L2VPN Working Group
Intended Status: Proposed Standard
Internet Draft
Expires: July 10, 2012

Himanshu Shah(Ciena)
Eric Rosen(Cisco)
Giles Heron(Cisco)
Vach Kompella(Alcatel-Lucent)

January 10 2012

ARP Mediation for IP Interworking of Layer 2 VPN
draft-ietf-l2vpn-arp-mediation-19.txt

Status of this Memo

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

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

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

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

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

This Internet-Draft will expire on July 10, 2012

Copyright Notice

Abstract

The Virtual Private Wire Service (VPWS) [RFC4664] provides point-to-point connections between pairs of Customer Edge (CE) devices. It does so by binding two Attachment Circuits (each connecting a CE device with a Provider Edge, PE, device) to a pseudowire (connecting the two PEs). In general, the Attachment Circuits must be of the same technology (e.g., both Ethernet, both ATM), and the pseudowire must carry the frames of that technology. However, if it is known that the frames' payload consists solely of IP datagrams, it is possible to provide a point-to-point connection in which the pseudowire connects Attachment Circuits of different technologies. This requires the PEs to perform a function known as "ARP Mediation". ARP Mediation refers to the process of resolving Layer 2 addresses when different resolution protocols are used on either Attachment Circuit. The methods described in this document are applicable even when the CEs run a routing protocol between them, as long as the routing protocol runs over IP.

Conventions used in this document

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].

Table of Contents

1. Introduction

Layer 2 Virtual Private Networks (L2VPN) are constructed over a Service Provider IP/MPLS backbone but are presented to the Customer Edge (CE) devices as Layer 2 networks. In theory, L2VPNs can carry any Layer 3 protocol, but in many cases, the Layer 3 protocol is IP. Thus it makes sense to consider procedures that are optimized for IP.

In a typical implementation, illustrated in the diagram below, the CE devices are connected to the Provider Edge (PE) devices via Attachment Circuits (AC). The ACs are Layer 2 circuits. In a pure L2VPN, if traffic sent from CE1 via AC1 reaches CE2 via AC2, both ACs would have to be of the same type (i.e., both Ethernet, both Frame Relay, etc.). However, if it is known that only IP traffic will be carried, the ACs can be of different technologies, provided that the PEs provide the appropriate procedures to allow the proper transfer of IP packets.

```
                                      +-----+
                          +------ -----| CE3 |
                          |AC3         +-----+
                       +-----+
                 ......| PE3 |...........
                 .     +-----+           .
                 .        |              .
                 .        |              .
+-----+ AC1 +-----+    Service      +-----+ AC2 +-----+
| CE1 |-----| PE1 |--- Provider ----| PE2 |-----| CE2 |
+-----+     +-----+    Backbone     +-----+     +-----+
                 .                      .
                 ........................
```

A CE, which is connected via a given type of AC, may use an IP Address Resolution procedure that is specific to that type of AC. For example, an Ethernet-attached IPv4 CE would use ARP [RFC826] and a Frame Relay-attached CE might use Inverse ARP [RFC2390]. If we are to allow the two CEs to have a Layer 2 connection between them, even though each AC uses a different Layer 2 technology, the PEs must intercept and "mediate" the Layer 2 specific address resolution procedures.

In this document, we specify the procedures for VPWS services, which the PEs MUST implement in order to mediate the IP address resolution mechanism. We call these procedures "ARP Mediation". Consider a Virtual Private Wire Service (VPWS) constructed between CE1 and CE2 in the diagram above. If AC1 and AC2 are of different technologies, e.g. AC1 is Ethernet and AC2 is Frame Relay (FR), then ARP requests coming from CE1 cannot be passed transparently to CE2. PE1 MUST interpret the meaning of the ARP requests and mediate the necessary information with PE2 before responding.

The document uses "ARP" terminology to mean any protocol that is used to resolve IP addresses to link layer addresses. For instance in IPv4, ARP and Inverse ARP protocols are used for address resolution while in IPv6 Neighbor Discovery [RFC4861] and Inverse Neighbor Discovery protocol [RFC3122] based on ICMPv6 are used for address resolution.

2. ARP Mediation (AM) function

The ARP Mediation (AM) function is an element of a PE node that deals with the IP address resolution for CE devices connected via a VPWS L2VPN. By placing this function in the PE node, ARP Mediation is transparent to the CE devices.

For a given point-to-point connection between a pair of CEs, the ARP Mediation procedure depends on whether the packets being forwarded are IPv4 or IPV6. A PE that is to perform ARP Mediation for IPv4 packets MUST perform the following logical steps:

1. Discover the IP address of the locally attached CE device
2. Terminate, do not forward ARP and Inverse ARP requests from the CE device at the local PE.
3. Distribute the IP Address to the remote PE using pseudowire control signaling.
4. Notify the locally attached CE of the IP address of the remote CE.
5. Respond appropriately to ARP and Inverse ARP requests from the local CE device, using IP address of the remote CE and the hardware address of the local PE.

A PE that is to perform ARP Mediation for IPv6 packets SHOULD perform the following logical steps:

1. Discover the IPv6 addresses of the locally attached CE device, together with those of the remote CE device.
2.
    a. Intercept Neighbor Discovery (ND) and Inverse Neighbor Discovery (IND) packets received from the local CE device.
    b. From these NB and IND packets learn the IPv6 configuration of the CE.

   c. Forward the ND and IND packets over the pseudowire to the remote PE.

3. Intercept Neighbor Discovery and Inverse Neighbor Discovery packets received over the pseudowire from the remote PE, possibly modifying them (if required for the type of outgoing AC) before forwarding to the local CE, and also learning information about the IPv6 configuration of the remote CE.

