Network Working Group   R. Aggarwal (Editor)
Internet Draft   Juniper Networks
Category: Proposed Standard
Expiration Date: January 2014   Y. Kamite
                                             NTT Communications

                                                        L. Fang
                                             Cisco Systems, Inc

                                                    July 2 2013

                         Multicast in VPLS

                  draft-ietf-l2vpn-vpls-mcast-14.txt

Abstract

   [RFC4761] and [RFC4762] describe a solution for Virtual Private LAN
   Service (VPLS) multicast that relies on the use of point-to-point or
   multipoint-to-point unicast Label Switched Paths (LSPs) for carrying
   multicast traffic. This solution has certain limitations for certain
   VPLS multicast traffic profiles. For example, it may result in highly
   non-optimal bandwidth utilization when large amount of multicast
   traffic is to be transported.

   This document describes a solution for overcoming a subset of the
   limitations of existing VPLS multicast solution. It describes
   procedures for VPLS multicast that utilize multicast trees in the
   service provider (SP) network. The solution described in this
   document allows sharing of one such multicast tree among multiple
   VPLS instances. Furthermore, the solution described in this document
   allows a single multicast tree in the SP network to carry traffic
   belonging only to a specified set of one or more IP multicast streams
   from one or more VPLS instances.

Table of Contents

1. Specification of requirements

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

2. Contributors

Yakov Rekhter
Juniper Networks
Chaitanya Kodeboniya

3. **Terminology**

   This document uses terminology described in [RFC4761] and [RFC4762].

   In this document we refer to various auto-discovery routes, as "A-D
   routes".

   This document uses the prefix 'C' to refer to the customer control or
   data packets and 'P' to refer to the provider control or data
   packets. An IP (multicast source, multicast group) tuple is
   abbreviated to (S, G).

4. **Introduction**

   [RFC4761] and [RFC4762] describe a solution for VPLS
   multicast/broadcast that relies on the use of pseudowires transported
   over unicast point-to-point (P2P) RSVP-TE or multipoint-to-point
   (MP2P) LDP Label Switched Paths (LSPs) ([RFC3209], [RFC3036]). In
   this document we refer to this solution as "ingress replication".

   With ingress replication when an ingress PE of a given VPLS instance
   receives a multicast/broadcast packet from one of the CEs that belong
   to that instance, the ingress PE replicates the packet for each
   egress PE that belong to that instance, and sends the packet to each
   such egress PE using unicast tunnels.

   The solution based on ingress replication has certain limitations for
   certain VPLS multicast/broadcast traffic profiles. For example, it
   may result in highly non-optimal bandwidth utilization in the MPLS
   network when large amount of multicast/broadcast traffic is to be
   transported (for more see [RFC5501]).

   Ingress replication may be an acceptable model when the bandwidth of
   the multicast/broadcast traffic is low and/or there is a small number
   of replications performed on each outgoing interface for a particular
   VPLS customer multicast stream. If this is not the case, it is
   desirable to utilize multicast trees in the SP network to transmit
   VPLS multicast and/or broadcast packets [RFC5501].

   This document describes procedures for overcoming the limitations of
   existing VPLS multicast solutions. It describes procedures for using
   MPLS point-to-multipoint (P2MP) LSPs in the Service Provider (SP)
   network to transport VPLS multicast and/or broadcast packets, where
   these LSPs are signaled by either P2MP RSVP-PE [RFC4875] or mLDP
   [RFC6388].

   The procedures described in this document are applicable to both

[RFC4761] and [RFC4762].

5. Overview

Procedures described in this document provide mechanisms that allow a
single multicast distribution tree in the Service Provider (SP)
network to carry all the multicast traffic from one or more VPLS
sites connected to a given PE, irrespective of whether these sites
belong to the same or different VPLS instances. We refer to such a
tree as an "Inclusive tree" if it carries multicast traffic from one
VPLS instance on a given PE. We refer to such a tree as an "Aggregate
Inclusive tree" if it carries multicast traffic from more than one
VPLS instance on a given PE. See section "Inclusive and Selective
Multicast Trees" for further discussion on Inclusive trees.

To further improve bandwidth utilization for IP multicast streams,
this document also provides procedures by which a single multicast
distribution tree in the SP network can be used to carry traffic
belonging only to a specified set of IP multicast streams, originated
in one or more VPLS sites connected to a given PE, irrespective of
whether these sites belong to the same or different VPLS instances.
We refer to such a tree as a "Selective tree" if it carries the IP
multicast stream(s) that belong to the same VPLS instance on a given
PE. We refer to such a tree as as an "Aggregate Selective tree" if it
carries the IP multicast streams that belong to different VPLS
instances on a given PE. Use of Selective and/or Aggregate Selective
trees allows multicast traffic, by default, to be carried on an
Inclusive tree, while traffic from some specific IP multicast
streams, e.g., high bandwidth streams, could be carried on one of the
Selective trees. See section "Inclusive and Selective Multicast
Trees" for further discussion on Selective trees.

Note that this document covers the use of Selective trees only for
carrying IP multicast streams. Any other use of such trees is outside
the scope of this document.

Unicast packets destined to unknown MAC addresses (i.e., not learned
yet at the ingress PE) in a given VPLS instance are flooded to remote
PEs participating in the same VPLS instance. This flooding MAY still
use ingress replication (as specified in [RFC4761], [RFC4762]), or
MAY use the procedures defined in this document to optimize flooding
across the SP core.

While the use of multicast trees in the SP network can be beneficial
when the bandwidth of the multicast traffic is high, or when it is
desirable to optimize the number of copies of a multicast packet
transmitted on a given link, this benefit comes at a cost of state in

the SP network to build multicast trees and overhead to maintain this state.

5.1. Inclusive and Selective Multicast Trees

Multicast trees used for VPLS can be of two types:

+ Inclusive trees. This option supports the use of a single
  multicast distribution tree, referred to as an Inclusive P-
  Multicast tree, in the SP network to carry all the multicast
  traffic from a specified set of VPLS sites connected to a given
  PE. There is no assumption made with respect to whether this
  traffic is IP encapsulated or not. A particular P-Multicast tree
  can be set up to carry the traffic originated by sites belonging
  to a single VPLS instance, or to carry the traffic originated by
  sites belonging to different VPLS instances. The ability to
  carry the traffic of more than one VPLS instance on the same tree
  is termed Aggregation. The tree needs to include every PE that is
  a member of any of the VPLS instances that are using the tree.
  This implies that a PE may receive multicast traffic for a
  multicast stream even if it doesn't have any receivers that are
  interested in receiving traffic for that stream.

  An Inclusive P-Multicast tree as defined in this document is a
  P2MP tree. Thus, a P2MP tree is used to carry traffic only from
  VPLS sites that are connected to the PE that is the root of the
  tree.

+ Selective trees. A Selective P-Multicast tree is used by a PE to
  send IP multicast traffic for one or more specific IP multicast
  streams, received by the PE over PE-CE interfaces that belong to
  the same or different VPLS instances, to a subset of the PEs that
  belong to those VPLS instances. Each of the PEs in the subset
  should be on the path to a receiver of one or more multicast
  streams that are mapped onto the tree. The ability to use the
  same tree for multicast streams that belong to different VPLS
  instances is termed Aggregation. The reason for having Selective
  P-Multicast trees is to provide a PE the ability to create
  separate SP multicast trees for specific multicast streams, e.g.
  high bandwidth multicast streams. This allows traffic for these
  multicast streams to reach only those PE routers that have
  receivers for these streams. This avoids flooding other PE
  routers in the VPLS instance.

A SP can use both Inclusive P-Multicast trees and Selective P-
Multicast trees or either of them for a given VPLS on a PE, based on
local configuration. Inclusive P-Multicast trees can be used for both
IP and non-IP data multicast traffic, while Selective P-Multicast
trees, as previously state, must be used only for IP multicast data
traffic. The use of Selective P-Multicast trees for non-IP multicast
traffic is outside the scope of this document.

P-Multicast trees in the SP network can be realized via a variety of
technologies. For both inclusive and selective P-Multicast trees,
these technologies include P2MP LSPs created by RSVP-TE or mLDP.
This document also describes the data plane encapsulations for
supporting these technologies. Other technologies for realizing P-
Multicast trees are outside the scope of this document.

5.2. BGP-Based VPLS Membership Auto-Discovery

In order to establish Inclusive P-Multicast trees for one or more
VPLS instances, when aggregation is performed (with either mLDP or
P2MP RSVP-TE as the tunneling technology), or when the tunneling
technology is P2MP RSVP-TE, the PE acting as a root of a P2MP LSP
must be able to discover the other PEs that have membership these
VPLS instances. Once the root PE discovers these other PEs, it
includes them as leaves in the P-multicast tree (i.e., P2MP LSP) This
document uses the BGP-based procedures described in [RFC4761] and
[RFC6074] for discovering the VPLS membership of all PEs. For more on
aggregation see section "Aggregation Considerations". When no
aggregation is performed and the tunneling technology is mLDP, then
the root of the P2MP LSP need not discover the other PEs that are the
leaves of that LSP.

The leaves of the Inclusive P-Multicast tree must also be able to
auto-discover the identifier of the tree (note that this applies when
the tree is established by either mLDP or P2MP RSVP-TE). Procedures
to accomplish this are described in section "Advertising P-Multicast
Tree to VPLS/C-Multicast Binding".

5.3. IP Multicast Group Membership Discovery

The setup of a Selective P-Multicast tree for one or more IP
multicast (C-S, C-G)s, requires the ingress PE to learn the PEs that
have receivers in one or more of these (C-S, C-G)s, in the following
cases:

+ When aggregation is used (with either mLDP or P2MP RSVP-TE as the
  tunneling technology), OR

+ When the tunneling technology is P2MP RSVP-TE

+ If ingress replication is used and the ingress PE wants to send
  traffic for (C-S, C-G)s to only those PEs that are on the path to
  receivers for the (C-S,C-G)s.

For more on aggregation see section "Aggregation Considerations".

For discovering the IP multicast group membership, this document
describes procedures that allow an ingress PE to enable explicit
tracking of IP multicast membership. Thus, an ingress PE can request
the IP multicast membership from egress PEs for one or more C-
multicast streams. These procedures are described in section
"Optimizing Multicast Distribution via Selective Trees".

These procedures are applicable when IGMP is used as the multicast
signaling protocol between the VPLS CEs. They are also applicable
when PIM as specified in [RFC4601] is used as the multicast routing
protocol between the VPLS CEs and PIM join suppression is disabled on
all the CEs. However these procedures do not apply when PIM is used
as the multicast routing protocol between the VPLS CEs and PIM join
suppression is not disabled on all the CEs. This is because when PIM
join suppression is not disabled on all the CEs, PEs connected to
these CEs can not rely on PIM to determine IP multicast membership of
the receivers behind these CEs. Procedures for this case are outside
the scope of this document.

The leaves of the Selective P-Multicast trees must also be able to
discover the identifier of these trees. Procedures to accomplish this
are described in section "Advertising P-Multicast Tree to VPLS/C-
Multicast Binding".

5.4. Advertising P-Multicast Tree to VPLS/C-Multicast Binding

This document describes procedures based on BGP VPLS Auto-Discovery
(A-D) routes ([RFC4761],[RFC6074]) that are used by the root of an
Aggregate P-Multicast tree to advertise the Inclusive or Selective P-
Multicast tree binding and the de-multiplexing information to the
leaves of the tree. This document uses the Provider Multicast Service
Interface (PMSI) Tunnel attribute defined [RFC6514] for this purpose.

Once an ingress PE decides to bind a set of VPLS instances or

customer multicast groups to an Inclusive P-Multicast tree or a
Selective P-Multicast tree, the PE needs to announce this binding to
other PEs in the network. This procedure is referred to as Inclusive
P-Multicast tree or Selective P-Multicast tree binding distribution
and is performed using BGP. The decision to bind a set of VPLS
instances or customer multicast group is local to the ingress PE
matter, and is controlled via provisioning/configuration on that PE.

