertises a route for a multi-homed ethernet segment, referred to as an Ethernet Segment Route. This allows the set of MESes connected to the same CE to discover each other automatically with minimal to no configuration. The procedures for constructing this route are described below. The usage of this route is described in the sections on "DF election" and "Split Horizon".

11.1. Constructing the Ethernet Segment Route

The NLRI format is described in section "BGP E-VPN NLRI".

The RD MUST be the RD of the E-VPN instance that is advertising the NLRI. The procedures for setting the RD for a given E-VPN are described in section 10.1.1.

The Ethernet Segment Identifier MUST be set to the eight octet ESI identifier described in section 6.

The MPLS label is referred to as an "ESI label". This label MUST be a downstream assigned MPLS label if the advertising MES is using ingress replication for sending multicast, broadcast or unknown unicast traffic, to other MESes. If the advertising MES is using P2MP MPLS LSPs for the same, then this label MUST be an upstream assigned MPLS label. The usage of this label is described in section "Split Horizon".

The Originating Router’s IP address MUST be set to an IP address of the PE. This address SHOULD be common for all the MVIs on the PE (e.,g., this address may be PE’s loopback address).

The Next Hop field of the MP_REACH_NLRI attribute of the route MUST be set to the same IP address as the one carried in the Originating Router’s IP Address field.

The BGP advertisement that advertises the Ethernet Segment route MUST also carry one Route Target (RT) attribute. The construction of this RT is specified below.

11.1.1. Ethernet Segment Route Target and Filtering

The Ethernet Segment Route Filtering should be done such that the Ethernet Segment Route is imported only by the MESes that are multi-homed to the Ethernet Segment. There are two mechanisms for doing this filtering.

11.1.1.1. ESI Import Extended Community

This approach applies only when it can be assumed that the system MAC addresses of the CEs are unique in the network.

Each MES that is connected to a particular ESI constructs an import filtering rule to import a route that carries the ES-Import extended community, described in section 9, constructed from the ESI.

Note that the new ES-Import extended community is not the same as the Route Target Extended Community. The Ethernet Segment route carries this new ES-Import extended community. The MESes apply filtering on this new extended community. As a result the Ethernet Segment route is imported only by the MESes that are connected to the ethernet segment.

This approach requires a new ES-Import extended community for filtering.

11.1.1.2. Route Target

If this approach is used then the Ethernet Segment route MUST carry one or more Route Target (RT) attributes. These RTs MUST be the set of RTs associated with all the E-VPN instances to which the Ethernet Segment, corresponding to the Ethernet Segment route, belongs.

This approach is to be used when the system MAC addresses of the CEs cannot be assumed to be unique.

11.2. Carrying LAG specific Information

This route will be enhanced to carry LAG specific information such as LACP parameters in the future.

12. Determining Reachability to Unicast MAC Addresses

MESes forward packets that they receive based on the destination MAC address. This implies that MESes must be able to learn how to reach a given destination unicast MAC address.

There are two components to MAC address learning, "local learning" and "remote learning":

12.1. Local Learning

A particular MES must be able to learn the MAC addresses from the CEs that are connected to it. This is referred to as local learning.

The MESes in a particular E-VPN MUST support local data plane learning using vanilla ethernet learning procedures. A MES must be capable of learning MAC addresses in the data plane when it receives packets such as the following from the CE network:

- DHCP requests

- gratuitous ARP request for its own MAC.

- ARP request for a peer.

Alternatively if a CE is a host then MESes MAY learn the MAC addresses of the host in the control plane.

In the case where a CE is a host or a switched network connected on ESI X to hosts, the MAC address that is reachable via a given MES may move such that it becomes reachable via the same MES on another MES on ESI Y. This is referred to as a "MAC Move" Procedures to support this are described in section "MAC Moves".

12.2. Remote learning

A particular MES must be able to determine how to send traffic to MAC addresses that belong to or are behind CEs connected to other MESes i.e. to remote CEs or hosts behind remote CEs. We call such MAC addresses as "remote" MAC addresses.

This document requires a MES to learn remote MAC addresses in the control plane. In order to achieve this each MES advertises the MAC addresses it learns from its locally attached CEs in the control plane, to all the other MESes in the E-VPN, using BGP.

12.2.1. Constructing the BGP E-VPN MAC Address Advertisement

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

The RD MUST be the RD of the E-VPN instance that is advertising the NLRI. The procedures for setting the RD for a given E-VPN are described in section 10.1.1.

The Ethernet Segment Identifier is set to the eight octet ESI identifier described in section "Ethernet Segment Identifier".

The Ethernet Tag ID may be zero or may represent a valid Ethernet Tag ID. This field may be non-zero in the following cases:

+ If there are multiple bridge domains in the E-VPN instance.

+ If qualified learning is used between the MESes and the CEs in the E-VPN.

When the the Ethernet Tag ID in the NLRI is set to a non-zero value, for a particular bridge domain, then this Ethernet TAG ID may either be the ethernet tag value associated with the CE or it may be the Ethernet Tag Identifier assigned by the E-VPN provider and mapped to the CE’s ethernet tag. The latter would be the case if the CE ethernet tags for a particular bridge domain are different on different CEs.

The MAC address length field is typically set to 48. However this specification enables specifying the MAC address as a prefix in which case the MAC address length field is set to the length of the prefix. This enables aggregation of MAC addresses if the deployment environment supports that. The encoding of a MAC address is the 6-octet MAC address specified by IEEE 802 documents [802.1D-ORIG] [802.1D-REV]. If the MAC address is advertised as a prefix then the trailing bits of the prefix MUST be set to 0 to ensure that the entire prefix is encoded as 6 octets.

The MPLS label field carries one or more labels (that corresponds to the stack of labels [MPLS-ENCAPS]). Each label is encoded as 3 octets, where the high-order 20 bits contain the label value, and the low order bit contains "Bottom of Stack" (as defined in [MPLS-ENCAPS]).

The MPLS label stack MUST be the downstream assigned E-VPN MPLS label stack that is used by the MES to forward MPLS encapsulated ethernet packets received from remote MESes, where the destination MAC address in the ethernet packet is the MAC address advertised in the above NLRI. The forwarding procedures are specified in section "Forwarding Unicast Packets" and "Load Balancing of Unicast Packets".

A MES may advertise the same single E-VPN label for all MAC addresses in a given E-VPN instance. This label assignment methodology is referred to as a per MVI label assigment. Or a MES may advertise a unique E-VPN label per <ESI, Ethernet Tag> combination. This label methodology is referred to as a per <ESI, Ethernet Tag> label assignment. Or a MES may advertise a unique E-VPN label per MAC address. All of these methodologies have their tradeoffs.

Per MVI label assignment requires the least number of E-VPN labels, but requires a MAC lookup in addition to a MPLS lookup on an egress MES for forwarding. On the other hand a unique label per <ESI, Ethernet Tag> or a unique label per MAC allows an egress MES to forward a packet that it receives from another MES, to the connected CE, after looking up only the MPLS labels and not having to do a MAC lookup.

A MES may also advertise more than one label for a given MAC address. For instance a MES may advertise two labels, one of which is for the ESI corresponding to the MAC address and the second is for the Etherent Tag on the ESI that the MAC address is learned on.

The Originating Router’s IP address MUST be set to an IP address of the PE. This address SHOULD be common for all the MVIs on the PE (e.,g., this address may be PE’s loopback address).

The Next Hop field of the MP_REACH_NLRI attribute of the route MUST be set to the same IP address as the one carried in the Originating Router’s IP Address field.

The BGP advertisement that advertises the MAC advertisement route MUST also carry one or more Route Target (RT) attributes. RTs may be configured (as in IP VPNs), or may be derived automatically from the Ethernet Tag ID, in the single VLAN case as described in section 13.1.1.1.

It is to be noted that this document does not require MESes 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.

13. Designated Forwarder Election

Consider a CE that is a host or a router that is multi-homed directly to more than one MES in a E-VPN on a given ethernet segment. One or more Ethernet Tags may be configured on the ethernet segment. In this scenario only one of the MESes, referred to as the Designated Forwarder (DF), is responsible for certain actions:

- Sending multicast and broadcast traffic, on a given Ethernet Tag on a particular ethernet segment, to the CE. Note that this behavior, which allows selecting a DF at the granularity of <ESI, Ethernet Tag> for multicast and broadcast traffic is the default behavior in this specification. Optional mechanisms, which will be specified in the future, will allow selecting a DF at the granularity of <ESI, Ethernet Tag, S, G>.

- Flooding unknown unicast traffic (i.e. traffic for which a MES does not know the destination MAC address), on a given Ethernet Tag on a particular ethernet segment to the CE, if the environment requires flooding of unknown unicast traffic.

Note that a CE always sends packets using a single link. For instance if the CE is a host then, as mentioned earlier, the host treats the multiple links that it uses to reach the MESes as a Link Aggregation Group (LAG).

If a bridge network is multi-homed to more than one MES in a E-VPN via switches, then the support of active-active points of attachments as described in this specification requires the bridge network to be connected to two or more MESes using a LAG. In this case the reasons for doing DF election are the same as those described above when a CE is a host or a router.

If a bridge network does not connect to the MESes using LAG, then only one of the links between a CE that is a switch and the MESes must be the active link. Procedures for supporting active-active points of attachments, when a bridge network does not connect to the MESes using LAG, are for further study.

The granularity of the DF election MUST be at least the ethernet segment via which the CE is multi-homed to the MESes. If the DF election is done at the ethernet segment granularity then a single MES MUST be elected as the DF on the ethernet segment.

If there are one or more Ethernet Tags (e.g., VLANs) on the ethernet segment then the granularity of the DF election SHOULD be the combination of the ethernet segment and Ethernet Tag on that ethernet segment. In this case the same MES MUST be elected as the DF for a particular Ethernet Tag on that ethernet segment.

There are two specified mechanisms for performing DF election.

13.1. DF Election Performed by each MES

The MESes perform a designated forwarder (DF) election, for an ethernet segment, or ethernet segment, Ethernet Tag combination using the Ethernet Tag A-D BGP route described in section "Auto-Discovery of Ethernet Tags on Ethernet Segments".

The DF election for a particular ESI or a particular <ESI, Ethernet Tag> combination proceeds as follows. First a MES constructs a candidate list of MESes. This comprises all the Ethernet Tag A-D routes with that particular ESI or <ESI, Ethernet Tag> tuple that a MES imports in a E-VPN instance, including the Ethernet Tag A-D route generated by the MES itself, if any. The DF MES is chosen from this candidate list. Note that DF election is carried out by all the MESes that import the DF route.

The default procedure for choosing the DF is the MES with the highest IP address, of all the MESes in the candidate list. This procedure MUST be implemented. It ensures that except during routing transients each MES chooses the same DF MES for a given ESI and Ethernet Tag combination.

Other alternative procedures for performing DF election are possible and will be described in the future.

13.2. DF Election Performed Only on Multi-Homed MESes

As a MES discovers other MESs that are members of the same multi-homed segment, using Ethernet Segment Routes, it starts building an ordered list based on the originating MES IP addresses. This list is used to select a DF and a backup DF (BDF) on a per group of Ethernet Tag basis. For example, the MES with the numerically highest identifier is considered the DF for a given group of VLANs for that ethernet segment and the next MES in the list is considered the BDF. To that end, the range of Ethernet Tags associated with the CE must be partitioned into disjoint sets. The size of each set is a function of the total number of CE Ethernet Tags and the total number of MESs that the ethernet segment is multi-homed to. The DF can employ any distribution function that achieves an even distribution of Ethernet Tags across the MESes that are multi-homed to the ethernet segment. The DF takes over the Ethernet Tag set of any MES encountering either a node failure or a link/ethernet segment failure causing that MES to be isolated from the multi-homed segment. In case of a failure that is affecting the DF, then the BDF takes over the DF VLAN set.

It should be noted that once all the MESs participating in an ethernet segment have the same ordered list for that site, then Ethernet Tag groups can be assigned to each member of that list deterministically without any need to explicitly distribute Ethernet Tags among the member MESs of that list. In other words, the DF election for a group of Ethernet Tags is a local matter and can be done deterministically. As an example, consider, that the ordered list consists of m MESs: (MES1, MES2,., MESm), and there are n Ethernet Tags for that site (V0, V1, V2, ., Vn-1). Then MES1 and MES2 can be the DF and the BDF respectively for all the Ethernet Tags corresponding to (i mod m) for i:1 to n. MES2 and MES3 can be the DF and the BDF respectively for all the Ethernet Tags corresponding to (i mod m) + 1 and so on till the last MES in the order list is reached. As a result MESm and MES1 is the DF and the BDF respectively for the all the VLANs corresponding to (i mod m) + m-1.

14. Handling of Multi-Destination Traffic

Procedures are required for a given MES to send broadcast or multicast traffic, received from a CE encapsulated in a given Ethernet Tag in a E-VPN, to all the other MESes that span that Ethernet Tag in the E-VPN. In certain scenarios, described in section "Processing of Unknown Unicast Packets", a given MES may also need to flood unknown unicast traffic to other MESes.

The MESes in a particular E-VPN may use ingress replication or P2MP LSPs or MP2MP LSPs to send unknown unicast, broadcast or multicast traffic to other MESes.

Each MES MUST advertise an "Inclusive Multicast Ethernet Tag Route" to enable the above. Next section provides procedures to construct the Inclusive Multicast Ethernet Tag route. Subsequent sections describe in further detail its usage.

14.1. Construction of the Inclusive Multicast Ethernet Tag Route

The RD MUST be the RD of the E-VPN instance that is advertising the NLRI. The procedures for setting the RD for a given E-VPN are described in section 10.1.1.

The Ethernet Segment Identifier MAY be set to the eight octet ESI identifier described in section "Ethernet Segment Identifier". Or it MAY be set to 0. It MUST be set to 0 if the Ethernet Tag is set to 0.

