PPVPN Working Group
Internet Draft

June 2003
Expires: December 2003

H. Shah	Ciena Networks
E. Rosen	Cisco Systems
W. Augustyn	consultant
G. Heron	PacketExchange,Ltd
T. Smith	Laurel Networks
A. Moranganti	ADC Telecom
S. Khandekar	Timetra Networks
V. Kompella	Timetra Networks
A. Malis	Vivace Networks
S. Wright	Bell South
V. Radoaca	Nortel Networks
A. Vishwanathan	Force10 Networks

ARP Mediation for IP Interworking of Layer 2 VPN

draft-shah-ppvpn-arp-mediation-02.txt

Status of this memo

Abstract

The VPWS service [L2VPN Framework] 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". This document specifies the ARP Mediation function, and specifies the encapsulation used to carry the IP datagrams on the Pseudowires when ARP mediation is used.

1.0 Introduction

Layer 2 Virtual Private Networks (L2VPN) are constructed with the use of a Service Provider IP 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 only Layer 3 protocol is IP. Thus it makes sense to consider procedures that are either optimized for IP or are outright dedicated to IP traffic only.

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 links. 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 FR, 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 |
                               |                    +-----+
                            +-----+
                    ........| 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 CE would use ARP, a FR-attached CE might use Inverse ARP. 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 technology-specific address resolution procedures.

In this draft, we specify the procedures 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.

2.0 ARP Mediation (AM) function

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

For a given point-to-point connection between a pair of CEs, a PE must perform three logical steps as part of the ARP Mediation procedure:

1. Discover the IP addresses of the locally attached CE device
2. Distribute those IP Addresses to the remote PE
3. Notify the locally attached CE of the remote CE's IP address.

This information is gathered using the mechanisms described in the following sections.

3.0 IP Layer 2 Interworking Circuits

The IP Layer 2 Interworking Circuits refer to Pseudowires that carry IP datagram as the payload. At ingress, data link header of an IP frame is removed and dispatched over the Pseudowire with or without the optional control word. At the egress, PE encapsulates the IP packet with the data link header used on the local Attachment Circuit.

The use of this encapsulation is determined by the exchange of value 0x000B as the VC type during Pseudowire establishment as described in [PWE3-Control].

4.0 Discovery of IP Addresses of Locally Attached CE Device

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

- . Discovery of locally attached CE IP device
- . Establishment of the PW

The establishment of PW occurs independently from local CE IP address discovery. During the period when (bi-directional) PW has been established but local CE IP device has not been detected, only datagrams inside of broadcast/multicast frames are propagated; IP datagrams inside unicast frames are dropped. The IP datagrams from

unicast frames flow only when IP end systems on both Attachment Circuits have been discovered, notified and proxy functions have completed.

4.1 Monitoring Local Traffic

The PE devices may learn the IP addresses of the locally attached CEs from any IP traffic, such as multicast/broadcast packets, that CE may generate irrespective of reacting to specific address resolution queries described below.

4.2 CE Devices Using ARP

If a CE device uses ARP to determine the IP address of its neighbor, the PE processes the ARP requests for the stated locally attached circuit and responds with ARP replies containing the remote CE's IP address, if the address is known. If the PE does not yet have the remote CE's IP address, it does not respond, but notes the IP address of the local CE and the circuit information, including related MAC address. Subsequently, when the IP address of the remote CE becomes available, the PE may initiate the ARP response as a means to notify the local CE, the IP address of the remote CE.

This is a typical operation for Ethernet attachment circuits. It is important to note that IP L2 Interworking circuit function is restricted to only one end station per Ethernet Attachment Circuit.

The PE may periodically generate ARP request messages to the CE's IP address as a means to verify the continued existence of the address and its binding to the stated MAC address. The absence of a response from the CE device for a given number of retries could be used as a cause for a withdrawal of the IP address advertisement to the remote PE and entering into the address resolution phase to rediscover the attached CE's IP address. Note that such "heartbeat" scheme is needed only for broadcast links, as a loss of CE may otherwise be undetectable.

4.3 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 for stated circuit and responds with an Inverse ARP reply containing the remote CE's IP address, if the address is known. If the PE does not yet have the remote CE's IP address, it does not respond, but notes 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 the Inverse ARP request as a means to notify the local CE, the IP address of the remote CE.

This is a typical operation for Frame Relay and ATM attachment circuits. In the cases where the CE does not use Inverse ARP, PE could still discover the CE as described in section 4.1 and 4.5.