When an Aggregated Inclusive P-Multicast tree is used by an ingress
PE, this binding distribution implies that (a) an ingress PE MUST
announce the binding of all VPLS instances bound to the Inclusive P-
Multicast tree, and (b) other PEs that have these instances receive
these announcements. The inner label assigned by the ingress PE for
each VPLS MUST be included, if more than one VPLS is bound to the
same P-Multicast tree. The Inclusive P-Multicast tree Identifier MUST
be included.

For a Selective P-Multicast tree this binding distribution implies
announcing all the specific <C-S, C-G> entries bound to this P-
Multicast tree along with the Selective P-Multicast tree Identifier.
The inner label assigned for each <C-S, C-G> MUST be included if <C-
S, C-G>s from different VPLS instances are bound to the same P-
Multicast tree. The labels MUST be distinct on a per VPLS basis and
MAY be distinct on a per <C-S, C-G> basis. The Selective P-Multicast
tree Identifier MUST be included.

[5.5](5.5). Aggregation

As described earlier in this document, the ability to carry the
traffic of more than one VPLS on the same P-Multicast tree is termed
aggregation.

Aggregation enables the SP to place a bound on the amount of
multicast tree forwarding and control plane state which the P routers
must have. Let us call the number of VPLS instances aggregated onto a
single P-Multicast tree as the "Aggregation Factor". When Inclusive
source P-Multicast trees are used the number of trees that a PE is
the root of is proportional to:

  + (Number of VPLS instances on the PE / Aggregation Factor).

In this case the state maintained by a P router, is proportional to:

```
  + ((Average number of VPLS instances on a PE / Aggregation Factor)
    * number of PEs) / (Average number of P-Multicast trees that
    transit a given P router)
```

Thus, the state does not grow linearly with the number of VPLS
instances.

Aggregation requires a mechanism for the egresses of the P-Multicast
tree to demultiplex the multicast traffic received over the P-
Multicast tree. To enable the egress nodes to perform this
demultiplexing, upstream-assigned labels [RFC5331] MUST be assigned
and distributed by the root of the aggregate P-multicast tree.

5.6. Inter-AS VPLS Multicast

This document defines four models of inter-AS VPLS service, referred
here as option (a), (b), (c), and (e). Options (a), (b), and (c)
defined in this document are very similar to the three methods,
method (a), method (b), and method (c), described in section "Multi-
AS VPLS" [RFC4761], which in turn extends the concepts of [RFC4364]
to inter-AS VPLS.

For option (a) and option (b) support, this document specifies a
model where Inter-AS VPLS service can be offered without requiring a
single P-Multicast tree to span multiple ASes. There are two variants
of this model and they are described in section "Inter-AS Inclusive
P-Multicast Tree A-D/Binding".

For option (c) support this document specifies a model where Inter-AS
VPLS service is offered by requiring a single P-Multicast tree to
span multiple ASes. This is because in the case of option (c) the
ASBRs do not exchange BGP-VPLS NLRIs or A-D routes.

In addition to options (a), (b), and (c), this document also
specifies option (e), which one may think of as a variant of option
(a).

For more on these inter-AS options see section "Inter-AS Inclusive P-
Multicast Tree A-D/Binding".

6. **Intra-AS Inclusive P-Multicast Tree Auto-discovery/Binding**

   This section specifies procedures for the intra-AS auto-discovery of
   VPLS membership and the distribution of information used to
   instantiate P-Multicast Tunnels.

   VPLS auto-discovery/binding consists of two components: intra-AS and
   inter-AS. The former provides VPLS auto-discovery/binding within a
   single AS. The latter provides VPLS auto-discovery/binding across
   multiple ASes. Inter-AS auto-discovery/binding is described in
   section "Inter-AS Inclusive P-Multicast Tree A-D/Binding".

   VPLS auto-discovery using BGP as described in [RFC4761, RFC6074]
   enables a PE to learn the VPLS instance membership of other PEs. A PE
   that belongs to a particular VPLS instance announces a BGP Network
   Layer Reachability Information (NLRI) that identifies the Virtual
   Switch Instance (VSI). This NLRI is constructed from the <Route-
   Distinguisher (RD), VPLS Edge Device Identifier (VE-ID)> tuple. The
   NLRI defined in [RFC4761] comprises the <RD, VE-ID> tuple and label
   blocks for pseudo-wire (PW) signaling. The VE-ID in this case is a
   two octet number. The NLRI defined in [RFC6074] comprises only the
   <RD, VE-ID> where the VE-ID is a four octet number.

   The procedures for constructing Inclusive intra-AS and inter-AS trees
   as specified in this document require the BGP A-D NLRI to carry only
   the <RD, VE-ID>. Hence these procedures can be used for both BGP-VPLS
   and LDP-VPLS with BGP A-D.

   It is to be noted that BGP A-D is an inherent feature of BGP-VPLS.
   However it is not an inherent feature of LDP-VPLS. In fact there are
   deployments and/or implementations of LDP-VPLS that require
   configuration to enable a PE in a particular VPLS to determine other
   PEs in the VPLS and exchange PW labels using FEC 128 [RFC4447]. The
   use of BGP A-D for LDP-VPLS [RFC6074], to enable automatic setup of
   PWs, requires FEC 129 [RFC4447]. However FEC 129 is not required in
   order to use procedures in this document for LDP-VPLS. An LDP-VPLS
   implementation that supports this document MUST support the BGP A-D
   procedures to setup P-Multicast trees, as described here, and it MAY
   support FEC 129 to automate the signaling of PWs.

6.1. **Originating intra-AS VPLS A-D routes**

   To participate in the VPLS auto-discovery/binding a PE router that
   has a given VSI of a given VPLS instance originates a BGP VPLS intra-
   AS A-D route and advertises this route in Multi-Protocol (MP) IBGP.
   The route is constructed as described in [RFC4761] and [RFC6074].

   The route carries a single L2VPN NLRI with the RD set to the RD of
   the VSI, and the VE-ID set to the VE-ID of the VSI. The route also
   carries one or more Route Targets (RTs) as specified in [RFC4761] and
   [RFC6074].

   If an Inclusive P-Multicast tree is used to instantiate the provider
   tunnel for VPLS multicast on the PE, the advertising PE MUST
   advertise the type and the identity of the P-Multicast tree in the
   PMSI Tunnel attribute. This attribute is described in section
   "Inclusive Tree/Selective Tree Identifier".

   A PE that uses an Inclusive P-Multicast tree to instantiate the
   provider tunnel MAY aggregate two or more VPLS instances present on
   the PE onto the same tree. If the PE decides to perform aggregation
   after it has already advertised the intra-AS VPLS A-D routes for
   these VPLS instances, then aggregation requires the PE to re-
   advertise these routes. The re-advertised routes MUST be the same as
   the original ones, except for the PMSI Tunnel attribute (the re-
   advertised route will replace of the previously advertised route). If
   the PE has not previously advertised intra-AS A-D routes for these
   VPLS instances, then the aggregation requires the PE to advertise
   (new) intra-AS A-D routes for these VPLS instances. The PMSI
   attribute in the newly advertised/re-advertised routes MUST carry the
   identity of the P-Multicast tree that aggregates the VPLS instances,
   as well as an MPLS upstream-assigned label [RFC5331]. Each re-
   advertised or newly advertised route MUST have a label that is
   distinct within the scope of the PE that advertises the route.

   Discovery of PE capabilities in terms of what tunnels types they
   support is outside the scope of this document. Within a given AS PEs
   participating in a VPLS are expected to advertise tunnel bindings
   whose tunnel types are supported by all other PEs that are
   participating in this VPLS and are part of the same AS.

6.2. Receiving intra-AS VPLS A-D routes

   When a PE receives a BGP Update message that carries an intra-AS A-D
   route such that (a) the route was originated by some other PE within
   the same AS as the local PE, (b) at least one of the Route Targets of
   the route matches one of the import Route Targets configured for a
   particular VSI on the local PE, (c) the BGP route selection
   determines that this is the best route with respect to the NLRI
   carried by the route, and (d) the route carries the PMSI Tunnel
   attribute, the PE performs the following:

     + If the Tunnel Type in the PMSI Tunnel attribute is set to LDP
       P2MP LSP, the PE SHOULD join the P-Multicast tree whose identity
       is carried in the PMSI Tunnel attribute.

     + If the Tunnel Type in the PMSI Tunnel attribute is set to RSVP-TE
       P2MP LSP, the receiving PE has to establish the appropriate state
       to properly handle the traffic received over that LSP. The PE
       that originated the route MUST establish an RSVP-TE P2MP LSP with
       the local PE as a leaf. This LSP MAY have been established before
       the local PE receives the route.

     + If the PMSI Tunnel attribute does not carry a label, then all
       packets that are received on the P-Multicast tree, as identified
       by the PMSI Tunnel attribute, are forwarded using the VSIs that
       have at least one of its import Route Targets that matches one of
       the Route Targets of the received A-D route.

     + If the PMSI Tunnel attribute has the Tunnel Type set to LDP P2MP
       LSP or RSVP-TE P2MP LSP, and the attribute also carries an MPLS
       label, then the egress PE MUST treat this as an upstream-assigned
       label, and all packets that are received on the P-Multicast tree,
       as identified by the PMSI Tunnel attribute, with that upstream
       label are forwarded using the VSIs that have at least one of its
       import Route Target that matches one of the Route Targets of the
       received intra-AS A-D route.

   Furthermore, if the local PE uses RSVP-TE P2MP LSP for sending
   (multicast) traffic, originated by VPLS sites connected to the PE, to
   the sites attached to other PEs then the local PE MUST use the
   Originating Router's IP address information carried in the intra-AS
   A-D route to add the PE, that originated the route, as a leaf node to
   the LSP. This MUST be done irrespective of whether the received
   Intra-AS A-D route carries the PMSI Tunnel attribute or not.

7. **Demultiplexing P-Multicast Tree Traffic**

   Demultiplexing received VPLS traffic requires the receiving PE to
   determine the VPLS instance the packet belongs to. The egress PE can
   then perform a VPLS lookup to further forward the packet. It also
   requires the egress PE to determine the identity of the ingress PE
   for MAC learning, as described in section "VPLS Data Packet
   Treatment".

7.1. **One P-Multicast Tree - One VPLS Mapping**

   When a P-Multicast tree is mapped to only one VPLS, determining the
   tree on which the packet is received is sufficient to determine the
   VPLS instance on which the packet is received. The tree is determined
   based on the tree encapsulation. If MPLS encapsulation is used,
   e.g.,: RSVP-TE P2MP LSPs, the outer MPLS label is used to determine
   the tree. Penultimate-hop-popping MUST be disabled on the MPLS LSP
   (RSVP-TE P2MP LSP or mLDP P2MP LSP).

7.2. **One P-Multicast Tree - Many VPLS Mapping**

   As traffic belonging to multiple VPLS instances can be carried over
   the same tree, there is a need to identify the VPLS the packet
   belongs to. This is done by using an inner label that determines the
   VPLS for which the packet is intended. The ingress PE uses this label
   as the inner label while encapsulating a customer multicast data
   packet. Each of the egress PEs must be able to associate this inner
   label with the same VPLS and use it to demultiplex the traffic
   received over the Aggregate Inclusive tree or the Aggregate Selective
   tree.

   If traffic from multiple VPLS instances is carried on a single tree,
   upstream-assigned labels [RFC5331] MUST be used. Hence the inner
   label is assigned by the ingress PE. When the egress PE receives a
   packet over an Aggregate tree, the outer encapsulation (in the case
   of MPLS P2MP LSPs, the outer MPLS label) specifies the label space to
   perform the inner label lookup. The same label space MUST be used by
   the egress PE for all P-Multicast trees that have the same root
   [RFC5331].

   If the tree uses MPLS encapsulation, as in RSVP-TE P2MP LSPs, the
   outer MPLS label and optionally the incoming interface provides the
   label space of the label beneath it. This assumes that penultimate-
   hop-popping is disabled. The egress PE MUST NOT advertise IMPLICIT
   NULL or EXPLICIT NULL for that tree once it is known to the egress PE
   that the tree is bound to one or more VPLS instances. Once the label

representing the tree is popped off the MPLS label stack, the next
label is the demultiplexing information that allows the proper VPLS
instance to be determined.