The Ethernet Tag ID is the identifier of the Ethernet Tag. It MAY be set to 0 in which case an egress MES MUST perform a MAC lookup to forward the packet.

The Originating Router's IP address MUST be set to an IP address of the PE. This address SHOULD be common for all the MVIs on the PE (e.,g., this address may be PE's loopback address).

The Next Hop field of the MP_REACH_NLRI attribute of the route MUST be set to the same IP address as the one carried in the Originating Router's IP Address field.

The BGP advertisement that advertises the Inclusive Multicast Ethernet Tag route MUST also carry one or more Route Target (RT) attributes. The assignemnt of RTs described in the section on "Constructing the BGP E-VPN MAC Address Advertisement" MUST be followed.

14.2. P-Tunnel Identification

In order to identify the P-Tunnel used for sending broadcast, unknown unicast or multicast traffic, the Inclusive Multicast Ethernet Tag route MUST carry a "PMSI Tunnel Attribute" specified in [BGP MVPN].

Depending on the technology used for the P-tunnel for the E-VPN on the PE, the PMSI Tunnel attribute of the Inclusive Multicast Ethernet Tag route is constructed as follows.

+ If the PE that originates the advertisement uses a P-Multicast tree for the P-tunnel for the E-VPN, the PMSI Tunnel attribute MUST contain the identity of the tree (note that the PE could create the identity of the tree prior to the actual instantiation of the tree).

+ A PE that uses a P-Multicast tree for the P-tunnel MAY aggregate two or more Ethernet Tags in the same or different E-VPNs present on the PE onto the same tree. In this case in addition to carrying the identity of the tree, the PMSI Tunnel attribute MUST carry an MPLS upstream assigned label which the PE has bound uniquely to the <ESI, Ethernet Tag> for E-VPN associated with this update (as determined by its RTs).

  If the PE has already advertised Inclusive Multicast Ethernet Tag routes for two or more Ethernet Tags that it now desires to aggregate, then the PE MUST re-advertise those routes. The re-advertised routes MUST be the same as the original ones, except for the PMSI Tunnel attribute and the label carried in that attribute.

+ If the PE that originates the advertisement uses ingress replication for the P-tunnel for the E-VPN, the route MUST include the PMSI Tunnel attribute with the Tunnel Type set to Ingress Replication and Tunnel Identifier set to a routable address of the PE. The PMSI Tunnel attribute MUST carry a downstream assigned MPLS label. This label is used to demultiplex the broadcast, multicast or unknown unicast E-VPN traffic received over a unicast tunnel by the PE.

+ The Leaf Information Required flag of the PMSI Tunnel attribute MUST be set to zero, and MUST be ignored on receipt.

14.3. Ethernet Segment Identifier and Ethernet Tag

As described above the encoding rules allow setting the Ethernet Segment Identifier and Ethernet Tag to either valid values or to 0. If the Ethernet Tag is set to a valid value, then an egress MES can forward the packet to the set of egress ESIs in the Ethernet Tag, in the E-VPN, by performing a MPLS lookup alone. Further if the ESI is also set to non zero then the egress MES does not need to replicate the packet as it is destined for a given ethernet segment. If both Ethernet Tag and ESI are set to 0 then an egress MES MUST perform a MAC lookup in the MVI determined by the MPLS label, after the MPLS lookup, to forward the packet.

If a MES advertises multiple Inclusive Ethernet Tag routes for a given E-VPN then the PMSI Tunnel Attributes for these routes MUST be distinct.

15. Processing of Unknown Unicast Packets

The procedures in this document do not require MESes to flood unknown unicast traffic to other MESes. If MESes learn CE MAC addresses via a control plane, the MESes can then distribute MAC addresses via BGP, and all unicast MAC addresses will be learnt prior to traffic to those destinations.

However, if a destination MAC address of a received packet is not known by the MES, the MES may have to flood the packet. Flooding must take into account "split horizon forwarding" as follows. The principles behind the following procedures are borrowed from the split horizon forwarding rules in VPLS solutions [RFC 4761, RFC 4762]. When a MES capable of flooding (say MESx) receives a broadcast Ethernet frame, or one with an unknown destination MAC address, it must flood the frame. If the frame arrived from an attached CE, MESx must send a copy of the frame to every other attached CE, as well as to all other MESs participating in the E-VPN. If, on the other hand, the frame arrived from another MES (say MESy), MESx must send a copy of the packet only to attached CEs. MESx MUST NOT send the frame to other MESs, since MESy would have already done so. Split horizon forwarding rules apply to broadcast and multicast packets, as well as packets to an unknown MAC address.

Whether or not to flood packets to unknown destination MAC addresses should be an administrative choice, depending on how learning happens between CEs and MESes.

The MESes in a particular E-VPN may use ingress replication using RSVP-TE P2P LSPs or LDP MP2P LSPs for sending broadcast, multicast and unknown unicast traffic to other MESes. Or they may use RSVP-TE P2MP or LDP P2MP or LDP MP2MP LSPs for sending such traffic to other MESes.

15.1. Ingress Replication

If ingress replication is in use, the P-Tunnel attribute, carried in the Inclusive Multicast Ethernet Tag routes for the E-VPN, specifies the downstream label that the other MESes can use to send unknown unicast, multicast or broadcast traffic for the E-VPN to this particular MES.

The MES that receives a packet with this particular MPLS label MUST treat the packet as a broadcast, multicast or unknown unicast packet. Further if the MAC address is a unicast MAC address, the MES MUST treat the packet as an unknown unicast packet.

15.2. P2MP MPLS LSPs

The procedures for using P2MP LSPs are very similar to VPLS procedures [VPLS-MCAST]. The P-Tunnel attribute used by a MES for sending unknown unicast, broadcast or multicast traffic for a particular ethernet segment, is advertised in the Inclusive Ethernet Tag Multicast route as described in section "Handling of Multi-Destination Traffic".

The P-Tunnel attribute specifies the P2MP LSP identifier. This is the equivalent of an Inclusive tree in [VPLS-MCAST]. Note that multiple Ethernet Tags, which may be in different E-VPNs, may use the same P2MP LSP, using upstream labels [VPLS-MCAST]. When P2MP LSPs are used for flooding unknown unicast traffic, packet re-ordering is possible.

The MES that receives a packet on the P2MP LSP specified in the PMSI Tunnel Attribute MUST treat the packet as a broadcast, multicast or unknown unicast packet. Further if the MAC address is a unicast MAC address, the MES MUST treat the packet as an unknown unicast packet.

16. Forwarding Unicast Packets

16.1. Forwarding packets received from a CE

When a MES receives a packet from a CE, on a given Ethernet Tag, it must first look up the source MAC address of the packet. In certain environments the source MAC address may be used to authenticate the CE and determine that traffic from the host can be allowed into the network.

If the MES decides to forward the packet the destination MAC address of the packet must be looked up. If the MES has received MAC address advertisements for this destination MAC address from one or more other MESes or learned it from locally connected CEs, it is considered as a known MAC address. Else the MAC address is considered as an unknown MAC address.

For known MAC addresses the MES forwards this packet to one of the remote MESes. The packet is encapsulated in the E-VPN MPLS label advertised by the remote MES, for that MAC address, and in the MPLS LSP label stack to reach the remote MES.

If the MAC address is unknown then, if the administrative policy on the MES requires flooding of unknown unicast traffic:

- The MES MUST flood the packet to other MESes. If the ESI over which the MES receives the packet is multi-homed, then the MES MUST first encapsulate the packet in the ESI MPLS label as described in section "Split Horizon". If ingress replication is used the packet MUST be replicated one or more times to each remote MES with the bottom label of the stack being a MPLS label determined as follows. This is the MPLS label advertised by the remote MES in a PMSI Tunnel Attribute in the Inclusive Multicast Ethernet Tag route for an <ESI, Ethernet Tag> combination. The Ethernet Tag in the route must be the same as the Ethernet Tag advertised by the ingress MES in its Ethernet Tag A-D route associated with the interface on which the ingress MES receives the packet. If P2MP LSPs are being used the packet MUST be sent on the P2MP LSP that the MES is the root of for the Ethernet Tag in the E-VPN. If the same P2MP LSP is used for all Ethernet Tags then all the MESes in the E-VPN MUST be the leaves of the P2MP LSP. If a distinct P2MP LSP is used for a given Ethernet Tag in the E-VPN then only the MESes in the Ethernet Tag MUST be the leaves of the P2MP LSP. The packet MUST be encapsulated in the P2MP

LSP label stack.

If the MAC address is unknown then, if the admnistrative policy on the MES does not allow flooding of unknown unicast traffic:

- The MES MUST drop the packet.

16.2. Forwarding packets received from a remote MES

16.2.1. Unknown Unicast Forwarding

When a MES receives a MPLS packet from a remote MES then, after processing the MPLS label stack, if the top MPLS label ends up being a P2MP LSP label associated with a E-VPN or the downstream label advertised in the P-Tunnel attribute and after performing the split horizon procedures described in section "Split Horizon":

- If the MES is the designated forwarder of unknown unicast, broadcast or multicast traffic, on a particular set of ESIs for the Ethernet Tag, the default behavior is for the MES to flood the packet on the ESIs. In other words the default behavior is for the MES to assume that the destination MAC address is unknown unicast, broadcast or multicast and it is not required to do a destination MAC address lookup, as long as the granularity of the MPLS label included the Ethernet Tag. As an option the MES may do a destination MAC lookup to flood the packet to only a subset of the CE interfaces in the Ethernet Tag. For instance the MES may decide to not flood an unknown unicast packet on certain ethernet segments even if it is the DF on the ethernet segment, based on administrative policy.

- If the MES is not the designated forwarder on any of the ESIs for the Ethernet Tag, the default behavior is for it to drop the packet.

16.2.2. Known Unicast Forwarding

If the top MPLS label ends up being a E-VPN label that was advertised in the unicast MAC advertisements, then the MES either forwards the packet based on CE next-hop forwarding information associated with the label or does a destination MAC address lookup to forward the packet to a CE.

17. Split Horizon

Consider a CE that is multi-homed to two or more MESes on an ethernet segment ES1. If the CE sends a multicast, broadcast or unknown unicast packet to a particular MES, say MES1, then MES1 will forward that packet to all or subset of the other MESes in the E-VPN. In this case the MESes, other than MES1, that the CE is multi-homed to MUST drop the packet and not forward back to the CE. This is referred to as "split horizon" in this document.

In order to accomplish this each MES distributes to other MESes that are connected to the ethernet segment an "Ethernet Segment Route".

17.1. ESI MPLS Label: Ingress Replication

An MES that is using ingress replication for sending broadcast, multicast or unknown unicast traffic, distributes to other MESes, that belong to the ethernet segment, a downstream assigned "ESI MPLS label" in the Ethernet Segment route. This label MUST be programmed in the platform label space by the advertising MES. Further the forwarding entry for this label must result in NOT forwarding packets received with this label onto the ethernet segment that the label was distributed for.

Consider MES1 and MES2 that are multi-homed to CE1 on ES1. Further consider that MES1 is using P2P or MP2P LSPs to send packets to MES2. Consider that MES1 receives a a multicast, broadcast or unknown unicast packet from CE1 on VLAN1 on ESI1.

First consider the case where MES2 distributes an unique Inclusive Multicast Ethernet Tag route for VLAN1, for each ethernet segment on MES2. In this case MES1 MUST NOT replicate the packet to MES2 for <ESI1, VLAN1>.

Next consider the case where MES2 distributes a single Inclusive Multicast Ethernet Tag route for VLAN1 for all ethernet segments on MES2. In this case when MES1 sends a multicast, broadcast or unknown unicast packet, that it receives from CE1, it MUST first push onto the MPLS label stack the ESI label that MES2 has distributed for ESI1. It MUST then push on the MPLS label distributed by MES2 in the Inclusive Ethernet Tag Multicast route for Ethernet Tag1. The resulting packet is further encapsulated in the P2P or MP2P LSP label stack required to transmit the packet to MES2. When MES2 receives this packet it determines the set of ESIs to replicate the packet to from the top MPLS label, after any P2P or MP2P LSP labels have been removed. If the next label is the ESI label assigned by MES2 then

MES2 MUST NOT forward the packet onto ESI1.

17.2. ESI MPLS Label: P2MP MPLS LSPs

An MES that is using P2MP LSPs for sending broadcast, multicast or unknown unicast traffic, distributes to other MESes, that belong to the ethernet segment, an upstream assigned "ESI MPLS label" in the Ethernet Segment route. This label is upstream assigned by the MES that advertises the route. This label MUST be programmed by the other MESes, that are connected to the ESI advertised in the route, in the context label space for the advertising MES. Further the forwarding entry for this label must result in NOT forwarding packets received with this label onto the ethernet segment that the label was distributed for.

Consider MES1 and MES2 that are multi-homed to CE1 on ES1. Further assume that MES1 is using P2MP MPLS LSPs to send broadcast, multicast or uknown unicast packets. When MES1 sends a multicast, broadcast or unknown unicast packet, that it receives from CE1, it MUST first push onto the MPLS label stack the ESI label that it has assigned for the ESI that the packet was received on. The resulting packet is further encapsulated in the P2MP MPLS label stack necessary to transmit the packet to the other MESes. Penultimate hop popping MUST be disabled on the P2MP LSPs used in the MPLS transport infrastructure for E-VPN. When MES2 receives this packet it decapsulates the top MPLS label and forwards the packet using the context label space determined by the top label. If the next label is the ESI label assigned by MES1 then MES2 MUST NOT forward the packet onto ESI1.

18. ESI MPLS Label: MP2MP LSPs

The procedures for ESI MPLS Label assignment and usage for MP2MP LSPs will be described in the next version.

19. Load Balancing of Unicast Packets

This section specifies how load balancing is achieved to/from a CE that has more than one interface that is directly connected to one or more MESes. The CE may be a host or a router or it may be a switched network that is connected via LAG to the MESes.

