Internet Draft Document
Category: Standards Track
Expires: August 2008

Vach Kompella
Joe Regan
Ron Haberman
Alcatel-Lucent

Shane Amante
Level 3 Communications

February 2008

Regional VPLS
draft-vkompella-l2vpn-rvpls-00.txt

Status of this Memo

This Internet-Draft will expire on August 21, 2008.

Copyright Notice

Abstract

This draft proposes an alternative signaling approach that improves the scaling of HVPLS, which is compatible with the basic HVPLS model. It reduces the learning requirements on the PE, for certain topologies.

1. Introduction

In this draft, we present an extension to HVPLS signaling that addresses the scalability issues arising when inter-regional VPLS's are connected.

[VPLS] defines how a hierarchical VPLS (H-VPLS) can be established in a single region or administrative domain. As the reach of a VPLS increases, a PE in the core of a flat VPLS can experience scaling issues in multiple dimensions: provisioning, signaling, flooding/replication, MAC addresses. An H-VPLS can alleviate the issues of provisioning, signaling and flooding/replication. This comes at the expense of an increased number of MAC addresses learned at an interior H-VPLS PE.

This draft proposes an approach that builds on [VPLS] to create a scalable inter-region VPLS. In order to achieve MAC address scalability, a gateway PE treats an entire region as a single PE. There is no visibility of MAC addresses beyond the gateway PE of another region. Instead, in an R-VPLS, the gateway PE will need to learn only the source MAC addresses of all locally originated customer packets which pass through the gateway PE. The scalability of the solution depends on the topology and distribution of MAC addresses.

2. The Scalability Issues

Consider the following network model:

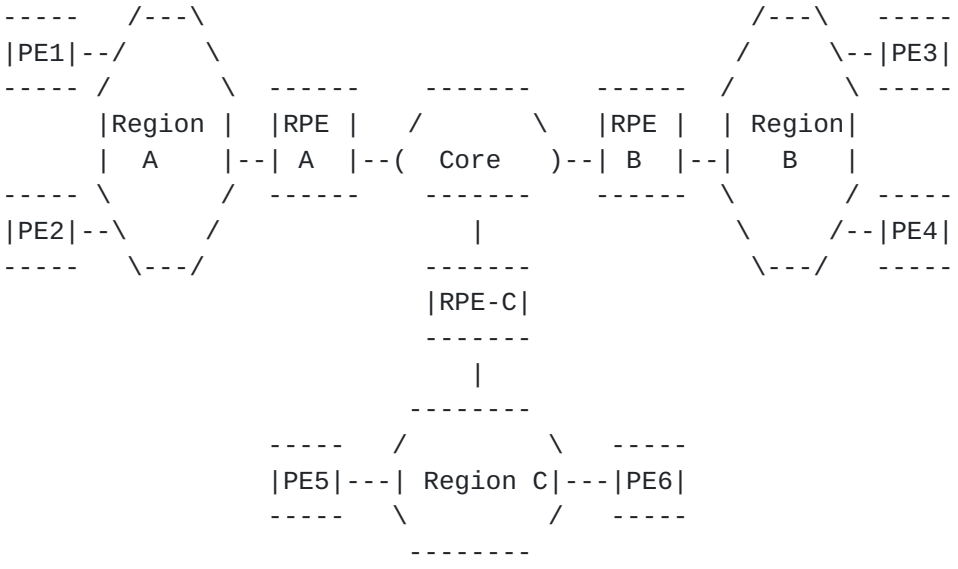


Figure 1: An R-VPLS with three regions

There are three regions A, B and C. In each region, a local VPLS instance is requested at each of PE1 through PE6, to which the customer attaches. The go through regional PEs (RPEs) which, in the conventional VPLS/MS-PW solutions below, would have just the properties of supporting PW cross-connects or HVPLS. However, in Section 3, the RPE is a regional PE supporting a modified forwarding and control plane behavior.

2.1. Full mesh VPLS

The first option for constructing this would be a full-mesh VPLS between the 6 nodes. That would be inefficient in several ways. Firstly, there are five sessions out of each PE. Secondly, when packets are flooded, multiple copies go through the regional PEs (RPEs). This is just the consequence of the topology of the network. However, the PEs are the only ones involved in the VPLS, and therefore, the only ones learning MAC addresses. This is the minimum set of nodes that would have to learn MAC addresses. The configuration overhead issue of adding another node to the VPLS involves touching the configuration on all the other members of the VPLS. (Auto-discovery [BGP-AD] does address this problem).