4.4 CE Devices Using PPP

If a CE device uses PPP to determine the IP address of its neighbor, a PE takes part in the IPCP [PPP-IPCP] exchange and supplies the IP address of the remote CE if the address is known. If the PE does not have the remote CE's IP address, it does not respond to the local CE's IPCP request but simply notes its IP address. Subsequently, when the IP address of the remote CE becomes available, the PE generates IPCP Configure-Request to the local CE.

The PE must deny configurations such as header compression and encryptions in the NCP packets with such options.

4.5 Proactive 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 [RFC 1256] 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.

The use of the router discovery mechanism by the PE is optional.

5.0 IP Address Distribution Between PE

5.1 When To Distribute IP Address

A PE device advertises the IP address of the attached CE only when the encapsulation type of the Pseudowire is IP L2 interworking. It is quite possible that the IP address of a CE device is not available at the time the PW labels are advertised. For example, in Frame Relay the CE device dispatches inverse ARP request only when the DLCI is active; if the PE signals the DLCI to be active only when it has received the IP address along with the VC-FEC from the remote PE, a chicken and egg situation arises. In order to avoid such problems, the PE must be prepared to advertise the VC-FEC before the CE's IP address is known. When the IP address of the CE device does become available, the PE re-advertises the VC-FEC along with the IP.

Similarly, if the PE detects invalidation of the CE's IP address (by methods described above) the PE must re-advertise the VC-FEC with null IP address to denote the withdrawal of the CE's IP address. The receiving PE then waits for the notification of remote IP address. During this period, propagation of unicast IP traffic is suspended while continuing to let multicast IP traffic flow.

If two CE devices are locally attached to the PE where, one CE is connected to an Ethernet data link and the other to a Frame Relay interface, for example, the IP addresses are learned in the same manner described above. However, since the CE devices are local, the distribution of IP addresses for these CE devices is a local step.

5.2 LDP Based Distribution

The [PWE3-CONTROL] uses Label Distribution Protocol (LDP) transport to exchange VC-FEC in the Label Mapping message in a downstream unsolicited mode. The VC-FEC comes in two flavors; Pwid and Generalized ID FEC elements and shares some fields that are common between them. The discussions below refer to these common fields for IP L2 Interworking Circuits.

The IP L2 Interworking uses IP datagram as payload over the Pseduowire. The use of such encapsulation is identified by VC type field of the VC-FEC as the value 0x000B [PWE3-Control].

In addition, this document defines an IP address TLV that must be included in the optional TLV field of the Label Mapping message when advertising VC-FEC for the IP L2 Interworking Circuit. Such use of optional TLV in the Label Mapping message to extend the attributes of the VC-FEC has also been specified in the [PWE3-Control].

