Network Working Group
Internet-Draft
Updates: 4761 (if approved)
Intended status: Standards Track
Expires: May 7, 2009

K. Kompella
B. Kothari
Juniper Networks
T. Spencer
AT&T

November 3, 2008

Multi-homing in BGP-based Virtual Private LAN Service
draft-kompella-l2vpn-vpls-multihoming-02.txt

Status of this Memo

Abstract

Virtual Private LAN Service (VPLS) is a Layer 2 Virtual Private Network (VPN) that gives its customers the appearance that their sites are connected via a Local Area Network (LAN). It is often required for the Service Provider (SP) to give the customer redundant connectivity to some sites, often called "multi-homing". This memo shows how multi-homing can be offered in the context of BGP-based VPLS.

Table of Contents

1. Introduction

Virtual Private LAN Service (VPLS) is a Layer 2 Virtual Private Network (VPN) that gives its customers the appearance that their sites are connected via a Local Area Network (LAN). It is often required for a Service Provider (SP) to give the customer redundant connectivity to one or more sites, often called "multi-homing". [RFC4761] explains how VPLS can be offered using BGP for auto-discovery and signaling; section 3.5 of that document describes how multi-homing can be achieved in this context. Implementation and deployment of multi-homing in BGP-based VPLS has suggested some refinement of the procedures described earlier; this memo details these changes.

Section 2 lays out some of the scenarios for multi-homing, other ways that this can be achieved, and some of the expectations of BGP-based multi-homing. Section 3 defines the components of BGP-based multi-homing, and the procedures required to achieve this. Section 7 may someday discuss security considerations.

1.1. General Terminology

Some general terminology is defined here; most is from [RFC4761] or [RFC4364]. Terminology specific to this memo is introduced as needed in later sections.

A "Customer Edge" (CE) device, typically located on customer premises, connects to a "Provider Edge" (PE) device, which is owned and operated by the SP. A "Provider" (P) device is also owned and operated by the SP, but has no direct customer connections. A "VPLS Edge" (VE) device is a PE that offers VPLS services.

A VPLS domain represents a bridging domain per customer. A Route Target community as described in [RFC4360] is typically used to identify all the PE routers participating in a particular VPLS domain. A VPLS site is a grouping of ports on a PE that belong to the same VPLS domain. Sites are referred to as local or remote depending on whether they are configured on the PE router in context or on one of the remote PE routers (network peers).

1.2. Conventions

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

2. Background

This section describes various scenarios where multi-homing may be required, and the implications thereof. It also describes some of the singular properties of VPLS multi-homing, and what that means from both an operational point of view and an implementation point of view. It describes briefly how the Spanning Tree Protocol can be used to achieve multi-homing, and how that compares with BGP-based multi-homing.

2.1. Scenarios

The most basic scenario is shown in Figure 1.

CE1 is a VPLS CE that is dual-homed to both PE1 and PE2 for redundant connectivity.

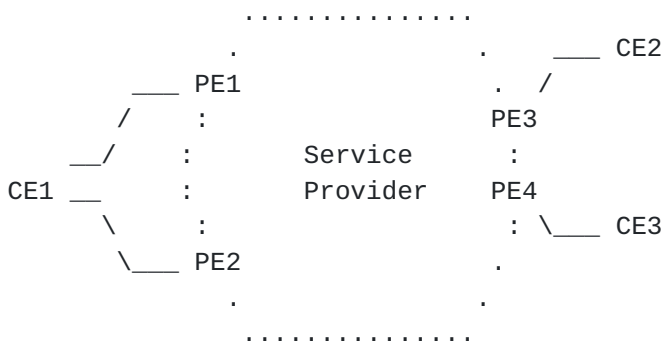


Figure 1: Scenario 1

CE1 is a VPLS CE that is dual-homed to both PE1 and PE2 for redundant connectivity. However, CE4, which is also in the same VPLS domain, is single-homed to just PE1.

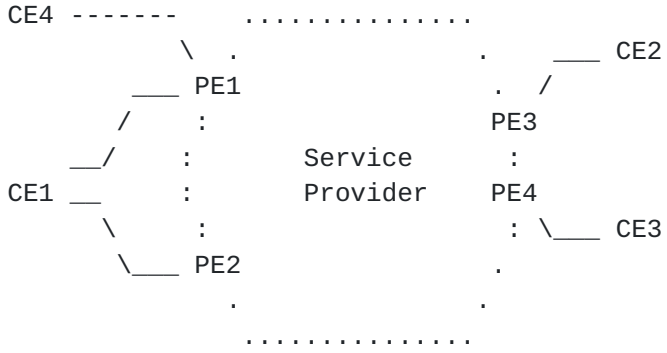


Figure 2: Scenario 2

2.2. VPLS Multi-homing Considerations

The first (perhaps obvious) fact about a multi-homed VPLS CE, such as CE1 in Figure 1 is that if CE1 is an Ethernet switch or bridge, a loop has been created in the customer VPLS. This is a dangerous situation for an Ethernet network, and the loop must be broken. Even if CE1 is a router, it will get duplicates every time a packet is flooded, which is clearly undesirable.