Details for the above-described procedures are given in the following sections.

3. IP Layer 2 Interworking Circuit

The IP Layer 2 interworking Circuit refers to interconnection of the Attachment Circuit with the IP Layer 2 Transport pseudowire that carries IP datagrams as the payload. The ingress PE removes the data link header of its local Attachment Circuit and transmits the payload (an IP packet) over the pseudowire with or without the optional control word. If the IP packet arrives at the ingress PE with multiple data link headers (for example in the case of bridged Ethernet PDU on an ATM Attachment Circuit), all data link headers MUST be removed from the IP packet before transmission over the PW. The egress PE encapsulates the IP packet with the data link header used on its local Attachment Circuit.

The encapsulation for the IP Layer 2 Transport pseudowire is described in [RFC4447]. The "IP Layer 2 interworking circuit" pseudowire is also referred to as "IP pseudowire" in this document.

In the case of an IPv6 L2 Interworking Circuit, the egress PE MAY modify the contents of Neighbor Discovery or Inverse Neighbor Discovery packets before encapsulating the IP packet with the data link header.

4. IP Address Discovery Mechanisms

An IP Layer 2 Interworking Circuit enters monitoring state immediately after configuration. During this state it performs two functions.

- Discovery of the CE IP device(s)
- Establishment of the PW

The establishment of the PW occurs independently from local CE IP address discovery. During the period when the PW has been established but the local CE IP device has not been discovered, only broadcast/multicast IP frames are propagated between the Attachment Circuit and pseudowire; unicast IP datagrams are dropped. The IP destination address is used to classify unicast/multicast packets.

Unicast IP frames are propagated between the AC and pseudowire only when CE IP devices on both Attachment Circuits have been discovered, notified and proxy functions have completed.

The need to wait for address resolution completion before unicast IP traffic can flow is simple.

. PEs do not perform routing operations
. The destination IP address in the packet is not necessarily that of the attached CE
. On a broadcast link, there is no way to find out the MAC address of the CE based on the Destination IP address of the packet.

4.1. Discovery of IP Addresses of Locally Attached IPv4 CE

A PE MUST support manual configuration of IPv4 CE addresses. This section also describes automated mechanisms by which a PE MAY also discover an IPv4 CE address.

4.1.1. Monitoring Local Traffic

The PE devices MAY learn the IP addresses of the locally attached CEs from any IP traffic, such as link local multicast packets (e.g., destined to 224.0.0.x), and are not restricted to the operations below.

4.1.2. CE Devices Using ARP

If a CE device uses ARP to determine the IP address to MAC address binding of its neighbor, the PE processes the ARP requests to learn the IP address of the local CE for the local Attachment Circuit.

This document mandates that there MUST be only one CE per Attachment Circuit. However, customer facing access topologies may exist whereby more than one CE appears to be connected to the PE on a single Attachment Circuit. For example, this could be the case when CEs are connected to a shared LAN that connects to the PE. In such case, the PE MUST select one local CE. The selection could be based on manual configuration or the PE MAY optionally use the following selection criteria. In either case, manual configuration of the IP address of the local CE (and its MAC address) MUST be supported.

- o Wait to learn the IP address of the remote CE (through PW signaling) and then select the local CE that is sending the request for IP address of the remote CE.
- o Augment cross checking with the local IP address learned through listening for link local multicast packets (as per section 4.1.1. above).
- o Augment cross checking with the local IP address learned through the Router Discovery protocol (as described below in section 4.1.5.).
- o There is still a possibility that the local PE may not receive an IP address advertisement from the remote PE and there may exist multiple local IP routers that attempt to 'connect' to remote CEs. In this situation, the local PE MAY use some other criteria to select one IP device from many (such as "the first ARP received"), or an operator MAY configure the IP address of the local CE. Note that the operator does not have to configure the IP address of the remote CE (as that would be learned through pseudowire signaling).

Once the local and remote CEs have been discovered for the given Attachment Circuit, the local PE responds with its own MAC address to any subsequent ARP requests from the local CE with a destination IP address matching the IP address of the remote CE.

The local PE signals the IP address of the local CE to the remote PE and MAY initiate an unsolicited ARP response to notify

the IP address to MAC address binding for the remote CE to the local CE (again using its own MAC address).

Once the ARP mediation function is completed (i.e. the PE device knows both the local and remote CE IP addresses), unicast IP frames are propagated between the AC and the established PW.

The PE MAY periodically generate ARP request messages for the IP address of the CE as a means of verifying the continued existence of the IP address and its MAC address binding. The absence of a response from the CE device for a given number of retries could be used as a trigger for withdrawal of the IP address advertisement to the remote PE. The local PE would then re-enter the address resolution phase to rediscover the IP address of the attached CE. Note that this "heartbeat" scheme is needed only where the failure of a CE device may otherwise be undetectable.

4.1.3. CE Devices Using Inverse ARP

If a CE device uses Inverse ARP to determine the IP address of its neighbor, the attached PE processes the Inverse ARP request from the Attachment Circuit and responds with an Inverse ARP reply containing the IP address of the remote CE, if the address is known. If the PE does not yet have the IP address of the remote CE, it does not respond, but records the IP address of the local CE and the circuit information. Subsequently, when the IP address of the remote CE becomes available, the PE MAY initiate an Inverse ARP request as a means of notifying the IP address of the remote CE to the local CE.

This is the typical mode of operation for Frame Relay and ATM Attachment Circuits. If the CE does not use Inverse ARP, the PE can still discover the IP address of the local CE using the mechanisms described in section 4.1.1. and 4.1.5.