2.2. Hierarchical VPLS

The second option would be to set up an H-VPLS, with PE1 and PE2 as spokes to RPE-A, PE3 and PE4 as spokes to RPE-B, and PE5 and PE6 as spokes to RPE-C. RPE-A, RPE-B and RPE-C are configured as the full-mesh VPLS. This option will reduce the number of sessions out of each PE, so no node has more than four sessions. The number of packets replicated at each service-aware node is a maximum of three. Configuration impacts are fairly minimal when adding another PE to a region because all that has to be configured is a spoke from the new PE to the regional PE. However, the MAC addresses are learned at the R-PEs as well as at the PEs. This introduces a new scaling problem, over the ones that are solved by introducing the hierarchy. Furthermore, PE1 has to hairpin through RPE-A to get to PE2 in the same region.

2.3. Multi-VPLS with multiple split horizon groups

A third option would be to set up a local VPLS in each region between the PEs and the regional RPE, and connect the regions through a core VPLS connecting the RPEs. This would require an implementation that can maintain multiple split horizon groups

in a single VPLS. This solution avoids the problem of hair-pinning through the RPE. It still keeps the replication and session counts low, but it does not address the MAC scalability problem.

2.4. Multi-segment VPLS

A fourth option would be to use multi-segment PWs between PEs. These MS-PWs would run through the RPEs acting as S-PEs. The system would behave like a full-mesh VPLS, but the session counts would stay low on the PEs, and the MAC addresses would only be learned at the PEs. However, the replication issue and the consequent traffic on the RPEs remain. In addition, the number of labels used in this service increases.

3. Regional VPLS

The R-VPLS solution tries to combine aspects of all the above in one solution. The PEs know their local RPE (perhaps through extensions to the Capability FEC TLV [LDP-Cap]). The rules are as follows:

Each PE sends a single PW label to the nodes in the region.
Each PE sends a PW label to its RPE for each remote RPE.
Each RPE sends a single PW label to each other RPE in the core.
Each RPE sends a PW label to each PE in its region for each RPE that it talks to.

Let's take an example to explain the signaling that could take place (this is an incomplete list of the labels exchanged, but enough to explain the flow of traffic for both unknown destination and known destination MACs):

1. PE1 sends out label 1011 to RPE-A for region B.
2. PE1 sends out label 1012 to RPE-A for region C.
3. PE1 sends out label 101 to PE2.
4. PE2 sends out label 1021 to RPE-A for region B.
5. PE2 sends out label 1022 to RPE-A for region C.
6. PE2 sends out label 102 to PE1.
7. RPE-A sends out label 1101 to PE1 for region B.
8. RPE-A sends out label 1102 to PE1 for region C.
9. RPE-A sends out label 1001 to RPE-B.
10. RPE-A sends out label 1002 to RPE-C.
11. RPE-B sends out label 2001 to RPE-A.
12. RPE-B sends out label 2002 to RPE-C.
13. RPE-C sends out label 3001 to RPE-A.
14. RPE-C sends out label 3002 to RPE-B.

15. PE3 sends out label 2031 to RPE-B for region A.
16. PE3 sends out label 2032 to RPE-B for region C.
17. PE3 sends out label 201 to PE4.
18. PE4 sends out label 2041 to RPE-B for region A.
19. PE4 sends out label 2042 to RPE-B for region C.
20. PE4 sends out label 102 to PE1.
21. RPE-B sends out label 2301 to PE3 for region A.
22. RPE-B sends out label 2302 to PE3 for region C.