The next is that (unlike the case of IP-based multi-homing) only one of PE1 and PE2 can be actively sending traffic, either towards CE1 or into the SP cloud. That is to say, load balancing techniques will not work. All other PEs MUST choose the same designated forwarder for a multi-homed site. Call the PE that is chosen to send traffic to/from CE1 the "designated forwarder".

In Figure 2, CE1 and CE4 must be dealt with independently, since CE1 is dual-homed, but CE4 is not.

2.3. Using the Spanning Tree Protocol for Multi-homing

It is quite common to have redundant links in Ethernet networks; here too, redundancy leads to loops, but these can be broken by the use of the Spanning Tree Protocol (STP). This technique can also be applied in the case of multi-homed CEs in a VPLS domain. One approach is to run STP on the multi-homed CE (say CE1 in Figure 1). CE1 would thus detect a potential loop in the virtual LAN, and "block" either the link to PE1 or to PE2, breaking the loop. Blocking the link to PE2 would effectively pick PE1 to be the designated forwarder, since (a) PE2 will not get any traffic from CE1 to forward; (b) PE2's traffic to CE1 will be ignored.

There are several operational disadvantages to the STP approach:

1. The SP has to trust the customer to run STP correctly and manage changes carefully. If the customer makes a mistake, the SP will pay for it by carrying the customer's "broadcast storm" across the SP network.

2. The choice of whether PE1 or PE2 will be the "designated forwarder" is made by the customer; however, the SP may feel that they should make this choice, and in fact may be in a better position to do so, as they know their network topology better.

3. STP has several characteristics that make it unsuitable for carrier networks.

Another approach is to run STP on the PEs. However, the whole point

of having a full mesh of PE-PE connections, and of "split horizon" forwarding (Section 4.2.5 [RFC4761]; Section 4.4 [RFC4762]) is so that STP is not needed on PEs. Furthermore, in Figure 2, PE1 must not block the pseudowires to PE3 and PE4 in order to break the loop.

2.4. Active/Backup Links

Another approach is to define "active" and "backup" links from a multi-homed CE to the PEs. For example, in Figure 1, CE1 could define the link to PE1 as active and the link to PE2 as backup. If the link to PE1, or PE1 itself, fails, the CE1 could detect this and switch to the backup. However, again, the SP has to trust the customer's staff to handle this correctly; also, the choice of whether to use PE1 or PE2 remains with the customer.

2.5. Comparisons

One of the above methods may be acceptable in some cases. The technique described in this memo is for those who are unsatisfied with these methods. This technique relies on BGP mechanisms; furthermore, the choice of "designated forwarder" is retained by the SP. Finally, this technique can be used in conjunction with STP to get further "insurance" against the possibility of loops.

3. Multi-homing Operation

This section describes procedures for electing a designated forwarder among the set of PEs that are multi-homed to a customer site. It is imperative that all VPLS PEs elect the same designated forwarder otherwise either a loop will be formed or traffic will be dropped. Thus, procedures defined here MUST be supported by all BGP speakers that are required to process VPLS NLRI advertisements.

3.1. VE ID Assignment

Figure 1 shows a customer site, CE1, multi-homed to two VPLS PEs, PE1 and PE2. In order for all VPLS PEs within the same VPLS domain to elect one of the multi-homed PEs as the designated forwarder, an indicator that the PEs are multi-homed is required. This is achieved by assigning the same VE ID on PE1 and PE2 for CE1. When remote VPLS PEs receive NLRI advertisement from PE1 and PE2 for CE1, the two NLRI advertisements for CE1 are identified as candidates for designated forwarder selection due to the same VE ID. Thus, same VE ID MUST be assigned on all VPLS PEs that are multi-homed to the same customer site.

Figure 2 shows two customer sites, CE1 and CE4, connected to PE1 and CE1 multi-homed to PE1 and PE2. In such a case, PE1 SHOULD assign different VE IDs to CE1 and CE4, but the VE ID for CE1 on both PE1 and PE2 MUST be same.

Note that a VE ID = 0 is invalid.