4.1.4. CE Devices Using PPP

The IP Control Protocol [RFC1332] describes a procedure to establish and configure IP on a point-to-point connection, including the negotiation of IP addresses. When such an Attachment Circuit is configured for IP interworking, PPP negotiation is not performed end-to-end between CE devices. Instead, PPP negotiation takes place between the CE and its local PE. The PE performs proxy PPP negotiation and informs the

attached CE of the IP address of the remote CE during IPCP negotiation using the IP-Address option (0x03).

When a PPP link completes LCP negotiations, the local PE MAY perform the following IPCP actions:

- The PE learns the IP address of the local CE from the Configure-Request received with the IP-Address option (0x03). If the IP address is non-zero, the PE records the address and responds with Configure-Ack. However, if the IP address is zero, the PE responds with Configure-Reject (as this is a request from the CE to assign it an IP address). Also, the IP address option is set with zero value in the Configure-Reject response to instruct the CE not to include that option in any subsequent Configure-Request.
- If the PE receives a Configure-Request without the IP-Address option, it responds with a Configure-Ack. In this case the PE is unable to learn the IP address of the local CE using IPCP and hence MUST rely on other means as described in sections 4.1.1. and 4.1.5. Note that in order to employ other learning mechanisms, the IPCP negotiations MUST have reached the open state.
- If the PE does not know the IP address of the remote CE, it sends a Configure-Request without the IP-Address option.
- If the PE knows the IP address of the remote CE, it sends a Configure-Request with the IP-Address option containing the IP address of the remote CE.

The IPCP IP-Address option MAY be negotiated between the PE and the local CE device. Configuration of other IPCP options MAY be rejected. Other NCPs, with the exception of the Compression Control Protocol (CCP) and Encryption Control Protocol (ECP), MUST be rejected. The PE device MAY reject configuration of the CCP and ECP.

4.1.5. Router Discovery method

In order to learn the IP address of the CE device for a given Attachment Circuit, the PE device MAY execute Router Discovery Protocol [RFC1256] whereby a Router Discovery Request (ICMP -

router solicitation) message is sent using a source IP address of zero. The IP address of the CE device is extracted from the Router Discovery Response (ICMP - router advertisement) message from the CE. It is possible that the response contains more than one router addresses with the same preference level; in which case, some heuristics (such as first on the list) are necessary. The use of the Router Discovery method by the PE is optional.

4.1.6. Manual Configuration

In some cases, it may not be possible to discover the IP address of the local CE device using the mechanisms described in sections 4.1 - 4.1.5 above. In such cases manual configuration MAY be used. All implementations of this document MUST support manual configuration of the IPv4 address of the local CE. This is the only REQUIRED mode for a PE to support.

The support for configuration of the IP address of the remote CE is OPTIONAL.

4.2. How a CE Learns the IPv4 address of a remote CE

Once the local PE has received the IP address information of the remote CE from the remote PE, it will either initiate an address resolution request or respond to an outstanding request from the attached CE device.

In the event that IPv4 address of the remote CE is manually configured, the address resolution can begin immediately as receipt of remote IP address of the CE becomes unnecessary.

4.2.1. CE Devices Using ARP

When the PE learns the IP address of the remote CE as described in section 5.1 below, it may or may not already know the IP address of the local CE. If the IP address is not known, the PE MUST wait until it is acquired through one of the methods described in sections 4.1.1, 4.1.2 and 4.1.5. If the IP address of the local CE is known, the PE MAY choose to generate an unsolicited ARP message to notify the local CE about the binding of the IP address of the remote CE with the PE's own MAC address.

When the local CE generates an ARP request, the PE MUST proxy the ARP response [RFC925] using its own MAC address as the source hardware address and the IP address of the remote CE as the source protocol address. The PE MUST respond only to those ARP requests whose destination protocol address matches the IP address of the remote CE.

4.2.2. CE Devices Using Inverse ARP

When the PE learns the IP address of the remote CE, it SHOULD generate an Inverse ARP request. If the Attachment Circuit requires activation (e.g. Frame Relay) the PE SHOULD activate it first before the Inverse ARP request. It should be noted, that the PE might never receive the response to its own request, nor see any Inverse ARP request from the CE, in cases where the CE is pre-configured with the IP address of the remote CE or where the use of Inverse ARP has not been enabled. In either case the CE has used other means to learn the IP address of its neighbor.

4.2.3. CE Devices Using PPP

When the PE learns the IP address of the remote CE, it SHOULD initiate a Configure-Request and set the IP-Address option to the IP address of the remote CE to notify the IP address of the remote CE to the local CE.

4.3. Discovery of IP Addresses of IPv6 CE Devices

4.3.1. Distinguishing Factors Between IPv4 and IPv6

IPv4 uses ARP and inverse ARP to resolve IP address and link layer associations. Since these are dedicated address resolution protocols, and not IP packets, they cannot be carried on an IP pseudowire. They MUST be processed locally and the IPv4 address information they carry signaled between the PEs using the pseudowire control plane. IPv6 uses ICMPv6 extensions to resolve IP address and link address associations. As these are IPv6

packets they can be carried on an IP pseudowire and therefore no IPv6 address signaling is required.

4.3.2. Requirements for PEs

A PE device that supports IPv6 MUST be capable of,

- Intercepting ICMPv6 Neighbor Discovery [RFC4861] and Inverse Neighbor Discovery [RFC3122] packets received over the AC as well as over the PW.
- Recording the IPv6 interface addresses and CE link-layer addresses present in these packets
- Possibly modifying these packets as dictated by the data link type of the egress AC (described in the following sections), and
- Forwarding them towards the original destination

The PE MUST also be capable of generating packets in order to interwork between Neighbor Discovery (ND) and Inverse Neighbor Discovery (IND). This is specified in Sections 4.3.3 to 4.3.6 below.