The ingress PE informs the egress PEs about the inner label as part
of the tree binding procedures described in section "BGP Extensions".

[8](8). Establishing P-Multicast Trees

This document supports only P2MP P-Multicast trees wherein it is
possible for egress PEs to identify the ingress PE to perform MAC
learning. Specific procedures are specified only for RSVP-TE P2MP
LSPs and mLDP P2MP LSPs. An implementation that supports this
document MUST support RSVP-TE P2MP LSPs and mLDP P2MP LSPs.

[8.1](8.1). Common Procedures

The following procedures apply to both RSVP-TE P2MP and mLDP P2MP
LSPs.

Demultiplexing the C-multicast data packets at the egress PE requires
that the PE must be able to determine the P2MP LSP that the packets
are received on. This enables the egress PE to determine the VPLS
instances that the packet belongs to. To achieve this the LSP MUST be
signaled with penultimate-hop-popping (PHP) off and a non special
purpose MPLS label off as described in section "Demultiplexing P-
Multicast Tree Traffic". In other words an egress PE MUST NOT
advertise IMPLICIT NULL or EXPLICIT NULL for a P2MP LSP that is
carrying traffic for one or more VPLS instances. This is because the
egress PE needs to rely on the MPLS label, that it advertises to its
upstream neighbor, to determine the P2MP LSP that a C-multicast data
packet is received on.

The egress PE also needs to identify the ingress PE to perform MAC
learning. When P2MP LSPs are used as P2MP trees, determining the P2MP
LSP that the packets are received on, is sufficient to determine the
ingress PE. This is because the ingress PE is the root of the P2MP
LSP.

The egress PE relies on receiving the PMSI Tunnel attribute in BGP to
determine the VPLS instance to P2MP LSP mapping.

8.2. **RSVP-TE P2MP LSPs**

   This section describes procedures that are specific to the usage of
   RSVP-TE P2MP LSPs for instantiating a P-Multicast tree. Procedures in
   [RFC4875] are used to signal the P2MP LSP. The LSP is signaled as the
   root of the P2MP LSP discovers the leaves. The egress PEs are
   discovered using the procedures described in section "Intra-AS
   Inclusive P-Multicast Tree Auto-discovery/Binding". Aggregation as
   described in this document is supported.

8.2.1. **P2MP TE LSP - VPLS Mapping**

   P2MP TE LSP to VPLS mapping is learned at the egress PEs using BGP
   based advertisements of the P2MP TE LSP - VPLS mapping. They require
   that the root of the tree include the P2MP TE LSP identifier as the
   tunnel identifier in the BGP advertisements. This identifier contains
   the following information elements:

        - The type of the tunnel is set to RSVP-TE P2MP LSP
        - RSVP-TE P2MP LSP's SESSION Object

   This Tunnel Identifier is described in section "Inclusive
   Tree/Selective Tree Identifier".

   Once the egress PE receives the P2MP TE LSP to VPLS mapping:

     + If the egress PE already has RSVP-TE state for the P2MP TE LSP,
       it MUST begin to assign an MPLS label from the non special
       purpose label range, for the P2MP TE LSP and signal this to the
       previous hop of the P2MP TE LSP. Further it MUST create
       forwarding state to forward packets received on the P2MP LSP.

     + If the egress PE does not have RSVP-TE state for the P2MP TE LSP,
       it MUST retain this mapping. Subsequently when the egress PE
       receives the RSVP-TE P2MP signaling message, it creates the RSVP-
       TE P2MP LSP state. It MUST then assign an MPLS label from the
       non-reserved label range, for the P2MP TE LSP, and signal this to
       the previous hop of the P2MP TE LSP.

       Note that if the signaling to set up an RSVP-TE P2MP LSP is
       completed before a given egress PE learns, via a PMSI Tunnel
       attribute, of the VPLS or set of VPLS instances to which the LSP
       is bound, the PE MUST discard any traffic received on that LSP
       until the binding is received. In order for the egress PE to be
       able to discard such traffic it needs to know that the LSP is
       associated with one or more VPLS instances and that the VPLS A-D
       route that binds the LSP to a VPLS has not yet been received.

This is provided by extending [RFC4875] with [RFC6511].

8.3. Receiver Initiated P2MP LSP

Receiver initiated P2MP LSPs can also be used. The mLDP procedures
([RFC6388]) MUST be used to signal such LSPs. The LSP is signaled
once the leaves receive the LDP FEC for the tree from the root, as
described in section "Intra-AS Inclusive P-Multicast Tree Auto-
discovery/Binding". When aggregation is used, an ingress PE is
required to discover the egress PEs (see section "Aggregation
Considerations" for the rationale), and this is achieved using the
procedures in section "Intra-AS Inclusive P-Multicast Tree Auto-
discovery/Binding".

8.3.1. P2MP LSP - VPLS Mapping

P2MP LSP to VPLS mapping is learned at the egress PEs using BGP based
advertisements of the P2MP LSP - VPLS mapping. They require that the
root of the tree include the P2MP LSP identifier as the tunnel
identifier in the BGP advertisements. This identifier contains the
following information elements:

        - The type of the tunnel is set to LDP P2MP LSP
        - LDP P2MP FEC which includes an identifier generated by the root.

Each egress PE SHOULD "join" the P2MP MPLS tree by sending LDP label
mapping messages for the LDP P2MP FEC, that was learned in the BGP
advertisement, using procedures described in [RFC6388].

8.4. Encapsulation of Aggregate P-Multicast Trees

An Aggregate Inclusive P-Multicast tree or an Aggregate Selective P-
Multicast tree MUST use MPLS encapsulation. The protocol type in the
data link header is as described in [RFC5332].

9. Inter-AS Inclusive P-Multicast Tree A-D/Binding

   As stated earlier, this document defines four models of inter-AS VPLS
   service, referred here as option (a), (b), (c) and (e). This section
   contains procedures to support these models.

   For supporting option (a), (b) and (e) this section specifies a model
   where inter-AS VPLS service can be offered without requiring a single
   P-Multicast tree to span multiple ASes. This allows individual ASes
   to potentially use different P-tunneling technologies. There are two
   variants of this model. One that requires MAC lookup on the ASBRs,
   and applies to option (a) and (e). The other is one that does not
   require MAC lookup on the ASBRs, and instead builds segmented inter-
   AS Inclusive or Selective trees. This applies only to option (b).

   For supporting option (c) this section specifies a model where Inter-
   AS VPLS service is offered by requiring a single Inclusive P-
   Multicast tree to span multiple ASes. This is referred to as a non-
   segmented P-Multicast tree. This is because in the case of option (c)
   the ASBRs do not exchange BGP-VPLS NLRIs or VPLS A-D routes.
   Selective inter-AS trees for option (c) support may be segmented or
   non-segmented.

   An implementation MUST support options (a), (b), and (c), and MAY
   support option(e). When there are multiple ways for implementing one
   of these options, this section specifies which one is mandatory.

9.1. VSIs on the ASBRs

   When VSIs are configured on ASBRs, the ASBRs MUST perform a MAC
   lookup, in addition to any MPLS lookups, to determine the forwarding
   decision on a VPLS packet. The P-Multicast trees are confined to an
   AS. An ASBR on receiving a VPLS packet from another ASBR is required
   to perform a MAC lookup to determine how to forward the packet. Thus
   an ASBR is required to keep a VSI for the VPLS instance and MUST be
   configured with its own VE ID for the VPLS instance. The BGP VPLS A-
   D routes generated by PEs in an AS MUST NOT be propagated outside the
   AS.

9.1.1. Option (a): VSIs on the ASBRs

   When VSIs are configured on ASBRs and option (a) is used, then an
   ASBR in one AS treats an adjoining ASBR in another AS as a CE and
   determines the VSI for packets received from that ASBR based on the
   incoming Ethernet interface or VLAN ID. In option (a) the ASBRs do
   not exchange VPLS A-D routes.

An implementation MUST support option (a).

9.1.2. Option (e): VSIs on the ASBRs

The VSIs on the ASBRs scheme can be used such that the interconnect
between the ASBRs is a PW and MPLS encapsulation is used between the
ASBRs. An ASBR in one AS determines the VSI for packets received
from an adjoining ASBR in another AS based on the incoming MPLS PW
label. This is referred to as option (e). The only VPLS A-D routes
that are propagated outside the AS are the ones originated by ASBRs.
This MPLS PW connects the VSIs on the ASBRs and MUST be signaled
using the procedures defined in [RFC4761] or [RFC4762].

The P-Multicast trees for a VPLS are confined to each AS and the VPLS
auto-discovery/binding MUST follow the intra-AS procedures described
in section "Demultiplexing Multicast Tree Traffic".

An implementation MAY support option (e).

9.2. Option (b) - Segmented Inter-AS Trees

In this model, an inter-AS P-Multicast tree, rooted at a particular
PE for a particular VPLS instance, consists of a number of
"segments", one per AS, which are stitched together at ASBRs. These
are known as "segmented inter-AS trees". Each segment of a segmented
inter-AS tree may use a different multicast transport technology. In
this model, an ASBR is not required to keep a VSI for the VPLS
instance, and is not required to perform a MAC lookup in order to
forward the VPLS packet. This implies that an ASBR is not required to
be configured with a VE ID for the VPLS.

An implementation MUST support option (b) using this model.

The construction of segmented Inter-AS trees requires the BGP-VPLS A-
D NLRI described in [RFC4761, RFC6074]. A BGP VPLS A-D route for a
<RD, VE ID> tuple advertised outside the AS, to which the originating
PE belongs, will be referred to as an inter-AS VPLS A-D route (though
this route is originated by a PE as an intra-AS route, and is
referred to as an inter-AS route outside the AS).

In addition to this, segmented inter-AS trees require support for the
PMSI Tunnel attribute described in section "Inclusive Tree/Selective
Tree Identifier". They also require additional procedures in BGP to
signal leaf A-D routes between ASBRs as explained in subsequent
sections.

9.2.1. **Segmented Inter-AS Trees VPLS Inter-AS A-D/Binding**

This section specifies the procedures for inter-AS VPLS A-D/binding
for segmented inter-AS trees.

An ASBR must be configured to support a particular VPLS as follows:

+ An ASBR MUST be configured with a set of (import) Route Targets
  (RTs) that specify the set of VPLS instances supported by the
  ASBR. These Route Targets control acceptance of BGP VPLS auto-
  discovery routes by the ASBR. Note that instead of being
  configured, the ASBR MAY obtain this set of (import) Route
  Targets (RTs) by using Route Target Constrain [RFC4684].

+ The ASBR MUST be configured with the tunnel types for the intra-
  AS segments of the VPLS instances supported by the ASBR, as well
  as (depending on the tunnel type) the information needed to
  create the PMSI Tunnel attribute for these tunnel types. Note
  that instead of being configured, the ASBR MAY derive the tunnel
  types from the intra-AS A-D routes received by the ASBR from the
  PEs in its own AS.

If an ASBR is configured to support a particular VPLS instance, the
ASBR MUST participate in the intra-AS VPLS auto-discovery/binding
procedures for that VPLS instance within the ASBR's own AS, as
defined in this document.

Moreover, in addition to the above the ASBR performs procedures
specified in section "Propagating BGP VPLS A-D routes to other ASes:
Overview".

9.2.2. **Propagating BGP VPLS A-D routes to other ASes: Overview**

A BGP VPLS A-D route for a given VPLS, originated by a PE within a
given AS, is propagated via BGP to other ASes. The precise rules for
distributing and processing the inter-AS A-D routes are given in
subsequent sections.

Suppose that an ASBR A receives and installs a BGP VPLS A-D route for
VPLS "X" and VE ID "V" that originated at a particular PE, PE1, that
is in the same AS as A. The BGP next hop of that received route
becomes A's "upstream neighbor" on a multicast distribution tree for
(X, V) that is rooted at PE1. Likewise, when A re-advertises this
route to ASBRs in A's neighboring ASes, from the perspective of these
ASBRs A becomes their "upstream neighbor" on the multicast
distribution tree for (X, V) that is rooted at PE1.