3.2. VE Preference

When multiple PEs are assigned the same VE ID for multi-homing, it is often desired to make a particular PE as the designated forwarder. A VE preference is introduced in this document that can be used to control the selection of the designated forwarder. A VE preference indicates a degree of preference for a particular customer site. Absence of this preference will still elect a designated forwarder based on the algorithm explained in Section 3.4.

Section 3.2.4 in [RFC4761] describes the Layer2 Info Extended Community that carries control information about the pseudowires. The last two octets that were reserved now carries VE preference as shown in Figure 3.

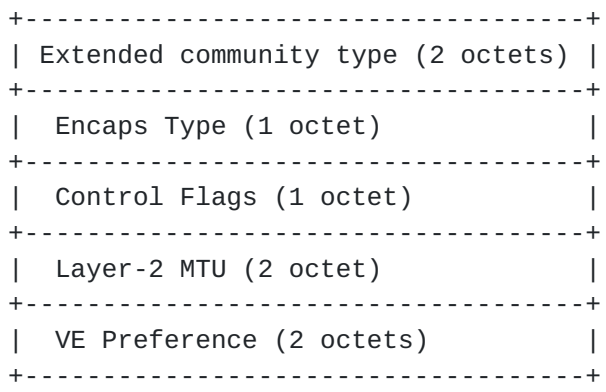


Figure 3: Layer2 Info Extended Community

A VE preference is a 2-octets unsigned integer. A value of zero indicates absence of VE preference and is not a valid preference value. This interpretation is required for backwards compatibility. Implementations using Layer2 Info Extended Community as described in (Section 3.2.4) [RFC4761] MUST set the last two octets as zero since it was a reserved field.

3.3. BGP Local Preference

Section 3.5 in [RFC4761] describes the use of BGP Local Preference in path selection to choose a particular NLRI, where Local Preference indicates the degree of preference for a particular VE. The use of Local Preference is inadequate when VPLS PEs are spread across multiple ASes as Local Preference is not carried across AS boundary.

For backwards compatibility, if VE preference as described in Section 3.2 is used, then BGP Local Preference MUST be set to the value of VE preference. Note that a Local Preference value of zero for a VE is not valid unless 'D' bit in the control flags is set (see [I-D.kothari-l2vpn-auto-site-id]). In addition, Local Preference value greater than or equal to 2^16 for VPLS advertisements is not valid.

3.4. Designated Forwarder Election

BGP-based multi-homing for VPLS relies on BGP path selection and VPLS path selection. BGP path selection as defined in this document for VPLS NLRIs MUST be done by any BGP speaker that is required to process VPLS NRLI advertisements. Thus, a Route Reflector, [RFC4456], MUST support the procedures defined in this document for BGP path selection for VPLS. Similarly, a BGP speaker that is also a VPLS PE MUST also do BGP path selection for VPLS advertisements.

VPLS path selection, however, is done only by a VPLS PE. The net result of doing both BGP and VPLS path selection is that of electing a single designated forwarder among the set of PEs to which a customer site is multi-homed.

In order to explain how these two path selection algorithms work, one must refer to the format of the VPLS NLRI. This NLRI contains: <Route Distinguisher, VE ID, VE Block Offset, VE Block Size, Label Base> (Section 3.2.2) [RFC4761]. These components are referred as RD, VE-ID, VBO, VBS and LB, respectively. In addition, a VPLS advertisement contains some attributes, among them the BGP nexthop (BNH), control flags (CF), VE Preference (VP), and Local Preference (LP). A VPLS advertisement might contain a Route Origin Attribute (RO). Finally, the VPLS domain (DOM) is needed; this is not carried explicitly in a VPLS advertisement, but is derived, typically from BGP policies applied on Route Targets carried in the advertisement. In addition to these fields in the advertisement, there are two derived fields called PE-ID and PREF. The Table 1 shows how to set the value of PREF based on VP and LP. The Table 2 shows how to set the value of PE-ID based on RO and BNH.

Valid values for VP	Valid values for LP	Valid values for PREF	Comment
0	0	0	malformed advertisement, unless CF:D=1
0	1 to (2^16-1)	LP	backwards compatibility
0	2^16 to (2^32-1)	(2^16-1)	backwards compatibility
>0	LP same as VP	VP	Implementation supports VP
>0	LP != VP	0	malformed advertisement

Table 1

RO Present	PE-ID	Comment
Yes	Global Administrator sub-field of RO	Source PE as specified in RO
No	BNH	Source PE as specified by BGP nexthop

Table 2

Taken all together, this yields:

<RD, VE-ID, VBO, VBS, LB; DOM, PE-ID, CF, PREF>

Both BGP and VPLS path selection algorithms are described in two stages. For each algorithm, the first stage divides all received VPLS advertisements into buckets of relevant and comparable advertisements. In this stage, advertisements may be discarded as not being relevant to path selection. The second stage picks a single "winner" from each bucket by repeatedly applying a tie-breaking algorithm on a pair of advertisements from that bucket. The tie-breaking rules are such that the order in which advertisements are picked from the bucket does not affect the final result. Note that this is a conceptual description of the process; an implementation MAY choose to realize this differently as long as the semantics are preserved.

