Internet Working Group
Internet Draft
Category: Standards Track

Ali Sajassi
Samer Salam
Keyur Patel
Pradosh Mohapatra
Clarence Filsfils
Sami Boutros
Cisco

Nabil Bitar
Verizon

Expires: January 7, 2011

July 7, 2010

Routed VPLS using BGP
draft-sajassi-l2vpn-rvpls-bgp-01.txt

Status of this Memo

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

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

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

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

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

This Internet-Draft will expire on December 7, 2010.

Copyright Notice

Abstract

VPLS, as currently defined, has challenges pertaining to the areas of redundancy and multicast optimization. In particular, multi-homing with all-active forwarding cannot be supported and there's no known solution to date for leveraging MP2MP MDTs for optimizing the delivery of multi-destination frames. This document defines an evolution of the current VPLS solution, referred to as Routed VPLS (R-VPLS), to address these shortcomings. In addition, this solution offers several benefits over current VPLS such as: ease of provisioning, per-flow load-balancing of traffic from/to multi-homed sites, optimum traffic forwarding to PEs with both single-homed and multi-homed sites, support for flexible multi-homing groups and fast convergence upon failures.

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 RFC 2119

Table of Contents

1. Introduction

VPLS, as defined in [RFC4664][RFC4761][RFC4762], is a proven and widely deployed technology. However, the existing solution has a number of challenges when it comes to redundancy and multicast optimization.

In the area of redundancy, 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 ACs cannot be supported without adding considerable complexity to the VPLS data-path.

In the area of multicast optimization, [VPLS-MCAST] describes how LSM MDTs can be used in conjunction with VPLS. However, this solution is limited to P2MP MDTs, as there's no known solution to date for leveraging MP2MP MDTs with VPLS. The lack of MP2MP support can create scalability issues for certain applications.

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 a new service interface type which is a hybrid combination of port-based and vlan-based service interfaces. This is referred to as 'VLAN-aware Port-Based' service interface.

This document defines an evolution of the current VPLS solution, to address the aforementioned shortcomings. The proposed solution is referred to as Routed VPLS (R-VPLS).

Section 2 provides a summary of the terminology used. Section 3 discusses the requirements for all-active resiliency and multicast optimization. Section 4 described the issues associated with the current VPLS solution in addressing the requirements. Section 5 offers an overview of R-VPLS and then Section 6 goes into the details of its components.

2. Terminology

CE: Customer Edge
DHD: Dual-homed Device
DHN: Dual-homed Network
LACP: Link Aggregation Control Protocol
LSM: Label Switched Multicast
MDT: Multicast Delivery Tree
MP2MP: Multipoint to Multipoint
P2MP: Point to Multipoint
P2P: Point to Point
PE: Provider Edge
PoA: Point of Attachment
PW: Pseudowire

3. Requirements

This section describes the requirements for all-active multi-homing, MP2MP MDT support, ease of provisioning and new service interface type.

3.1. All-Active Multi-homing

3.1.1. Flow-based Load Balancing

A customer network or a customer device can be multi-homed to a provider network using IEEE link aggregation standard -[802.1AX]. In [802.1AX], 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
i
i) Layer 3: Source IP Address, Destination IP Address
i
i
i) 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 VPLS dual-homing is the ability to accommodate flexible flow-based load-balancing from the CE node based on L2, L3 and/or L4 header fields.

3.1.2. Flow-based Multi-pathing

[PWE3-FAT-PW] defines a mechanism that allows PE nodes to exploit equal-cost multi-paths (ECMPs) in the MPLS core network by identifying traffic flows within a PW, and associating these flows with a Flow Label. The flows can be classified based on any arbitrary combination of L2, L3 and/or L4 headers. Any active/active VPLS dual-homing mechanism should seamlessly interoperate and leverage the mechanisms defined in [PWE3-FAT-PW].

3.1.3. Geo-redundant PE Nodes

The PE nodes offering dual-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 VPLS dual-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 dual-homing mechanism, is not appealing from cost standpoint. Furthermore, the IGP cost from remote PEs to the pair of PEs in the dual-homed setup cannot be assumed to be the same when those latter PEs are geo-redundant.

3.1.4.
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 VPLS multi-homing solution should support optimal forwarding of unicast traffic for all the following scenarios:

i) single-homed CE to single-homed CE
i
i) single-homed CE to dual-homed CE
i
i
i) dual-homed CE to single-homed CE
iv) dual-homed CE to dual-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.

3.1.5.
Flexible Redundancy Grouping Support

In order to simplify service provisioning and activation, the VPLS 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 dual-homed to any one of these three redundancy groups.

3.2.
Multi-homed Network

Supporting all-active multi-homing of an Ethernet network (a.k.a. Multi-homed Network or MHN) to several VPLS PEs poses a number of challenges.

First, some resiliency mechanism needs to be in place between the MHN and the PEs offering multi-homing, in order to prevent the

formation of L2 forwarding loops. Two options are possible here: either the PEs participate in the control plane protocol of the MHN (e.g. MST or ITU-T G.8032), or some auxiliary mechanism needs to run between the CE nodes and the PEs. The latter must be complemented with an interworking function, at the CE, between the auxiliary mechanism and the MHN's native control protocol. However, unless the PEs participate directly in the control protocol of the MHN, fast control-plane re-convergence and fault recovery cannot be guaranteed. Secondly, all existing Ethernet network resiliency mechanisms operate at best at the granularity of VLANs. Hence, any load-balancing would be limited to L2 flows at best if not at the VLAN granularity level. Depending on the applications at hand, this coarse flow granularity may not have enough entropy to provide proper link/node utilization distribution within the provider's network. Thirdly, an open issue remains with the handling of MHN partitioning: the PEs need to reliably detect the situation where the MHN has been partitioned and each PE needs to handle inbound/outbound traffic for only those customers (or hosts) connected to the local partition.

As described above, all-active load balancing for L3 and L4 flows is not feasible for MHNs. Although all-active load balancing for L2 flows is possible, it comes at the cost of requiring the locally attached PEs to perform local switching for a subset of the traffic within the MHN - e.g., using service provider resources to perform intra-site traffic forwarding and switching. Therefore, all-active load balancing for MHNs is not considered as a requirement; however, what is considered as a requirement for MHNs is for the PEs to auto detect the resiliency protocol used in a MHN and to auto-provision themselves to perform load balancing at the VLAN granularity without participating in the MHN's resiliency protocol.