If an IP PW is used to interconnect CEs that use IPv6 Router Discovery [RFC4861], a PE device MUST also be capable of intercepting and processing those Router Discovery packets. This is required in order to translate between different link layer addresses. If a Router Discovery message contains a link layer address, then the PE MAY also use this message to discover the link layer address and IPv6 interface address. This is described in more detail in Section 4.3.7 and Section 4.3.8.

The PE device MUST learn a list of CE IPv6 interface addresses for its directly-attached CE and another list of CE IPv6 interface addresses for the far-end CE. The PE device MUST also learn the link-layer address of the local CE and be able to use it when forwarding traffic between the local and far-end CEs. The PE MAY also wish to monitor the source link-layer address of data packets received from the CE, and discard packets not matching its learned CE link-layer address.

4.3.3. Processing of Neighbor Solicitations

A Neighbor Solicitation received on an AC from a local CE SHOULD be inspected to determine and learn an IPv6 interface address (if provided, this will not be the case for Duplicate Address Detection) and any link-layer address provided. The packet MUST then be forwarded over the pseudowire unmodified. A Neighbor Solicitation received over the pseudowire SHOULD be inspected to determine and learn an IPv6 interface address for the far-end CE. If a source link-layer address option is present, the PE MUST remove it. The PE MAY substitute an appropriate link-layer address option, specifying the link-layer address of the PE interface attached to the local AC. Note that if the local AC is Ethernet, failure to substitute a link-layer address option may mean that the CE has no valid link-layer address with which to transmit data packets.

When a PE with a local AC, which is of the type point-to-point layer 2 circuit e.g. FR, ATM or PPP, receives a Neighbor Solicitation from a far end PE over the pseudowire, after learning the IP address of the far-end CE, the PE MAY use one of the following procedures:

1. Forward the Neighbor Solicitation to the local CE after replacing the source link-layer address with the link-layer address of the local AC.
2. Send an Inverse Neighbor Solicitation to the local CE, specifying the far-end CE's IP address and the link-layer address of the local AC.
3. Reply to the far end PE with a Neighbor Advertisement, using the IP address of the local CE as the source address and an appropriate link-layer address option that specifies the link-layer address of the local AC. As described later, the IP address of the local CE is learned through IPv6CP in the case of PPP and through Neighbor Solicitation in other cases.

4.3.4. Processing of Neighbor Advertisements

A Neighbor Advertisement received on an AC from a local CE SHOULD be inspected to determine and learn an IPv6 interface address and any link-layer address provided. The packet MUST then be forwarded over the IP pseudowire unmodified.

A Neighbor Advertisement received over the pseudowire SHOULD be inspected to determine and learn an IPv6 interface address for the far-end CE. If a source link-layer address option is present, the PE MUST remove it. The PE MAY substitute an appropriate link-layer address option, specifying the link-layer address of the local AC. Note that if the local AC is Ethernet, failure to substitute a link-layer address option may mean that the local AC has no valid link-layer address with which to transmit data packets.

When a PE with a local AC which is of the type point-to-point layer 2 circuit, such as ATM, FR or PPP, receives a Neighbor Advertisement over the pseudowire, in addition to learning the remote CE's IPv6 address, it SHOULD perform the following steps:

- If the AC supports Inverse Neighbor Discovery (IND) and the PE had already processed an Inverse Neighbor Solicitation (INS) from local CE, it SHOULD send an Inverse Neighbor Advertisement (INA) on the local AC using source IP address information received in ND-ADV and its own local AC link layer information.
- If the PE has not received any Inverse Neighbor Solicitation (INS) from the local CE, and the AC supports Inverse Neighbor Discovery (IND), it SHOULD send an INS on the local AC using source IP address information received in the INA together with its own local AC link layer information.

4.3.5. Processing Inverse Neighbor Solicitations (INS)

An INS received on an AC from a local CE SHOULD be inspected to determine and learn the IPv6 addresses and the link-layer addresses. The packet MUST then be forwarded over the pseudowire unmodified.

An INS received over the pseudowire SHOULD be inspected to determine and learn one or more IPv6 addresses for the far-end CE. If the local AC supports IND (e.g., a switched Frame Relay

AC), the packet SHOULD be forwarded to the local CE, after modifying the link-layer address options to match the type of the local AC.

If the local AC does not support IND, processing of the packet depends on whether the PE has learned at least one interface address for its directly-attached CE.

. If it has learned at least one IPv6 address for the CE, the PE MUST discard the Inverse Neighbor Solicitation (INS) and generate an Inverse Neighbor Advertisement (INA) back into the pseudowire. The destination address of the INA is the source address from the INS, the source address is one of the local CE's interface addresses, and all the local CE's interface addresses that have been learned so far SHOULD be included in the Target Address List. The Source and Target Link-Layer addresses are copied from the INS. In addition, the PE SHOULD generate ND advertisements on the local AC using the IPv6 address of the remote CE and link-layer address of the local PE.

. If it has not learned at least one IPv6 and link-layer address of its directly-connected CE, the INS MUST be continued to be discarded until the PE learns an IPv6 and link-layer address from the local CE (through receiving, for example, a Neighbor Solicitation). After this has occurred, the PE will be able to respond to INS messages received over the pseudowire as described above.

4.3.6. Processing of Inverse Neighbor Advertisements (INA)

An INA received on an AC from a local CE SHOULD be inspected to determine and learn one or more IPv6 addresses for the CE. It MUST then be forwarded unmodified over the pseudowire.