19.1. Load balancing of traffic from a MES to remote CEs

Whenever a remote MES imports a MAC advertisement for a given <ESI, Ethernet Tag> in a E-VPN instance, it MUST consider the MAC as reachahable via all the MESes from which it has imported Ethernet Tag A-D routes for that <ESI, Ethernet Tag>. Further the remote MES MUST use these MAC advertisement and Ethernet Tag A-D routes to constuct the set of next-hops that it can use to send the packet to the destination MAC. Each next-hop comprises a MPLS label stack, that is to be used by the egress MES to forward the packet. This label stack is determined as follows. If the next-hop is constructed as a result of a MAC route which has a valid MPLS label stack, then this label stack MUST be used. However if the MAC route doesn’t exist or if it doesn’t have a valid MPLS label stack then the next-hop and MPLS label stack is constructed as a result of one or more corresponding Ethernet Tag A-D routes as follows. Note that the following description applies to determining the label stack for a particular next-hop to reach a given MES, from which the remote MES has received and imported one or more Ethernet Tag A-D routes that have the matching ESI and Ethernet Tag as the one present in the MAC advertisement. The Ethernet Tag A-D routes mentioned in the following description refer to the ones imported from this given MES.

If there is a corresponding Ethernet Tag A-D route for that <ESI, Ethernet Tag> then that label stack MUST be used. If such an Ethernet Tag A-D route doesn’t exist but Ethernet Tag A-D routes exist for <ESI, Ethernet Tag = 0> and <ESI = 0, Ethernet Tag> then the label stack must be constructed by using the labels from these two routes. If this is not the case but an Ethernet Tag A-D route exists for <ESI, Ethernet Tag = 0> then the label from that route must be used. Finally if this is also not the case but an Ethernet Tag A-D route exists for <ESI = 0, Ethernet Tag = 0> then the label from that route must be used.

The following example explains the above when Ethernet Tag A-D routes are advertised per <ESI, Ethernet Tag>.

Consider a CE, CE1, that is dual homed to two MESes, MES1 and MES2 on a LAG interface, ES1, and is sending packets with MAC address MAC1 on VLAN1. Based on E-VPN extensions described in sections "Determining

Reachability of Unicast Addresses" and "Auto-Discovery of Ethernet Tags on Ethernet Segments", a remote MES say MES3 is able to learn that a MAC1 is reachable via MES1 and MES2. Both MES1 and MES2 may advertise MAC1 in BGP if they receive packets with MAC1 from CE1. If this is not the case and if MAC1 is advertised only by MES1, MES3 still considers MAC1 as reachable via both MES1 and MES2 as both MES1 and MES2 advertise a Ethernet Tag A-D route for <ESI1, VLAN1>.

The MPLS label stack to send the packets to MES1 is the MPLS LSP stack to get to MES1 and the E-VPN label advertised by MES1 for CE1's MAC.

The MPLS label stack to send packets to MES2 is the MPLS LSP stack to get to MES2 and the MPLS label in the Ethernet Tag A-D route advertised by MES2 for <ES1, VLAN1>, if MES2 has not advertised MAC1 in BGP.

We will refer to these label stacks as MPLS next-hops.

The remote MES, MES3, can now load balance the traffic it receives from its CEs, destined for CE1, between MES1 and MES2. MES3 may use the IP flow information for it to hash into one of the MPLS next-hops for load balancing for IP traffic. Or MES3 may rely on the source and destination MAC addresses for load balancing.

Note that once MES3 decides to send a particular packet to MES1 or MES2 it can pick from more than path to reach the particular remote MES using regular MPLS procedures. For instance if the tunneling technology is based on RSVP-TE LSPs, and MES3 decides to send a particular packet to MES1 then MES3 can choose from multiple RSVP-TE LSPs that have MES1 as their destination.

When MES1 or MES2 receive the packet destined for CE1 from MES3, if the packet is a unicast MAC packet it is forwarded to CE1. If it is a multicast or broadcast MAC packet then only one of MES1 or MES2 must forward the packet to the CE. Which of MES1 or MES2 forward this packet to the CE is determined by default based on which of the two is the DF. An alternate procedure to load balance multicast packets will be described in the future.

If the connectivity between the multi-homed CE and one of the MESes that it is multi-homed to fails, the MES MUST withdraw the MAC address from BGP. This enables the remote MESes to remove the MPLS next-hop to this particular MES from the set of MPLS next-hops that can be used to forward traffic to the CE. For further details and procedures on withdrawl of E-VPN route types in the event of MES to CE failures please section "MES to CE Network Failures".

19.2. Load balancing of traffic between a MES and a local CE

A CE may be configured with more than one interface connected to different MESes or the same MES for load balancing. The MES(s) and the CE can load balance traffic onto these interfaces using one of the following mechanisms.

19.2.1. Data plane learning

Consider that the MESes perform data plane learning for local MAC addresses learned from local CEs. This enables the MES(s) to learn a particular MAC address and associate it with one or more interfaces. The MESes can now load balance traffic destined to that MAC address on the multiple interfaces.

Whether the CE can load balance traffic that it generates on the multiple interfaces is dependent on the CE implementation.

19.2.2. Control plane learning

The CE can be a host that advertises the same MAC address using a control protocol on both interfaces. This enables the MES(s) to learn the host’s MAC address and associate it with one or more interfaces. The MESes can now load balance traffic destined to the host on the multiple interfaces. The host can also load balance the traffic it generates onto these interfaces and the MES that receives the traffic employs E-VPN forwarding procedures to forward the traffic.

20. MAC Moves

In the case where a CE is a host or a switched network connected to hosts, the MAC address that is reachable via a given MES on a particular ESI may move such that it becomes reachable via another MES on another ESI. This is referred to as a "MAC Move".

Remote MESes must be able to distinguish a MAC move from the case where a MAC address on an ESI is reachable via two different MESes and load balancing is performed as described in section "Load Balancing of Unicast Packets". This distinction can be made as follows. If a MAC is learned by a particular MES from multiple MESes, then the MES performs load balancing only amongst the set of MESes that advertised the MAC with the same ESI. If this is not the case then the MES chooses only one of the advertising MESes to reach the MAC as per BGP path selection.

There can be traffic loss during a MAC move. Consider MAC1 that is advertised by MES1 and learned from CE1 on ESI1. If MAC1 now moves behind MES2, on ESI2, MES2 advertises the MAC in BGP. Until a remote MES, MES3, determines that the best path is via MES2, it will continue to send traffic destined for MAC1 to MES1. This will not occur deterministially until MES1 withdraws the advertisement for MAC1.

One recommended optimization to reduce the traffic loss during MAC moves is the following option. When an MES sees a MAC update from a CE on an ESI, which is different from the ESI on which the MES has currently learned the MAC, the corresponding entry in the local bridge forwarding table SHOULD be immediately purged causing the MES to withdraw its own E-VPN MAC advertisement route and replace it with the update.

A future version of this specification will describe other optimized procedures to minimize traffic loss during MAC moves.

21. Multicast

The MESes in a particular E-VPN may use ingress replication or P2MP LSPs to send multicast traffic to other MESes.

21.1. Ingress Replication

The MESes may use ingress replication for flooding unknown unicast, multicast or broadcast traffic as described in section "Handling of Multi-Destination Traffic". A given unknown unicast or broadcast packet must be sent to all the remote MESes. However a given multicast packet for a multicast flow may be sent to only a subset of the MESes. Specifically a given multicast flow may be sent to only those MESes that have receivers that are interested in the multicast flow. Determining which of the MESes have receivers for a given multicast flow is done using explicit tracking described below.

21.2. P2MP LSPs

A MES may use an "Inclusive" tree for sending an unknown unicast, broadcast or multicast packet or a "Selective" tree. This terminology is borrowed from [VPLS-MCAST].

A variety of transport technologies may be used in the SP network. For inclusive P-Multicast trees, these transport technologies include point-to-multipoint LSPs created by RSVP-TE or mLDP. For selective P-

Multicast trees, only unicast MES-MES tunnels (using MPLS or IP/GRE encapsulation) and P2MP LSPs are supported, and the supported P2MP LSP signaling protocols are RSVP-TE, and mLDP.

21.3. MP2MP LSPs

The root of the MP2MP LDP LSP advertises the Inclusive Multicast Tag route with the PMSI Tunnel attribute set to the MP2MP Tunnel identifier. This advertisement is then sent to all MESes in the EVPN. Upon receiving the Inclusive Multicast Tag routes with a PMSI Tunnel attribute that contains the MP2MP Tunnel identifier, the receiving MESes initiate the setup of the MP2MP tunnel towards the root using the procedures in [MLDP].

21.3.1. Inclusive Trees

An Inclusive Tree allows the use of a single multicast distribution tree, referred to as an Inclusive P-Multicast tree, in the SP network to carry all the multicast traffic from a specified set of E-VPN instances on a given MES. A particular P-Multicast tree can be set up to carry the traffic originated by sites belonging to a single E-VPN, or to carry the traffic originated by sites belonging to different E-VPNs. The ability to carry the traffic of more than one E-VPN on the same tree is termed ’Aggregation’. The tree needs to include every MES that is a member of any of the E-VPNs that are using the tree. This implies that a MES may receive multicast traffic for a multicast stream even if it doesn’t have any receivers that are interested in receiving traffic for that stream.

An Inclusive P-Multicast tree as defined in this document is a P2MP tree. A P2MP tree is used to carry traffic only for E-VPN CEs that are connected to the MES that is the root of the tree.

The procedures for signaling an Inclusive Tree are the same as those in [VPLS-MCAST] with the VPLS-AD route replaced with the Inclusive Multicast Ethernet Tag route. The P-Tunnel attribute [VPLS-MCAST] for an Inclusive tree is advertised in the Inclusive Ethernet Tag A-D route as described in section "Handling of Multi-Destination Traffic". Note that a MES can "aggregate" multiple inclusive trees for different E-VPNs on the same P2MP LSP using upstream labels. The procedures for aggregation are the same as those described in [VPLS-MCAST], with VPLS A-D routes replaced by E-VPN Inclusive Multicast Ethernet Tag A-D routes.

21.3.2. Selective Trees

A Selective P-Multicast tree is used by a MES to send IP multicast traffic for one or IP more specific multicast streams, originated by CEs connected to the MES, that belong to the same or different E-VPNs, to a subset of the MESs that belong to those E-VPNs. Each of the MESs in the subset should be on the path to a receiver of one or more multicast streams that are mapped onto the tree. The ability to use the same tree for multicast streams that belong to different E-VPNs is termed a MES the ability to create separate SP multicast trees for specific multicast streams, e.g. high bandwidth multicast streams. This allows traffic for these multicast streams to reach only those MES routers that have receivers in these streams. This avoids flooding other MES routers in the E-VPN.

A SP can use both Inclusive P-Multicast trees and Selective P-Multicast trees or either of them for a given E-VPN on a MES, based on local configuration.

The granularity of a selective tree is <RD, MES, S, G> where S is an IP multicast source address and G is an IP multicast group address or G is a multicast MAC address. Wildcard sources and wildcard groups are supported. Selective trees require explicit tracking as described below.

A E-VPN MES advertises a selective tree using a E-VPN selective A-D route. The procedures are the same as those in [VPLS-MCAST] with S-PMSI A-D routes in [VPLS-MCAST] replaced by E-VPN Selective A-D routes. The information elements of the E-VPN selective A-D route are similar to those of the VPLS S-PMSI A-D route with the following differences. A E-VPN Selective A-D route includes an optional Ethernet Tag field. Also a E-VPN selective A-D route may encode a MAC address in the Group field. The encoding details of the E-VPN selective A-D route will be described in the next revision.

Selective trees can also be aggregated on the same P2MP LSP using aggregation as described in [VPLS-MCAST].

21.4. Explicit Tracking

[VPLS-MCAST] describes procedures for explicit tracking that rely on Leaf A-D routes. The same procedures are used for explicit tracking in this specification with VPLS Leaf A-D routes replaced with E-VPN Leaf A-D routes. These procedures allow a root MES to request multicast membership information for a given (S, G), from leaf MESs. Leaf MESs rely on IGMP snooping or PIM snooping between the MES and the CE to determine the multicast membership information. Note that the procedures in [VPLS-MCAST] do not describe how explicit tracking is performed if the CEs are enabled with join suppression. The procedures for this case will be described in a future version.

22. Convergence

This section describes failure recovery from different types of network failures.

22.1. Transit Link and Node Failures between MESes

The use of existing MPLS Fast-Reroute mechanisms can provide failure recovery in the order of 50ms, in the event of transit link and node failures in the infrastructure that connects the MESes.

22.2. MES Failures

Consider a host host1 that is dual homed to MES1 and MES2. If MES1 fails, a remote MES, MES3, can discover this based on the failure of the BGP session. This failure detection can be in the sub-second range if BFD is used to detect BGP session failure. MES3 can update its forwarding state to start sending all traffic for host1 to only MES2. It is to be noted that this failure recovery is potentially faster than what would be possible if data plane learning were to be used. As in that case MES3 would have to rely on re-learning of MAC addresses via MES2.

22.2.1. Local Repair

It is possible to perform local repair in the case of MES failures. Details will be specified in the future.

22.3. MES to CE Network Failures

When an ethernet segment connected to a MES fails or when a Ethernet Tag is deconfigured on an ethernet segment, then the MES MUST withdraw the Ethernet Tag A-D route(s) announced for the <ESI, Ethernet Tags> that are impacted by the failure or de-configuration. In addition the MES MUST also withdraw the MAC advertisement routes that are impacted by the failure or de-configuration.

The Ethernet Tag A-D routes should be used by an implementation to optimize the withdrawal of MAC advertisement routes. When a MES receives a withdrawl of a particular Ethernet Tag A-D route it SHOULD consider all the MAC advertisement routes, that are learned from the same <ESI, Ethernet Tag> as in the Ethernet Tag A-D route, as having been withdrawn. This optimizes the network convergence times in the event of MES to CE failures.