3.3. Multicast Optimization with MP2MP MDT

In certain applications, multiple multicast sources may exist for a given VPLS instance, and these sources are dispersed over the various PEs. For these applications, relying on P2MP MDTs for VPLS can result in an increase in the number of states in the core relative to the use of MP2MP MDTs by a factor of O(n); where n is the average number of PEs per VPLS instance. In scenarios where the average number of PEs per VPLS instance is large, then the use of MDT rooted on every PE can result in two or more orders of magnitude more states in the core relative to the use of MP2MP MDTs. By using MP2MP MDTs, it is possible to scale multicast states in the core better by eliminating the above O(n) factor all together. Therefore, the scalability of multicast becomes no longer a function of the number of sites or number of PEs.

3.4. 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.

3.5. 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.

R-VPLS is required to support the above three service types plus one additional service type which is primarily intended for hosted data center applications and it is 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 a new service interface type, which will be referred to as the 'VLAN-aware Port-based' service interface. The characteristics of this service interface are as follows:

- The service interface must provide all-to-one 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).
- 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.

Since this is a port-based service, customer VLAN translation is not allowed over this service interface. If VLAN translation is required, then VLAN-based service MUST be used.

The main difference, in terms of service provider resource allocation, between this new service type and the previously defined three types is that the new service requires 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.

3.6. Fast Convergence

A key driver for multi-homing is providing protection against node as well as link and port failures. The R-VPLS solution should ensure fast convergence upon the occurrence of these failures, in order to minimize the disruption of traffic flowing from/to a multi-homed site. Here, two cases need to be distinguished depending on whether a device or a network is being multi-homed. This is primarily because a different set of convergence time characteristics can be guaranteed by the core network operator in each case.

For the case of a multi-homed device with all-active forwarding, the convergence of site-to-core traffic upon attachment circuit or PE node failure is a function of how quickly the CE node can

redistribute the traffic flows over the surviving member links of the multi-chassis Ethernet link aggregation group. For managed services, where the CE is owned by the Service Provider, the latter

can offer convergence time guarantees for such failures. Whereas, for non-managed services the SP has no control over the CE's capabilities and cannot provide any guarantees. For multi-homed device with all-active forwarding, the convergence of core-to-site traffic is a function of how quickly the protocol running between the PEs can detect and react to the topology change. The key requirement here is to have the convergence time be independent (to the extent possible) of the number of MAC addresses affected by the topology change, and the number of service instances emanating from the affected site. Given that all this is under the control of the core network operator, strict convergence time guarantees can be delivered by the operator.

For the case of a multi-homed network, the convergence time of site-to-core traffic upon attachment circuit or PE node failures is a function of two components: first, how quickly the MHN's control protocol detects and reacts to the topology change (this may involve blocking/unblocking VLANs on ports as well as propagating MAC address flush indications); and second, the reaction time of the locally attached PE(s) in order to update their forwarding state as necessary. The first component is outside the control of the core network operators, therefore it is not possible for them to make any convergence time guarantees except under tightly controlled conditions. For a multi-homed network, the convergence time of core-to-site traffic upon failures is a function of the inter-PE protocol if the PEs don't participate in the MHN control protocol. Otherwise, the convergence time is a function of both the inter-PE protocol in addition to the MHN's control protocol convergence time. In the latter scenario, again no guarantees can be made by the core operator as far as the convergence time is concerned except under tightly controlled conditions.

3.7. 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.

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.

4. VPLS Issues

This section describes issues associated with the current VPLS

solution in meeting the above requirements. The current solution for VPLS, as defined in [RFC4761]and [RFC4762], relies on establishing a full-mesh of pseudowires among participating PEs, and data-plane learning for the purpose of building the MAC forwarding tables. This

learning is performed on traffic received over both the attachment circuits as well as the pseudowires.
Supporting an all-active multi-homing solution with current VPLS is subject to three fundamental problems: the formation of forwarding loops, duplicate delivery of flooded frames and MAC Forwarding Table instability. These problems will be described next in the context of the example network shown in figure 1 below.

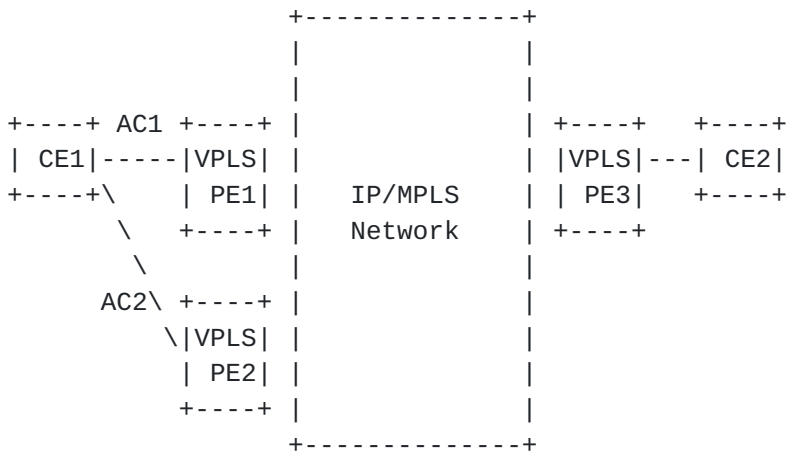


Figure 1: VPLS Multi-homed Network

In the network of Figure 1, it is assumed that CE1 has both attachment circuits AC1 & AC2 active towards PE1 and PE2, respectively. This can be achieved, for example, by running a multi-chassis Ethernet link aggregation group from CE1 to the pair of PEs.

4.1. Forwarding Loops

Consider the case where CE1 sends a unicast frame over AC1, destined to CE2. If PE1 doesn't have a forwarding entry in its MAC address table for CE2, it will flood the frame to all other PEs in the VPLS instance (namely PE3 & PE2) using either ingress replication over the full-mesh of pseudowires, or alternatively over an LSM tree [VPLS-MCAST]. When PE2 receives the flooded traffic, and assuming it doesn't know the destination port to CE2, it will flood the traffic over the ACs for the VFI in question, including AC2. Hence, a forwarding loop is created where CE1 receives its own traffic.

4.2. Duplicate Frame Delivery

Examine the scenario where CE2 sends a multi-destination frame (unknown unicast, broadcast or multicast) to PE3. PE3 will then flood the frame to both PE1 & PE2, using either ingress replication over the pseudowire full-mesh or an LSM tree. Both PE1 and PE2 will

receive copies of the frame, and both will forward the traffic on to CE1. Net result is that CE1 receives duplicate frames.