When the BGP VPLS A-D routes have been distributed to all the
necessary ASes, they define a "reverse path" from any AS that
supports VPLS X and VE ID V back to PE1. For instance, if AS2
supports VPLS X, then there will be a reverse path for VPLS X and VE
ID V from AS2 to AS1. This path is a sequence of ASBRs, the first of
which is in AS2, and the last of which is in AS1. Each ASBR in the
sequence is the BGP next hop of the previous ASBR in the sequence.

This reverse path information can be used to construct a
unidirectional multicast distribution tree for VPLS X and VE ID V,
containing all the ASes that support X, and having PE1 at the root.
We call such a tree an "inter-AS tree". Multicast data originating in
VPLS sites for VPLS X connected to PE1 will travel downstream along
the tree which is rooted at PE1.

The path along an inter-AS tree is a sequence of ASBRs. It is still
necessary to specify how the multicast data gets from a given ASBR to
the set of ASBRs which are immediately downstream of the given ASBR
along the tree. This is done by creating "segments": ASBRs in
adjacent ASes will be connected by inter-AS segments, ASBRs in the
same AS will be connected by "intra-AS segments".

For a given inter-AS tree and a given AS there MUST be only one ASBR
within that AS that accepts traffic flowing on that tree. Further
for a given inter-AS tree and a given AS there MUST be only one ASBR
in that AS that sends the traffic flowing on that tree to a
particular adjacent AS. The precise rules for accomplishing this are
given in subsequent sections.

An ASBR initiates creation of an intra-AS segment when the ASBR
receives an inter-AS A-D route from an EBGP neighbor. Creation of the
segment is completed as a result of distributing, via IBGP, this
route within the ASBR's own AS.

For a given inter-AS tunnel each of its intra-AS segments could be
constructed by its own independent mechanism. Moreover, by using
upstream-assigned labels within a given AS multiple intra-AS segments
of different inter-AS tunnels of either the same or different VPLS
instances may share the same P-Multicast tree.

If the P-Multicast tree instantiating a particular segment of an
inter-AS tunnel is created by a multicast control protocol that uses
receiver-initiated joins (e.g, mLDP), and this P-Multicast tree does
not aggregate multiple segments, then all the information needed to
create that segment will be present in the inter-AS A-D routes
received by the ASBR from the neighboring ASBR. But if the P-
Multicast tree instantiating the segment is created by a protocol
that does not use receiver-initiated joins (e.g., RSVP-TE, ingress

unicast replication), or if this P-Multicast tree aggregates multiple
segments (irrespective of the multicast control protocol used to
create the tree), then the ASBR needs to learn the leaves of the
segment. These leaves are learned from A-D routes received from other
PEs in the AS, for the same VPLS as the one that the segment belongs
to.

The following sections specify procedures for propagation of inter-AS
A-D routes across ASes in order to construct inter-AS segmented
trees.

9.2.2.1. **Propagating Intra-AS VPLS A-D routes in EBGP**

For a given VPLS configured on an ASBR when the ASBR receives intra-
AS A-D routes originated by PEs in its own AS, the ASBR MUST
propagate each of these route in EBGP. This procedure MUST be
performed for each of the VPLS instances configured on the ASBR.
Each of these routes is constructed as follows:

   + The route carries a single BGP VPLS A-D NLRI with the RD and VE
     ID being the same as the NLRI in the received intra-AS A-D route.

   + The Next Hop field of the MP_REACH_NLRI attribute is set to a
     routable IP address of the ASBR.

   + The route carries the PMSI Tunnel attribute with the Tunnel Type
     set to Ingress Replication; the attribute carries no MPLS labels.

   + The route MUST carry the export Route Target used by the VPLS.

9.2.2.2. **Inter-AS A-D route received via EBGP**

When an ASBR receives from one of its EBGP neighbors a BGP Update
message that carries an inter-AS A-D route, if (a) at least one of
the Route Targets carried in the message matches one of the import
Route Targets configured on the ASBR, and (b) the ASBR determines
that the received route is the best route to the destination carried
in the NLRI of the route, the ASBR re-advertises this inter-AS A-D
route to other PEs and ASBRs within its own AS. The best route
selection procedures MUST ensure that for the same destination, all
ASBRs in an AS pick the same route as the best route. The best route
selection procedures are specified in [RFC4761] and clarified in
[MULTI-HOMING]. The best route procedures ensure that if multiple
ASBRs, in an AS, receive the same inter-AS A-D route from their EBGP
neighbors, only one of these ASBRs propagates this route in IBGP.
This ASBR becomes the root of the intra-AS segment of the inter-AS

   tree and ensures that this is the only ASBR that accepts traffic into
   this AS from the inter-AS tree.

   When re-advertising an inter-AS A-D route the ASBR MUST set the Next
   Hop field of the MP_REACH_NLRI attribute to a routable IP address of
   the ASBR.

   Depending on the type of a P-Multicast tunnel used to instantiate the
   intra-AS segment of the inter-AS tunnel, the PMSI Tunnel attribute of
   the re-advertised inter-AS A-D route is constructed as follows:

     + If the ASBR uses ingress replication to instantiate the intra-AS
       segment of the inter-AS tunnel, the re-advertised route MUST NOT
       carry the PMSI Tunnel attribute.

     + If the ASBR uses a P-Multicast tree to instantiate the intra-AS
       segment of the inter-AS tunnel, the PMSI Tunnel attribute MUST
       contain the identity of the tree that is used to instantiate the
       segment (note that the ASBR could create the identity of the tree
       prior to the actual instantiation of the segment). If in order to
       instantiate the segment the ASBR needs to know the leaves of the
       tree, then the ASBR obtains this information from the A-D routes
       received from other PEs/ASBRs in ASBR's own AS.

     + An ASBR that uses a P-Multicast tree to instantiate the intra-AS
       segment of the inter-AS tunnel MAY aggregate two or more VPLS
       instances present on the ASBR onto the same tree. If the ASBR
       already advertises inter-AS A-D routes for these VPLS instances,
       then aggregation requires the ASBR to re-advertise these routes.
       The re-advertised routes MUST be the same as the original ones,
       except for the PMSI Tunnel attribute. If the ASBR has not
       previously advertised inter-AS A-D routes for these VPLS
       instances, then the aggregation requires the ASBR to advertise
       (new) inter-AS A-D routes for these VPLS instances. The PMSI
       Tunnel attribute in the newly advertised/re-advertised routes
       MUST carry the identity of the P-Multicast tree that aggregates
       the VPLS instances, as well as an MPLS upstream-assigned label
       [RFC5331]. Each newly advertised or re-advertised route MUST have
       a label that is distinct within the scope of the ASBR.

   In addition the ASBR MUST send to the EBGP neighbor, from whom it
   receives the inter-AS A-D route, a BGP Update message that carries a
   leaf A-D route. The exact encoding of this route is described in
   section "BGP Extensions". This route contains the following
   information elements:

+ The route carries a single NLRI with the Route Key field set to
  the <RD, VE ID> tuple of the BGP VPLS A-D NLRI of the inter-AS A-
  D route received from the EBGP neighbor. The NLRI also carries
  the IP address of the ASBR (this MUST be a routable IP address).

+ The leaf A-D route MUST include the PMSI Tunnel attribute with
  the Tunnel Type set to Ingress Replication, and the Tunnel
  Identifier set to a routable address of the advertising router.
  The PMSI Tunnel attribute MUST carry a downstream assigned MPLS
  label that is used to demultiplex the VPLS traffic received over
  a unicast tunnel by the advertising router.

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

+ To constrain the distribution scope of this route the route MUST
  carry the NO_ADVERTISE BGP community ([RFC1997]).

+ The ASBR constructs an IP-based Route Target extended community
  by placing the IP address carried in the next hop of the received
  Inter-AS VPLS A-D route in the Global Administrator field of the
  community, with the Local Administrator field of this community
  set to 0, and sets the Extended Communities attribute of the leaf
  A-D route to that community. Note that this Route Target is the
  same as the ASBR Import RT of the EBGP neighbor from which the
  ASBR received the inter-AS VPLS A-D route.

9.2.2.3. Leaf A-D Route received via EBGP

When an ASBR receives via EBGP a leaf A-D route, the ASBR accepts the
route only if (a) at least one of the Route Targets carried in the
message matches one of the import Route Targets configured on the
ASBR, and (b) the ASBR determines that the received route is the best
route to the destination carried in the NLRI of the route.

If the ASBR accepts the leaf A-D route, the ASBR looks for an
existing A-D route whose BGP-VPLS A-D NLRI has the same value as the
<RD, VE-ID> field of the leaf A-D route just accepted. If such an A-D
route is found, then the MPLS label carried in the PMSI Tunnel
attribute of the leaf A-D route is used to stitch a one hop ASBR-ASBR
LSP to the tail of the intra-AS tunnel segment associated with the
found A-D route.

9.2.2.4. **Inter-AS A-D Route received via IBGP**

   In the context of this section we use the term "PE/ASBR router" to
   denote either a PE or an ASBR router.

   Note that a given inter-AS A-D route is advertised within a given AS
   by only one ASBR as described above.

   When a PE/ASBR router receives from one of its IBGP neighbors a BGP
   Update message that carries an inter-AS A-D route, if (a) at least
   one of the Route Targets carried in the message matches one of the
   import Route Targets configured on the PE/ASBR, and (b) the PE/ASBR
   determines that the received route is the best route to the
   destination carried in the NLRI of the route, the PE/ASBR performs
   the following operations. The best route determination is based as
   described in [RFC4761] and clarified in [MULTI-HOMING].

   If the router is an ASBR then the ASBR propagates the route to its
   EBGP neighbors. When propagating the route to the EBGP neighbors the
   ASBR MUST set the Next Hop field of the MP_REACH_NLRI attribute to a
   routable IP address of the ASBR.

   If the received inter-AS A-D route carries the PMSI Tunnel attribute
   with the Tunnel Type set to LDP P2MP LSP, the PE/ASBR SHOULD join the
   P-Multicast tree whose identity is carried in the PMSI Tunnel
   attribute.

   If the received inter-AS A-D route carries the PMSI Tunnel attribute
   with the Tunnel Identifier set to RSVP-TE P2MP LSP, then the ASBR
   that originated the route MUST establish an RSVP-TE P2MP LSP with the
   local PE/ASBR as a leaf. This LSP MAY have been established before
   the local PE/ASBR receives the route, or MAY be established after the
   local PE receives the route.

   If the received inter-AS A-D route carries the PMSI Tunnel attribute
   with the Tunnel Type set to LDP P2MP LSP, or RSVP-TE P2MP LSP, but
   the attribute does not carry a label, then the P-Multicast tree, as
   identified by the PMSI Tunnel attribute, is an intra-AS LSP segment
   that is part of the inter-AS Tunnel for the <VPLS, VE ID> advertised
   by the inter-AS A-D route and rooted at the PE that originated the A-
   D route. If the PMSI Tunnel attribute carries a (upstream-assigned)
   label, then a combination of this tree and the label identifies the
   intra-AS segment. If the receiving router is an ASBR, this intra-AS
   segment may further be stitched to ASBR-ASBR inter-AS segment of the
   inter-AS tunnel. If the PE/ASBR has local receivers in the VPLS,
   packets received over the intra-AS segment must be forwarded to the
   local receivers using the local VSI.

9.3. Option (c): Non-Segmented Tunnels

   In this model, there is a multi-hop EBGP peering between the PEs (or
   BGP Route Reflector) in one AS and the PEs (or BGP Route Reflector)
   in another AS. The PEs exchange BGP-VPLS NLRI or BGP-VPLS A-D NLRI,
   along with PMSI Tunnel attribute, as in the intra-AS case described
   in section "Demultiplexing P-Multicast Tree Traffic".

   The PEs in different ASes use a non-segmented inter-AS P2MP tunnel
   for VPLS multicast. A non-segmented inter-AS tunnel is a single
   tunnel which spans AS boundaries. The tunnel technology cannot change
   from one point in the tunnel to the next, so all ASes through which
   the tunnel passes must support that technology. In essence, AS
   boundaries are of no significance to a non-segmented inter-AS P2MP
   tunnel.