23. LACP State Synchronization

This section requires review and discussion amongst the authors and will be revised in the next version.

To support CE multi-homing with multi-chassis Ethernet bundles, the MESes connected to a given CE should synchronize [802.1AX] LACP state amongst each other. This ensures that the MESes can present a single LACP bundle to the CE. This is required for initial system bring-up and upon any configuration change.

This includes at least the following LACP specific configuration parameters:

- System Identifier (MAC Address): uniquely identifies a LACP speaker.
- System Priority: determines which LACP speaker’s port priorities are used in the Selection logic.
- Aggregator Identifier: uniquely identifies a bundle within a LACP speaker.
- Aggregator MAC Address: identifies the MAC address of the bundle.
- Aggregator Key: used to determine which ports can join an Aggregator.
- Port Number: uniquely identifies an interface within a LACP speaker.
- Port Key: determines the set of ports that can be bundled.
- Port Priority: determines a port’s precedence level to join a bundle in case the number of eligible ports exceeds the maximum number of links allowed in a bundle.

Furthermore, the MESes should also synchronize operational (run-time) data, in order for the LACP Selection logic state-machines to execute. This operational data includes the following LACP operational parameters, on a per port basis:

- Partner System Identifier: this is the CE System MAC address.
- Partner System Priority: the CE LACP System Priority
- Partner Port Number: CE’s AC port number.
- Partner Port Priority: CE’s AC Port Priority.
- Partner Key: CE’s key for this AC.
- Partner State: CE’s LACP State for the AC.

- Actor State: PE’s LACP State for the AC.
- Port State: PE’s AC port status.

The above state needs to be communicated between MESes forming a multi-chassis bundle during LACP initial bringup, upon any configuration change and upon the occurrence of a failure.

It should be noted that the above configuration and operational state is localized in scope and is only relevant to MESes which connect to the same multi-homed CE over a given Ethernet bundle.

Furthermore, the communication of state changes, upon failures, must occur with minimal latency, in order to minimize the switchover time and consequent service disruption. The protocol details for synchronizing the LACP state will be described in the following version.

24. Acknowledgements

We would like to thank Yakov Rekhter, Pedro Marques, Kaushik Ghosh, Nischal Sheth, Robert Raszuk and Amit Shukla for discussions that helped shape this document. We would also like to thank Han Nguyen for his comments and support of this work.

25. References

[E-VPN-REQ] A. Sajassi, R. Aggarwal, et. al., "Requirements for Ethernet VPN", draft-sajassi-raggarwa-l2vpn-evpn-req-00.txt

[RFC4364] "BGP/MPLS IP VPNs", Rosen, Rekhter, et. al., February 2006

[VPLS-MCAST] "Multicast in VPLS". R. Aggarwal et.al., draft-ietf-l2vpn-vpls-mcast-04.txt

[RFC4761] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.

[RFC4762] Lasserre, M. and V. Kompella, "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.

[VPLS-MULTIHOMING] "BGP based Multi-homing in Virtual Private LAN Service", K. Kompella et. al., draft-ietf-l2vpn-vpls-multihoming-00.txt

[PIM-SNOOPING] "PIM Snooping over VPLS", V. Hemige et. al., draft-ietf-l2vpn-vpls-pim-snooping-01

[IGMP-SNOOPING] "Considerations for Internet Group Management Protocol (IGMP) and Multicast Listener Discovery (MLD) Snooping Switches", M. Christensen et. al., RFC4541,

[RT-CONSTRAIN] P. Marques et. al., "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, November 2006

26. Author’s Address

Rahul Aggarwal
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089 US
Email: rahul@juniper.net

Ali Sajassi
Cisco
170 West Tasman Drive
San Jose, CA 95134, US
Email: sajassi@cisco.com

Wim Henderickx
Alcatel-Lucent
e-mail: wim.henderickx@alcatel-lucent.com

Aldrin Isaac
Bloomberg
Email: aisaac71@bloomberg.net

James Uttaro
AT&T
200 S. Laurel Avenue
Middletown, NJ 07748
USA
Email: uttaro@att.com

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

Ravi Shekhar
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089 US

Florin Balus
Alcatel-Lucent
e-mail: Florin.Balus@alcatel-lucent.com

Keyur Patel
Cisco
170 West Tasman Drive
San Jose, CA 95134, US
Email: keyupate@cisco.com

Sami Boutros
Cisco
170 West Tasman Drive
San Jose, CA 95134, US
Email: sboutros@cisco.com

Network Working Group
Internet Draft
Intended Status: Standards Track
Expiration Date: January 7, 2011

Kamran Raza
Cisco Systems

Sami Boutros
Cisco Systems

July 8, 2010

LDP Typed Wildcard PW FEC Elements

draft-raza-l2vpn-pw-typed-wc-fec-01.txt

Status of this Memo

Copyright Notice

Abstract

An extension to the Label Distribution Protocol (LDP) defines the general notion of a "Typed Wildcard Forwarding Equivalence Class (FEC) Element". This can be used when it is desired to request all label bindings for a given type of FEC Element, or to release or withdraw all label bindings for a given type of FEC element. However, a typed wildcard FEC element must be individually defined for each type of FEC element. This specification defines the typed wildcard FEC elements for the Pseudowire Identifier (PW Id) and Generalized Pseudowire Identifier (Gen. PW Id) FEC types.

Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

Table of Contents

1. Introduction

An extension [TYPED-WC] to the Label Distribution Protocol (LDP) [RFC5036] defines the general notion of a "Typed Wildcard Forwarding Equivalence Class (FEC) Element". This can be used when it is desired to request all label bindings for a given type of FEC Element, or to release or withdraw all label bindings for a given type of FEC element. However, a typed wildcard FEC element must be individually defined for each type of FEC element.

[RFC4447] defines the "PWid FEC Element" and "Generalized PWid FEC Element" but it does not specify Typed Wildcard format for these elements. This document specifies the format of the Typed Wildcard FEC for the "PWid FEC Element" and the "Generalized PWid FEC Element" defined in [RFC4447]. The procedures for Typed Wildcard processing for PWid and Generalized PWid FEC Elements are same as described in [TYPED-WC] for any typed wildcard FEC Element type.

2. Typed Wildcard for PWid FEC Element

The format of the PWid FEC Typed Wildcard FEC is:

```
 0                   1                   2
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Typed Wcard   | Type = PWid   |   Len = 0     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 1: Format of PWid Typed Wildcard FEC Element

Where:

Typed Wcard (one octet): as specified in [TYPED-WC]

FEC Element Type (one octet): PWid FEC Element (type 0x80 [RFC4447])

Len FEC Type Info (one octet): Zero. (There is no additional FEC info)

3. Typed Wildcard for Generalized PWid FEC Element

The format of the Generalized PWid FEC Typed Wildcard FEC is:

```
 0                   1                   2
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Typed Wcard   | Type=Gen.PWid |   Len = 0     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2: Format of Generalized PWid Typed Wildcard FEC Element

Where:

Typed Wcard (one octet): as specified in [TYPED-WC]

FEC Element Type (one octet): Generalized PWid FEC Element (type 0x81 [RFC4447])

Len FEC Type Info (one octet): Zero. (There is no additional FEC info)

When Generalized PWid FEC Typed Wildcard is used, "PW Grouping ID TLV" [RFC4447] MUST NOT be present in the same message.

4. Operation

The use of Typed Wildcard FEC elements for PW can be useful under several scenarios. This section describes two use cases to illustrate their usage. The following use cases consider two LSR nodes, A and B, with LDP session between them to exchange L2VPN PW bindings.

4.1. PW Consistency Check

A user may request a control plane consistency check at LSR A for the PWid FEC and Generalized PWid FEC bindings that it had learnt from LSR B over LDP session. To perform this consistency check, LSR A marks all its learnt PW bindings from LSR B as stale, and then sends a Label Request message towards LSR B with Typed Wildcard FEC element for PWid FEC element and Generalized PWid FEC element. Upon receipt of such request, LSR B replays its database related to PWid FEC elements and Generalized PWid FEC element in Label Mapping message. As a PW binding is received at LSR A, the associated binding state is marked as refreshed (no stale). When replay completes for a given type of FEC, LSR B sends End-of-LIB Notification [END-OF-LIB] to mark the end of update for the given FEC type. Upon receipt of this Notification at LSR A, any remaining stale PW binding of given FEC type learnt from the peer LSR B, is cleaned up and removed from the database. This completes consistency check with LSR B at LSR A for given FEC type.

4.2. PW Graceful Shutdown

It may be desirable to perform shutdown/removal of existing PW bindings advertised towards a peer in a graceful manner - - i.e. all advertised PW bindings to be removed from a peer without session flap. For example, to request a graceful delete of the PWid FEC and Generalized PWid FEC bindings at LSR A learnt from LSR B, LSR A would send a Label Withdraw message towards LSR B with Typed Wildcard FEC elements pertaining to PWid FEC element and Generalized PWid FEC element. Upon receipt of such message, LSR B will delete all PWid and Generalized PWid bindings learnt from LSR A. Afterwards, LSR B would send Label Release message corresponding to recieved Label Withdraw with Typed FEC element.

5. Security Considerations

No new security considerations beyond that apply to the base LDP specification [RFC5036], [RFC4447] and [MPLS_SEC] apply to the use of the PW Typed Wildcard FEC Element types described in this document.

6. IANA Considerations

This document defines no new element for IANA Consideration.

7. Acknowledgments

The authors would like to thank Eric Rosen, M. Siva, and Zafar Ali for their valuable comments.

This document was prepared using 2-Word-v2.0 template.dot.

8. References

8.1. Normative References

[RFC5036] Andersson, L., Menei, I., and Thomas, B., Editors, "LDP Specification", RFC 5036, September 2007.

[TYPED-WC] Thomas, B., Asati, R., and Minei, I., "LDP Typed Wildcard FEC", draft-ietf-mpls-ldp-typed-wildcard-07.txt, Work in Progress, March 2010.

[END-OF-LIB] Asati, R., Mohapatra, P., Chen, E., and Thomas, B., "Signaling LDP Label Advertisement Completion", draft-ietf-mpls-ldp-end-of-lib-04.txt, Work in Progress, June 2010.

[RFC4447] L. Martini, Editor, E. Rosen, El-Aawar, T. Smith, G. Heron, "Pseudowire Setup and Maintenance using the Label Distribution Protocol", RFC 4447, April 2006.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC2119, March 1997.

8.2. Informative References

[MPLS_SEC] Fang, L. et al., "Security Framework for MPLS and GMPLS Networks", draft-ietf-mpls-mpls-and-gmpls-security-framework-05.txt, Work in Progress, March 2009.

Author’s Address

Syed Kamran Raza
Cisco Systems, Inc.,
2000 Innovation Drive,
Kanata, ON K2K-3E8, Canada.
E-mail: skraza@cisco.com

Sami Boutros
Cisco Systems, Inc.
3750 Cisco Way,
San Jose, CA 95134, USA.
E-mail: sboutros@cisco.com

Internet Working Group
Internet Draft
Category: Standards Track

Ali Sajassi(Editor)
Samer Salam
Clarence Filsfils
Cisco

R. Aggarwal(Editor)
Juniper Networks

Nabil Bitar
Verizon

Jim Uttaro
AT&T

Aldrin Isaac
Bloomberg

Wim Henderickx
Alcatel-Lucent

Expires: April 17, 2011 October 17, 2010

Requirements for Ethernet VPN (E-VPN)
draft-sajassi-raggarwa-l2vpn-evpn-req-00.txt

Status of this Memo

Copyright and License Notice

Abstract

The widespread adoption of Ethernet L2VPN services and the advent of new applications for the technology (e.g. data center interconnect) have culminated in a new set of requirements that are not readily addressable by the current VPLS solution. In particular, multi-homing with all-active forwarding is not supported and there's no existing solution to leverage MP2MP LSPs for optimizing the delivery of multi-destination frames. Furthermore, the provisioning of VPLS, even in the context of BGP-based auto-discovery, requires network operators to specify various network parameters on top of the access configuration. This document specifies the requirements for an Ethernet VPN (E-VPN) solution which addresses the above issues.

Table of Contents

1. Specification of Requirements

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 [RFC2119].

2. Introduction

VPLS, as defined in [RFC4664][RFC4761][RFC4762], is a proven and widely deployed technology. However, the existing solution has a number of limitations when it comes to redundancy, multicast optimization and provisioning simplicity. Furthermore, new applications are driving several new requirements for a VPLS service.

In the area of multi-homing current VPLS can only support multi-homing with active/standby resiliency model, for e.g. as described in [VPLS-BGP-MH]. Flexible multi-homing with all-active Attachment Circuits (ACs) cannot be supported by current VPLS solution.

In the area of multicast optimization, [VPLS-MCAST] describes how multicast LSPs can be used in conjunction with VPLS. However, this solution is limited to P2MP LSPs, as there's no defined solution for leveraging MP2MP LSPs with VPLS.

In the area of provisioning simplicity, current VPLS does offer a mechanism for single-sided provisioning by relying on BGP-based service auto-discovery [RFC4761][L2VPN-Sig]. This, however, still requires the operator to configure a number of network-side parameters on top of the access-side Ethernet configuration.

Furthermore, data center interconnect applications are driving the need for new service interface types which are a hybrid combination of VLAN Bundling and VLAN-based service interfaces. These are referred to as "VLAN-aware Bundling" service interfaces.

Also virtualization applications are fueling an increase in the volume of MAC addresses that are to be handled by the network, which gives rise to the requirement for having the network re-convergence upon failure be independent of the number of MAC addresses learned by the PE.

In addition, there are requirements for minimizing the amount of flooding of multi-destination frames and localizing the flooding to the confines of a given site.