Assume that a CE with MAC address M1 is connected to PE1 and wishes to send data to a CE with MAC address M2 that is connected to PE3. The data flows as follows:

1. PE1 looks up M2 in its VSI, and doesn't find a match.
2. PE1 floods the packet with label 101 to PE2.
3. PE1 floods the packet to the other regions through RPE-A by sending two copies of the packet, with labels 1101 and 1102.
4. RPE-A learns M1 is at PE1.
5. RPE-A has a PW cross-connect to send packets labeled with 1101 to RPE-B with label 2001.
6. RPE-A has a PW cross-connect to send packets labeled with 1102 to RPE-C with label 3001.
7. RPE-B looks in its VSI for M2. Since it doesn't find it, it replicates the packet to PE3 with label 2031 since the packet came from region A.
8. RPE-B also replicates the packet to PE4 with label 2041.
9. PE3 learns that M1 is in region A.
10. PE3 sends the packet to M2.
11. Eventually, M2 responds, sending a packet back to M1 through P3.
12. PE3 knows that M1 is in region A, so it sends RPE-B the packet labeled with 2301.
13. RPE-B learns M2 is at PE3.
14. RPE-B has a PW cross-connect to send packets labeled with 2301 to RPE-A with label 1001.
15. RPE-A looks up M1 in its VSI, and knows that the packet belongs to PE1, and labels it with 1011 to inform PE1 that this packet originated in region B.
16. PE1 learns that M2 is in region B.
17. Learning is now complete, and unicast flows can now take place.
18. PE1 uses its VSI to figure that M2 is in region B, and sends packets to RPE-A using label 1101.
19. RPE-A cross-connects 1101 to RPE-B with label 2001.
20. RPE-B looks up M2 in its VSI and sends the packet to PE2.

3.1. How to interpret an R-VPLS

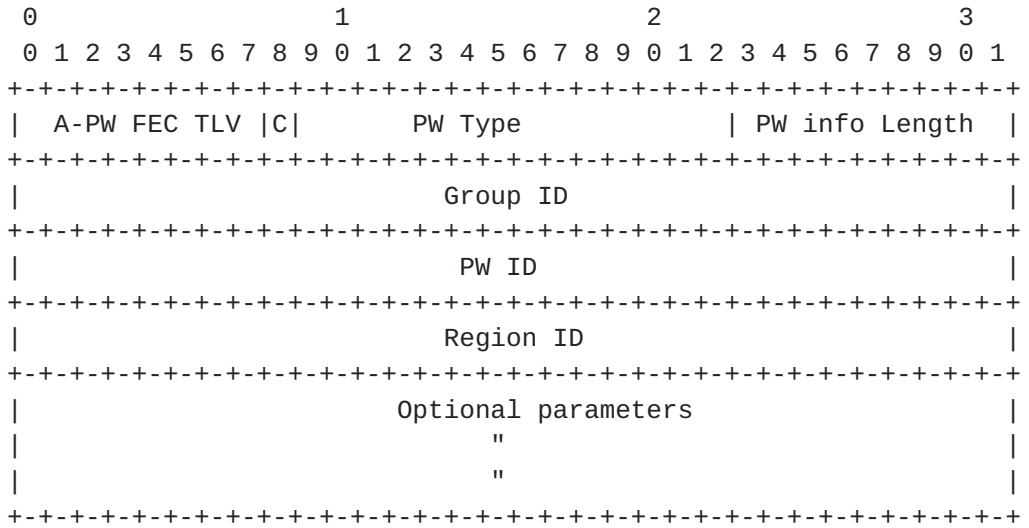
There are two aspects to the operation of an R-VPLS. One aspect is the forwarding: the PE effectively learns the destination region in its learning process. In that sense, the forwarding process of learning and the construction of the forwarding database are identical with a conventional VPLS.

At the RPE, when receiving a packet from the local region, the forwarding is modeled like a multi-segment PW to the remote RPE. The remote RPE uses its VSI to forward to its local PEs.

The second aspect is where the learning is done. The learning has to be done in the PEs. The RPEs also perform learning, but only when packets arrive from their local region, so that they only learn local MAC addresses, i.e., source MAC addresses originating within their region.

3.2. Protocol Format

The signaling between PE and RPE is a FEC with the conventional VPLS identification augmented with a region ID, which we call an Augmented PW FEC. The signaling between PEs in a region and between RPEs remains the conventional VPLS FEC.



Region ID.

The router ID or loopback address of the remote RPE. For signaling the F-label, the A-PW FEC is used, but the region ID is set to the local RPE's router ID.

3.3. Optimized Regional Flooding

It is possible to eliminate replication to multiple regions if we use a special region flooding label from the local PE to the local RPE. The local RPE signals a label per remote region, which is used for unicast forwarding. However, it signals a special F-label for use in optimized flooding, using the A-PW FEC, with its own router ID as the region ID.