3.4.1. BGP Path Selection

3.4.1.1. Bucketization

An advertisement

AD = <RD, VE-ID, VBO, VBS, LB; DOM, PE-ID, CF, PREF>

is discarded if DOM is not of interest to the BGP speaker. Otherwise, AD is put into the bucket for <RD, VE-ID, VBO>. In other words, the prefix to use for comparison in BGP path selection consists of <RD, VE-ID, VBO> and only advertisements with exact same <RD, VE-ID, VBO> are candidates for path selection.

3.4.1.2. Tie-breaking Rules

Given two advertisements AD1 and AD2 as below, the following tie-breaking rules MUST be applied in the given order (note that the RDs,

VE-IDs and VBOs are the same):

AD1 = <RD, VE-ID, VBO, VBS1, LB1; DOM, PE-ID1, CF1:D, PREF1>
AD2 = <RD, VE-ID, VBO, VBS2, LB2; DOM, PE-ID2, CF2:D, PREF2>

where CF:D is the 'D' bit in the control flags and PREF is derived as shown in Table 1

1. if (CF1:D != 1) AND (CF2:D == 1) AD1 wins; stop
   if (CF1:D == 1) AND (CF2:D != 1) AD2 wins; stop
   else continue

2. if (PREF1 > PREF2) AD1 wins; stop;
   else if (PREF1 < PREF2) AD2 wins; stop;
   else continue

3. if (PE-ID1 < PE-ID2) AD1 wins; stop;
   else if (PE-ID1 > PE-ID2) AD2 wins; stop;
   else AD1 and AD2 are equivalent; BGP will consider this as an update

For VPLS advertisements, the above rules supercede the tie breaking rules described in (Section 9.1.2.2) [RFC4271]

3.4.2. VPLS Path Selection

3.4.2.1. Bucketization

An advertisement

AD = <RD, VE-ID, VBO, VBS, LB; DOM, PE-ID, CF, PREF>

is discarded if DOM is not of interest to the VPLS PE. Otherwise, AD is put into the bucket for <DOM, VE-ID>. In other words, all advertisements for a particular VPLS domain that have the same VE-ID are candidates for VPLS path selection.

3.4.2.2. Tie-breaking Rules

Given two advertisements AD1 and AD2 as below, the following tie-breaking rules MUST be applied in the given order (note that VE-IDs are same).

AD1 = <RD, VE-ID, VBO, VBS1, LB1; DOM, PE-ID1, CF1:D, PREF1>
AD2 = <RD, VE-ID, VBO, VBS2, LB2; DOM, PE-ID2, CF2:D, PREF2>

where CF:D is the 'D' bit in the control flags

1. if (CF1:D != 1) AND (CF2:D == 1) AD1 wins; stop
   if (CF1:D == 1) AND (CF2:D != 1) AD2 wins; stop
   else continue

2. if (PREF1 > PREF2) AD1 wins; stop;
   else if (PREF1 < PREF2) AD2 wins; stop;
   else continue

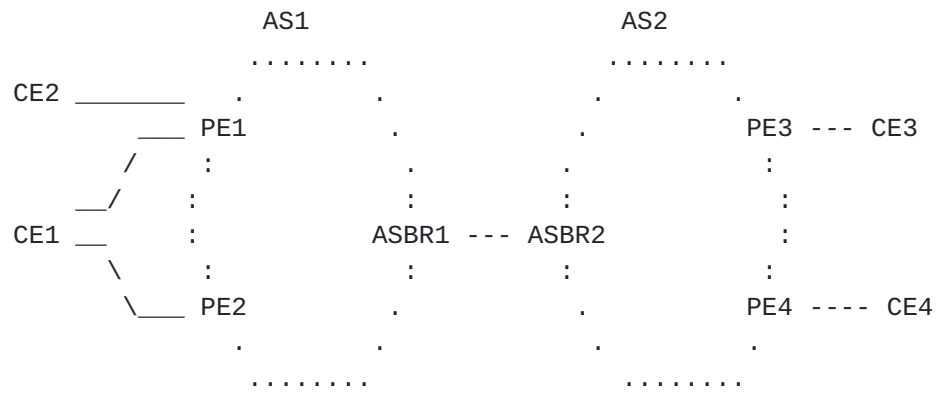
3. if (PE-ID1 < PE-ID2) AD1 wins; stop;
   else if (PE-ID1 > PE-ID2) AD2 wins; stop;
   else AD1 and AD2 are from the same VPLS PE; AD1 and AD2 should
   both be retained and an implementation MAY sort the
   advertisements by other criteria such as VBO

If the final "winning" advertisement has VE-ID = 0 OR VBO = 0 OR VBS = 0, it is discarded.