Moreover, there are requirements for supporting flexible VPN topologies and policies beyond those currently covered by (H-)VPLS.

The focus of this document is on defining the requirements for a new solution, namely Ethernet VPN (E-VPN), which addresses the above issues.

Section 2 provides a summary of the terminology used. Section 3 discusses the redundancy requirements. Section 4 describes the multicast optimization requirements. Section 5 articulates the ease of provisioning requirements. Section 6 focuses on the new service interface requirements. Section 7 highlights the fast convergence requirements. Section 8 describes the flood suppression requirement, and finally section 9 discusses the requirements for supporting flexible VPN topologies and policies.

3. Terminology

CE: Customer Edge
E-VPN: Ethernet Virtual Private Network
MHD: Multi-homed Device
MHN: Multi-homed Network
LACP: Link Aggregation Control Protocol
LSP: Label Switched Path
PE: Provider Edge
PoA: Point of Attachment
PW: Pseudowire

4. Redundancy Requirements

4.1. Flow-based Load Balancing

A common mechanism for multi-homing a CE node to a set of PE nodes involves leveraging multi-chassis Ethernet link aggregation groups based on [802.1AX] LACP. [PWE3-ICCP] describes one such scheme. In Ethernet link aggregation, the load-balancing algorithms by which a

CE distributes traffic over the Attachment Circuits connecting to the PEs are quite flexible. The only requirement is for the algorithm to ensure in-order frame delivery for a given traffic flow. In typical implementations, these algorithms involve selecting an outbound link within the bundle based on a hash function that identifies a flow based on one or more of the following fields:

i. Layer 2: Source MAC Address, Destination MAC Address, VLAN
ii. Layer 3: Source IP Address, Destination IP Address
iii. Layer 4: UDP or TCP Source Port, Destination Port
iv. Combinations of the above.

A key point to note here is that [802.1AX] does not define a standard load-balancing algorithm for Ethernet bundles, and as such different implementations behave differently. As a matter of fact, a bundle operates correctly even in the presence of asymmetric load-balancing over the links. This being the case, the first requirement for active/active multi-homing is the ability to accommodate flexible flow-based load-balancing from the CE node based on L2, L3 and/or L4 header fields.

A solution MUST be capable of supporting flexible flow-based load balancing from the CE as described above. Further the MPLS network MUST be able to support flow-based load-balancing of traffic destined to the CE, even when the CE is connected to more than one PE. Thus the solution MUST be able to exercise multiple links connected to the CE, irrespective of the number of PEs that the CE is connected to.

4.2. Flow-based Multi-pathing

Any solution that meets the active-active flow based load balancing requirement described in section 3.1 MUST also be able to exercise multiple paths between a given pair of PEs. For instance if there are multiple RSVP-TE LSPs between a pair of PEs then the solution MUST be capable of load balancing traffic between those LSPs on a per flow basis. Similarly if LDP is being used as the transport LSP protocol, then the solution MUST be able to leverage LDP ECMP capabilities. The solution MUST also be able to leverage work in the MPLS WG that is in progress to improve the load balancing capabilities of the network based on entropy labels.

It is worth pointing out that flow-based multi-pathing complements flow-based load balancing described in the previous section.

4.3. Geo-redundant PE Nodes

The PE nodes offering multi-homed connectivity to a CE or access network may be situated in the same physical location (co-located), or may be spread geographically (e.g. in different COs or POPs). The latter is desirable when offering a geo-redundant solution that ensures business continuity for critical applications in the case of power outages, natural disasters, etc. An active/active multi-homing mechanism SHOULD support both co-located as well as geo-redundant PE placement. The latter scenario often means that requiring a dedicated link between the PEs, for the operation of the multi-homing mechanism, is not appealing from cost standpoint. Furthermore, the IGP cost from remote PEs to the pair of PEs in the multi-homed setup cannot be assumed to be the same when those latter PEs are geo-redundant.

4.4. Optimal Traffic Forwarding

In a typical network, and considering a designated pair of PEs, it is common to find both single-homed as well as multi-homed CEs being connected to those PEs. An active/active multi-homing solution SHOULD support optimal forwarding of unicast traffic for all the following scenarios:

i. single-homed CE to single-homed CE
ii. single-homed CE to multi-homed CE
iii. multi-homed CE to single-homed CE
iv. multi-homed CE to multi-homed CE

This is especially important in the case of geo-redundant PEs, where having traffic forwarded from one PE to another within the same multi-homed group introduces additional latency, on top of the inefficient use of the PE node’s and core nodes’ switching capacity. A multi-homed group (also known as a multi-chassis LACP group) is a group of PEs supporting a multi-homed CE.

4.5. Flexible Redundancy Grouping Support

In order to simplify service provisioning and activation, the multi-homing mechanism SHOULD allow arbitrary grouping of PE nodes into redundancy groups where each redundancy group represents all multi-homed groups that share the same group of PEs. This is best explained with an example: consider three PE nodes - PE1, PE2 and PE3. The multi-homing mechanism MUST allow a given PE, say PE1, to be part of multiple redundancy groups concurrently. For example, there can be a group (PE1, PE2), a group (PE1, PE3), and another group (PE2, PE3) where CEs could be multi-homed to any one of these three redundancy groups.

4.6. Multi-homed Network

There are applications which require an Ethernet network, rather than a single device, to be multi-homed to a group of PEs. The Ethernet network would typically run a resiliency mechanism such as MST or [G.8032] Ring Automated Protection Switching. The PEs may or may not participate in the control protocol of the Ethernet network.

A solution MUST support multi-homed network connectivity with active/standby redundancy.

A solution MUST also support multi-homed network with active/active VLAN-based load-balancing (i.e. disjoint VLAN sets active on disparate PEs).

A solution MAY support multi-homed network with active/active MAC-based load-balancing (i.e. different MAC addresses on a VLAN are reachable via different PEs).

5. Multicast Optimization Requirements

There are environments where the usage of MP2MP LSPs may be desirable for optimizing multicast, broadcast and unknown unicast traffic. [VPLS-LSM] precludes the usage of MP2MP LSPs since current VPLS solutions require an egress PE to perform learning when it receives unknown uncast packets over a LSP. This is challenging when MP2MP LSPs are used as MP2MP LSPs do not have inherent mechanisms to identify the sender. The usage of MP2MP LSPs for multicast optimization becomes tractable if the need to identify the sender for performing learning is lifted. A solution MUST be able to provide a mechanism that does not require learning when packets are received over a MP2MP LSP. Further a solution MUST be able to provide procedures to use MP2MP LSPs for optimizing delivery of multicast, broadcast and unknown unicast traffic.

6. Ease of Provisioning Requirements

As L2VPN technologies expand into enterprise deployments, ease of provisioning becomes paramount. Even though current VPLS has auto-discovery mechanisms which allow for single-sided provisioning, further simplifications are required, as outlined below:

- Single-sided provisioning behavior MUST be maintained
- For deployments where VLAN identifiers are global across the MPLS network (i.e. the network is limited to a maximum of 4K services), it is required that the devices derive the MPLS specific attributes (e.g. VPN ID, BGP RT, etc.) from the VLAN identifier. This way, it is sufficient for the network operator to configure the VLAN identifier(s) on the access circuit, and all the MPLS and BGP parameters required for setting up the service over the core network would be automatically derived without any need for explicit configuration.

- Implementations SHOULD revert to using default values for parameters as and where applicable.

7. New Service Interface Requirements

[MEF] and [IEEE 802.1Q] have the following services specified:

- Port mode: in this mode, all traffic on the port is mapped to a single bridge domain and a single corresponding L2VPN service instance. Customer VLAN transparency is guaranteed end-to-end.

- VLAN mode: in this mode, each VLAN on the port is mapped to a unique bridge domain and corresponding L2VPN service instance. This mode allows for service multiplexing over the port and supports optional VLAN translation.

- VLAN bundling: in this mode, a group of VLANs on the port are collectively mapped to a unique bridge domain and corresponding L2VPN service instance. Customer MAC addresses must be unique across all VLANs mapped to the same service instance.

For each of the above services a single bridge domain is assigned per service instance on the PE supporting the associated service. For example, in case of the port mode, a single bridge domain is assigned for all the ports belonging to that service instance regardless of number of VLANs coming through these ports.

It is worth noting that the term 'bridge domain' as used above refers to a MAC forwarding table as defined in the IEEE bridge model, and does not denote or imply any specific implementation.

[RFC 4762] defines two types of VPLS services based on "unqualified and qualified learning" which in turn maps to port mode and VLAN mode respectively.

A solution is required to support the above three service types plus two additional service types which are primarily intended for hosted data center applications and are described below.

For hosted data center interconnect applications, network operators require the ability to extend Ethernet VLANs over a WAN using a single L2VPN instance while maintaining data-plane separation between the various VLANs associated with that instance. This gives rise to two new service interface types: VLAN-aware Bundling without Translation, and VLAN-aware Bundling with Translation.

The VLAN-aware Bundling without Translation service interface has the following characteristics:

- The service interface MUST provide bundling of customer VLANs into a single L2VPN service instance.

- The service interface MUST guarantee customer VLAN transparency end-to-end.
- The service interface MUST maintain data-plane separation between the customer VLANs (i.e. create a dedicated bridge-domain per VLAN).
- In the special case of all-to-one bundling, the service interface MUST not assume any a priori knowledge of the customer VLANs. In other words, the customer VLANs shall not be configured on the PE, rather the interface is configured just like a port-based service.

The VLAN-aware Bundling with Translation service interface has the following characteristics:
- The service interface MUST provide bundling of customer VLANs into a single L2VPN service instance.
- The service interface MUST maintain data-plane separation between the customer VLANs (i.e. create a dedicated bridge-domain per VLAN).
- The service interface MUST support customer VLAN translation to handle the scenario where different VLAN Identifiers (VIDs) are used on different interfaces to designate the same customer VLAN.

The main difference, in terms of service provider resource allocation, between these new service types and the previously defined three types is that the new services require several bridge domains to be allocated (one per customer VLAN) per L2VPN service instance as opposed to a single bridge domain per L2VPN service instance.

8. Fast Convergence

A solution MUST provide the ability to recover from PE-CE attachment circuit failures as well as PE node failure for the case of both multi-homed device and multi-homed network. The recovery mechanism(s) MUST provide convergence time that is independent of the number of MAC addresses learned by the PE. This is particularly important in the context of virtualization applications which are fueling an increase in the number of MAC addresses to be handled by the Layer 2 network.
Furthermore, the recovery mechanism(s) SHOULD provide convergence time that is independent of the number of service instances associated with the attachment circuit or PE.

9. Flood Suppression

The solution SHOULD allow the network operator to choose whether unknown unicast frames are to be dropped or to be flooded. This attribute need to be configurable on a per service instance basis.

In addition, for the case where the solution is used for data-center interconnect, it is required to minimize the flooding of broadcast frames outside the confines of a given site. Of particular interest is periodic ARP traffic.

Furthermore, it is required to eliminate any unnecessary flooding of unicast traffic upon topology changes, especially in the case of multi-homed site where the PEs have a priori knowledge of the backup paths for a given MAC address.

10. Supporting Flexible VPN Topologies and Policies

A solution MUST be capable of supporting flexible VPN topologies that are not constrained by the underlying mechanisms of the solution. One example of this is hub and spoke where one or more sites in the VPN are hubs and the others as spokes. The hubs are allowed to send traffic to other hubs and to spokes, while spokes can communicate only with other hubs. The solution MUST provide the ability to support hub and spoke. Further the solution MUST provide the ability to apply policies at the MAC address granularity to control which PEs in the VPN learn which MAC address and how a specific MAC address is forwarded. It MUST be possible to apply policies to allow only some of the member PEs in the VPN to send or receive traffic for a particular MAC address.

11. Security Considerations

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

12. IANA Considerations

None.

13. Normative References

[RFC4664] "Framework for Layer 2 Virtual Private Networks (L2VPNs)", September 2006.

[RFC4761] "Virtual Private LAN Service (VPLS) Using BGP for Auto-discovery and Signaling", January 2007.

[RFC4762] "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", January 2007.

[802.1AX] IEEE Std. 802.1AX-2008, "IEEE Standard for Local and metropolitan area networks - Link Aggregation", IEEE Computer Society, November, 2008.

14. Informative References

[VPLS-BGP-MH] Kothari et al., "BGP based Multi-homing in Virtual Private LAN Service", draft-ietf-l2vpn-vpls-multihoming-00, work in progress, November, 2009.

[VPLS-MCAST] Aggarwal et al., "Multicast in VPLS", draft-ietf-l2vpn-vpls-mcast-06.txt, work in progress, March, 2010.

[PWE3-ICCP] Martini et al., "Inter-Chassis Communication Protocol for L2VPN PE Redundancy", draft-ietf-pwe3-iccp-02.txt, work in progress, Octoer, 2009.

[PWE3-FAT-PW] Bryant et al., "Flow Aware Transport of Pseudowires over an MPLS PSN", draft-ietf-pwe3-fat-pw-03.txt, work in progress, January 2010.

15. Authors' Addresses

Ali Sajassi
Cisco
170 West Tasman Drive
San Jose, CA 95134, USA
Email: sajassi@cisco.com

Samer Salam
Cisco
595 Burrard Street, Suite 2123
Vancouver, BC V7X 1J1, Canada
Email: ssalam@cisco.com

Rahul Aggarwal
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089, USA
Email: rahul@juniper.net

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

James Uttaro
AT&T
200 S. Laurel Avenue
Middletown, NJ 07748, USA
Email: uttaro@att.com

Aldrin Isaac
Bloomberg
Email: aisaac71@bloomberg.net

Clarence Filsfils
Cisco
Email: cfilsfil@cisco.com

Wim Henderickx
Alcate-lLucent
Email: wim.henderickx@alcatel-lucent.be

INTERNET-DRAFT So et al
Intended Status: Proposed Standard Verizon
Expires: November 2011 October 17, 2010

VPN Extensions for Private Clouds
draft-so-vepc-00.txt

Status of this Memo

Copyright and License Notice

Abstract

This contribution addresses the service providers requirements to support Cloud services interworking with the existing MPLS-based L2 and L3 VPN services. Maintenance of virtual separation of the traffic, data, and queries must be supported for the VPN customers that are conscious of end-to-end security features and functions that VPN technologies provide today.