4.3. MAC Forwarding Table Instability

Assume that both PE1 and PE2 have learnt that CE2 is reachable via PE3. Now, CE1 starts sending unicast traffic to CE2. Given that CE1 has its ACs configured in an Ethernet link aggregation group, it will forward traffic over both ACs using some load-balancing technique as described in section 3.1 above. Both PE1 and PE2 will forward frames from CE1 to PE3. Consequently, PE3 will see the same MAC address for CE1 constantly moving between its pseudowire to PE1 and its pseudowire to PE2. The MAC table entry for CE1 will keep flip-flopping indefinitely depending on traffic patterns. This MAC table instability on PE3 may lead to frame mis-ordering for traffic going from CE2 back to CE1.

Shifting focus towards the requirement to support MP2MP MDT, the problem facing VPLS here is performing MAC learning over MP2MP MDT, as discussed next.

4.4. Identifying Source PE in MP2MP MDT

In the solution described in [VPLS-MCAST], a PE must perform MAC learning on traffic received over an LSM MDT. To that end, the receiving PE must be able to identify the source PE transmitting the frame, in order to associate the MAC address with the p2p pseudowire leading back to the source. With P2MP MDT, the MDT label uniquely identifies the source PE. For inclusive trees, the MDT label also identifies the VFI; whereas, for aggregate inclusive trees, a second upstream-assigned label identifies the VFI.

However, when it comes to MP2MP MDT, the MDT label identifies the root of the tree (which most likely is not the source PE), and the second label (if present) identifies the VFI. There is no known solution to date for dynamic label allocation among the VPLS PEs to identify the source PE since neither upstream nor downstream label assignment can work among the VPLS PEs.

From the above, it should be clear that with the current VPLS solution it is not possible to support all-active multi-homing or MP2MP MDTs. In the sections that follow, we will explore a new solution that meets the requirements identified in section 3 and addresses the problems highlighted in this section.

5. Solution Overview: Routed VPLS (R-VPLS)

R-VPLS follows a conceptually simple model where customer MAC

addresses are treated as routable addresses over the MPLS core, and distributed using BGP. In a sense, the R-VPLS solution represents an evolution of VPLS where data-plane learning over pseudowires is

replaced with control-plane based MAC distribution and learning over the MPLS core.

MAC addresses are learnt in the data-plane over the access attachment circuits (ACs) using native Ethernet bridging capabilities as is the case in current VPLS. MAC addresses learnt by a PE over its ACs are advertised in BGP along with a downstream-assigned MPLS label identifying the bridge-domain (this is analogous to L3VPNs where the label identifies the VRF). The BGP route is advertised to all other PEs in the same service instance. Remote PEs receiving these BGP NLRIs install the advertised MAC addresses in their forwarding tables with the associated MPLS/IP adjacency information. When multiple PE nodes advertise the same MAC address with the same BGP Local Preference, then the receiving PEs create multiple adjacencies for the same MAC address. This allows for load-balancing of Ethernet traffic among multiple disposition PEs when the AC is part of a multi-chassis Link Aggregation Group. The imposition PE can select one of the available adjacencies for forwarding traffic based on any hashing of Layer 2, 3 or 4 fields. Multicast and broadcast traffic can be forwarded using ingress replication per current VPLS, or over a P2MP LSM tree leveraging the model described in [VPLS-MCAST] or using a MP2MP MDT. The latter is possible since no MAC address learning is performed for traffic received from the core. Forwarding of unknown unicast traffic over the MPLS/IP core is optional and if the default mode is set to not forward it, it is still flooded over the local ACs per normal bridging operations.

Auto-discovery in R-VPLS involves identifying the set of PEs belonging to a given service instance and also discovering the set of PEs that are connected to the same multi-homed site. After auto-discovery is complete, an inclusive MP2MP MDT is set up per [MPLS-MDT]. Optionally, a set of P2MP MDTs per [VPLS-MCAST] can be set up or if ingress replication is required, a set of MP2P tunnels can be used. The purpose of the MP2MP MDT or the set of P2MP MDTs, or the set of MP2P tunnels, is for transporting customer multicast/ broadcast frames and optionally for customer unknown unicast frames. No MAC address learning is needed for frames received over the MDT(s)or the MP2P tunnels.

The mapping of customer Ethernet frames to a service instance for qualified learning and unqualified learning, is performed as in VPLS. Furthermore, the setup of any additional MDT per user multicast group or groups is also performed per [VPLS-MCAST].

Figure 2 below shows the model of a PE participating in R-VPLS. The modules in this figure will be used to explain the components of R-VPLS.

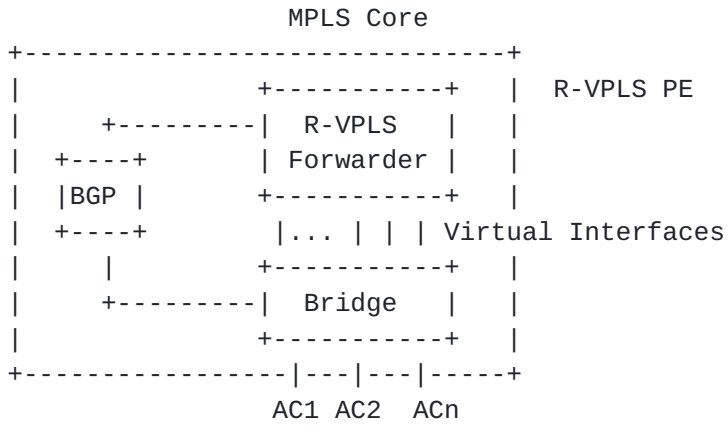


Figure 2: R-VPLS PE Model

5.1.
MAC Learning & Forwarding in Bridge Module

The Bridge module within an R-VPLS PE performs basic bridging operations as before and is responsible for:

i) Learning the source MAC address on all frames received over the ACs, and dynamically building the bridge forwarding database.

ii) Forwarding known unicast frames to local ACs for local destinations or the Virtual interface(s) for remote destinations.

iii) Flooding unknown unicast frames over the local ACs and optionally over the Virtual interface(s).

iv) Flooding multicast/broadcast frames to the local ACs and to the Virtual interface(s).

v) Informing the BGP module of all MAC addresses learnt over the local ACs. Also informing the BGP module when a MAC entry ages out, or is flushed due to a topology change.

vi) Enforcing the filtering rules described in section 7.3.