An INA received over the pseudowire SHOULD be inspected to determine and learn one or more IPv6 addresses for the far-end CE.

If the local AC supports IND (e.g., a Frame Relay AC), the packet MAY be forwarded to the local CE, after modifying the link-layer address options to match the type of the local AC.

If the local AC does not support IND, the PE MUST discard the INA and generate a Neighbor Advertisement (NA) towards its local CE. The source IPv6 address of the NA is the source IPv6 address

from the INA, the destination IPv6 address is the destination IPv6 address from the INA and the link-layer address is that of the local AC on the PE.

4.3.7. Processing of Router Solicitations

A Router Solicitation received on an AC from a local CE SHOULD be inspected to determine and learn an IPv6 address for the CE, and, if present, the link-layer address of the CE. It MUST then be forwarded unmodified over the pseudowire.

A Router Solicitation received over the pseudowire SHOULD be inspected to determine and learn an IPv6 address for the far-end CE. If a source link-layer address option is present, the PE MUST remove it. The PE MAY substitute a source link-layer address option specifying the link-layer address of its local AC. The packet is then forwarded to the local CE.

4.3.8. Processing of Router Advertisements

A Router Advertisement received on an AC from a local CE SHOULD be inspected to determine and learn an IPv6 address for the CE, and, if present, the link-layer address of the CE. It MUST then be forwarded unmodified over the pseudowire.

A Router Advertisement received over the pseudowire SHOULD be inspected to determine and learn an IPv6 address for the far-end CE. If a source link-layer address option is present, the PE MUST remove it. The PE MAY substitute a source link-layer address option specifying the link-layer address of its local AC. If an MTU option is present, the PE MAY reduce the specified MTU if the MTU of the pseudowire is less than the value specified in the option. The packet is then forwarded to the local CE.

4.3.9. Duplicate Address Detection

Duplicate Address Detection [RFC4862] allows IPv6 hosts and routers to ensure that the addresses assigned to interfaces are unique on a link. As with all Neighbor Discovery packets, those used in Duplicate Address Detection will simply flow through the pseudowire, being inspected at the PEs at each end, processing

is performed as above. However, the source IPv6 address of Neighbor Solicitations used in Duplicate Address Detection is the unspecified address, so the PEs cannot learn the CE's IPv6 interface address (nor would it make sense to do so, given that at least one address is tentative at that time).

4.3.10. CE address discovery for CEs attached using PPP

The IPv6 Control Protocol (IPv6CP) [RFC5072] describes a procedure to establish and configure IPv6 on a point-to-point connection, including the negotiation of a link-local interface identifier. As in the case of IPv4, when such an AC is configured for IP interworking, PPP negotiation is not performed end-to-end between CE devices. Instead, PPP negotiation takes place between the CE and its local PE. The PE performs proxy PPP negotiation and informs the attached CE of the link-local identifier of its local interface using the Interface-Identifier option (0x01). This local interface identifier is used by stateless address auto configuration [RFC4862].

When a PPP link completes IPv6CP negotiations and the PPP link is open, a PE MAY discover the IPv6 unicast address of the CE using any of the mechanisms described above.

5. CE IPv4 Address Signaling between PEs

5.1. When to Signal an IPv4 address of a CE

A PE device advertises the IPv4 address of the attached CE only when the encapsulation type of the pseudowire is IP Layer2 Transport (the value 0x0000B, as defined in [RFC4446]). The IP Layer2 transport PW is also referred to as IP PW and is used interchangeably in this document. It is quite possible that the IPv4 address of a CE device is not available at the time the PW labels are signaled. For example, in Frame Relay the CE device sends an inverse ARP request only when the DLCI is active. If the PE signals the DLCI to be active only when it has received the IPv4 address along with the PW FEC from the remote PE, a deadlock situation arises. In order to avoid such problems, the PE MUST be prepared to advertise the PW FEC before the IPv4 address of the CE is known and hence uses IPv4 address value zero. When the IPv4 address of the CE device does become

available, the PE re-advertises the PW FEC along with the IPv4 address of the CE.

Similarly, if the PE detects that an IP address of a CE is no longer valid (by methods described above), the PE MUST re-advertise the PW FEC with null IP address to denote the withdrawal of IP address of the CE. The receiving PE then waits for notification of the remote IP address. During this period, propagation of unicast IPv4 traffic is suspended, but multicast IPv4 traffic can continue to flow between the AC and the pseudowire.

If two CE devices are locally attached to the PE on disparate AC types (for example, one CE connected to an Ethernet port and the other to a Frame Relay port), the IPv4 addresses are learned in the same manner as described above. However, since the CE devices are local, the distribution of IPv4 addresses for these CE devices is a local step.

Note that the PEs discover the IPv6 addresses of the remote CE by intercepting Neighbor Discovery and Inverse Neighbor Discovery packets that have been passed in-band through the pseudowire. Hence, there is no need to communicate the IPv6 addresses of the CEs through LDP signaling.

If the pseudowire is carrying both IPv4 and IPv6 traffic, the mechanisms used for IPV6 and IPv4 SHOULD NOT interact. In particular, just because a PE has learned a link-layer address for IPv6 traffic by intercepting a Neighbor Advertisement from its directly-connected CE, it SHOULD NOT assume that it can use that link-layer address for IPv4 traffic until that fact is confirmed by reception of, for example, an IPv4 ARP message from the CE.

5.2. LDP Based Distribution of CE IPv4 Addresses

[RFC4447] uses Label Distribution Protocol (LDP) transport to exchange PW FECs in the Label Mapping message in the Downstream Unsolicited (DU) mode. The PW-FEC comes in two flavors; PWid and Generalized ID FEC elements and has some common fields between them. The discussions below refer to these common fields for IP L2 Interworking encapsulation.