4. Multi-AS VPLS

Section 3.4 in [RFC4761] describes three methods (a, b and c) to connect sites in a VPLS to PEs that are across multiple AS. Since VPLS advertisements in method (a) do not cross AS boundaries, multi-homing operations for method (a) remain exactly the same as they are within as AS. However, both for method (b) and (c), VPLS advertisements do cross AS boundary. This section describes the VPLS operations for method (b) and method (c). Consider Figure 4 for inter-AS VPLS with multi-homed customer sites.

4.1. Inter-AS Method (b): EBGP Redistribution of VPLS Information between ASBRs



Assume VE IDs to be:
CE1: 1
CE2: 2
CE3: 3
CE4: 4

Figure 4: Inter-AS VPLS

A customer has four sites, CE1, CE2, CE3 and CE4. CE1 is multi-homed to PE1 and PE2 in AS1. CE2 is single-homed to PE1. CE3 and CE4 are also single homed to PE3 and PE4 respectively in AS2. After running path selection algorithm, all four VPLS PEs must elect the same set of designated forwarder for all customer sites. Since BGP Local Preference is not carried across AS boundary, VE preference as described in Section 3.2 MUST be used for carrying site preference in inter-AS VPLS operations.

As explained in (Section 3.4.2) [RFC4761], ASBR1 will send a VPLS

NLRI received from PE1 to ASBR2 with new labels and itself as the BGP nexthop. ASBR2 will send the received NLRI from ASBR1 to PE3 and PE4 with new labels and itself as the BGP nexthop. Since VPLS PEs use BGP Local Preference in path selection, for backwards compatibility, ASBR2 MUST set the Local Preference value in the VPLS advertisements it sends to PE3 and PE4 to the VE preference value contained in the VPLS advertisement it receives from ASBR1. ASBR1 MUST do the same for the NLRIs it sends to PE1 and PE2. If ASBR1 receives a VPLS advertisement without a valid VE preference from a PE within its AS, then ASBR1 MUST set the VE preference in the advertisements to the Local Preference value before sending it to ASBR2. Similarly, ASBR2 must do the same for advertisements without VE Preference it receives from PEs within its AS. Thus, in method (b), ASBRs MUST update the VE and Local Preference based on the advertisements they receive either from a PE within their AS or an ASBR.

Since ASBR rewrites the BGP nexthop for VPLS advertisements it receives from other ASes, the VPLS PEs no longer have the visibility of the remote end PEs. In Figure 4, both PE3 and PE4 receives VPLS NLRIs from ASBR2 for VE IDs 1 and 2, with BGP nexthop of ASBR2. However, the VPLS PEs that originated the advertisements for VE IDs 1 and 2 are PE1 and PE2. Due to lack of information about the PEs that originated the VPLS NLRIs, both PE3 and PE4 will only create one PW to ASBR2 as to PE3 and PE4, the customer sites CE1 and CE2 appear connected to ASBR2. However, two PWs are required in this case, one for CE1 and another one for CE2. This can only be achieved if PE3 and PE4 know the originator PE for each advertisement received. For this purpose, Route Origin Extended Community [RFC4360] is used to carry the source PE's IP address.

To use Route Origin Extended Community for carrying the originator VPLS PE's loopback address, the type field of the community MUST be set to 0x01 and the Global Administrator sub-field MUST be set to the PE's loopback IP address.

If a PE receives a VPLS NLRI with Route Origin Extended Community, then the PE MUST use the IP address contained in the community as the source PE. Otherwise, BGP nexthop for the VPLS advertisements MUST be used as the source PE IP address. A VPLS PE MUST create one active PW per remote PE. In Figure 4, PE1 will send the VPLS advertisements with Route Origin Extended Community containing its loopback address. PE2 will do the same. Even though PE3 receives the VPLS advertisements for VE ID 1 and 2 from the same BGP nexthop, ASBR2, the source PE address contained in the Route Origin Extended Community is different for the CE1 and CE2 advertisements, and thus, PE3 creates two PWs, one for CE1 (for VE ID 1) and another one for CE2 (for VE ID 2).