5.2.
MAC Address Distribution in BGP

The BGP module within an R-VPLS PE is responsible for two main functions:

First, advertising all MAC addresses learnt over the local ACs (by

the Bridge module) to all remote PEs participating in the R-VPLS instance in question. This is done using a new BGP NLRI as defined in the next section. The BGP module should withdraw the advertised

NLRIs for MAC addresses as they age out, or when the bridge table is flushed due to a topology change. Since no MAC address learning is performed for traffic received from the MPLS core, these BGP NLRI advertisements are used to build the forwarding entries for remote MAC addresses reachable over the MPLS network.

This brings the discussion to the second function of the BGP module, namely: programming entries in the forwarding table (in the R-VPLS Forwarder module) using the information in the received BGP NLRIs. These entries will be used for forwarding traffic over the MPLS core to remotely reachable MAC addresses. Of course, the BGP module must remove the forwarding entries corresponding to withdrawn NLRIs. Note that these entries are not subject to timed aging (as they follow a control-plane learning paradigm rather than data-plane learning).

6. BGP Encoding

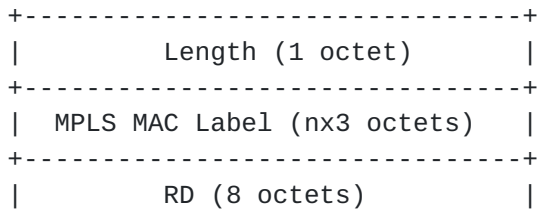
This section describes the new BGP Routes and Attributes that are required for R-VPLS. Three new BGP Routes (NLRIs) are defined below for the R-VPLS solution. All these R-VPLS NLRIs are carried in BGP using BGP Multiprotocol Extensions [RFC4760] with the existing L2VPN AFI but with different new SAFIs.

In order for two BGP speakers to exchange these NLRIs, they must use BGP Capabilities Advertisement to ensure that they both are capable of properly processing such NLRIs. This is done as specified in [RFC4760], by using capability code 1 (multiprotocol BGP) with an AFI of L2VPN and the corresponding SAFI for that NLRI.

6.1. R-VPLS MAC NLRI

This Layer-2 BGP route is used for distribution of MAC addresses over MPLS/IP network and has dual purposes:

1. For auto-discovery of member PEs in a given R-VPLS instance for the purpose of setting up an MP2MP MDT, a set of P2MP MDTs, or a set of MP2P tunnels among these PEs
2. For distribution of host MAC addresses to other remote PEs in a given R-VPLS instance




VLAN (2 octets)

```
        |    MAC address (6 octets)      |
        +--------------------------------+
```

Figure 1: R-VPLS MAC NLRI Format

Length: This field indicates the length in octets for this NLRI.

MPLS Label: This is a downstream assigned MPLS label that typically identifies the R-VPLS instance on the downstream PE (this label can be considered analogous to L3VPN label associated with a given VRF). The downstream PE may assign more than one label per RFC 3107. If this label is NULL, it means the VPN label (for this R-VPLS instance) that was previously advertised as part of auto-discovery MUST be used. If this label is not NULL, then it MUST be used by the remote PEs for traffic forwarding destined to the associated MAC address.

RD: This field is encoded as described in [RFC4364]. The RD MUST be the RD of the R-VPLS instance that is advertising this NLRI.

VLAN: This field may be zero or may represent a valid VLAN ID associated with the host MAC. If it is zero, then it means that there is only one bridge domain per R-VPLS instance (the most typical case) and the forwarding lookup on the egress PE should be performed based on bridge-domain ID (derived from R-VPLS instance) and MAC address. If this field is non-zero, then it means that there can be multiple bridge domains per R-VPLS instance (for the new VLAN-aware port-based service) and the forwarding lookup on the egress PE should be performed based on bridge-domain ID (derived from <R-VPLS instance, VLAN-ID>) and MAC address.

MAC: This MAC address can be either unicast or broadcast MAC address. If it is an unicast address, then it represents a host MAC address being distributed for the purpose of control plane learning via BGP. However, if it is a broadcast address, then it is used during auto-discovery phase of R-VPLS instance so that an inclusive MDT or a set of MP2P tunnels can be setup among participant PEs for that R-VPLS instance.

A new SAFI known as R-VPLS-MAC SAFI pending IANA assignment will be used for this NLRI. The NLRI field in the MP_REACH_NLRI/ MP_UNREACH_NLRI attribute contains the R-VPLS MAC NLRI encoded as specified in the above.

6.2. R-VPLS RG NLRI

This Layer-2 BGP route is used for distribution of a common site ID among member PEs of a redundancy group. For MHD scenarios, this route is used for auto-discovery of member PEs connected to an MHD and Designated Forwarder (DF) election among these PEs.

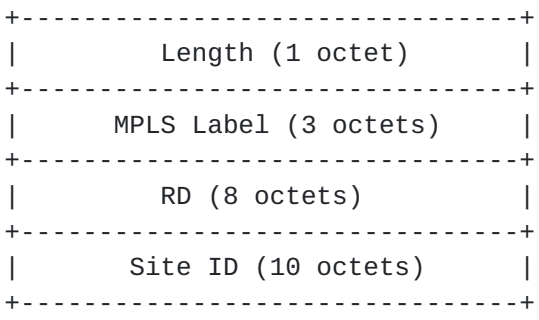


Figure 2: R-VPLS RG NLRI Format

Length: This field indicates the length in octets for this NLRI.

MPLS Label: This label basically identifies the site of origin and it is used for filtering purposes on egress PEs so that multi-destination frames that are sourced by a site are not sent back to the same site. This filtering action is commonly referred to as split-horizon. When multi-destination frames are sent using P2MP MDT, then this label is upstream assigned. When multi-destination frames are sent using ingress replication over a set of MP2P tunnels, then this label is downstream assigned. When multi-destination frames are sent using MP2MP tunnel, then this label needs to be scoped uniquely within the MP2MP tunnel context.

RD: This field is encoded as described in [RFC4364]. The RD MUST be the RD of the R-VPLS instance that is advertising this NLRI.

Site ID: This field uniquely represent a multi-homed site or a device connected to a set of PEs. In case of MHD scenarios, this ID consists of the CE's LAG system ID (MAC address), the CE's LAG system priority, and the CE's LAG Aggregator Key.

A new SAFI known as R-VPLS-RG SAFI pending IANA assignment will be used for this NLRI. The NLRI field in the MP_REACH_NLRI/ MP_UNREACH_NLRI attribute contains the R-VPLS RG NLRI encoded as specified in the above.