Table of Contents

1 Introduction

Data center, WAN/MAN, and the end user are three of the components that make up the Cloud in the vision of Cloud Computing. However, the existing technologies often treat each component as black boxes, detached from each other. This fact limits the overall cohesiveness of an end-to-end service. For example, the network often views the data center as a black box, meaning the network has no control or visibility (from a standards point-of-view) into the data center.

As a network provider, a Cloud-service product may be offered across multiple data centers globally, some of which may be owned by a network provider while others may be owned by a partner/vendor. In addition, multiple Cloud-Service products can be offered in the same data centers. A list of the problems that this situation is causing the network provider/operator, especially for the existing VPN customers, is presented below. These must be addressed immediately, in order for service providers to persuade the existing VPN customers to leverage the deployed Cloud-based services.

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 Cloud Customer End to End Separation

2.1. VPN Traffic Segregation Requirements

The success of VPN services in the enterprise and the government world is largely due to its ability to virtually segregate the customer traffic at layer 2 and layer 3. The lower the layer that segregation can be maintained, the safer it is for the customers from security and privacy perspectives. Today data centers segregate the customer traffic at layer 7 (application), and there is no standard for extending the VPN into data center. Network service providers view the VPN extension into data center, allowing traffic segregation per VPN, an essential necessity to the success of Cloud-Services in the enterprises and government markets. Cloud-Applications (or the virtualization function) SHOULD have the ability to get access to VPN (including Layer 2/3 VPN) services, to segregate different Cloud-Services traffic trough the network.

2.2. Potential Solution

2.2.1. VPN Gateway Managed Connection Segregation

One possible way to achieve this is to have each Cloud-Application setup connections with the VPN gateways, while the gateways attach each connection to corresponding VPN. Each Cloud-Application SHALL be transmitted over a pre-defined set of connections, and each VPN utilizing the application SHALL be transmitted over a sub-set of application connections. In this case, each Cloud-Application SHALL maintain its own independent routing table. This is possible for some current operating systems, which already support multiple routing instances for its TCP/IP stack.

2.2.2 solution using Provider Backbone Bridging (PBB) and Shortest Path Bridging (SPB)

Ethernet and VLANS are the standard L2 connectivity model throughout the data center environment. As such the IEEE has been working on numerous projects to simplify and extend traditional Ethernet models for scale and flexibility. Additionally the IEEE has projects looking at new attachments models for Virtual Machines (VM’s) to become more autonomic and secure for environments that include wholly owned and multi-tenant.

Although VLAN and PPPoE are different types of connections, the two methods described above are fundamentally the same. Consequently, it is possible to generalize the descriptions above to cover both the cases.

2.2.3 VPN Gateway Controlled Traffic Flow

It is also possible for each Cloud-Application to acquire access to L2/L3 VPN with one shared routing table supported on the server. One way to do that is to have the VPN gateway manage the traffic flow instead of other way around. In that case, the VPN gateway has the VRF table and the destination server connection address. Once the server receives the traffic, it determines intra-data center destination based on the application. So the control sequence is VPN first, and then application. The control sequence for the first two methods described above is application first, and then VPN.

2.2.4 Inter-VPN Interworking

L2/L3 VPN based MPLS network can also be deployed in the data center to manage the intra-data center traffic flow. The data center VPN structure can be set up in such a way that each external VPN can be mapped to a unique internal VPN.

2.3. Cloud Services Virtualization

2.3.1. Cloud Virtualization Requirements

Today data center virtualization is totally handled by data center servers and hypervisors. The entire process is invisible to the underlying networks. The virtualization function including application server and virtual machine (VM) allocation and assignment, disk space allocation, traffic loading and balancing, QoS assignments, and so on. There shall be a way that the network can influence some virtualization functions that are important to the concept and spirit of the VPN.

- The Private Cloud provisioning and management system SHALL have the ability to dedicate a specific block of disk space per services per VPN.

- Each VPN SHALL have the exclusive access to the dedicated block of disk space.

- Each VPN SHALL have the ability to indicate the mechanism used to prevent the unwanted data retrieval for the block of disk space after it is no longer used by the VPN, before it can be re-used by other parties.

- Each VPN SHALL have the ability to request a dedicated VM with certainly CPU capability, amount of memory and disk space.

- The VPN SHALL have the ability to request dedicated L2/3 network resources within the data center such as bandwidth, priorities, and so on.

- The VPN SHALL have the ability to hold the requested resources without sharing with any other parties.

2.4. Cloud Services Restoration

Today, data center restoration and diversity designs are not necessarily linked to the network restoration and diversity design. This results in over-redundant design, wasting money and resources, and may cause traffic oscillation and service and performance degradation. This problem is particularly important to the VPN traffic, which is usually highly performance sensitive. The VPN extension SHOULD be able to indicate how the restoration is handled across layers, so that a unified end-to-end design and optimization can be achieved.

Furthermore the restoration capability awareness needs to be scalable, meaning problems occur in one area of the Cloud SHALL NOT affect all other areas of the Cloud involved. This way each component of the Cloud can scale independently without causing systemic failures and/or allowing a single failure to cascade across the Cloud.

2.5 Other Non-VPN Specific Areas

There are a number of known technology gaps preventing the data centers, networks, and the end users from interworking together in providing optimized and seamless end-to-end services. Although those areas are beyond VPN, they impact the VPN-based cloud services significantly. Those areas are listed below, but they are beyond the scope of this draft.

2.5.1. Cloud Traffic Load-Balancing and Congestion Avoidance

Todays Cloud traffic balancing and congestion avoidance is purely data center based. The network condition is not taken into consideration. The VPN extension SHOULD support the network condition to be used for the traffic balancing and congestion avoidance decision-making.

2.5.2. QoS Synchronization

It is required that the virtualization functions QoS requirement SHOULD be synchronized with VPN service.

2.5.3 Cross Layer Optimization

The VPN resource requested by the server CAN be optimized by statistical multiplexing of the resource. For example, for each VPN resource, it is possible to configure committed BW for each QoS resources and peak BW for best effort traffic, and the peak BW resources CAN be shared by different VPN service.

2.5.4 Automation end to end Configuration

The automatic end-to-end network configuration will reduce the operational cost and also the probability of occurrence of mis-configuration. The VPN Extension SHALL support the automatic end-to-end network configuration.

2.6. End-to-End Quality of Experience (ETE-QoE)

Quality of experience (QoE) management refers to maintaining a set of application /service layer parameters within certain threshold with an objective to retain the user experience for a specific service. Very often when new underlying technologies and/or mechanisms are introduced for implementing the same services (voice, data, video, messaging, etc.), opportunities exist to improve the user experiences. Conversely the user experience may suffer unless the appropriate transport level parameters that significantly impact the QoE are monitored and managed.

2.7. OAM Considerations

The VPN Extension solution MUST have sufficient OAM mechanisms in place to allow consistent end-to-end management of the solution in existing deployed networks. The solution SHOULD use existing protocols (802.3ag, Y.1731, BFD) wherever possible to help with interoperability of existing OAM deployments.

2.8. Work Item Considerations in IETF Clouds

In VPN extension to private Clouds, various application level parameters, protocol level parameter, and service monitoring parameters may need to be defined, and the results of monitoring may need to be exchanged periodically. In private cloud environment, since the resources exist in one or co-operative administrative domain, it is easier to monitor and manage the application and transport level parameters for the underlying resources. In some cases, proactive mechanisms can be readily implemented before user experiences degrade to the level of annoyance. In public and hybrid (a smart combination of private and public) clouds it is required to derive a list of mutually agreed upon monitoring and management parameters. Active monitoring using virtual agents and resources is also possible. However, allocation of resources and placement of the virtual agent including the amount of traffic generated for QoE management, and the exchange of the desired information back and forth need to be achieved.

3 Security Considerations

The VPN extension SHOULD support variety of security measures in securing tenancy of virtual resources such as resource locking, containment, authentication, access control, encryption, integrity measure, and etc. The VPN extension SHOULD allow the security to be configure end-to-end on a per VPN per-user bases. For example, the Virtual Systems MUST resource lock resources such as memory, but must also provide a cleaning function to insure confidentiality, before being reallocated.

VPN extension for private Clouds SHOULD specify an authentication mechanism based on an authentication algorithms (MD5, HMAC-SHA-1)for both header and payload. Encryption MAY also be use to provide confidentiality.

Security boundaries MAY also be create to maintain domains of TRUSTED, UNTRUSTED, and Hybrid. Within each domain access control techniques MAY be uses to secure resource and administrative domains.

4 IANA Considerations

None

5 References

5.1 Normative References

[RFC2119] S. Bradner, "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

5.2 Informative References

None

Author’s Addresses

Ning So
Verizon Inc.
2400 N. Glenville Rd.,
Richardson, TX 75082

ning.so@verizonbusiness.com

Henry Yu
tw telecom
10475 Park Meadows Dr.
Littleton, CO 80124
henry.yu@twtelecom.com

John M. Heinz
CenturyLink
Phone: 913-533-2115
john.m.heinz@centurylink.com

Paul Unbehagen
Alcatel-Lucent
8742 Lucent Boulevard
Highlands Ranch, CO 80129
paul.unbehagen@alcatel-lucent.com

Mike Mangino
Alcatel-Lucent
8742 Lucent Boulevard
Highlands Ranch, CO 80129
mike.mangino@alcatel-lucent.com

Bhumip Khasnabish
ZTE USA, Inc.
33 Wood Ave. S., 2nd Flr
Iselin, NJ, USA
Tel.:1-781-752-8003
Email: vumip1@gmail.com

Lizhong Jin
ZTE Corporation
889, Bibo Road
Shanghai, 201203, China
Email: lizhong.jin@zte.com.cn

Network working group
Internet Draft
Category: Standard Track
Expires: March 2011

X. Xu
Huawei Technologies

September 1, 2010

Virtual Subnet: A Scalable Data Center Network Architecture

draft-xu-virtual-subnet-03

Status of this Memo

Copyright Notice

Abstract

This document proposes a scalable data center network architecture which, as an alternative to the Spanning Tree Protocol Bridge network, uses a Layer 3 routing infrastructure based on BGP/MPLS IP VPN technology [RFC4364] with some extensions, together with some other proven technologies including ARP proxy [RFC925][RFC1027] to provide scalable virtual Layer 2 network connectivity services.

Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC-2119 [RFC2119].

Table of Contents

1. Problem Statement

With the popularity of cloud services, the scale of today’s data centers expands larger and larger. In addition, virtual machine migration technology, which allows a virtual machine to be able to migrate to any physical server while keeping the same IP address, is becoming more and more prevalent for achieving service agility in data centers. As a result, large Layer 2 networks are needed for server-to-server connectivity. Meanwhile, due to the huge-volume traffic exchanged between servers, the Layer 2 networks SHOULD provide enough capacity for server-to-server interconnections.

Unfortunately, today’s data center network using the Spanning-Tree Protocol (STP) Bridge technology, can not address the above challenges facing today’s large-scale data centers in several ways. First, STP can calculate out only one single forwarding tree for all connected servers of a particular Virtual Local Area Network (VLAN) and it can not support multi-path routing, e.g., Equal Cost Multi-Path (ECMP), hence the available network capacity in data center networks can’t be highly utilized so as to provide enough bandwidth between servers; Second, since the Bridge forwarding is based on the flat MAC addresses, the scalability of the Bridge forwarding table would become a big issue, especially when the existing large Layer 2 network scales even larger; Third, broadcast storm impacts imposed by some protocols, e.g., Address Resolution Protocol (ARP) and the flooding of unknown destination unicast frames become much more serious and unpredictable in the continually growing large-scale STP Bridge networks.

2. Terminology

This memo makes use of the terms defined in [RFC4364], [MVPN], [RFC2236] and [RFC2131]. Below are provided terms specific to this document:

- Service Domain: A group of servers which are dedicated for a given service and are usually located on a separate IP subnet.

3. Design Goals

To overcome the limitations of the STP Bridge networks as mentioned above, this document describes Virtual Subnet (VS), a practical data center network architecture, which aims to meet the following objectives:

- Bandwidth Utilization Maximization

To provide enough bandwidth between servers, the server-to-server traffic SHOULD always be delivered along the shortest path while multi-path routing is used for load-balancing purpose.

- Layer 2 Connectivity

To be compatible with the applications running in today's data centers (e.g., virtual machine migration), servers of a given service domain SHOULD be connected as if they were on a Local Area Network (LAN) or an IP subnet.

- Domain Isolation

To achieve performance and security isolation, servers belonging to different service domains SHOULD be isolated just as if they were located on separate Virtual LANs (VLAN) or IP subnets.

- Forwarding Table Scalability

To accommodate tens to hundreds of thousands of servers in a single data center network, the forwarding tables of those forwarding devices (e.g., routers or switches) SHOULD be scalable enough.

- Broadcast Storm Suppression

To alleviate the serious impacts on network performances which are imposed by broadcast storms, broadcast domains SHOULD be limited to their smallest scopes.

4. Architecture Description

VS actually uses BGP/MPLS IP VPN technology [RFC4364] with some extensions, together with other proven technologies including ARP proxy [RFC925][RFC1027] to build scalable large IP subnets across the MPLS/IP backbone of the data center network. As a result, VS can be deployed today as a practical and scalable data center network.