4.2. Method (c): Multi-Hop EBGP Redistribution of VPLS Information between ASes

In this method, there is a multi-hop E-BGP peering between the PEs or Route Reflectors in AS1 and the PEs or Route Reflectors in AS2. There is no VPLS state in either control or data plane on the ASBRs. The multi-homing operations on the PEs in this method are exactly the same as they are in intra-AS scenario. However, since Local Preference is not carried across AS boundary, the translation of LP to VP and vice versa MUST be done by RR, if RR is used to reflect VPLS advertisements to other ASes. This is exactly the same as what a ASBR does in case of method (b). A RR must set the VP to the LP value in an advertisement before sending it to other ASes and must set the LP to the VP value in an advertisement that it receives from other ASes before sending to the PEs within the AS.

5. VPLS Operation with multiple VE Identifiers

VE Identifiers uniquely identifies a particular customer site in a VPLS domain. Even when multiple customer sites are attached to the same VPLS PE as in Figure 5, a single VE ID is sufficient to represent the two customer sites, A and B, as both are connected to the same PE.

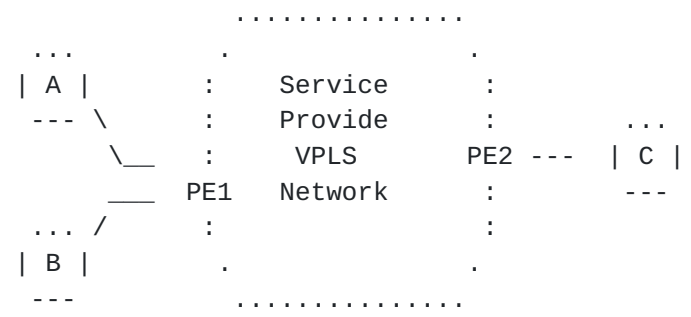


Figure 5: Multiple Customer sites with single VE ID

However, if sites of a customer are multi-homed to different set of PEs, such as in Figure 6, and redundancy per site is desired, then PEs MUST advertise a unique VE ID for each site that requires redundancy.

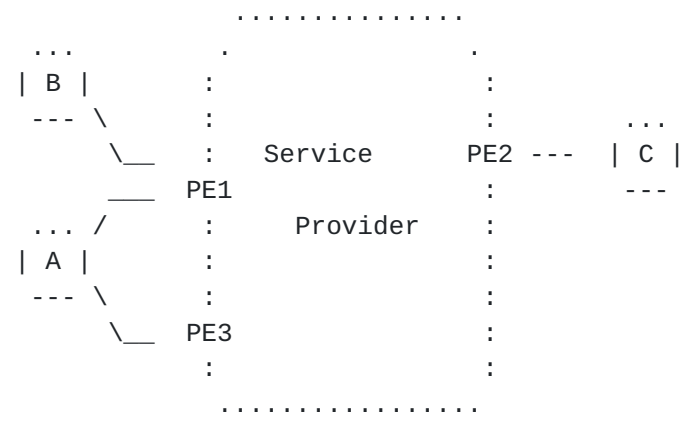


Figure 6: Multiple Customer sites with different VE ID

In Figure 6, site A is multi-homed to PE1 and PE3, but site B is single homed to PE1. Since redundancy for both A and B is desired, PE1 and PE3 MUST assign the same VE ID for site A, and PE1 MUST assign a different VE ID for B.

This section describes the VPLS operations, both for intra and inter

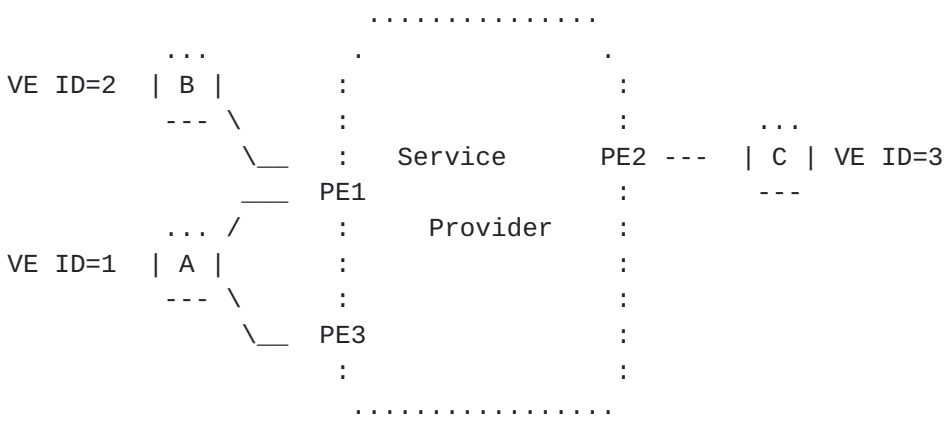
AS scenarios, when there are multiple sites with different VE IDs, as in Figure 6.