6.3. R-VPLS MH-ID NLRI

This Layer-2 BGP route is used for distribution of a site ID to the remote PEs that have VPNs participating in that site. This route is primarily used by remote PEs for the creation of the path list for a

given site and load balancing of traffic destined to that site among its member PEs.

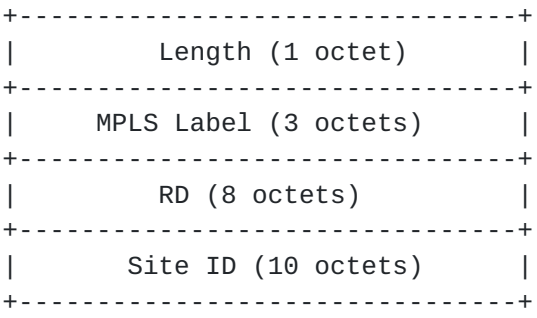


Figure 3: R-VPLS MH-ID NLRI Format

MPLS Label: This is a downstream assigned label that identifies the Site ID (and subsequently the AC) on the disposition PE. This label is used for forwarding of known unicast L2 frames in the disposition PE when MPLS forwarding is used in lieu of MAC lookup. When MAC lookup is used, this label MUST be set to NULL.

Length: This field indicates the length in octets for this NLRI.

RD: This field is encoded as described in [RFC4364]. The RD MUST be the RD of the R-VPLS instance that is advertising this NLRI.

Site ID: This field uniquely represent a multi-homed site or a device connected to a set of PEs. In case of MHD scenarios, this ID consists of the CE's LAG system ID (MAC address), the CE's LAG system priority, and the CE's LAG Aggregator Key.

A new SAFI known as R-VPLS-MH-ID SAFI pending IANA assignment will be used for this NLRI. The NLRI field in the MP_REACH_NLRI/ MP_UNREACH_NLRI attribute contains the R-VPLS MH-ID NLRI encoded as specified in the above.

6.4. BGP Route Targets

Each BGP R-VPLS NLRI will have one or more route-target extended communities to associate a R-VPLS NLRI with a given VSI. These route-targets control distribution of the R-VPLS NLRIs and thereby will control the formation of the overlay topology of the network that constitutes a particular VPN. This document defines the following route-targets for R-VPLS:

6.4.1. VPN-RT

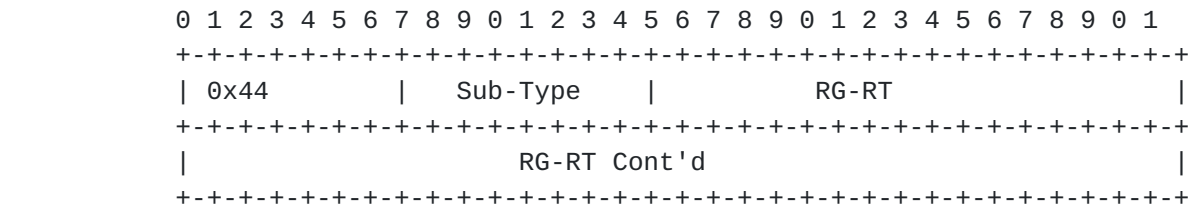
This RT includes all the PEs in a given R-VPLS service instance. It is used to distribute R-VPLS MAC NLRIs and it is analogous to RT used for VPLS instance in [RFC 4671] or [RFC 4672].

In data center applications where the network is limited to supporting only 4K VLANs, then this VPN-RT can be derived automatically from the VLAN itself (e.g., the VLAN ID can be used as the VPN ID). Such RT auto-derivation is applicable to both Port mode and VLAN mode services. In case of Port mode service, the default VLAN for the port is used to derive the RT automatically and in case of the VLAN mode service, the S-VLAN (service VLAN) is used to derive the RT automatically.

6.4.2. RG-RT

This RT is a transitive RT extended community and it includes all the PEs in a given Redundancy Group, i.e. connected to the same multi-homed site. It is used to distribute R-VPLS RG NLRIs. This RT is derived automatically from the Site ID by encoding the 6-byte system MAC address of the Site ID 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 service provider network (e.g., the CE is a managed CE or the customer doesn't fiddle with the CE's system MAC address).

Each RG specific RT extended community is encoded as a 8-octet value as follows:



6.4.3. MH-RT

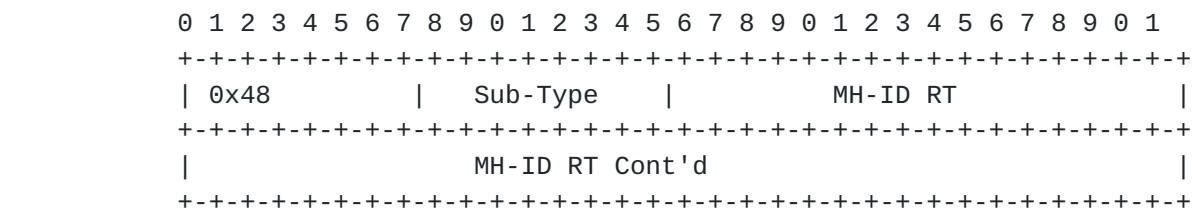
This RT is a transitive RT extended community and it includes all the PEs whose R-VPLS service instances are part of the same multi-homed site. It is used to distribute R-VPLS MH-ID NLRIs. This RT is derived automatically from the MH-ID by encoding the 6-byte system MAC address of the MH-ID in this RT. For a given multi-homed site, this RT and RG-RT correspond to the same Site ID; however, the reason for having two different RTs is to have exact filtering and to differentiate between filtering needed among member PEs of a multi-homed site versus among member PEs of all R-VPLS instances

participating in a multi-homed site. The former is needed for DF election in a multi-homed site; whereas, the latter is needed for

load-balancing of the unicast traffic by the remote PEs toward the multi-homed site.

In order to derive this RT automatically, it is assumed that the system MAC address of the CE is unique in the service provider network.

Each MH-ID specific RT extended community is encoded as a 8-octet value as follows:



7. Operation

This section describes the detailed operation of R-VPLS.