Since VS constructs large-scale IP subnets, rather than real LANs, the non-IP traffic would not be supported in VS anymore. However, given that IP traffic is the predominant type of traffic in today's data center networks and the non-IP traffic will disappear from the data center networks with the elapse of time, we believe that VS can be used as a practical data center network solution in most cases.

The following sections describe VS in detail.

4.1. Unicast IP traffic

4.1.1. Unicast IP Traffic inside a Service Domain

```
                          +--------------------+
    +-----------------+   |                    |   +------------------+
    |VPN_A:10/8       |   |                    |   |VPN_A:10/8        |
    |                 |   |                    |   |                  |
    |    +------+    ++---+-+                +-+---++    +------+     |
    |    |Host A+----+ PE-1 |                | PE-2 +----+Host B|     |
    |    +------+    ++-+-+-+                +-+-+-++    +------+     |
    |   10.1.1.1/8    | | |  IP/MPLS Backbone  | | |    10.1.1.2/8    |
    +-----------------+ | |                    | | +------------------+
                        | +--------------------+ |
                        |                        |
                        |                        |
                        V                        V
    +-------+------------+--------+     +-------+------------+--------+
    |VRF ID |Destination |Next Hop|     |VRF ID |Destination |Next Hop|
    +-------+------------+--------+     +-------+------------+--------+
    | VPN_A |10.1.1.1/32 |  Local |     | VPN_A |10.1.1.2/32 |  Local |
    +-------+------------+--------+     +-------+------------+--------+
    | VPN_A |10.1.1.2/32 |  PE-2  |     | VPN_A |10.1.1.1/32 |  PE-1  |
    +-------+------------+--------+     +-------+------------+--------+
```

Figure 1: Unicast IP Traffic inside a Service Domain

As shown in Figure 1, BGP/MPLS IP VPN technology with some extensions is deployed in a data center network. To maintain proper isolation of one service domain from another, each service domain is mapped to a distinct VPN and servers of a given service domain, as Customer Edge (CE) hosts, are attached to Provider Edge (PE) routers directly or through one or more Ethernet switches. In addition, to build large IP subnets across the MPLS/IP backbone, different sites of a particular VPN are associated with an identical IP subnet, in another words, PE routers attached to a given VPN are configured with distinct IP addresses of the same IP subnet on their VRF attachment circuits. PE routers create host routes for local CE hosts automatically according to their corresponding ARP entries. Instead of distributing the routes for the configured VPN subnets, PE routers distribute host routes for local CE hosts to each other. In addition, APR proxy is implemented on PE routers for every attached VPN, thus, upon receiving from a local CE host an ARP request for a remote CE host, the PE as an ARP proxy returns its own MAC address as a response.

Assume host A broadcasts an ARP request for host B before communicating with B, upon the receipt of this ARP request, PE-1 lookups the associated VRF to find the host route for B. If found and the route is learnt from a remote PE, PE-1 acting as an ARP proxy returns its own MAC address in the response to that ARP request. Otherwise, no ARP reply SHOULD be sent. After obtaining the ARP reply from PE-1, A sends an IP packet to B with destination MAC address of PE-1's MAC address. Upon receiving this packet, PE-1 acting as an ingress PE, tunnels the packet towards PE-2 which in turn, as an egress PE, forwards the packet to B.

4.1.2. Unicast IP Traffic between Service Domains

```
  +-------+-----------+--------+   +-------+-----------+--------+
  |VRF_ID |Destination |Next Hop|   |VRF_ID |Destination |Next Hop|
  +-------+-----------+--------+   +-------+-----------+--------+
  | VPN_A |10.1.1.2/32 |  PE-1  |   | VPN_B |20.1.1.2/32 |  PE-2  |
  +-------+-----------+--------+   +-------+-----------+--------+
  | VPN_A |10.1.1.1/32 | Local  |   | VPN_B |20.1.1.1/32 | Local  |
  +-------+-----------+--------+   +-------+-----------+--------+
  | VPN_A | 0.0.0.0/0  |10.1.1.1|   | VPN_B | 0.0.0.0/0  |20.1.1.1|
  +-------+-----------+--------+   +-------+-----------+--------+
          ^                                               ^
          |            +--------------------+             |
          |            |     IP Network     |             |
          |            +----+----------+---+              |
          |             +---+--+    +---+--+              |
          |             | GW-1 |    | GW-2 |              |
          |             +---+--+    +--+---+              |
          |VPN A:10.1.1.1/8 |          |VPN B:20.1.1.1/8 |
          |                 |          |                  |
          +-------------+---+--+    +--+---+-------------+
                       +-+ PE-3 +----+ PE-4 +-+
 +------------------+  | +------+    +------+ |  +------------------+
 |VPN A:10/8        |  |                      |  |VPN_B:20/8        |
 |                  |  |                      |  |                  |
 |    +------+    ++--+--+          +--+--++    +------+     |
 |    |Host A+----+ PE-1 |          | PE-2 +----+Host B|     |
 |    +------+    ++-++--+          +--++-++    +------+     |
 |    10.1.1.2/8  | ||   IP/MPLS Backbone   || |    20.1.1.2/8     |
 +------------------+ ||                      || +------------------+
                      |+----------------------+|
                      |                        |
                      V                        V
 +-------+-----------+--------+   +-------+-----------+--------+
 |VRF ID |Destination |Next Hop|   |VRF ID |Destination |Next Hop|
 +-------+-----------+--------+   +-------+-----------+--------+
```

```
| VPN_A |10.1.1.2/32 |  Local |   | VPN_B |20.1.1.2/32 |  Local |
+-------+------------+--------+   +-------+------------+--------+
| VPN_A |10.1.1.1/32 |  PE-3  |   | VPN_B |20.1.1.1/32 |  PE-4  |
+-------+------------+--------+   +-------+------------+--------+
| VPN_A | 0.0.0.0/0  |  PE-3  |   | VPN_B | 0.0.0.0/0  |  PE-4  |
+-------+------------+--------+   +-------+------------+--------+
```

Figure 2: Unicast IP Traffic between Service Domains

For servers of different VPNs (i.e., service domains) to communicate with each other, these VPNs SHOULD not be configured with any overlapping address spaces, besides, each VPN SHOULD be configured with at least one default route towards the gateway router (i.e. a CE router). As shown in Figure 2, PE-1 and PE-3 are attached to one VPN (i.e. VPN A) while PE-2 and PE-4 are attached to another VPN (i.e., VPN B). Host A and its gateway router (i.e., GW-1) are connected to PE-1 and PE-3, respectively. PE-3 is configured with a default route for VPN A and this default route is advertised to other PE routers. Similarly, host B and its gateway router (i.e., GW-2) are connected to PE-2 and PE-4, respectively. PE-4 is configured with a default route for VPN B and this default route is advertised to other PE routers. Now A sends an ARP request for its gateway (i.e., 10.1.1.1) before communicating with B. Upon receiving this ARP request, PE-1 lookups the associated VRF to find the host route for the gateway. If found and the route is learnt from a remote PE, PE-1 as an ARP proxy, returns its own MAC address in the ARP reply. After obtaining the ARP reply, A sends an IP packet for B with destination MAC address of PE-1's MAC. Upon receiving this packet, PE-1 as an ingress PE, tunnels it towards PE-3 according to the best-match route for that packet (i.e., the default route). PE-3 as an egress PE, in turn, forwards this packet towards the gateway router (i.e., GW-1). After the packet arrives at the gateway router for B (i.e., GW-2) after traversing through an IP network, GW-2 forwards the packet to PE-4 with destination MAC address of PE-4's MAC address if it has learnt an ARP for B from PE-4. Otherwise, GW-2 SHOULD broadcast an APR request for B. Upon receiving this packet, PE-4 as an ingress PE, tunnels it towards PE-2 which in turn, forwards it towards B.

4.2. Multicast/Broadcast IP Traffic

The MVPN technology [MVPN], in particular, the Protocol-Independent-Multicast (PIM) tree option with some extensions, is partially reused here to support IP multicast and broadcast between CE hosts of the same VPN. For example, PE routers attached to a given VPN join a default provider multicast distribution tree which is dedicated for that VPN. PE routers receiving customer multicast or broadcast traffic from local CE hosts forward such traffic to other remote PE routers over the corresponding default provider multicast distribution tree. When customer multicast or broadcast traffic is received from a provider multicast distribution tree, PE routers forward such traffic to the associated VRF attachment circuits.

For the customer multicast group of a particular VPN which carries high-volume traffic and not all sites of that VPN need the traffic of that customer multicast group, a dedicated provider multicast distribution tree other than the default provider multicast distribution tree for that VPN can be assigned optionally. As a result, those PE routers of that VPN that have no local CE hosts interested in that customer multicast group will not receive such traffic from remote PE routers anymore.

More details about how to support multicast and broadcast traffic in VS will be explored in a later version of this document.

4.3. CE Host Discovery

To discover all local CE hosts including gateway routers, PE routers SHOULD perform at least once ARP scan on the attached VPN subnet after rebooting. For example, a PE broadcasts an ARP request for each IP address within the subnet of each attached VPN. Alternatively, this PE could also broadcast an ARP request for a directed broadcast address (i.e., 255.255.255.255) or an ALL-Systems multicast group address (i.e., 224.0.0.1), that is to say, the target protocol address field is filled with 2555.255.255.255 or 224.0.0.1. Any CE host receiving this ARP request SHOULD respond with an ARP reply containing its IP and MAC addresses. After a round of such ARP scan, the PE will discover all local CE hosts and cache their ARP entries in its ARP table. After that, the PE could send ARP requests in unicast to each already-learnt local CE host periodically so as to check whether the CE host is still present on the subnet. Using unicast ARP requests has the advantage that it is quieter than using the broadcast because it won’t be received by all CE hosts on the subnet. When receiving a gratuitous ARP from a local CE host, the PE SHOULD cache the ARP entry of that CE host in its ARP table immediately if no ARP entry for that CE host exists yet. Otherwise, the PE SHOULD just update the corresponding ARP entry of that CE host. Most operating systems generate a gratuitous ARP request when the host boots up, the host’s network interface or links comes up, or an address assigned to the interface changes. In the scarce scenarios where a host does not generate a gratuitous ARP, the PE would have to perform ARP scan periodically.

4.4. CE Host Multi-homing and Mobility

When a given PE receives a host route for one of its local CE hosts from a remote PE, it SHOULD immediately send an ARP request for that CE host to the attached VPN subnet so as to determine whether that CE host is still connected locally. If an ARP reply is received in a short amount of time (imaging the CE host multi-homing scenario), the PE just needs to update the ARP entry for that CE host as normal. Otherwise (considering the virtual machine migration scenario), the PE SHOULD delete the ARP entry corresponding to that host from its APR table. Meanwhile, the PE SHOULD broadcast a gratuitous ARP on the attached VPN subnet on behalf of that CE host, with the sender hardware address field being filled with one of its own MAC addresses. As a result, the ARP entry for that CE host which has been cached on other local CE hosts is updated.

4.5. APR Proxy

The PE, acting as an ARP proxy, SHOULD only respond to the ARP requests for those CE hosts which have been learnt from other remote PE routers. Especially, the PE SHOULD not respond to ARP requests for local CE hosts. Otherwise, in case that the ARP reply from the PE covers that from the requested CE host, the packet for that local CE host which is sent from another local CE would be unnecessarily relayed by the PE.

When Virtual Router Redundancy Protocol (VRRP) [RFC2338], together with ARP proxy is enabled on multiple PE routers which are attached to the same VPN site, only the PE acting as VRRP master is delegated to perform ARP proxy function on the shared VPN subnet. In addition, it SHOULD use the virtual MAC address of that VRRP group in any ARP packet it sends, e.g., an APR reply to the ARP request from a local CE hosts.

4.6. DHCP Relay Agent

To avoid flooding Dynamic Host Configuration Protocol (DHCP) [RFC2131] broadcast messages through the data center network, DHCP Relay Agent can be implemented on PE routers for each attached VPN. Thus, DHCP broadcast messages received from DHCP clients on local CE hosts would be relayed by DHCP Relay Agents on PE routers to DHCP servers in unicast.

5. VS vs VPLS

Virtual Private LAN Service (VPLS) [RFC4761, RFC4762] provides private LAN services for IP as well as other protocols. To some extent, PE routers in VPLS work much similar as STP Bridges. As a result, the broadcast storm issues are intactly inherited from traditional STP bridge networks to VPLS.

At the cost of being lacking in support for non-IP traffic, VS alleviates the broadcast storm issues by using CE host route based Layer 3 routing and ARP proxy technologies on PE routers.

In addition, if CE hosts of multiple VPNs are attached to a PE router through an intermediate Ethernet bridge, in VPLS, this intermediate bridge would have to learn the MAC addresses of both local CE hosts and remote CE hosts of these attached VPNs. However, in VS, such intermediate bridge only needs to learn MAC addresses of local CE hosts and local PE routers due to the ARP proxy implemented on PE routers.

6. VS vs IPLS

Both VS and IP LAN Service (IPLS) [IPLS] are IP only L2VPN technologies.

In IPLS, ARP packets even including the unicast ARP reply packets are forwarded from attachment circuits to "multicast" PWs (although ARP request broadcast packets can be suppressed by PEs on which there are matching ARP entries for the ARP requests in their ARP caches). Besides, the received APR packets from the "multicast" PWs will be flooded to all CEs. As a result, the broadcast storm imposed by ARP traffic is worsen rather than being alleviated. Besides, as said in IPLS, "An IP frame received over a unicast PW is prepended with a MAC header before transmitting it on the appropriate attachment circuits and the source MAC address is the PE’s own local MAC address or a MAC address which has been specially configured on the PE for this use." As a result, the intermediary Ethernet switches between the PE and CEs can not keep the MAC entries of the remote CEs from expiring even there is continuous traffic between these CEs. Note that the destination MAC address of the packet to the remote CE which is sent from a local CE is the MAC of the remote CE, rather than the local PE’s MAC. Thus, flooding unknown destination unicast frames would not be avoided anymore on the above Ethernet switches unless these intermediary switches are configured to not age out the learned MAC entries (whether such configuration has any side-effects is uncertain). Third, IPLS prohibits connection of a common LAN or VLAN to more than one PE. In other words, IPLS can not support CE hosts being multi-homed to multiple PE Routers to achieve redundancy and load-balancing.