5.1. Pseudowire Establishment

This section explains how PWs are established between the PEs, when more than one customer site with different VE ID is connected to the same PE. Procedures described in this section are in context of one VPLS domain. Route Target, as explained in [RFC4360], identifies a VPLS domain.

When a PE receives VPLS NLRIs for multiple VE IDs for the same VPLS instance from a remote PE, it MUST create an active PW by selecting one of the VE IDs as primary and SHOULD create standby PWs for other VE IDs. The setting up of PWs follow existing procedures defined in RFC 4761. To select a site for setting up primary PW, an advertisement with the lowest VE ID is selected.

In Figure 7,since VE preference of PE1 is better than PE3 for VE ID 1, PE1 wins the designated forwarder election based on Section 3.4. Thus, PE1 is the designated forwarder for site A, and since site B is single homed to PE1, PE1 will always be the forwarder for site B.



VPLS NLRIs advertised by each PE:

PE1: VE_ID=1, LB=11, OFF=1, Pref=200 (for site A)
VE_ID=2, LB=11, OFF=1, Pref=100 (for site B)

PE2: VE_ID=3, LB=21, OFF=1 Pref=100 (for site B)

PE3: VE_ID=1, LB=101, OFF=1 Pref=100 (for site C)

where 'LB' is label base and 'OFF' is VE block offset as carried in VPLS NLRI. 'Pref' is VE Preference as carried in Layer2 Info Extended Community

Figure 7: Multiple Customer sites with different VE ID

PE2 MUST select VE ID of 1 for setting up active PW to PE1. Thus, PE2 MUST use labels 21 and 22 to accept traffic from PE1, label 21 being the active PW and label 22 for standby PW. Note that PE2 MUST associate MAC addresses from both PWs to remote PE, PE1, and should not associate MAC addresses to individual PWs. In case the active PW from PE1 to PE2 goes down, all MAC addresses learned from PE1 on PE2 will remain associated unless age out occurs or PE1 sends a FLUSH-MESSAGE to PE2 [I-D.kothari-l2vpn-vpls-flush]. More details on flushing of MAC addresses are explained in Section 6.

When a PE has multiple sites, it MUST advertise the same label base, block offset and range for all its sites. In Figure 7, PE1 is advertising label base of 11, block offset of 1 and block range of 8 for both sites A (VE ID 1) and B (VE ID 2). When PE1's advertisements reach PE2, PE2 will always send traffic to PE1 with label 13, irrespective of which site A or B is active. This eliminates the need for PE2 to have multiple PWs to PE1.

5.2. Handling Link Failures

In Figure 7, when link connectivity between site A and PE1 goes down, PE1 MUST immediately send traffic to PE2 with label 22 instead of label 21 that it was previously using. It MUST also send a BGP update with 'D' bit set in the control flags. PE1 is no longer the designated forwarder for site A.

Since PE2 has both labels, 21 and 22, there is no disruption of traffic to PE2 when PE1 switches to label 22 from label 21. There should be no MAC movement or re-learning on PE2 for traffic from PE1 for site B as a result of label switch by PE1. In addition, PE2 will continue to use label 13 for traffic to PE1 and thus, connectivity failure between A and PE1 has no impact on the traffic from PE2 to PE1.

When PE3 receives the advertisement from PE1 with the 'D' bit set, it MUST elect itself as the designated forwarder for site A based on the multi-homing path selection rules. Similarly, PE2 elects PE3 as the designated forwarder for the site A. PE3 creates PWs to PE2 using normal procedures and starts using label advertised by PE2 to send traffic to PE2. Due to the change in designated forwarder, MAC addresses received on PE2 with label 21 are associated with PE3. Note that if MAC addresses for site A were not flushed on PE2 when link between A and PE1 went down, then PE2 can see MAC movement as it is now learning site A's MAC addresses from PE3.

Instead of connectivity between site A and PE1, if link connectivity between site B and PE1 goes down, there is no impact on traffic as PE1 is already using label 21 for traffic to PE2. PE2 will continue to use label 13 for traffic to PE1 with no change. Unless explicitly flushed or age out occurs, MAC addresses for site B will remain as is on PE2.