7.1. Auto-discovery

The set of PEs participating in a given R-VPLS instance need to discover each other for the purpose of setting up the tree(s) or the tunnels which will be used for the delivery of multi-destination frames. To that end, every PE advertises, on a per R-VPLS instance basis, an R-VPLS MAC NLRI as follows:

- MAC Address field is set to the broadcast MAC (FFFF.FFFF.FFFF)
- VLAN ID field is set to zero
- RD is set as described previously
- MPLS Label field is set to a downstream-assigned label which uniquely identifies the R-VPLS service instance on the originating PE. This will be referred to as the VPN label.

The above NLRI is advertised along with the RT Extended Community attribute corresponding to the R-VPLS service instance and the PMSI Tunnel attribute per [MCAST-BGP]. The default operation of R-VPLS is to use a unique MP2MP MDT per service instance. Therefore, in the PMSI Tunnel attribute, the Tunnel Type field is set to "mLDP MP2MP LSP" (value 7) and the MPLS Label field is set to zero. If there is a need to multiplex more than one R-VPLS instance over the same MDT, then a non-zero label value can be used in the PMSI Tunnel attribute.

Optionally, the network operator may choose to use P2MP MDTs instead. If so, then the Tunnel Type field is set to "mLDP P2MP LSP" (value 2) and the MPLS label field is set as described above.

If the MPLS network does not support LSM, then ingress replication is used instead. In this case, the PMSI Tunnel attribute would have the Tunnel Type field set to "Ingress Replication" (value 6) and the MPLS Label field is set to the same value as the MPLS Label field in the associated R-VPLS MAC NLRI.

7.2. Setup of Multicast Tunnels

In order to automate the setup of the default MP2MP MDT, the following procedure is to be followed: The first PE to come up in an R-VPLS instance advertises an R-VPLS MAC NLRI (as described in section 7.1) with the Tunnel-id field of the PMSI Tunnel attribute set to NULL. The BGP Route Reflector chooses a root (based on some policy) and re-advertises the NLRI with the PMSI Tunnel attribute modified to include the selected Tunnel-id. This advertisement is then sent to all PEs in the R-VPLS instance. To ensure that the original advertising PE receives the assigned Tunnel-ID, BGP Route Reflector shall modify its route advertisement procedure such that the Originator attribute shall be set to the router-id of the Route Reflector and the Next-hop attribute shall be set to the local address of the BGP session for such R-VPLS MAC NLRI announcements. Upon receiving the NLRIs with non-NULL Tunnel-id, the PEs initiate the setup of the MP2MP tunnel towards the root using the procedures in [MLDP].

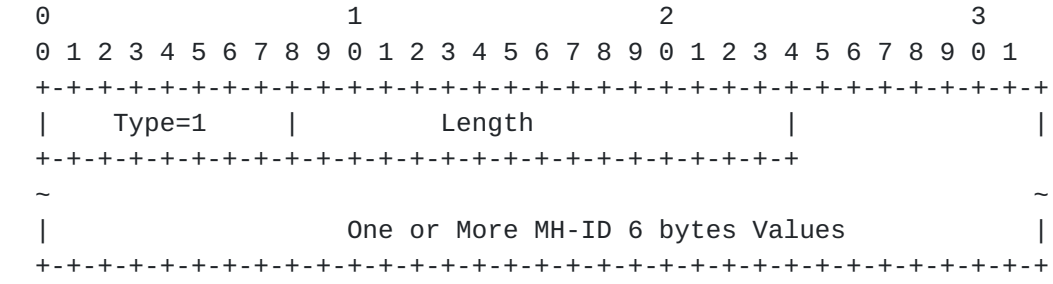
If the PEs are configured to use the optional P2MP MDT instead of MP2MP MDT, then the PE itself sets the Tunnel-id field in the PMSI Tunnel attribute associated with the R-VPLS MAC NLRI described in section 7.1.

7.3. Host MACs Distribution over Core

Upon learning a host MAC in its bridge module, the PE advertises the newly learned MAC over MPLS core to other remote PEs using the MAC NLRI. If the MAC address is originated from a multi-homed site, then the MPLS label field in the MAC NLRI is set to NULL because the remote PEs know that they MUST use the MPLS label associated with the broadcast MAC, which is advertised during auto-discovery phase, as the VPN label. Furthermore, the MH-ID is set as part of a separate new MH-ID attribute for this MAC NLRI to indicate that this MAC is associated with that site ID. However, if the MAC address is originated from a single-homed site, then the MPLS label field in the MAC NLRI is set to the downstream assigned label representing the R-VPLS instance and the MH-ID is not set for that MAC NLRI

(indicating to the remote PEs that this MAC is associated with this advertising PE only).

The MH-ID attribute is a new optionally transitive attribute of type [TBD] and is defined as:



7.4. Device Multi-homing

7.4.1. Special Considerations for Multi-homing

In the case where a set of VPLS PEs offer flexible multi-homing for a number of CEs, special considerations are required to prevent the creation of forwarding loops and delivery of duplicate frames when forwarding multi-destination frames.

Consider the example network shown in figure 3 below. In this network, it is assumed that the ACs from all CEs to their corresponding PEs are active and forwarding, i.e. all-active redundancy model.

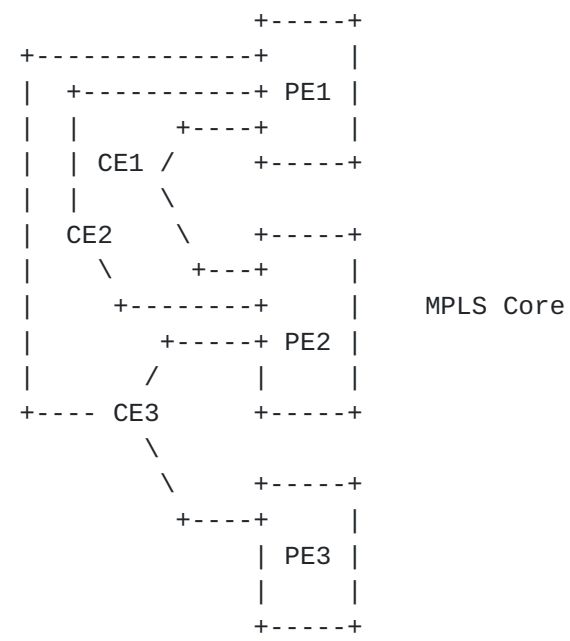


Figure 3: VPLS with Flexible Multi-homing

Take, for instance, the scenario where CE1 transmits a broadcast frame toward PE1. PE1 will attempt to flood the frame over all its local ACs and to all remote PEs (PE2 and PE3) in the same VPLS

instance. The R-VPLS solution ensures that these broadcast frames do not loop back to CE1 by way of PE2. The solution also ensures that CE2 and CE3 do not receive duplicates of the broadcast, via PE1/PE2 and PE2/PE3, respectively. This is achieved by enforcing the following behavior:

7.4.1.1. Filtering Based on Site ID

Every R-VPLS PE is configured with a Site ID on the AC connecting to a multi-homed CE per [VPLS-BGP-DH]. The PE forwarding a multi-destination frame tags the flooded traffic with the Site ID that identifies the originating site, so that traffic from a multi-homed CE is not re-forwarded back to that CE upon receipt from the MPLS core. This filtering action is commonly referred to as split-horizon. This tagging can be achieved by embedding a 'source label' as the end-of-stack label in the MPLS packets. The source label is set to the value of the MPLS label field in the RG NLRI for that site. This source label is matched against the Site-ID label of a given AC, for traffic received from the MPLS core. If there is a match, then traffic is filtered on that AC. If there's no match, then the traffic is allowed to egress that AC, as long as that AC is the Designated Forwarder for that site.

7.4.1.2. Defining a Designated Forwarder

A Designated Forwarder (DF) PE is elected for handling all multi-destination frames received from the MPLS core towards a given multi-homed device. Only the DF PE is allowed to forward traffic received from the MPLS core (over the multipoint LSP or full-mesh of PWs) towards a given MHD. The DF is elected dynamically using the procedures in [VPLS-BGP-DH]. There can be transient duplicate frames and loops. The DF election procedure to avoid transient duplicate frames and loops will be described in the future revision.

7.4.2. Multi-homed Site Topology Discovery

Given that one of the requirements of R-VPLS is ease of provisioning, the set of PEs connected to the same CE must discover each other automatically with minimal to no configuration. To that end, each PE extracts the following from the [802.1AX] LACPDUs transmitted by the CE on a given port:

- CE LACP System Identifier comprised of 6 bytes MAC Address and 2 bytes System Priority
- CE LACP Port Key (2 bytes)

The PE uses this information to construct the Site ID associated with the port, and advertises an R-VPLS RG NLRI for every unique Site ID. The NLRI is tagged with the RG-RT extended community discussed in section 6.4.2 above. Furthermore, the PE automatically

enables the import of BGP routes tagged with said RT which is derived from the Site ID. This allows the PEs connected to the same CE to discover each other.

As a PE discovers the other members of the RG, it starts building an ordered list based on PE identifiers (e.g. IP addresses). This list is used to select a DF and a backup DF (BDF) on a per group of VLAN basis. For example, the PE with the numerically highest (or lowest) identifier is considered the DF for a given group of VLANs for that site and the next PE in the list is considered the BDF. To that end, the range of VLANs associated with the CE must be partitioned into disjoint sets. The size of each set is a function of the total number of CE VLANs and the total number of PEs in the RG. The DF can employ any distribution function that achieves an even distribution of VLANs across RG members. The BDF takes over the VLAN set of any PE encountering either a node failure or a link/port failure causing that PE to be isolated from the multi-homed site.

It should be noted that once all the PEs participating in a site have the same ordered list for that site, then VLAN groups can be assigned to each member of that list deterministically without any need to explicitly distribute VLAN IDs among the member PEs of that list. In other words, the DF election for a group of VLANs is a local matter and can be done deterministically. As an example, consider, that the ordered list consists of m PEs: (PE1, PE2,..., PEm), and there are n VLANs for that site (V0, V1, V2, ..., Vn-1). The PE1 and PE2 can be the DF and the BDF respectively for all the VLANs corresponding to (i mod m) for i:1 to n. PE2 and PE3 can be the DF and the BDF respectively for all the VLANs corresponding to (i mod m) + 1 and so on till the last PE in the order list is reached and we have PEm and PE1 is the DF and the BDF respectively for the all the VLANs corresponding to (i mod m) + m-1.

While the discovery of the multi-homed topology is in progress, different PEs may have inconsistent views of the network. This could lead to having duplicate packets temporarily delivered to the multi-homed CE. Procedures for preventing temporary packet duplication and/or loops will be covered in future revisions of this document.

7.4.3. Dynamic Assignment of Site-ID Label

In order to automate the assignment of the Site-ID label used as 'source label' for the Site-ID split-horizon filtering, the following procedure is to be followed: During the multi-homed site topology discovery, the first PE to come up in a multi-homed site advertises an RG NLRI (as described in section 6.2) with the MPLS Label field set to NULL. The BGP Route Reflector chooses a label and re-advertises the RG NLRI with the MPLS Label field modified to

include the selected value. This advertisement is then sent to all PEs in that multi-homed site. To ensure that the original advertising PE receives the assigned label, filtering based on the

node-origin on the Route Reflector is disabled. Upon receiving the RG NLRIs with non-NULL label, the PEs use that label as the source label for split-horizon filtering of that site.

It should be noted that this procedure for dynamic assignment of Site-ID label only assigns a single label per site (and not per site per PE) which simplifies the implementation of split-horizon filtering. Furthermore, it is independent from multi-destination tunnel type and can be equally applied across all different tunnel types: MP2MP MDT, P2MP MDT, and MP2P ingress replication tunnels.

7.4.4. Load-balancing

Consider the case where a given CE is multi homed to a set of PEs {PE1, PE2, ... PEn} over a multi-chassis LAG. For a specific MAC address M1, the CE may hash the active traffic flow to some PEi (1<=i<=n) in the set, and there could be a (possibly indefinite) lapse of time before any traffic from M1 is hashed to the other PEs in the set. In such a scenario, any remote PE in the same R-VPLS instance would have received an R-VPLS MAC NLRI for M1 only from PEi. However, it is desirable to be able to load-balance traffic from the remote PE (PEr) destined to M1 over the entire set of the multi-homed site PEs {PE1, PE2, ... PEn}. To facilitate that, R-VPLS makes use of site routes (MH-ID NLRIs) in addition to MAC routes (MAC NLRIs). All PEs that are connected to the same multi-homed CE advertise R-VPLS MH-ID NLRIs, with the CE's Site ID, to all PEs in the R-VPLS instances that said CE is part of. When any of the PEs in the RG learns a new MAC address for traffic coming from the CE, it advertises an R-VPLS MAC NLRI with the Next-Hop attribute set to the corresponding Site ID. The combination of the MAC route and site route advertisements allows all the remote PEs to build a BGP path-list comprising of the set of PEs that have reachability to a given MAC address via a given multi-homed CE. The remote PEs use the Site ID in the Next-Hop attribute of the MAC NLRI to determine the list of member PEs for that site. Furthermore, they retrieve the VPN label corresponding to the R-VPLS instance on a given PE from the previously advertised broadcast MAC NLRI as part of auto-discovery. From the combination of the two, the remote PEs can create a list of label tuples corresponding to the member PEs of that site for a give VPN: {(Lt1, Lv1), (Lt2, Lv2), ... (Ltm, Lvm)}; where Lti and Lvi represent the tunnel and the VPN labels respectively for PEi. The remote PEs can use this path-list to perform flow-based load-balancing for traffic destined to that given MAC address. This works even if only a single PE within the RG learns a given MAC address from the CE.