   This model requires no VPLS A-D routes in the control plane or VPLS
   MAC address learning in the data plane on the ASBRs. The ASBRs only
   need to participate in the non-segmented P2MP tunnel setup in the
   control plane, and do MPLS label forwarding in the data plane.

   When the tunneling technology is P2MP LSP signaled with mLDP, and one
   does not use [RFC6512], the setup of non-segmented inter-AS P2MP
   tunnels requires the P-routers in one AS to have IP reachability to
   the loopback addresses of the PE routers in another AS. That is, the
   reachability to the loopback addresses of PE routers in one AS MUST
   be present in the IGP in another AS.

   The data forwarding in this model is the same as in the intra-AS case
   described in section "Demultiplexing P-Multicast Tree Traffic".

   An implementation MUST support this model.

10. Optimizing Multicast Distribution via Selective Trees

   Whenever a particular multicast stream is being sent on an Inclusive
   P-Multicast tree, it is likely that the data of that stream is being
   sent to PEs that do not require it as the sites connected to these
   PEs may have no receivers for the stream. If a particular stream has
   a significant amount of traffic, it may be beneficial to move it to a
   Selective P-Multicast tree which has at its leaves only those PEs,
   connected to sites that have receivers for the multicast stream (or
   at least includes fewer PEs that are attached to sites with no
   receivers compared to an Inclusive tree).

   A PE connected to the multicast source of a particular multicast
   stream may be performing explicit tracking - i.e., it may know the

PEs that have receivers in the multicast stream. Section "Receiving
S-PMSI A-D routes by PEs" describes procedures that enable explicit
tracking. If this is the case Selective P-Multicast trees can also be
triggered on other criteria. For instance there could be a "pseudo
wasted bandwidth" criteria: switching to a Selective tree would be
done if the bandwidth multiplied by the number of "uninterested" PEs
(PEs that are receiving the stream but have no receivers) is above a
specified threshold. The motivation is that (a) the total bandwidth
wasted by many sparsely subscribed low-bandwidth groups may be large,
and (b) there's no point to moving a high-bandwidth group to a
Selective tree if all the PEs have receivers for it.

Switching a (C-S, C-G) stream to a Selective P-Multicast tree may
require the root of the tree to determine the egress PEs that need to
receive the (C-S, C-G) traffic. This is true in the following cases:

  + If the tunnel is a P2MP tree, such as a RSVP-TE P2MP Tunnel, the
    PE needs to know the leaves of the tree before it can instantiate
    the Selective tree.

  + If a PE decides to send traffic for multicast streams, belonging
    to different VPLS instances, using one P-Multicast Selective
    tree, such a tree is termed an Aggregate tree with a selective
    mapping. The setting up of such an Aggregate tree requires the
    ingress PE to know all the other PEs that have receivers for
    multicast groups that are mapped onto the tree (see section
    "Aggregation Considerations" for the rationale).

  + If ingress replication is used and the ingress PE wants to send
    traffic for (C-S, C-G)s to only those PEs that are on the path to
    receivers to the (C-S,C-G)s.

For discovering the IP multicast group membership, for the above
cases, this document describes procedures that allow an ingress PE to
enable explicit tracking. Thus an ingress PE can request the IP
multicast membership from egress PEs for one or more C-multicast
streams. These procedures are described in section "Receiving S-PMSI
A-D routes by PEs".

The root of the Selective P-Multicast tree MAY decide to do explicit
tracking of the IP multicast stream only after it has determined to
move the stream to a Selective tree, or it MAY have been doing
explicit tracking all along. This document also describes explicit
tracking for a wildcard source and/or group in section "Receiving S-
PMSI A-D routes by PEs", which facilitates a Selective P-Multicast
tree only mode in which IP multicast streams are always carried on a
Selective P-Multicast tree. In the description on Selective P-

Multicast trees the notation C-S, is intended to represent either a
specific source address or a wildcard. Similarly C-G is intended to
represent either a specific group address or a wildcard.

The PE at the root of the tree MUST signal the leaves of the tree
that the (C-S, C-G) stream is now bound to the Selective Tree. Note
that the PE could create the identity of the P-Multicast tree prior
to the actual instantiation of the tunnel.

If the Selective tree is instantiated by a RSVP-TE P2MP LSP the PE at
the root of the tree MUST establish the P2MP RSVP-TE LSP to the
leaves. This LSP MAY have been established before the leaves receive
the Selective tree binding, or MAY be established after the leaves
receive the binding. A leaf MUST NOT switch to the Selective tree
until it receives the binding and the RSVP-TE P2MP LSP is setup to
the leaf.

10.1. Protocol for Switching to Selective Trees

Selective trees provide a PE the ability to create separate P-
Multicast trees for certain <C-S, C-G> streams. The source PE, that
originates the Selective tree, and the egress PEs, MUST use the
Selective tree for the <C-S, C-G> streams that are mapped to it. This
may require the source and egress PEs to switch to the Selective tree
from an Inclusive tree if they were already using an Inclusive tree
for the <C-S, C-G> streams mapped to the Selective tree.

Once a source PE decides to setup a Selective tree, it MUST announce
the mapping of the <C-S, C-G> streams (which may be in different VPLS
instances) that are mapped to the tree to the other PEs using BGP.
After the egress PEs receive the announcement they setup their
forwarding path to receive traffic on the Selective tree if they have
one or more receivers interested in the <C-S, C-G> streams mapped to
the tree. Setting up the forwarding path requires setting up the
demultiplexing forwarding entries based on the top MPLS label (if
there is no inner label) or the inner label (if present) as described
in section "Establishing P-Multicast Trees".

When the P2MP LSP is established using mLDP, the egress PEs MAY
perform this switch to the Selective tree once the announcement from
the ingress PE is received, or MAY wait for a preconfigured timer to
do so, after receiving the announcement.

When the P2MP LSP protocol is P2MP RSVP-TE, an egress PE MUST perform
this switch to the Selective tree only after the announcement from
the ingress PE is received and the RSVP-TE P2MP LSP has been setup to
the egress PE. This switch MAY be done after waiting for a

preconfigured timer after these two steps have been accomplished.

A source PE MUST use the following approach to decide when to start
transmitting data on the Selective tree, if it is currently using an
Inclusive tree. After announcing the <C-S, C-G> stream mapping to a
Selective tree, the source PE MUST wait for a "switch-over" delay
before sending <C-S, C-G> stream on the Selective tree. It is
RECOMMENDED to allow this delay to be configurable. Once the "switch-
over" delay has elapsed, the source PE MUST send <C-S, C-G> stream on
the Selective tree. In no case is any <C-S, C-G> packet sent on both
Selective and Inclusive trees.

When a <C-S, C-G> stream is switched from an Inclusive to a Selective
tree, the purpose of running a switch-over timer is to minimize
packet loss without introducing packet duplication. However, jitter
may be introduced due to the difference in transit delays between the
Inclusive and Selective trees.

For best effect, the switch-over timer should be configured to a
value that is "just long enough" (a) to allow all the PEs to learn
about the new binding of <C-S, C-G> to a Selective tree, and (b) to
allow the PEs to construct the P-tunnel associated with the Selective
tree, if it doesn't already exist.

[10.2](10.2). Advertising (C-S, C-G) Binding to a Selective Tree

The ingress PE informs all the PEs that are on the path to receivers
of the (C-S, C-G) of the binding of the Selective tree to the (C-S,
C-G), using BGP. The BGP announcement is done by sending update for
the MCAST-VPLS address family using what we referred to as an "S-PMSI
A-D route". The format of the NLRI of this route is described in
section "Inclusive Tree/Selective Tree Identifier". The NLRI MUST be
constructed as follows:

   + The Route Distinguisher (RD) MUST be set to the RD configured
      locally for the VPLS. This is required to uniquely identify the
      <C-S, C-G> as the addresses could overlap between different VPLS
      instances. This MUST be the same RD value used in the VPLS auto-
      discovery process.

   + The Multicast Source field MUST contain the source address
      associated with the C-multicast stream, and the Multicast Source
      Length field is set appropriately to reflect this. If the source
      address is a wildcard the source address is set to 0.

+ The Multicast Group field MUST contain the group address
   associated with the C-multicast stream, and the Multicast Group
   Length field is set appropriately to reflect this. If the group
   address is a wildcard the group address is set to 0.

+ The Originating Router's IP Address field MUST be set to the IP
   address that the (local) PE places in the BGP next-hop of the
   BGP-VPLS A-D routes. Note that the <RD, Originating Router's IP
   address> tuple uniquely identifies a given VPLS instance on a PE.

The PE constructs the rest of the Selective A-D route as follows.

Depending on the type of a P-Multicast tree used for the P-tunnel,
the PMSI tunnel attribute of the S-PMSI A-D route is constructed as
follows:

+ The PMSI tunnel attribute MUST contain the identity of the P-
   Multicast tree (note that the PE could create the identity of the
   tree prior to the actual instantiation of the tree).

+ If in order to establish the P-Multicast tree the PE needs to
   know the leaves of the tree within its own AS, then the PE
   obtains this information from the leaf A-D routes received from
   other PEs/ASBRs within its own AS (as other PEs/ASBRs originate
   leaf A-D routes in response to receiving the S-PMSI A-D route) by
   setting the Leaf Information Required flag in the PMSI Tunnel
   attribute to 1. This enables explicit tracking for the multicast
   stream(s) advertised by the S-PMSI A-D route.

+ If a PE originates S-PMSI A-D routes with the Leaf Information
   Required flag in the PMSI Tunnel attribute set to 1, then the PE
   MUST be (auto)configured with an import Route Target, which
   controls acceptance of leaf A-D routes by the PE. (Procedures for
   originating leaf A-D routes by the PEs that receive the S-PMSI A-
   D route are described in section "Receiving S-PMSI A-D routes by
   PEs.")

   This Route Target is IP address specific. The Global
   Administrator field of this Route Target MUST be set to the IP
   address carried in the Next Hop of all the S-PMSI A-D routes
   advertised by this PE (if the PE uses different Next Hops, then
   the PE MUST be (auto)configured with multiple import RTs, one per
   each such Next Hop). The Local Administrator field of this Route
   Target MUST be set to 0.

   If the PE supports Route Target Constrain [RFC4684], the PE
   SHOULD advertise this import Route Target within its own AS using

Route Target Constrains. To constrain distribution of the Route
Target Constrain routes to the AS of the advertising PE these
routes SHOULD carry the NO_EXPORT Community ([RFC1997]).

+ A PE MAY aggregate two or more S-PMSIs originated by the PE onto
   the same P-Multicast tree. If the PE already advertises S-PMSI A-
   D routes for these S-PMSIs, then aggregation requires the PE to
   re-advertise these routes. The re-advertised routes MUST be the
   same as the original ones, except for the PMSI tunnel attribute.
   If the PE has not previously advertised S-PMSI A-D routes for
   these S-PMSIs, then the aggregation requires the PE to advertise
   (new) S-PMSI A-D routes for these S-PMSIs. The PMSI Tunnel
   attribute in the newly advertised/re-advertised routes MUST carry
   the identity of the P-Multicast tree that aggregates the S-PMSIs.
   If at least some of the S-PMSIs aggregated onto the same P-
   Multicast tree belong to different VPLS instances, then all these
   routes MUST carry an MPLS upstream assigned label [RFC5331]. If
   all these aggregated S-PMSIs belong to the same VPLS, then the
   routes MAY carry an MPLS upstream assigned label [RFC5331]. The
   labels MUST be distinct on a per VPLS instance basis, and MAY be
   distinct on a per route basis.

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

By default the set of Route Targets carried by the route MUST be the
same as the Route Targets carried in the BGP-VPLS A-D route
originated from the VSI. The default could be modified via
configuration.

10.3. Receiving S-PMSI A-D routes by PEs

Consider a PE that receives an S-PMSI A-D route. If one or more of
the VSIs on the PE have their import Route Targets that contain one
or more of the Route Targets carried by the received S-PMSI A-D
route, then for each such VSI the PE performs the following.

