Network Working Group
Internet-Draft
Updates: 4761 (if approved)
Intended status: Standards Track
Expires: January 14, 2009

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

July 13, 2008

Multi-homing in BGP-based Virtual Private LAN Service
draft-kompella-l2vpn-vpls-multihoming-01.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 3
 1.1. General Terminology 3
 1.2. Conventions 3
2. Background 4
 2.1. Scenarios 4
 2.2. VPLS Multi-homing Considerations 5
 2.3. Using the Spanning Tree Protocol for Multi-homing 5
 2.4. Active/Backup Links 6
 2.5. Comparisons 6
3. Multi-homing Operation 7
 3.1. VE ID Assignment 7
 3.2. VE Preference 7
 3.3. BGP Local Preference 8
 3.4. Designated Forwarder Election 8
  3.4.1. BGP Path Selection 9
  3.4.2. VPLS Path Selection 10
4. Multi-AS VPLS 12
5. Security Considerations 13
6. IANA Considerations 14
7. Acknowledgments 15
8. References 16
 8.1. Normative References 16
 8.2. Informative References 16
Authors' Addresses 17
Intellectual Property and Copyright Statements 18

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 5 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.

```
                          ...............
                         .               .    ___ CE2
                   ___ PE1                .  /
                  /    :                  PE3
               __/    :       Service      :
           CE1 __     :       Provider    PE4
                 \     :                   : \___ CE3
                  \___ PE2                .
                         .               .
                          ...............
```

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.

```
           CE4 -------    ...............
                      \  .               .    ___ CE2
                   ___ PE1                .  /
                  /    :                  PE3
               __/    :       Service      :
           CE1 __     :       Provider    PE4
                 \     :                   : \___ CE3
                  \___ PE2                .
                         .               .
                          ...............
```

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.

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.

```
+-----------------------------------+
| Extended community type (2 octets) |
+-----------------------------------+
|  Encaps Type (1 octet)            |
+-----------------------------------+
|  Control Flags (1 octet)          |
+-----------------------------------+
|  Layer-2 MTU (2 octet)            |
+-----------------------------------+
|  VE Preference (2 octets)         |
+-----------------------------------+
```

Figure 3: Layer2 Info Extended Community

A VE preference 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. A VPLS advertisement with a higher VE preference MUST be preferred.

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])

3.4. Designated Forwarder Election

BGP-based multi-homing for VPLS relies on BGP path selection and VPLS path selection. BGP path selection 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). 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. Taken all together, this yields:

<RD, VE-ID, VBO, VBS,LB; DOM, BNH, CF, VP, LP>

Note that an advertisement with VE-ID = 0 is invalid.

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, BNH, CF, VP, LP>

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, BNH1, CF1:D, VP1, LP1>
AD2 = <RD, VE-ID, VBO, VBS2, LB2; DOM, BNH2, CF2:D, VP2, LP2>
```

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 (VP1 == 0) OR (VP2 == 0) continue
    else if (VP1 > VP2) AD1 wins; stop
    else if (VP1 < VP2) AD2 wins; stop
    else continue

3.  if (LP1 > LP2) AD1 wins; stop;
    else if (LP1 < LP2) AD2 wins; stop;
    else continue

4.  if (BNH1 < BNH2) AD1 wins; stop;
    else if (BNH1 > BNH2) AD2 wins; stop;
    else AD1 and AD2 are equivalent; BGP will consider this as an
    update
```

Note that all other BGP path selection criteria, such as IGP metric, MUST be ignored while doing path selection for VPLS advertisements.

3.4.2. VPLS Path Selection

3.4.2.1. Bucketization

An advertisement

```
AD = <RD, VE-ID, VBO, VBS, LB; DOM, BNH, CF, VP, LP>
```

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, BNH1, CF1:D, VP1, LP1>
AD2 = <RD, VE-ID, VBO, VBS2, LB2; DOM, BNH2, CF2:D, VP2, LP2>
```

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 (VP1 == 0) OR (VP2 == 0) continue
    else if (VP1 > VP2) AD1 wins; stop
    else if (VP1 < VP2) AD2 wins; stop
    else continue

3.  if (LP1 > LP2) AD1 wins; stop;
    else if (LP1 < LP2) AD2 wins; stop;
    else continue

4.  if (BNH1 < BNH2) AD1 wins; stop;
    else if (BNH1 > BNH2) 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. Consider Figure 4 for inter-AS VPLS with multi-homed customer sites.

```
                         AS1                    AS2
                       ........               ........
                      .        .             .        .
                ___ PE1         .           .          PE3 ___
               /    :            .         .            :     \
            __/    :             :         :             :     \__
        CE1 __     :           ASBR1 --- ASBR2           :      __ CE2
              \     :            :         :            :      /
               \___ PE2         .           .          PE4 ___/
                      .        .             .         .
                       ........                ........
```

Figure 4: Inter-AS VPLS

A customer has two sites, CE1 and CE2. CE1 is multi-homed to PE1 and PE2 in AS1. CE2 is multi-homed to PE3 and PE4 in AS2. After running path selection algorithm, all four VPLS PEs must elect the same set of designated forwarder for CE1 and CE2. 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.

In method (b), there is control plane VPLS state on the ASBRs. 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. Thus, in method (b), ASBRs MUST set the BGP Local Preference in VPLS advertisements to the VE preference value, if specified in the NLRIs received from other ASBRs.

In method (c), there is no state of any kind on the ASBRs. Thus, multi-homing operations do not apply to ASBRs in this method.

5. Security Considerations

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

6. IANA Considerations

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

7. Acknowledgments

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

8. References

8.1. 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-00 (work in progress), November 2007.

8.2. 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.

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