In contrast, all the above three issues with IPLS do not exist in VS while supporting IP only L2VPN services.

7. Conclusion

By using Layer 3 routing on the backbone of the data center network to replace the STP Bridge forwarding, traffic between any two servers is forwarded along the shortest path between them and multi-path routing is easily achieved. Thus, the total network bandwidth of the data center network is utilized to maximum extent.

By reusing the BGP/MPLS IP VPN technology to build large IP subnets across the backbones of data center networks, servers of a given VPN are allowed to communicate with each other just as if they are on the same subnet.

Due to the BGP/MPLS IP VPN technology, forwarding tables of P routers is sized to the number of PE routers rather than the total number of servers. Meanwhile, forwarding tables of PE routers can also scale well by distributing VPN instances and their corresponding routing table entries among multiple PE routers. Especially, thanks to the Outbound Route Filtering (ORF) capability of BGP, PE routers only needs to maintain the routing tables of their attached VPNs. Thus, the forwarding table scalability issue with data center networks is largely alleviated.

By enabling APR proxy function on PE routers, ARP broadcast messages from local CE hosts are blocked by local PE routers. Thus, APR broadcast messages will not flood the whole data center network. Besides, by enabling DHCP Relay Agent function on PE routers, DHCP broadcast messages from local CE hosts would be transformed into unicast messages by the DHCP Relay Agents and then be forwarded to DHCP servers in unicast. Thus, the broadcast storms in the data center networks are largely suppressed.

8. Future work

How to support IPv6 CE hosts in VS is for future study.

9. Security Considerations

TBD.

10. IANA Considerations

There is no requirement for IANA.

11. Acknowledgements

Thanks to Dino Farinacci for his valuable comments.

12. References

12.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

12.2. Informative References

[RFC4364] Rosen. E and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

[MVPN] Rosen. E and Aggarwal. R, "Multicast in MPLS/BGP IP VPNs", draft-ietf-l3vpn-2547bis-mcast-10.txt (work in progress), Janurary 2010.

[MVPN-BGP] R. Aggarwal, E. Rosen, T. Morin, Y. Rekhter, C. Kodeboniya, "BGP Encodings for Multicast in MPLS/BGP IP VPNs", draft-ietf-l3vpn-2547bis-mcast-bgp-08.txt (work in progress), September 2009.

[RFC826] Plummer, D., "An Ethernet Address Resolution Protocol or Converting Network Protocol Addresses to 48-bit Ethernet Addresses for Transmission on Ethernet Hardware", RFC-826, Symbolics, November 1982.

[RFC925] Postel, J., "Multi-LAN Address Resolution", RFC-925, USC Information Sciences Institute, October 1984.

[RFC1027] Smoot Carl-Mitchell, John S. Quarterman, "Using ARP to Implement Transparent Subnet Gateways", RFC 1027, October 1987.

[RFC2338] Knight, S., et. al., "Virtual Router Redundancy Protocol", RFC 2338, April 1998.

[RFC2131] Droms, R., "Dynamic Host Configuration Protocol", RFC 2131, March 1997.

[RFC2236] Fenner, W., "Internet Group Management Protocol, Version 2", RFC 2236, November 1997.

[RFC4761] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.

[RFC4762] Lasserre, M. and V. Kompella, "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.

[IPLS] H. Shah., et. al., "IP-Only LAN Service (IPLS)", draft-ietf-l2vpn-ipls-09.txt (work in progress), February 2010.

Authors' Addresses

Xiaohu Xu
Huawei Technologies,
No.3 Xinxi Rd., Shang-Di Information Industry Base,
Hai-Dian District, Beijing 100085, P.R. China
Phone: +86 10 82882573
Email: xuxh@huawei.com

Network Working Group
Internet Draft
Intended status: Standard Track
Expires: April 2011

Yong Lucy
Huawei
October 8, 2010

Flow-Aware Pseudowire for the Virtual Private LAN Service
draft-yong-l2vpn-fat-pw-4-vpls-00.txt

Status of this Memo

Copyright Notice

Abstract

A pseudowire (PW) is used in Virtual Private LAN Service (VPLS) solutions to form any-to-any connections and provide service demuliplexing among Provider Edge routers (PEs), and is normally transported over one single network path. Flow-aware PW enable a PW to take advantage of Equal Cost Multipath (ECMP) and/or Link Aggregation Groups (LAG) in a packet switched network (PSN). PW packets with a flow label can be transported over multi-paths. This method can apply to the PWs in a VPLS service.

This document describes how VPLS solutions utilize a PW with a flow label, and defines protocol extension for the provisioning of such PWs.

Table of Contents

1. Introduction

[RFC4664] specifies the Layer 2 virtual private network (L2VPN) framework. The L2VPN framework uses a point-to-point (P2P) pseudowire (PW) between any pair of Provider Edge routers (PEs) to provide connection and service demultiplexing. Each P2P PW is mapped to a traffic engineered (TE) tunnel traversing a packet switched network (PSN).

Two popular L2VPN services are Virtual Private Wire Service (VPWS) and Virtual Private LAN Service (VPLS). VPWS is a P2P transport service. VPLS is multi-point emulated LAN service. Two standard VPLS solutions are specified in [RFC4761] and [RFC4762]. One is BGP-based auto-discovery and singling; the other is LDP-based signaling.

Flow-aware PW [FAT-PW] is developed recently in IETF. It adds a flow label on the label stack and enables the distinction of the flows within a PW being carried over equal cost multipath (ECMP) and/or a link aggregation group (LAG) in a packet switched network (PSN). The target application of a PW with a flow label (i.e., of a PW split across multiple parallel paths) is to transport large volumes of IP traffic between two routers, for example, when providing a VPWS.

Service Providers use VPLS to provide an emulated LAN service and to transport customer Layer 2 frames between Customer Edge routers (CEs). Many L2VPN services carry Layer 2 frames that contain IP payloads. There is an incentive for a service provider to use the PW with a flow label in a VPLS service to support large volumes of data between points on the emulated LAN. This document describes a VPLS solution that uses a PW with a flow label and defines protocol extension for provisioning the PW.

2. Conventions Used in This Document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

3. VPLS Supported by PW with a Flow Label

A VPLS is a layer 2 service that emulates an Ethernet LAN across a Wide Area Network (WAN). It is a multipoint service. Although VPLS service frames are Ethernet frames, and frame forwarding is based on the destination MAC address, CE devices may be routers where the frame payload is IP data.

[RFC4761] specifies BGP based auto-discovery and signaling for VPLS configuration and operation procedures. [RFC4762] specifies LDP based VPLS configuration and operation procedures. Both configure PWs to support a VPLS instance, and in both cases the PW cannot be split across multiple paths. The extensions in this document are necessary for provisioning the PWs with a flow label that can be split across multiple paths in a VPLS instance.

Note: a VPLS service that uses PW with a flow label may have some different service aspects in a PSN from the one that does not. It is outside the scope of this document how a service provider differentiates the service profile between these two cases and how a PE classifies the flows for a PW with a flow label.

3.1. RFC4761 Extension for PW with a Flow Label

[RFC4761] uses BGP to discover PEs in a VPLS instance and configure a PW between any pair of PEs in the VPLS instance. [RFC4761] uses VPLS BGP Network Layer Reachability Information (NLRI) to exchange VPLS membership and demultiplexers, and uses the "Layer 2 Info Extended Community" to signal control information about the PW to be setup for a VPLS edge device (VE).

In this case, if a flow label is to be used to allow the PW to be carried over ECMP or a LAG, it is necessary to coordinate the use of the flow label between the ingress and egress PE. To signal the presence of the flow label in a PW, this document suggests using two bits in Control Flags. [RFC4761] has specified "Layer 2 Info Extended Community" and Control Flags Bit Vector. This document suggests using two bits in Control Flags Bit Vector to signal flow label present between the ingress and egress PEs. The suggested format is shown below.

```
  0 1 2 3 4 5 6 7
 +-+-+-+-+-+-+-+-+
 |   MBZ |T|R|C|S|      (MBZ = MUST Be Zero)
 +-+-+-+-+-+-+-+-+
```

Name C and S remain the same meaning as [RFC4761]. Name T and R are defined here.

- o When T=1 the PE is requesting the ability to send a PW packet that includes a flow label. When T=0, the PE is indicating that it will not send a PW packet containing a flow label.

- o When R=1 the PE is able to receive a PW packet with a flow label present. When R=0 the PE is unable to receive a PW packet with the flow label present.

The two new bits in Control Flags are used to synchronize the flow label state between the ingress and egress PEs. If PE does not support flow label, these two bits MUST be set to zero according to [RFC4761], which preserves backward compatibility. A PE that uses BGP signaling and does not set T bit to 1 MUST NOT include a flow label in the PW packet. This preserves backward compatibility with existing PW specifications.

A PE sending a Control Flag with T = 1 to a peer PE and receiving a Flow Label Flag with R = 1 from the peer PE SHOULD include a flow label in the PW packet. Under all other combinations of Flow Label Flag in signaling a PE MUST NOT include a flow label in the PW packet.

The signaling process allows that some PWs in a VPLS instance use a flow label on PW packets and other PWs in the same VPLS instance do not use flow labels. This provides the flexibility to support network migration.

What is signaled is the desire to include the flow label in the label stack. As [FAT-PW] mentions, the value of the label is a local matter for the ingress PE, and the label value itself is not signaled.

3.2. RFC4762 Extension for PW with Flow Label

[RFC4762] uses LDP to provision a VPLS instance and configure an Ethernet PW [RFC4448] between every pair of PEs to form a full mesh topology among PEs.

[RFC4762] identified three relevant interface parameters for a VPLS. This document adds the Flow Label Sub-TLV defined in [FAT-PW] as a relevant interface parameter for a VPLS. When Flow Label Sub-TLV is presented in label mapping message, ingress and egress PE MUST perform the same procedures described in section 4 of [FAT-PW].

If LDP signaling [RFC4762] is not in use for PW setup, then whether the flow label is used or not MUST be identically provisioned in both PEs at the PW endpoints. If there is no provisioning support for this option, the default behavior is not to include the flow label.

Data forwarding on an Ethernet PW MUST follow the procedures described in [RFC4762].

3.3. Virtual Service Instance (VSI) Forwarder

Each VPLS forms a full mesh among PEs in the VPLS. Every VSI at a PE in a given VPLS has exactly one point-to-point PW to every other VSI in the same VPLS. MAC address learning is done per PW association, i.e., the FIB keeps the mapping of the customer MAC address and PW association.[RFC4664]

When a PW in a VPLS is used with a flow label, the PW still appears as one single PW to the VSI. The VSI forwarder function is the same as the PW without flow label.[RFC4762] It is worth mentioning the case that, if ECMP is used at PEs, the ingress PE may distribute PW packets with the flow label to different tunnels; so the egress PE gets the packets from different tunnels and pass them to the same PW forwarder. The distribution method is local and outside the scope of document.

Flow recognition is discussed in Section 3.4. The VSI forwarder SHOULD be able to generate a flow label and process PW encapsulation as described in section 3.1 or 3.2.

It is possible that a VPLS uses point-to-multipoint (P2MP) PWs for traffic optimization [P2MP-PW-REQ], [BCAST-EXT]. A P2MP PW with a flow label requires that all the egress points can process that flow label, which makes harder to synchronize the decision. The solution for P2MP is for further study.

3.4. Flow Identification

A VPLS service transports customer Ethernet frames. When using a PW with a flow label, it requires that ingress PE identifies a flow or a group of flows within the service so that all frames from any one flow are given the same flow label and treated the same way in the network. This can be done by parsing the ingress Ethernet traffic and considering all of the IP traffic. Source and destination IP address, source and destination port, and protocol type may be used to identify the flow. Whether the ingress PE uses a PE bridge element or VSI forwarder to recognize the flow is a local implementation matter and is outside the scope of document.

4. Security Considerations

The protocol extension in this draft does not introduce any new security risk to the services and network beyond the analysis in [FAT-PW].

5. IANA Considerations

Not Any.

6. References

6.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC4448] Martini, L., Rosen, E., El-Aawar, N., and G. Heron, "Encapsulation Methods for Transport of Ethernet over MPLS Networks", RFC 4448, April 2006.

[RFC4761] Kompella, K., Rekhter, Y., "BGP Auto-Discovery and Signaling for VPLS", RFC 4761, January 2007.

[RFC4762] Lasserre & Kompella, Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling,

6.2. Informative References

[RFC4664] Andersson, L., Rosen, E.,"Framework for Layer 2 Virtual Private Networks (L2VPNs)", RFC 4664, 2006.

[FAT-PW] Bryan, S., et. Al, "Flow Aware Transport of Pseudowire over an MPLS PSN", draft-ietf-pwe3-fat-pw-04, Work in progress.

[P2MP-PW-REQ] Jounay, F., et. al, "Requirements for Point to Multipoint Pseudowire", draft-ietf-pwe3-p2mp-pw-requirements-03.txt, Work in progress.

[BCAST-EXT] Delord, S. et. Al, Extension to LDP-VPLS for Ethernet Broadcast and Multicast, draft-delord-l2vpn-ldp-vpls-broadcast-exten-03-txt, work in progress.

7. Acknowledgments

Author would like to thank Adrian Farrel, Mach Chen for the review and valuable suggestions.

Authors' Addresses

Lucy Yong
Huawei USA
1700 Alma Dr. Suite 500
Plano, TX 75075
Phone: +1 469-229-5387
Email: lucyyong@huawei.com