7.4.5. Auto-Derivation of MH-ID RTs

The MH-ID NLRIs corresponding to a given multi-homed CE need to reach any PE that participates in at least one of the R-VPLS instances that said CE is part of. Therefore, the choice of the RT Extended Community used to tag that NLRI must accommodate that. In order to avoid any manual configuration of this RT, referred to as MH-RT (section 6.4.3), the remote PEs need to automatically discover its value from at least one of the PEs in the RG. This is done as follows: Upon discovering all the connected CEs, a PE starts the service auto-discovery procedures outlined in section 7.1 above. In the MAC NLRI sent for discovery, the sending PE embeds the Site IDs of all CEs that are part of the associated service instance in the SNPA field of the Next-Hop attribute. When a remote PE receives the MAC NLRI, it first derives the MH-RT extended communities based on these Site IDs and then automatically starts importing MH-ID routes tagged with these MH-RTs extended community attributes.

7.4.6. Site-ID Label for Single-Homed Sites

For a single-homed site, we shouldn't need to assign a site-ID label; however, it makes the processing at the disposition PE simpler if the packet is encapsulated with a site-ID label with a NULL value. If a site-ID label is not used and the packet is sourced from a single-homed site and destined to a multi-homed site, then at the disposition PE, a NULL label needs to get injected into the packet for frames received over multicast MDT(s) so that the 'source label' check can be performed on the egress AC. Furthermore, if ingress replication is used and the use of flow label is optional, then it is difficult to identify the label that follows the VPN label - it is difficult to discern between a flow label and a 'source label'. Therefore, in order to avoid such complications on the disposition PE, we mandate the use of 'source label' with the value of NULL for packets originating from the single-homed sites.

7.4.7. LACP State Synchronization

To support CE multi-homing with multi-chassis Ethernet bundles, the R-VPLS PEs connected to a given CE should synchronize [802.1AX] LACP state amongst each other. 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.

The above information must be synchronized between the R-VPLS PEs wishing to form a multi-chassis bundle with a given CE, in order for the former to convey a single LACP peer to that CE. This is required for initial system bring-up and upon any configuration change. Furthermore, the PEs must 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 R-VPLS PEs 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 PEs within a given Redundancy Group, i.e. 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. [PWE3-ICCP] defines a mechanism for synchronizing LACP state, using LDP, which can be leveraged for R-VPLS. The use of BGP for synchronization of LACP state is left for further study.

7.5. Frame Forwarding over MPLS Core

The VPLS Forwarder module is responsible for handling frame transmission and reception over the MPLS core. The processing of the frame differs depending on whether the destination is a unicast or multicast/broadcast address. The two cases are discussed next.

7.5.1. Unicast

For known unicast traffic, the VPLS Forwarder sends frames into the MPLS core using the forwarding information received by BGP from remote PEs. The frames are tagged with an LSP tunnel label and a VPN

label. If per flow load-balancing over MPLS core is required between ingress and egress PEs, then a flow label is added after the VPN label.

For unknown unicast traffic, an R-VPLS PE can optionally forward these frames over MPLS core; however, the default is not to forward. If these frames are to be forwarded, then the same set of options used for forwarding multicast/broadcast frames (as described in next section) are also used for forwarding these unknown unicast frames.

7.5.2. Multicast/Broadcast

For multi-destination frames (multicast and broadcast) delivery, R-VPLS provides the flexibility of using a number of options:

Option 1: the R-VPLS Forwarder can perform ingress replication over a set of MP2P tunnel LSPs.

Option 2: the R-VPLS Forwarder can use P2MP MDT per the procedures defined in [VPLS-MCAST].

Option 3: the R-VPLS Forwarder can use MP2MP MDT per the procedures described in section 6.4. This option is considered as default mode.

7.6. MPLS Forwarding at Disposition PE

The general assumption for forwarding frames to customer sites at disposition PEs is that the received packet from MPLS core is terminated on the bridge module and a MAC lookup is performed to forward the frame to the right AC. This requires that the MPLS encapsulation to carry the VPN label which in turn identifies the right VSI for forwarding the frame.

It is sometimes desirable to be able to forward L2 frames to the right AC at the disposition PE without any MAC lookup (e.g., using only MPLS forwarding). In such scenarios, the MPLS encapsulation needs to carry a label associated with the egress AC. In vlan-mode service, this AC label needs to be in addition to the VPN label. Therefore, for consistency one may want to use both the AC and the VPN labels for all types of services when doing MPLS forwarding at the disposition PE. The VPN label is retrieved from the MAC route during auto-discovery phase and the Site label is retrieved from the MH-ID route. From these labels, the remote PEs can create a list of label tuples corresponding to the member PEs of that site for a give VPN: {(Lt1, Lv1, Ls1), (Lt2, Lv2, Ls2), ... (Ltm, Lvm, Lsm)}; where Lti, Lvi, and Lsi represent the tunnel, the VPN, and the AC labels respectively for PEi.

8. Acknowledgements

The authors would like to acknowledge the valuable inputs received from Pedro Marques and Robert Raszuk.

9. Security Considerations

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

10. IANA Considerations

This document requires IANA to assign a new SAFI value for L2VPN_MAC SAFI.

11. Intellectual Property Considerations

This document is being submitted for use in IETF standards discussions.

12. Normative References

[RFC4664] "Framework for Layer 2 Virtual Private Networks (L2VPNs)", RFC4664, 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", RFC4762, January 2007.

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

13. 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.

14. Authors' Addresses

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

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

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

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

Pradosh Mohapatra
Cisco
170 West Tasman Drive
San Jose, CA 95134, US
Email: pmohapat@cisco.com

Clarence Filsfils
Cisco
170 West Tasman Drive
San Jose, CA 95134, US
Email: cfilsfil@cisco.com

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