3.4. Packet processing details

The following describes the processing of a packet in the forwarding plane at each service-aware node (PE and RPE).

3.4.1. Packet from customer to PE

When a PE receives a packet from the customer, it learns the source MAC address. Then depending on whether it knows the region of the destination MAC or not, it takes the following actions.

3.4.1.1. Destination MAC unknown

When a PE receives a packet from the customer and doesn't know the destination region, it replicates the packet with the region PW label for each region to the local RPE.

Alternatively, if the local PE has an F-label from its local RPE, it will send only one copy of the packet to the local RPE with the F-label.

Finally, the local PE will replicate packets to each PE in the region, using the PW label received from them.

3.4.1.2. Destination MAC known

When a PE receives a packet from the customer and has an entry in its VSI, it forwards the packet to the PW endpoint, with the appropriate PW label. This could be either to the local RPE, using the region label for the destination region that the MAC resides in, or to a local PE within the region, using its PW label.

3.4.2. Packet from PE to local RPE

When the local RPE receives a packet from the local PE, it learns the location of the source MAC against that PE. Then depending on the label received, it takes one of two actions.

If the received label was a remote region label, then the forwarding plane already has a PW cross-connect with the remote RPE and outgoing label. In case the PE is not using an F-label, and it needs to flood the packet, the RPE sees the flood as a set of unicasts, and no particular action has to be taken other than MS-PW style forwarding.

If the PE is flooding the packet using the F-label, then the RPE needs to replicate the incoming F-label to the appropriate label towards each remote RPE.

3.4.3. Packet from RPE to RPE

When a packet is received from another RPE, no MAC learning needs to be performed. Based on whether the RPE knows where the destination MAC or not, it takes the following action.

3.4.4. Destination MAC unknown

When a packet is received from a remote RPE, if the destination MAC is unknown, the packet is flooded to each PE with the appropriate region label that identifies which region the packet originated.

3.4.5. Destination MAC known

When a packet is received from a remote RPE, if the destination MAC is known, the packet is sent to the destination PE with the appropriate region label that identifies which region the packet originated.

3.4.6. Packet from RPE to PE

When a packet is received at a PE from its local RPE, the PE associates the source MAC with the region it originated from (which it can tell from the region label used). The packet is then forwarded according to whether the destination MAC address is known or not.

3.4.6.1. Destination MAC unknown

When a packet is received from the local RPE, if the destination MAC is unknown, the packet is flooded on all attachment circuits belonging to the VPLS.

3.4.6.2. Destination MAC known

When a packet is received from the local RPE, if the destination MAC is known, the packet is sent to the appropriate attachment circuit.

3.5. Improvements for later consideration

There are a number of questions that have already risen regarding. These will be dealt with either in this draft or in follow-on drafts:

- scalability comparisons between the solutions in Section 2
- dual homing/redundancy of RPEs
- cascading regions
- discovery of regions and RPEs

4. References

Normative References

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

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

Informative References

[LDP-Cap] "LDP Capabilities," R. Thomas et al, draft-ietf-mpls-ldp-capabilities-01.txt, work in progress, February 2008.

5. Security Considerations

No new security issues arise out of the extensions proposed here than exist in the base VPLS standards.

6. IANA Considerations

No IANA allocations have been specified yet (but a new FEC type will be forthcoming, as well as changes to the LDP Capability FEC TLV).

7. Authors' Addresses

Vach Kompella
Alcatel-Lucent
vach.kompella@alcatel-lucent.com

Joe Regan
Alcatel-Lucent
joe.regan@alcatel-lucent.com

Ron Haberman
Alcatel-Lucent
ron.haberman@alcatel-lucent.com

Shane Amante
Level 3 Communications
shane@castlepoint.net

8. Full Copyright Statement

9. Intellectual Property

BCP 79.

Copies of IPR disclosures made to the IETF Secretariat and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the IETF on-line IPR repository at http://www.ietf.org/ipr.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement this standard. Please address the information to the IETF at ietf-ipr@ietf.org.

10. Acknowledgments

Funding for the RFC Editor function is provided by the IETF Administrative Support Activity (IASA).