Procedures for receiving an S-PMSI A-D route by a PE (both within and
outside of the AS of the PE that originates the route) are the same
as specified in section "Inter-AS A-D route received via IBGP" except
that (a) instead of Inter-AS A-D routes the procedures apply to S-
PMSI A-D routes, and (b) the rules for determining whether the
received S-PMSI A-D route is the best route to the destination
carried in the NLRI of the route, are the same as BGP path selection

rules and may be modified by policy, and (c) a PE performs procedures
specified in that section only if in addition to the criteria
specified in that section the following is true:

+ If as a result of multicast state snooping on the PE-CE
   interfaces, the PE has snooped state for at least one multicast
   join that matches the multicast source and group advertised in
   the S-PMSI A-D route. Further if the oifs (outgoing interfaces)
   for this state contains one or more interfaces to the locally
   attached CEs. When the multicast signaling protocol among the CEs
   is IGMP, then snooping and associated procedures are defined in
   [RFC4541]. The snooped state is determined using these
   procedures. When the multicast signaling protocol among the CEs
   is PIM, the procedures in [RFC4541] are not sufficient to
   determine the snooped state. The additional details required to
   determine the snooped state when CE-CE protocol is PIM are for
   further study. When such procedures are defined it is expected
   that the procedures in this section will apply to the snooped
   state created as a result of PIM as PE-CE protocol.

The snooped state is said to "match" the S-PMSI A-D route if any of
the following is true:

+ The S-PMSI A-D route carries (C-S, C-G) and the snooped state is
   for (C-S, C-G) or for (C-*, C-G), OR

+ The S-PMSI A-D route carries (C-*, C-G) and (a) the snooped
   state is for (C-*, C-G) OR (b) the snooped state is for at least
   one multicast join with the multicast group address equal to C-G
   and there doesn't exist another S-PMSI A-D route that carries (C-
   S, C-G) where C-S is the source address of the snooped state.

+ The S-PMSI A-D route carries (C-S, C-*) and (a) the snooped
   state is for at least one multicast join with the multicast
   source address equal to C-S, and (b) there doesn't exist another
   S-PMSI A-D route that carries (C-S, C-G) where C-G is the group
   address of the snooped state.

+ The S-PMSI A-D route carries (C-*, C-*) and there is no other S-
   PMSI A-D route that matches the snooped state as per the above
   conditions.

Note if the above conditions are true, and if the received S-PMSI A-D
route has a PMSI Tunnel attribute with the Leaf Information Required
flag set to 1, then the PE originates a leaf A-D route, constructed
as follows:

   + The route carries a single MCAST-VPLS NLRI with the Route Key
      field set to the MCAST-VPLS NLRI of the received S-PMSI A-D
      route.

   + The Originating Router's IP address set to the IP address of the
      PE (this MUST be a routable IP address).

   + The PE constructs an IP-based Route Target Extended Community by
      placing the IP address carried in the Next Hop of the received S-
      PMSI A-D route in the Global Administrator field of the
      Community, with the Local Administrator field of this Community
      set to 0 and setting the Extended Communities attribute of the
      leaf A-D route to that Community.

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

   + To constrain the distribution scope of this route, the route
      MUST carry the NO_EXPORT Community [RFC1997], except for the
      inter-AS scenario with option (c).

   Once the leaf A-D route is constructed, the PE advertises this route
   into IBGP.

   In addition to the procedures specified in section "Inter-AS A-D
   route received via IBGP" the PE MUST set up its forwarding path to
   receive traffic, for each multicast stream in the matching snooped
   state, from the tunnel advertised by the S-PMSI A-D route (the PE
   MUST switch to the Selective tree).

   When a new snooped state is created by a PE then the PE MUST first
   determine if there is a S-PMSI A-D route that matches the snooped
   state as per the conditions described above. If such a S-PMSI A-D
   route is found, then the PE MUST follow the procedures described in
   this section, for that particular S-PMSI A-D route. If later on the
   snooped state ages out and is deleted from the PE, the PE SHOULD
   withdraw the leaf A-D route that it had originated in response to the
   S-PMSI A-D route.

[10.4](10.4). Inter-AS Selective Tree

Inter-AS Selective trees support all three options of inter-AS VPLS service, option (a), (b) and (c), that are supported by Inter-AS Inclusive trees. They are constructed in a manner that is very similar to Inter-AS Inclusive trees.

For option (a) and option (b) support inter-AS Selective trees are constructed without requiring a single P-Multicast tree to span multiple ASes. This allows individual ASes to potentially use different P-tunneling technologies. There are two variants of this. One that requires MAC and IP multicast lookup on the ASBRs and another that does not require MAC/IP multicast lookup on the ASBRs and instead builds segmented inter-AS Selective trees.

Segmented Inter-AS Selective trees can also be used with option (c), unlike Segmented Inter-AS Inclusive trees. This is because the S-PMSI A-D routes can be exchanged via ASBRs (even though BGP VPLS A-D routes are not exchanged via ASBRs).

In the case of Option (c) an Inter-AS Selective tree may also be a non-segmented P-Multicast tree that spans multiple ASes.

[10.4.1](10.4.1). VSIs on the ASBRs

The requirements on ASBRs, when VSIs are present on the ABSRs, include the requirements presented in section "Inter-AS Inclusive P-Multicast Tree A-D/Binding". The source ASBR (that receives traffic from another AS) may independently decide whether it wishes to use Selective trees or not. If it uses Selective trees the source ASBR MUST perform a MAC lookup to determine the Selective tree to forward the VPLS packet on.

[10.4.1.1](10.4.1.1). VPLS Inter-AS Selective Tree A-D Binding

The mechanisms for propagating S-PMSI A-D routes are the same as the intra-AS case described in section "MCAST-VPLS NLRI". The BGP Selective tree A-D routes generated by PEs in an AS MUST NOT be propagated outside the AS.

10.4.2. **Inter-AS Segmented Selective Trees**

   Inter-AS Segmented Selective trees MUST be used when option (b) is
   used to provide the inter-AS VPLS service. They MAY be used when
   option (c) is used to provide the inter-AS VPLS service.

   A Segmented inter-AS Selective Tunnel is constructed similar to an
   inter-AS Segmented Inclusive Tunnel. Namely, such a tunnel is
   constructed as a concatenation of tunnel segments. There are two
   types of tunnel segments: an intra-AS tunnel segment (a segment that
   spans ASBRs within the same AS), and inter-AS tunnel segment (a
   segment that spans adjacent ASBRs in adjacent ASes). ASes that are
   spanned by a tunnel are not required to use the same tunneling
   mechanism to construct the tunnel - each AS may pick up a tunneling
   mechanism to construct the intra-AS tunnel segment of the tunnel, in
   its AS.

   The PE that decides to set up a Selective tree, advertises the
   Selective tree to multicast stream binding using an S-PMSI A-D route
   as per procedures in section "Advertising (C-S, C-G) Binding to a
   Selective Tree", to the routers in its own AS.

   An S-PMSI A-D route advertised outside the AS, to which the
   originating PE belongs, will be referred to as an inter-AS Selective
   Tree A-D route (although this route is originated by a PE as an
   intra-AS route it is referred to as an inter-AS route outside the
   AS).

10.4.2.1. **Handling S-PMSI A-D routes by ASBRs**

   Procedures for handling an S-PMSI A-D route by ASBRs (both within and
   outside of the AS of the PE that originates the route) are the same
   as specified in section "Propagating VPLS BGP A-D routes to other
   ASes", except that instead of Inter-AS BGP-VPLS A-D routes and the
   BGP-VPLS A-D NLRI these procedures apply to S-PMSI A-D routes and the
   S-PMSI A-D NLRI.

   In addition to these procedures an ASBR advertises a leaf A-D route
   in response to an S-PMSI A-D route only if:

     + The S-PMSI A-D route was received via EBGP from another ASBR and
        the ASBR merges the S-PMSI A-D route into an Inter-AS BGP VPLS A-
        D route as described in the next section. OR

> + The ASBR receives a leaf A-D route from a downstream PE or ASBR
>   in response to the S-PMSI A-D route, received from an upstream PE
>   or ASBR, that the ASBR propagated inter-AS to downstream ASBRs
>   and PEs.
>
> + The ASBR has snooped state from local CEs that matches the NLRI
>   carried in the S-PMSI A-D route as per the following rules:
>
>   i) The NLRI encodes (C-S, C-G) which is the same as the snooped
>   (C-S, C-G)
>
>   ii) The NLRI encodes (*, C-G) and there is snooped state for at
>   least one (C-S, C-G) and there is no other matching S-PMSI A-D
>   route for (C-S, C-G) OR there is snooped state for (*, C-G)
>
>   iii) The NLRI encodes (*, *) and there is snooped state for at
>   least one (C-S, C-G) or (*, C-G) and there is no other matching
>   S-PMSI A-D route for that (C-S, C-G) or (*, C-G) respectively.

The C-multicast data traffic is sent on the Selective tree by the
originating PE. When it reaches an ASBR that is on the Inter-AS
segmented tree, it is delivered to local receivers, if any. It is
then forwarded on any inter-AS or intra-AS segments that exist on the
Inter-AS Selective Segmented tree. If the Inter-AS Segmented
Selective Tree is merged onto an Inclusive tree, as described in the
next section, the data traffic is forwarded onto the Inclusive tree.

10.4.2.1.1. Merging Selective Tree into an Inclusive Tree

Consider the situation where:

> + An ASBR is receiving (or expecting to receive) inter-AS (C-S, C-
>   G) data from upstream via a Selective tree.
>
> + The ASBR is sending (or expecting to send) the inter-AS (C-S, C-
>   G) data downstream via an Inclusive tree.

This situation may arise if the upstream providers have a policy of
using Selective trees but the downstream providers have a policy of
using Inclusive trees. To support this situation, an ASBR MAY, under
certain conditions, merge one or more upstream Selective trees into a
downstream Inclusive tree. Note that this can be the case only for
option (b) and not for option (c) as for option (c) the ASBRs do not

have Inclusive tree state.

A Selective tree (corresponding to a particular S-PMSI A-D route) MAY
be merged by a particular ASBR into an Inclusive tree (corresponding
to a particular Inter-AS BGP VPLS A-D route) if and only if the
following conditions all hold:

+ The S-PMSI A-D route and the Inter-AS BGP VPLS A-D route
   originate in the same AS. The Inter-AS BGP VPLS A-D route carries
   the originating AS in the AS_PATH attribute of the route. The S-
   PMSI A-D route carries the originating AS in the AS_PATH
   attribute of the route.

+ The S-PMSI A-D route and the Inter-AS BGP VPLS A-D route have
   exactly the same set of RTs.

An ASBR performs merging by stitching the tail end of the P-tunnel,
as specified in the PMSI Tunnel attribute of the S-PMSI A-D route
received by the ASBR, to the head of the P-tunnel, as specified in
the PMSI Tunnel attribute of the Inter-AS BGP VPLS A-D route re-
advertised by the ASBR.

An ASBR that merges an S-PMSI A-D route into an Inter-AS BGP VPLS A-D
route MUST NOT re-advertise the S-PMSI A-D route.

10.4.3. Inter-AS Non-Segmented Selective trees

Inter-AS Non-segmented Selective trees MAY be used in the case of
option (c).

In this method, there is a multi-hop EBGP peering between the PEs (or
a Route Reflector) in one AS and the PEs (or Route Reflector) in
another AS. The PEs exchange BGP Selective tree A-D routes, along
with PMSI Tunnel attribute, as in the intra-AS case described in
section "Option (c): Non-Segmented Tunnels".

The PEs in different ASes use a non-segmented Selective inter-AS P2MP
tunnel for VPLS multicast.

This method requires no VPLS information (in either the control or
the data plane) on the ASBRs. The ASBRs only need to participate in
the non-segmented P2MP tunnel setup in the control plane, and do MPLS
label forwarding in the data plane.

The data forwarding in this model is the same as in the intra-AS case
described in section "Establishing P-Multicast Trees".

11. BGP Extensions

This section describes the encoding of the BGP extensions required by
this document.

11.1. Inclusive Tree/Selective Tree Identifier

Inclusive P-Multicast tree and Selective P-Multicast tree
advertisements carry the P-Multicast tree identifier.