In addition to PW-FEC, this document uses an IP Address List TLV (as defined in [RFC5036]) that is to be included in the optional parameter field of the Label Mapping message when advertising the PW FEC for the IP Layer2 Transport. The use of optional

parameters in the Label Mapping message to extend the attributes of the PW FEC is specified in [RFC4447].

As defined in [RFC4447], when processing a received PW FEC, the PE matches the PW ID and PW type with the locally configured PW ID and PW Type. If there is a match and if the PW Type is IP Layer2 Transport, the PE further checks for the presence of an Address List TLV [RFC5036] in the optional parameter TLVs. The processing of the Address List TLV is as follows.

- o If a PE is configured for an AC to a CE enabled for IPv4 or dual-stack IPv4/IPv6, the PE SHOULD advertise an Address List TLV with address family type of IPv4 address. The PE SHOULD process the IPv4 Address List TLV as described in this document. The PE MUST advertise and process IPv6 capability using the procedures described in Section 6 below.
- o If a PE does not receive any IPv4 address in the Address List TLV it MAY assume IPv4 behavior. The address resolution for IPv4 MUST then depend on local manual configuration. In the case of mis-matched configuration whereby one PE has manual configuration while other does not, the IP address to Link Layer address mapping remains unresolved resulting into unsuccessful propagation of IPv4 traffic to the local CE.
- o If a PE is configured for an AC to a CE enabled for IPv6 only, the PE MUST advertise IPv6 capability using the procedures described in Section 6 below. In addition, by virtue of not setting the manual configuration for IPv4 support, an IPv6 only support is realized.

We use the Address List TLV [RFC5036] to signal the IPv4 address of the local CE. This IP Address List TLV is included in the optional parameter field of the Label Mapping message.

The Address List TLV is only used for IPv4 addresses.

The fields of the IP Address List TLV are set as follows:

Length
: Set to 6 to encompass 2 bytes of Address Family field and 4 bytes of Addresses field (because a single IPv4 address is used).

Address Family

Set to 1 to indicate IPv4 as defined in [RFC5036].

Addresses
Contains a single IPv4 address that is the address of the CE attached to the advertising PE.

The address in the Addresses field is set to all zeros to denote that the advertising PE has not learned the IPv4 address of its local CE. Any non-zero address value denotes the IPv4 address of the advertising PE's attached CE device.

The IPv4 address of the CE is also supplied in the optional parameters field of the LDP Notification message along with the PW FEC. The LDP Notification message is used to signal any change in the status of the CE's IPv4 address.

The encoding of the LDP Notification message is as follows.

```
0                   1                   2                   3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|0|   Notification (0x0001)     |      Message Length           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       Message ID                              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       Status (TLV)                            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                 IP Address List TLV (as defined above)        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                 PWId FEC or Generalized ID FEC                |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

The Status TLV status code is set to 0x0000002C "IP address of CE", to indicate that an IP Address update follows. Since this notification does not refer to any particular message the Message ID field is set to 0.

The PW FEC TLV SHOULD NOT include the interface parameters as they are ignored in the context of this message.

6. IPv6 Capability Advertisement

A 'Stack Capability' Interface Parameter sub-TLV is signaled by the two PEs so that they can agree which network protocol(s) they SHOULD be using. As discussed earlier, the use of Address-List TLV signifies the support for IPv4 stack, so the 'Stack

Capability' sub-TLV is used to indicate whether support for IPv6 stack is required on a given IP PW.

The 'Stack Capability' sub-TLV is part of the interface parameters. The proposed format for the Stack Capability Interface Parameter sub-TLV is as follows:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Parameter ID  |     Length    |       Stack Capability        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Parameter ID = 0x16

Length = 4

The Stack Capability field is a bit field. Only one bit is defined in this document. When bit zero (the least significant bit with bitmask 0x0001) is set, it indicates IPv6 stack capability.

The presence of stack capability TLV is relevant only when the PW type is IP PW. A PE that supports the IPv6 on an IP PW MUST signal the Stack Capability sub-TLV in the initial Label Mapping message for the PW. The PE nodes compare the value advertised by the remote PE with the local configuration and only use a capability which is supported by both.

The behavior of a PE that does not understand an Interface Parameter sub-TLV is specified in section 5.5 of RFC 4447 [RFC4447].

In some deployment scenarios, it may be desirable to take a PW operationally down if there is a mismatch of the Stack Capability between the PEs. In other deployment scenarios, an operator may wish the IP version supported by both PEs to fall-back to IPv4 if one of the PEs does not support IPv6. The following procedures MUST be followed for each of these cases.

6.1. PW Operational Down on Stack Capability Mis-Match

If a PE that supports IPv6 and has not yet sent a Label Mapping, receives an initial Label Mapping message from the far end PE that does not include the 'Stack Capability' sub-TLV, or one is received but it is not set to 'IPv6 Stack Capability' value, then the PE supporting this procedure MUST NOT send a Label Mapping for this PW.

If a PE that supports IPv6 has already sent an initial Label Mapping message for the PW and does not receive a 'Stack Capability' sub-TLV in the Label Mapping message from the far-end PE, or one is received but it is not set to 'IPv6 Stack Capability', the PE supporting this procedure MUST withdraw its PW label with the LDP status code meaning "IP Address type mismatch" (Status Code 0x0000004A). However, subsequently if the configuration was to change at the far-end PE and a 'Stack Capability' sub-TLV in the Label Mapping message is received from the far-end PE, the local PE MUST re-advertise the Label Mapping message for the PW.

6.2. Stack Capability Fall-back

If a PE that supports IPv6 and has not yet sent a Label Mapping, receives an initial Label Mapping from the far end PE that does not include the 'Stack Capability' sub-TLV, or one is received but it is not set to the 'IPv6 Stack Capability' value, then it MAY send a Label Mapping for this PW but MUST NOT include the Stack Capability sub-TLV.

If a PE that supports IPv6 and has already sent a Label Mapping for the PW with the 'Stack Capability' sub-TLV, but does not receive a 'Stack Capability' sub-TLV from the far-end PE in the initial Label Mapping message, or one is received but it is not set to the 'IPv6 Stack Capability' value, the PE following this procedure MUST send a Label Withdraw for its PW label with the LDP status code meaning "Wrong IP Address type" (Status Code 0x000004B) followed by a Label Mapping message that does not include the 'Stack Capability' sub-TLV.
If a Label Withdraw message with the "Wrong IP Address Type" status code is received by a PE, it SHOULD treat this as a normal Label Withdraw, but MUST NOT respond with a Label Release. It MUST continue to wait for the next control message for the PW as specified in section 6.2 of RFC 4447 [RFC4447].

7. IANA Considerations

7.1. LDP Status messages

This document uses new LDP status codes, IANA already maintains a registry of name "STATUS CODE NAME SPACE" defined by [RFC5036]. The following values are suggested for assignment:

```
0x0000002C "IP Address of CE"
0x0000004A "IP Address Type Mismatch"
0x0000004B "Wrong IP Address Type"
```

7.2. Interface Parameters

This document proposes a new Interface Parameters sub-TLV, to be assigned from the 'Pseudowire Interface Parameters Sub-TLV type Registry'. The following value is suggested for the Parameter ID:

```
0x16   "Stack Capability"
```

IANA is also requested to set up a registry of "L2VPN PE stack capabilities". This is a 16 bit field. Stack Capability bitmask 0x0001 is specified in Section 6 of this document. The remaining bitfield values (0x0002,..,0x8000) are to be assigned by IANA using the "IETF Review" policy defined in [RFC5226].

L2VPN PE Stack Capabilities:

```
Bit (Value)       Description
===============   ========================================
Bit 0  (0x0001) - IPv6 stack capability
Bit 1  (0x0002) - Reserved
Bit 2  (0x0004) - Reserved
         .
         .
         .

Bit 14 (0x4000) - Reserved
Bit 15 (0x8000) - Reserved
```

8. Security Considerations

The security aspect of this solution is addressed for two planes; control plane and data plane.

8.1. Control Plane Security

Control plane security pertains to establishing the LDP connection, and to pseudowire signaling and CE IP address distribution over that LDP connection. For greater security the LDP connection between two trusted PEs MUST be secured by each PE verifying the incoming connection against the configured address of the peer and authenticating the LDP messages using MD5 authentication, as described in section 2.9 of [RFC5036]. Pseudowire signaling between two secure LDP peers does not pose a security issue but mis-wiring could occur due to configuration error. However, the fact that the pseudowire will only be established if the two PEs have matching configurations (e.g. PW ID, PW type, and MTU) provides some protection against mis-wiring due to configuration errors.

Learning the IP address of the appropriate CE can be a security issue. It is expected that the Attachment Circuit to the local CE will be physically secured. If this is a concern, the PE MUST be configured with IP and MAC address of the CE when connected with Ethernet or IP and virtual circuit information (DLCI or VPI/VCI) when connected over Frame Relay or ATM and IP address only when connected over PPP. During ARP/inverse ARP frame processing, the PE MUST verify the received information against local configuration before forwarding the information to the remote PE to protect against hijacking of the connection.

For IPv6, the preferred means of security is Secure Neighbor Discovery (SEND) [RFC3971]. SEND provides a mechanism for securing Neighbor Discovery packets over media (such as wireless links) that may be insecure and open to packet interception and substitution. SEND is based upon cryptographic signatures of Neighbor Discovery packets. These signatures allow the receiving node to detect packet modification and confirm that a received packet originated from the claimed source node. SEND is incompatible with the Neighbor Discovery packet modifications described in this document. As such, SEND cannot be used for Neighbor Discovery across an ARP Mediation pseudowire. PEs taking part in IPv6 ARP Mediation MUST remove all SEND packet options from Neighbor Discovery packets before forwarding into

the pseudowire. If the CE devices are configured to accept only SEND Neighbor Discovery packets, this will lead to Neighbor Discovery failing. Thus, the CE devices MUST be configured to accept non-SEND packets, even if they treat them with lower priority than SEND packets. Because SEND cannot be used in combination with IPv6 ARP Mediation, it is suggested that IPv6 ARP Mediation is only used with secure Attachment Circuits. An exception to this recommendation applies to an implementation that supports the SEND Proxy [SPROXY] experimental draft which allows a device such as PEs to act as an ND proxy as described in [SPROXY].

8.2. Data plane security

The data traffic between CE and PE is not encrypted and it is possible that in an insecure environment, a malicious user may tap into the CE to PE connection and generate traffic using the spoofed destination MAC address on the Ethernet Attachment Circuit. In order to avoid such hijacking, the local PE may verify the source MAC address of the received frame against the MAC address of the admitted connection. The frame is forwarded to the PW only when authenticity is verified. When spoofing is detected, the PE MUST sever the connection with the local CE, tear down the PW and start over.

9. Acknowledgements

The authors would like to thank Yetik Serbest, Prabhu Kavi, Bruce Lasley, Mark Lewis, Carlos Pignataro and other folks who participated in the discussions related to this document.

10. References

10.1. Normative References

[RFC826] RFC 826, STD 37, D. Plummer, "An Ethernet Address Resolution protocol: Or Converting Network Protocol Addresses to 48.bit Ethernet Addresses for Transmission on Ethernet Hardware".