6. MAC Flush Operations

In a service provider VPLS network, customer MAC learning is confined to PE devices and any intermediate nodes, such as a Route Reflector, do not have any for state for MAC addresses.

Topology changes either in the service provider's network or in customer's network can result in the movement of MAC addresses from one PE device to another. Such events can result into traffic being dropped due to stale state of MAC addresses on the PE devices. Age out timers that clear the stale state will resume the traffic forwarding, but age out timers are typically in minutes, and convergence of the order of minutes can severely impact customer's service. To handle such events and expedite convergence of traffic, flushing of affected MAC addresses is highly desirable.

A VPLS PE uses VPLS FLush Capability [I-D.kothari-l2vpn-vpls-flush] to negotiate the use of VPLS-FLUSH message for MAC flush operations. This section describes the scenarios where VPLS flush is desirable and the specific VPLS Flush TLVs that provide capability to flush the affected MAC addresses on the PE devices. All operations described in this section are in context of a particular VPLS domain and not across multiple VPLS domains.

6.1. MAC List FLush

If multiple customer sites are connected to the same PE, PE1 as shown in Figure 7, and redundancy per site is desired when multi-homing procedures described in this document are in affect, then it is desired to flush just the relevant MAC addresses from a particular site when the site connectivity is lost.

To flush particular set of MAC addresses, a PE SHOULD originate a VPLS-FLUSH message with MAC list TLV (TLV type 0) that contains a list of MAC addresses that needs to be flushed. In Figure 7, if connectivity between A and PE1 goes down and if PE1 was the designated forwarder for A, PE1 SHOULD send a list of MAC addresses that belong to A to all its BGP peers.

If connectivity to both site A and B are down on PE1, then PE1 SHOULD not send a VPLS-FLUSH message as the remote PEs will flush all MAC addresses that belong to PE1, as described in Section 6.2.

If a single customer site is connected to a PE, and the connectivity to the site is lost, then the PE SHOULD not send a VPLS-FLUSH message as the remote PEs will flush all MAC addresses that they learned from the source PE (see Section 6.2).

It is RECOMMENDED that in case of excessive link flap of customer attachment circuit in a short duration, a PE should have a means to throttle advertisements of VPLS-FLUSH messages so that excessive flooding of such advertisements do not occur.

6.2. Implicit MAC Flush

If a PE detects that all PWs from a source PE for a VPLS domain are down, then the PE should flush all MAC addresses learned from that source PE. Need for a VPLS-FLUSH message is only for cases when a primary PW is torn down and standby PWs are in operational state. Thus, a PE should not advertise VPLS-FLUSH message for cases when an implicit flush due to loss of all PWs is sufficient.

When a connectivity to a customer site is lost, a PE either withdraws the VPLS NLRI that it previously advertised for the site or it sends a BGP update message for the site's VPLS NLRI with the 'D' bit set. In either case, remote PEs learn that a particular site is no longer reachable.

In Figure 7, if PE1 withdraws VPLS NLRIs for both site A and B or sends BGP update with VPLS NLRIs for both A and B with 'D' bit set, then PE2 SHOULD flush all MAC addresses that it learned from PE1. PE1 should not send a VPLS-FLUSH message in this case.

If PE1 withdraws VPLS NLRIs for just site A or sends an update for site A NLRIs with 'D' bit set, then PE2 SHOULD not flush MAC addresses that it learned from PE1, unless PE2 has no standby PWs to PE1.

7. Security Considerations

No new security issues are introduced beyond those that are described in [RFC4761].

8. IANA Considerations

At this time, this memo includes no request to IANA.

9. Acknowledgments

The authors would like to thank Chaitanya Kodeboyina, Yakov Rekhter, Nischal Sheth and Amit Shukla for their insightful comments and probing questions.

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.

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

[I-D.kothari-l2vpn-auto-site-id] Kothari, B., Kompella, K., and T. IV, "Automatic Generation of Site IDs for Virtual Private LAN Service", draft-kothari-l2vpn-auto-site-id-01 (work in progress), October 2008.

[I-D.kothari-l2vpn-vpls-flush] Kothari, B. and R. Fernando, "VPLS Flush in BGP-based Virtual Private LAN Service", draft-kothari-l2vpn-vpls-flush-00 (work in progress), October 2008.

10.2. Informative References

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, February 2006.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

[RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, April 2006.

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

[RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

Authors' Addresses

Kireeti Kompella
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: kireeti@juniper.net

Bhupesh Kothari
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: bhupesh@juniper.net

Tom Spencer
AT&T

Email: tsiv@att.com

Full Copyright Statement

Intellectual Property