This document reuses the BGP attribute, called PMSI Tunnel attribute
that is defined in [RFC6514].

This document supports only the following Tunnel Types when PMSI
Tunnel attribute is carried in VPLS A-D or VPLS S-PMSI A-D routes:

   + 0 - No tunnel information present
   + 1 - RSVP-TE P2MP LSP
   + 2 - LDP P2MP LSP
   + 6 - Ingress Replication

11.2. MCAST-VPLS NLRI

This document defines a new BGP NLRI, called the MCAST-VPLS NLRI.

Following is the format of the MCAST-VPLS NLRI:

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

The Route Type field defines encoding of the rest of MCAST-VPLS NLRI
(Route Type specific MCAST-VPLS NLRI).

The Length field indicates the length in octets of the Route Type
specific field of MCAST-VPLS NLRI.

This document defines the following Route Types for A-D routes:

```
    + 3 - Selective Tree A-D route;
    + 4 - Leaf A-D route.
```

   The MCAST-VPLS NLRI is carried in BGP using BGP Multiprotocol
   Extensions [RFC4760] with an AFI of 25 (L2VPN AFI), and an SAFI of
   MCAST-VPLS. The NLRI field in the MP_REACH_NLRI/MP_UNREACH_NLRI
   attribute contains the MCAST-VPLS NLRI (encoded as specified above).

   In order for two BGP speakers to exchange labeled MCAST-VPLS NLRI,
   they must use BGP Capabilities Advertisement to ensure that they both
   are capable of properly processing such NLRI. This is done as
   specified in [RFC4760], by using capability code 1 (multiprotocol
   BGP) with an AFI of 25 and a SAFI of MCAST-VPLS.

   The following describes the format of the Route Type specific MCAST-
   VPLS NLRI for various Route Types defined in this document.

11.2.1. S-PMSI A-D route

   An S-PMSI A-D route type specific MCAST-VPLS NLRI consists of the
   following:

```
                +-----------------------------------+
                |      RD   (8 octets)              |
                +-----------------------------------+
                | Multicast Source Length (1 octet) |
                +-----------------------------------+
                |  Multicast Source (Variable)      |
                +-----------------------------------+
                |  Multicast Group Length (1 octet) |
                +-----------------------------------+
                |  Multicast Group   (Variable)     |
                +-----------------------------------+
                |   Originating Router's IP Addr    |
                +-----------------------------------+
```

   The RD is encoded as described in [RFC4364].

   The Multicast Source field contains the C-S address i.e the address
   of the multicast source. If the Multicast Source field contains an
   IPv4 address, then the value of the Multicast Source Length field is
   32. If the Multicast Source field contains an IPv6 address, then the
   value of the Multicast Source Length field is 128. The value of the
   Multicast Source Length field may be set to 0 to indicate a wildcard.

   The Multicast Group field contains the C-G address i.e. the address
   of the multicast group. If the Multicast Group field contains an IPv4

   address, then the value of the Multicast Group Length field is 32.
   If the Multicast Group field contains an IPv6 address, then the value
   of the Multicast Group Length field is 128. The Multicast Group
   Length field may be set to 0 to indicate a wildcard.

   Whether the Originating Router's IP Address field carries an IPv4 or
   IPv6 address is determined from the value of the Length field of the
   MCAST-VPLS NLRI. If the Multicast Source field contains an IPv4
   address and the Multicast Group field contains an IPv4 address, then
   the value of the Length field is 22 bytes if the Originating Router's
   IP address carries an IPv4 address and 34 bytes if it is an IPv6
   address. If the Multicast Source and Multicast Group fields contain
   IPv6 addresses, then the value of the Length field is 46 bytes if the
   Originating Router's IP address carries an IPv4 address and 58 bytes
   if it is an IPv6 address. The following table summarizes the above.

Multicast Source	Multicast Group	Originating Router's IP Address	Length
IPv4	IPv4	IPv4	22
IPv4	IPv4	IPv6	34
IPv6	IPv6	IPv4	46
IPv6	IPv6	IPv6	58

   Usage of Selective Tree A-D routes is described in section
   "Optimizing Multicast Distribution via Selective Trees".

[11.2.2](11.2.2). Leaf A-D route

   A leaf A-D route type specific MCAST-VPLS NLRI consists of the
   following:

```
                +-----------------------------------+
                |       Route Key (variable)        |
                +-----------------------------------+
                |    Originating Router's IP Addr   |
                +-----------------------------------+
```

   Whether the Originating Router's IP Address field carries an IPv4 or
   IPv6 address is determined from the Length field of the MCAST-VPLS
   NLRI and the length of the Route Key field. From these two length
   fields one can compute the length of the Originating Router's IP
   Address. If this computed length is 4 then the address is an IPv4
   address and if its 16 then the address is an IPv6 address.

   Usage of leaf A-D routes is described in sections "Inter-AS Inclusive

P-Multicast tree A-D/Binding" and "Optimizing Multicast Distribution
via Selective trees".

[12](12). Aggregation Considerations

This document does not specify the mandatory implementation of any
particular set of rules for determining whether or not the Inclusive
or Selective trees of two particular VPLS instances are to be
instantiated by the same Aggregate Inclusive/Selective Tree. This
determination can be made by implementation-specific heuristics, by
configuration, or even perhaps by the use of offline tools.

This section discusses potential methodologies with respect to
aggregation.

In general the heuristic used to decide which VPLS instances or (C-S,
C-G) entries to aggregate is implementation dependent. It is also
conceivable that offline tools can be used for this purpose. This
section discusses some tradeoffs with respect to aggregation.

The "congruency" of aggregation is defined by the amount of overlap
in the leaves of the client trees that are aggregated on an SP tree.
For Aggregate Inclusive trees the congruency depends on the overlap
in the membership of the VPLS instances that are aggregated on the
Aggregate Inclusive tree. If there is complete overlap aggregation is
perfectly congruent. As the overlap between the VPLS instances that
are aggregated reduces, the congruency reduces.

From the above definition of "congruency" it follows that in order
for a given PE to determine the congruency of the client trees that
this PE could aggregate, the PE has to know the leaves of these
client trees. This is irrespective of whether the aggregated SP tree
is established using mLDP or RSVP-TE.

If aggregation is done such that it is not perfectly congruent a PE
may receive traffic for VPLS instances to which it doesn't belong. As
the amount of multicast traffic in these unwanted VPLS instantes
increases aggregation becomes less optimal with respect to delivered
traffic. Hence there is a tradeoff between reducing multicast state
in the core and delivering unwanted traffic.

An implementation should provide knobs to control aggregation based
on the congruency of the tree to be aggregated. This will allow an SP
to deploy aggregation depending on the VPLS membership and traffic
profiles in its network. If different PEs are setting up Aggregate
Inclusive trees this will also allow an SP to engineer the maximum
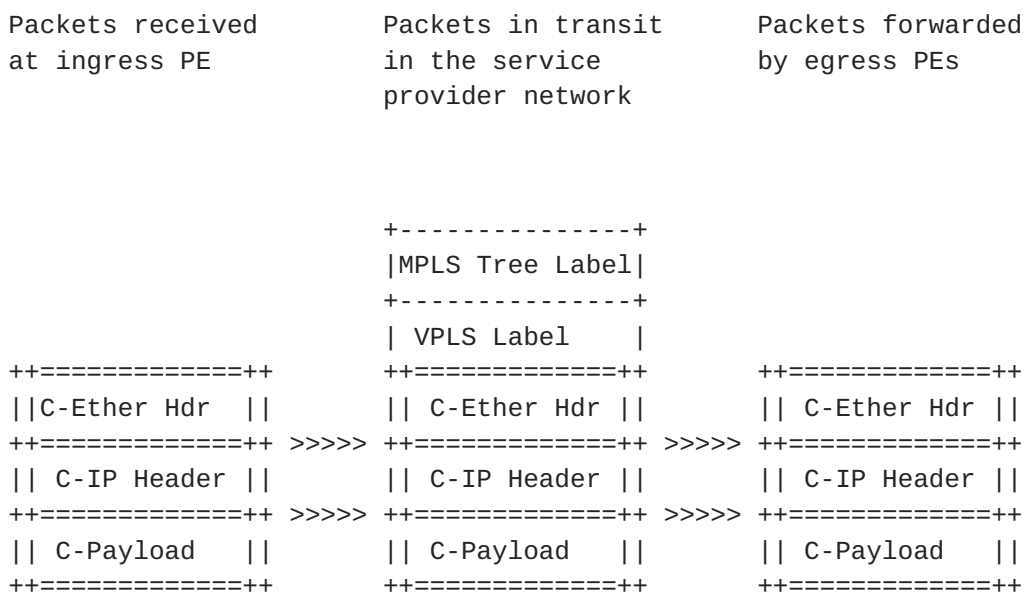amount of unwanted VPLS instances that a particular PE may receive

traffic for.

The state/bandwidth optimality trade-off can be further improved by
having a versatile many-to-many association between client trees and
provider trees. Thus a VPLS instance can be mapped to multiple
Aggregate trees. The mechanisms for achieving this are for further
study. Also it may be possible to use both ingress replication and an
Aggregate tree for a particular VPLS. Mechanisms for achieving this
are also for further study.

13. Data Forwarding

13.1. MPLS Tree Encapsulation

13.1.1. Mapping multiple VPLS instances to a P2MP LSP

The following diagram shows the progression of the VPLS multicast
packet as it enters and leaves the SP network when MPLS trees are
being used for multiple VPLS instances. RSVP-TE P2MP LSPs are
examples of such trees.

```
   Packets received          Packets in transit        Packets forwarded
   at ingress PE             in the service            by egress PEs
                             provider network



                             +---------------+
                             |MPLS Tree Label|
                             +---------------+
                             | VPLS Label    |
   ++=============++         ++=============++         ++=============++
   ||C-Ether Hdr  ||         || C-Ether Hdr ||         || C-Ether Hdr ||
   ++=============++ >>>>> ++=============++ >>>>> ++=============++
   || C-IP Header ||         || C-IP Header ||         || C-IP Header ||
   ++=============++ >>>>> ++=============++ >>>>> ++=============++
   || C-Payload   ||         || C-Payload   ||         || C-Payload   ||
   ++=============++         ++=============++         ++=============++
```

When an ingress PE receives a packet, the ingress PE using the
procedures defined in [RFC4261] [RFC4262] determines the VPLS
instance associated with the packet. If the packet is an IP multicast
packet, and the ingress PE uses an Aggregate Selective tree for the
(C-S, C-G) carried in the packet, then the ingress PE pushes the VPLS
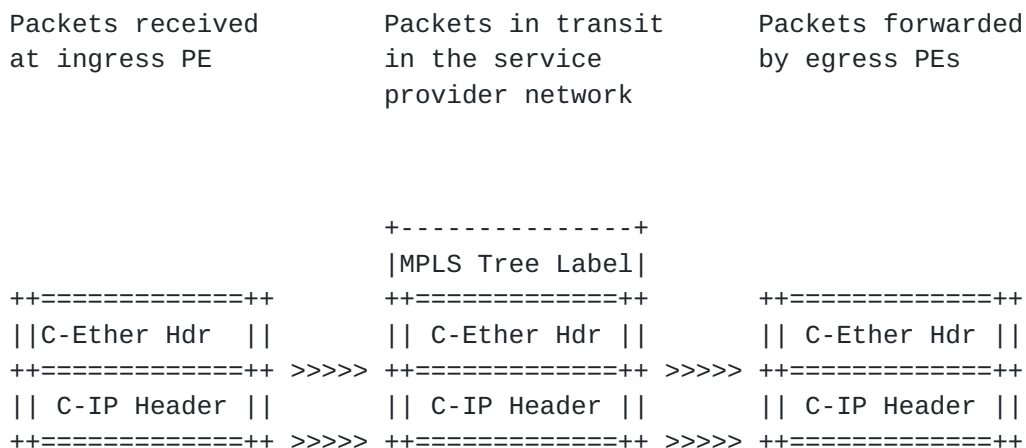Label associated with the VPLS instance on the ingress PE and the

MPLS Tree Label associate with the Aggregate Selective tree, and
sends the packet over the P2MP LSP associated with the Aggregate
Selective tree. Otherwise, if the ingress PE does not use an
Aggregate Selective tree for the (C-S, C-G), or the packet is either
non-IP multicast or broadcast, the ingress PE pushes the VPLS label
associated with the VPLS instance on the ingress PE and the MPLS Tree
Label associated with the Aggregate Inclusive tree, and sends the
packet over the P2MP LSP associated with the Aggregate Inclusive
tree.