[RFC2390] RFC 2390, T. Bradley et al., "Inverse Address Resolution Protocol".

[RFC4447] L. Martini et al., "Pseudowire Setup and Maintenance using LDP", RFC 4447.

[RFC4446] L. Martini et al,. "IANA Allocations for pseudo Wire Edge to Edge Emulation (PWE3)", RFC 4446.

[RFC2119] S.Bradner, "Key words for use in RFCs to indicate requirement levels", RFC 2119.

[RFC5036] L.Anderseen et al., "LDP Specification", RFC 5036.

[RFC4861] Narten, T., Nordmark, E. and W.Simpson, "Neighbor Discovery for IP Version 6 (IPv6)", RFC 4861.

[RFC3122] Conta, A., "Extensions to IPv6 Neighbor Discovery for Inverse Discovery Specification", RFC 3122.

[RFC4862] Thomson, S. and Narten, T., "IPv6 Stateless Address Autoconfiguration", RFC 4862.

[RFC3971] Arkko, J. et al., "Secure Neighbor Discovery (SEND)", RFC 3971.

[RFC5226] Narten, T et al., "Guidelines for Writing an IANA Considerations Section in RFCs", RFC 5226.

10.2. Informative References

[RFC4664] L. Andersson et al., "Framework for L2VPN", RFC 4664.

[RFC1332] G. McGregor, "The PPP Internet Protocol Control Protocol (IPCP)", RFC 1332.

[RFC5072] D. Haskin, "IP Version 6 over PPP", RFC 5072.

[RFC925] J.Postel, "Multi-LAN Address Resolution", RFC 925.

[RFC1256] S.Deering, "ICMP Router Discovery Messages", RFC 1256.

[RFC5309] Shen and Zinin, "Point-to-point operation over LAN in Link State Routing Protocols", RFC 5309.

[SPROXY] S.Krishnan et al., "Secure Proxy ND support for SEND", draft-ietf-csi-proxy-send-05.txt

11. Authors' Addresses

This document is the combined effort of many who have contributed, carefully reviewed and provided the technical clarifications for the document.

Himanshu Shah (editor)

Ciena
Email: hshah@ciena.com

Eric Rosen (editor)
Cisco Systems
Email: erosen@cisco.com

Giles Heron
Cisco Systems (editor)
Email: giheron@cisco.com

Vach Kompella (editor)
Alcatel-Lucent
Email: vach.kompella@alcatel-lucent.com

Matthew Bocci
Alcatel-Lucent
Email: Mathew.bocci@alcatel-lucent.com

Tiberiu Grigoriu
Alcatel-Lucent
Email: Tiberiu.Grigoriu@alcatel-lucent.com

Neil Hart
Alcatel-Lucent
Email: Neil.Hart@alcatel-lucent.com

Andrew Dolganow
Alcatel-Lucent
Email: Andrew.Dolganow@alcatel-lucent.com

Shane Amante
Level 3
Email: Shane@castlepoint.net

Toby Smith
Google
EMail: tob@google.com

Andrew G. Malis
Verizon
EMail: Andy.g.Malis@verizon.com

Steven Wright
Bell South Corp
Email: steven.wright@bellsouth.com

Waldemar Augustyn

Consultant
Email: waldemar@wdmsys.com

Arun Vishwanathan
Juniper Networks
Email: arunvn@juniper.net

Ashwin Moranganti
IneoQuest Technologies
Email: Ashwin.Moranganti@Ineoquest.com

APPENDIX A:

A.1. Use of IGPs with IP L2 Interworking L2VPNs

In an IP L2 interworking L2VPN, when an IGP on a CE connected to a broadcast link is cross-connected with an IGP on a CE connected to a point-to-point link, there are routing protocol related issues that MUST be addressed. The link state routing protocols are cognizant of the underlying link characteristics and behave accordingly when establishing neighbor adjacencies, representing the network topology, and passing protocol packets. The point to point operations of the routing protocols over a LAN is discussed in [RFC5309].

A.1.1. OSPF

The OSPF protocol treats a broadcast link type with a special procedure that engages in neighbor discovery to elect a designated and a backup designated router (DR and BDR respectively) with which each other router on the link forms adjacencies. However, these procedures are neither applicable nor understood by OSPF running on a point-to-point link. By cross-connecting two neighbors with disparate link types, an IP L2 interworking L2VPN may experience connectivity issues.

Additionally, the link type specified in the router LSA will not match for the two cross-connected routers.

Finally, each OSPF router generates network LSAs when connected to a broadcast link such as Ethernet, receipt of which by an OSPF router which believes itself to be connected to a point-to-point link further adds to the confusion.

Fortunately, the OSPF protocol provides a configuration option (ospfIfType), whereby OSPF will treat the underlying physical broadcast link as a point-to-point link.

It is strongly recommended that all OSPF protocols on CE devices connected to Ethernet interfaces use this configuration option when attached to a PE that is participating in an IP L2 Interworking VPN. The point-to-point operation of the routing protocol over

A.1.2. RIP

RIP protocol broadcasts RIP advertisements every 30 seconds. If the multicast/broadcast traffic snooping mechanism is used as described in section 4.1, the attached PE can learn the local CE router's IP address from the IP header of its advertisements. No special configuration is required for RIP in this type of Layer 2 **IP Interworking L2VPN.**

A.1.3. IS-IS

The IS-IS protocol does not encapsulate its PDUs in IP, and hence cannot be supported in IP L2 Interworking L2VPNs.