When processing a received VC-FEC, the PE matches the VC-Id and VC-type with the locally configured VC-Id to determine if the VC-FEC is of type IP L2 Interworking. If matched, it further checks the presence of IP address TLV. If an IP address TLV is absent, a Label Release message is issued to reject the PW establishment.

```
   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
  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
  |1|0|  IP address TLV (TBD)     |           Length              |
  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
  |                          IP Address                           |
  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

The Length field is defined as the length of the IP address and is set to value 4.

The IP address field is set to value null to denote that advertising PE has not learned the IP address of his local CE device. The non-zero value of the IP address field denotes IP address of advertising PEÆs attached CE device.

The IP address TLV is also used in the LDP notification message along with the VC-FEC. The IP address TLV in Notification message is

used as an update mechanism to notify the changes in the IP address of the local CE device as described in [SHAH-CONTROL].

5.3 Out-of-band Distribution, Manual Configuration

In some cases, it may not be possible to deduce the IP addresses from the VPN traffic nor induce remote PEs to supply the necessary information on demand. For those cases, out-of-band methods, such as manual configuration, could be used. The use of these types of methods is useful only to handle corner cases.

6.0 How CE Learns The Remote CE's IP address

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

6.1 CE Devices Using ARP

When the PE learns the remote CE's IP address as described in section 5.1 and 5.2, it may or may not know the local CE's IP address. If the local CE's IP address is not known, the PE must wait until it is acquired through one of the methods described in sections 4.1, 4.3 and 4.5. 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 remote CE's IP address with the PE's own MAC address.

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

6.2 CE Devices Using Inverse ARP

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

6.3 CE Devices Using PPP

When the PE learns the remote CE's IP address, it must initiate the Configure-Request using the remote CE's IP address or respond to pending Configure-Request from the local CE. As noted earlier, all

other configuration options related to compression, encryptions, etc., should be rejected.

7.0 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.

7.1 OSPF

The OSPF protocol treats 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 it 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 has the potential to experience connectivity issues.

Additionally, the link type specified in the router LSA will not match for two routers that are supposedly sharing the same link type. Finally, each OSPF router generates network LSAs when connected to a broadcast link such as Ethernet, receipt of which by an OSPF router on the 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.

7.2 IS-IS

The IS-IS protocol sends a LAN Hello PDU (IIH packet) with the MAC address and the IP address of the intermediate system (i.e., CE device) when attached to Ethernet links. The CE device expects its neighbor to insert its own MAC and IP address in the response. If the neighbor is connected via a point-to-point link type, the LAN Hello PDU will be silently discarded. Similarly, Hello PDUs on the point-to-point link do not contain any MAC address, which will confuse a neighbor on an Ethernet link, if these two neighbors were cross-connected via above described mechanisms.

Thus, use of the IS-IS protocol on CE devices presents problems when interconnected by disparate data link types in an IP L2 Interworking VPN environment. There are some mechanisms defined in draft-ietf-isis-igp-p2p-over-lan-00.txt to accommodate point-to-point behavior over broadcast networks. The feasibility of such techniques to solve this problem is under review.

It is important to note that the use of the IS-IS protocol in enterprise networks (i.e., CE routers) is less common. The IS-IS related difficulties for IP L2 Interworking VPNs, hence are minimized.

7.3 RIP

RIP protocol broadcasts RIP advertisements every 30 seconds. If the group/broadcast address snooping mechanism is used as described above, the attached PE can learn the advertising (CE) router's IP address from the IP header of the advertisement. No special configuration is required for RIP in this type of Layer 2 IP Interworking L2VPN.

8.0 Security Considerations

The security aspects of this solution will be discussed at a later time.

9.0 Acknowledgements

The authors would like to thank Prabhu Kavi, Bruce Lasley and other folks who participated in the discussions related to this draft.

10.0 Intellectual Property Considerations

Tenor/Enterasys Networks may seek patent or other intellectual property protection for some of all of the technologies disclosed in this document. If any standards arising from this document are or become protected by one or more patents assigned to Tenor/Enterasys Networks, Tenor/Enterasys intends to disclose those patents and license them on reasonable and non-discriminatory terms.

11.0 References

11.1 Normative References

[ARP] 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".

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

11.2 Informative References

[L2VPN-REQ] W. Augustyn et al., "Service Requirements for Layer 2 Provider Provisioned Virtual Private Networks", February 2003, work in progress.

[L2VPN-FRM] L. Andersson et al., "L2VPN Framework", January 2003, work in progress.

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

[L2VPN-Kompella] K. Kompella et al., "Layer 2 VPNs Over Tunnels", June 2002, work in progress.

[PWE3-CONTROL] L. Martini et al., "Transport of Layer 2 Frames Over MPLS", November 2002, work in progress.

[L2VPN-Signaling] E. Rosen et al., "LDP-based Signaling for L2VPNs", September 2002, work in progress.

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

[SHAH-CONTROL] H. Shah et al., "Dynamic Parameters Signaling for MPLS-based Pseudowires", June 2003, work in progress

12.0 Authors' Addresses

Himanshu Shah
35 Nagog Park,

Acton, MA 01720
Email: hshah@ciena.com

Eric Rosen
Cisco Systems
1414 Massachusetts Avenue,
Boxborough, MA 01719
Email: erosen@cisco.com

Waldemar Augustyn
Email: waldemar@nxp.com

Giles Heron
PacketExchange Ltd.
The Truman Brewery
91 Brick Lane
LONDON E1 6QL
United Kingdom
Email: giles@packetexchange.net

Sunil Khandekar and Vach Kompella
TiMetra Networks
274 Ferguson Dr.
Mountain View, CA 94043
Email: sunil@timetra.com
Email: vkompella@timetra.com

Toby Smith
Laurel Networks
Omega Corporate Center
1300 Omega drive
Pittsburgh, PA 15205
Email: jsmith@laurelnetworks.com

Arun Vishwanathan
Force10 Networks
1440 McCarthy Blvd.,
Milpitas, CA 95035
Email: arun@force10networks.com

Ashwin Moranganti
Appian Communications
35 Nagog Park,
Acton, MA 01720
Email: amoranganti@appiancom.com

Andrew G. Malis
Vivace Networks, Inc.
2730 Orchard Parkway

San Jose, CA 95134
Email: Andy.Malis@vivacenetworks.com

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

Vasile Radoaca
Nortel Networks
Email: vasile@nortelnetworks.com

Full Copyright Statement