The egress PE does a lookup on the outer MPLS tree label, and
determines the MPLS forwarding table in which to lookup the inner
MPLS label (VPLS label). This table is specific to the tree label
space (as identified by the MPLS Tree label). The inner label (VPLS
label) is unique within the context of the root of the tree (as it is
assigned by the root of the tree, without any coordination with any
other nodes). Thus it is not unique across multiple roots. So, to
unambiguously identify a particular VPLS one has to know the VPLS
label, and the context within which that label is unique. The context
is provided by the outer MPLS label (MPLS Tree label) [RFC5331].

The outer MPLS label is popped. The lookup of the resulting MPLS
label determines the VSI in which the egress PE needs to do the C-
multicast data packet lookup. It then pops the inner MPLS label and
sends the packet to the VSI for multicast data forwarding.

13.1.2. Mapping one VPLS instance to a P2MP LSP

The following diagram shows the progression of the VPLS multicast
packet as it enters and leaves the SP network when a given MPLS tree
is being used for a single VPLS instance. RSVP-TE P2MP LSPs are
examples of such trees.

```
   Packets received         Packets in transit        Packets forwarded
   at ingress PE            in the service            by egress PEs
                            provider network



                            +---------------+
                            |MPLS Tree Label|
   ++============++         ++============++         ++============++
   ||C-Ether Hdr ||         || C-Ether Hdr ||        || C-Ether Hdr ||
   ++============++ >>>>>    ++============++ >>>>>   ++============++
   || C-IP Header ||        || C-IP Header ||        || C-IP Header ||
   ++============++ >>>>>    ++============++ >>>>>   ++============++
```

```
    || C-Payload   ||        || C-Payload   ||        || C-Payload   ||
    ++=============++        ++=============++        ++=============++
```

When an ingress PE receives a packet, the ingress PE using the
procedures defined in [RFC4261] [RFC4262] determines the VPLS
instance associated with the packet. If the packet is an IP multicast
packet, and the ingress PE uses a Selective tree for the (C-S, C-G)
carried in the packet, then the ingress PE pushes the MPLS Tree Label
associate with the Selective tree, and sends the packet over the P2MP
LSP associated with the Selective tree. Otherwise, if the ingress PE
does not use a Selective tree for the (C-S, C-G), or the packet is
either non-IP multicast or broadcast, the ingress PE pushes the MPLS
Tree Label associated with the Inclusive tree, and sends the packet
over the P2MP LSP associated with the Inclusive tree.

The egress PE does a lookup on the MPLS tree label and determines the
VSI in which the receiver PE needs to do the C-multicast data packet
lookup. It then pops the MPLS label and sends the packet to the VSI
for multicast data forwarding.

14. VPLS Data Packet Treatment

If the destination MAC address of a VPLS packet received by an
ingress PE from a VPLS site is a multicast address, a P-Multicast
tree SHOULD be used to transport the packet, if possible. If the
packet is an IP multicast packet and a Selective tree exists for that
multicast stream, the Selective tree MUST be used. Else if a (C-*,
C-*) Selective tree exists for the VPLS it SHOULD be used. Else if an
Inclusive tree exists for the VPLS, it SHOULD be used.

If the destination MAC address of a VPLS packet is a broadcast
address, it is flooded. If a (C-*, C-*) Selective tree exists for the
VPLS the PE SHOULD flood over it. Else if Inclusive tree exists for
the VPLS the PE SHOULD flood over it. Else the PE MUST flood the
packet using the procedures in [RFC4761] or [RFC4762].

If the destination MAC address of a packet is a unicast address and
it has not been learned, the packet MUST be sent to all PEs in the
VPLS. Inclusive P-Multicast trees or a Selective P-Multicast tree
bound to (C-*, C-*) SHOULD be used for sending unknown unicast MAC
packets to all PEs. When this is the case the receiving PEs MUST
support the ability to perform MAC address learning for packets
received on a multicast tree. In order to perform such learning, the
receiver PE MUST be able to determine the sender PE when a VPLS
packet is received on a P-Multicast tree. This further implies that

the MPLS P-Multicast tree technology MUST allow the egress PE to
determine the sender PE from the received MPLS packet.

When a receiver PE receives a VPLS packet with a source MAC address,
that has not yet been learned, on a P-Multicast tree, the receiver PE
determines the PW to the sender PE. The receiver PE then creates
forwarding state in the VPLS instance with a destination MAC address
being the same as the source MAC address being learned, and the PW
being the PW to the sender PE.

It should be noted that when a sender PE that is sending packets
destined to an unknown unicast MAC address over a P-Multicast tree
learns the PW to use for forwarding packets destined to this unicast
MAC address, it might immediately switch to transport such packets
over this particular PW. Since the packets were initially being
forwarded using a P-Multicast tree, this could lead to packet
reordering. This constraint should be taken into consideration if
unknown unicast frames are forwarded using a P-Multicast tree,
instead of multiple PWs based on [RFC4761] or [RFC4762].

An implementation SHOULD support the ability to transport unknown
unicast traffic over Inclusive P-Multicast trees. Furthermore, an
implementation MUST support the ability to perform MAC address
learning for packets received on a P-Multicast tree.

15. Security Considerations

Security considerations discussed in [RFC4761] and [RFC4762] apply to
this document. This section describes additional considerations.

As mentioned in [RFC4761], there are two aspects to achieving data
privacy and protect against denial-of-service attacks in a VPLS:
securing the control plane and protecting the forwarding path.
Compromise of the control plane could result in a PE sending
multicast data belonging to some VPLS to another VPLS, or black-
holing VPLS multicast data, or even sending it to an eavesdropper;
none of which are acceptable from a data privacy point of view. In
addition, compromise of the control plane could result in black-
holing VPLS multicast data.

The mechanisms in this document use BGP for the control plane. Hence
techniques such as in [RFC5925] help authenticate BGP messages,
making it harder to spoof updates (which can be used to divert VPLS
traffic to the wrong VPLS) or withdrawals (denial-of-service
attacks). In the multi-AS methods (b) and (c) described in section
"Inter-AS Inclusive P-Multicast Tree A-D/Binding" this also means
protecting the inter-AS BGP sessions, between the ASBRs, the PEs, or

the Route Reflectors.

Note that [RFC5925] will not help in keeping MPLS labels, associated
with P2MP LSPs or the upstream MPLS labels used for aggregation,
private -- knowing the labels, one can eavesdrop on VPLS traffic.
However, this requires access to the data path within a Service
Provider network, which is assumed to be composed of trusted
nodes/links.

One of the requirements for protecting the data plane is that the
MPLS labels are accepted only from valid interfaces. This applies
both to MPLS labels associated with P2MP LSPs and also applies to the
upstream assigned MPLS labels. For a PE, valid interfaces comprise
links from other routers in PE's own AS. For an ASBR, valid
interfaces comprise links from other routers in ASBR's own AS, and
links from other ASBRs in ASes that have instances of a given VPLS.
It is especially important in the case of multi-AS VPLS instances
that one accept VPLS packets only from valid interfaces.

16. IANA Considerations

This document defines a new NLRI, called MCAST-VPLS, to be carried in
BGP using multiprotocol extensions. It requires assignment of a new
SAFI. This is to be assigned by IANA.

This document defines a BGP optional transitive attribute, called
PMSI attribute. This is the same attribute as the one defined in
[RFC6514] and the code point for this attribute has already been
assigned by IANA as 22 [BGP-IANA]. Hence no further action is
required from IANA regarding this attribute.

17. Acknowledgments

Many thanks to Thomas Morin for his support of this work.

We would also like to thank authors of [RFC6514] and [RFC6513], as
the details of the inter-AS segmented tree procedures in this
document, as well as some text that describes these procedures have
benefited from those in [RFC6514] and [RFC6513]. The same applies to
the notion of Inclusive and Selective trees, as well as the
procedures for switching from Inclusive to Selective trees.

We would also like to thank Nabil Bitar, Stewart Bryant, Wim
Henderickx, and Eric Rosen for their review and comments.

18. Normative References

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

[RFC3036] L. Andersson, et. al., "LDP Specification", RFC 3036, January 2001

[RFC3209] D. Awduche, et. al., "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001

[RFC4760] T. Bates, et. al., "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.

[RFC4761] K. Kompella, Y. Rekther, "Virtual Private LAN Service using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.

[RFC4762] M. Lasserre, V. Kompella, "Virtual Private LAN Services over MPLS Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.

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

[RFC6511] Z. Ali, G. Swallow, R. Aggarwal, "Non PHP behavior and out-of-band mapping for RSVP-TE LSPs", RFC 6511, February 2012.

[RFC6512] IJ. Wijnands, et. al., "Using Multipoint LDP When the Backbone Has No Route to the Root", RFC6512, February 2012

19. Informative References

[RFC6514] R. Aggarwal, E. Rosen, Y. Rekhter, T. Morin, "BGP Encodings for Multicast in MPLS/BGP IP VPNs". RFC 6514, February 2012.

[RFC6513] E. Rosen, R. Aggarwal, "Multicast in MPLS/BGP IP VPNs", RFC 6513, February 2012.

[RFC6388] I. Minei et. al, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", RFC 6388, November 2011.

[RFC6074] E. Rosen et. al., "Provisioning, Autodiscovery, and Signaling in L2VPNs", RFC 6074, January 2011.

[RFC5501] Y. kamite, et. al., "Requirements for Multicast Support in

Virtual Private LAN Services", RFC 5501, January 2009.

[RFC5332] T. Eckert, E. Rosen, R. Aggarwal, Y. Rekhter, "MPLS
Multicast Encapsulations", RFC 5332, August 2008.

[RFC4875] R. Aggarwal et. al, "Extensions to RSVP-TE for Point to
Multipoint TE LSPs", RFC 4857, May 2007.

[RFC4684] P. Marques et. al., "Constrained Route Distribution for
Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS)
Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684,
November 2006.

[RFC4601] B. Fenner, et. al., "PIM-SM Protocol Specification", RFC
4601, August 2006.

[RFC4541] M. Christensen et. al., "Considerations for Internet Group
Management Protocol (IGMP) and Multicast Listener Discovery (MLD)
Snooping Switches", RFC 4541, May 2006.

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

[RFC4364] "BGP MPLS VPNs", E. Rosen, Y.Rekhter, RFC4364, February
2006.

[RFC5925] J. Touch, et. al., "The TCP Authentication Option",
RFC5925, June 2010.

[RFC1997] R. Chandra, et. al., "BGP Communities Attribute", RFC 1997,
August 1996.

[MULTI-HOMING] K. Kompella et. al., "Multi-homing in BGP-based
Virtual Private LAN Service", draft-ietf-l2vpn-vpls-multihoming Work
in Progress.

[BGP-IANA] http://www.iana.org/assignments/bgp-parameters

20. Author's Address

Rahul Aggarwal

998 Lucky Avenue
Menlo Park, CA 94025
USA
Phone: +1-415-806-5527
Email: raggarwa_1@yahoo.com

Yuji Kamite
NTT Communications Corporation
Tokyo Opera City Tower
3-20-2 Nishi Shinjuku, Shinjuku-ku,
Tokyo 163-1421,
Japan

Email: y.kamite@ntt.com

Luyuan Fang
Cisco Systems
111 Wood Avenue South
Iselin, NJ 08830
USA

Email: lufang@cisco.com

Yakov Rekhter
Juniper Networks
1194 North Mathilda Ave.
Sunnyvale, CA 94089
USA

Email: yakov@juniper.net

Chaitanya Kodeboniya
Cisco Systems
Email: ckodeboy@cisco.com