Network Working Group
INTERNET DRAFT

Nishit Vasavada
Amber Networks, Inc.

July 2001

Layer 3 VPNs using Encapsulation Services Protocol
<draft-vasavada-ppvpn-es-l3vpn-00.txt>

1. Status of this Memo

This document is an Internet-Draft and is in full conformance with all provisions of Section 10 of RFC 2026.

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

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

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

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

2. Abstract

[RFC2547bis] defines a way to implement Layer 3 VPNs using BGP and MPLS. [GRE_IP_MPLS] shows a method to implement RFC 2547 style VPNs across a non-MPLS network. This document shows an alternative way of implementing Layer 3 VPNs in a non-MPLS network. Unlike [RFC2547bis], it does not require BGP either to be running on the PE.

3. 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 RFC 2119 [6].

4. Introduction

[RFC2547bis] discusses in great detail the motivation and requirements for a Layer 3 Virtual Private Network (VPN). The goal

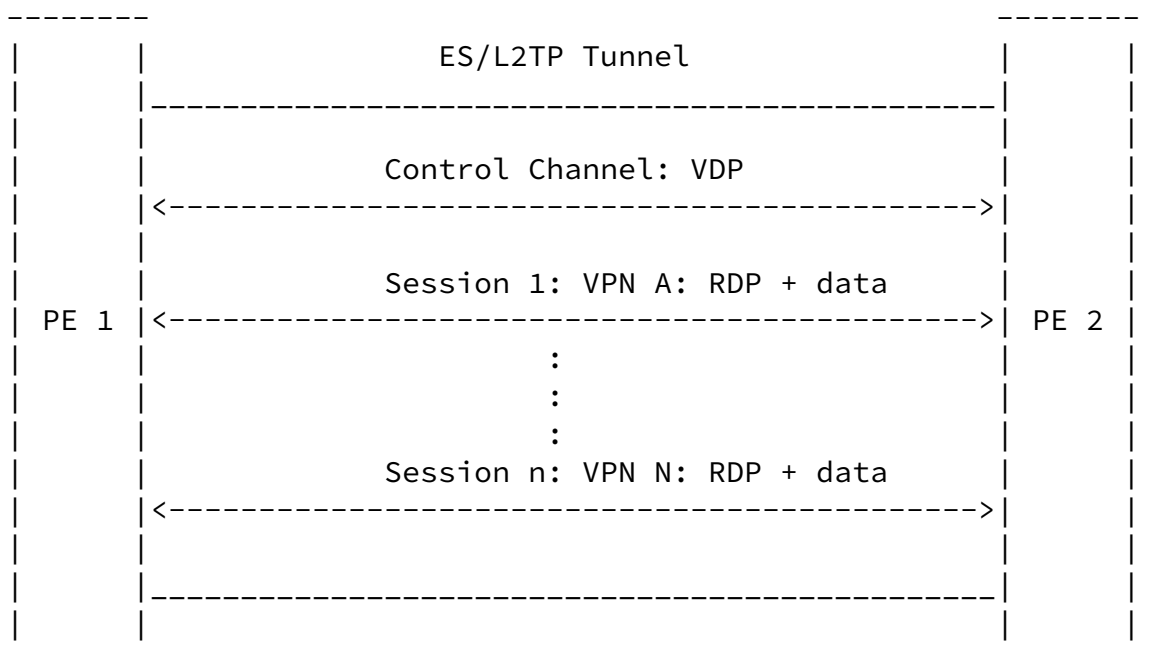
of this document is to accomplish the same as that of [RFC2547bis], and therefore the common details are not repeated here.

[RFC2547bis] requires that the Service Provider(SP)'s network be MPLS-enabled. This means all the routers in the SP's core network MUST be able to support MPLS. [RFC2547bis] uses BGP for route distribution, which requires BGP to be running in the SP's core network at least with an overlay topology. While this may be the

case in fair number of networks, a technology not requiring to run BGP for route distribution may be more suitable for networks not running BGP.

[ES] defines a generic protocol to emulate and encapsulate Layer 1 and Layer 2 circuits over a core network. We extend [ES] to carry VPN Discovery Protocol (VDP) and Route Distribution Protocol (RDP). VDP is an auto-discovery mechanism for discovering other Provider Edge (PE) routers which are connected to a site belonging to a VPN that has a site connected to the PE router running VDP. RDP is an extensible mechanism to distribute route information for each VPN to all other PEs with sites belonging to that specific VPN.

A special ES tunnel is set up between two PEs to carry L3 VPN traffic. It carries control and data traffic for all VPNs which are common to the two PEs. Inside each tunnel, each ES session represents a specific VPN.



-------- --------

Figure 1. Two PEs running ES based L3 VPNs

The ES tunnel is set up by following the process outlined in [ES]. The access link type is "L3VPN". Service attributes are chosen according to the tunnel guiding parameters - e.g. it may indicate the remote PE. The service type capability negotiated is ES, again as specified in [ES]. Session 0 is reserved for carrying VDP traffic and will be referred to as control session in rest of the document. The control session is set up as soon as the tunnel comes up. The numerically lower IP address initiates and numerically higher IP address passively awaits for the session. VPN related route information (through RDP) and data traffic is carried in individual sessions - one session per VPN. Thus, one session of VPD runs once per pair of PE (one per tunnel), while one session of RDP runs once per VPN per tunnel.

The sessions map VPNs to ES tunnel/session with the use of VPN-IDs. Each VPN is assigned an 8-byte ID known as VPN-ID. The VPN-ID is

passed to the remote PE during L2TP session set up through the end-identifier AVP specified in [L2TPES]. VPN-ID with all zeros is reserved, while all 1's is used for the control session. Rest of the sessions carry a specific VPN-ID during session set-up, and all the traffic through that session is mapped to the VPN represented by that VPN-ID.

5. VPN Discovery Protocol (VDP)

The aim of VPN Discovery Protocol is to dynamically determine VPN membership at the remote end of IP-VPN tunnels. The protocol communicates with its remote peer using the VPN control session (a special session) inside a VPN tunnel.

The protocol is a simple, reasonably state-less Query-Response based protocol. It makes the following assumptions:

- It runs over an unreliable transport (an ES/L2TP session in this case).
- Fragmentation of protocol PDUs are handled at underlying IP layer
- L2TP signaling is asymmetric in nature. VPN Discovery protocol assumes that the PE with numerically lower IP address will always initiate the establishment of underlying tunnels and sessions. The other end (with numerically higher IP address) will passively wait

for the incoming tunnel/session requests.

5.1. VDP Operation

- The protocol gets triggered as soon as the VPN control session becomes active between two PEs.
- For each remote PE, the local PE maintains a state for each configured VPN on the system. The state of the VPN for the remote PE changes through the operation of the VDP (the state transition is described in the next section).
- Initially, all the VPNs on all remote PEs are in a "Dirty" state. This state means that the membership information has not been conveyed to the remote PE.
- A periodic timer, with a configurable timeout value, is used to send the list of dirty VPNs to the remote PE, as a VPN-Query-Request PDU.
- The remote PE, upon receipt of the query request, sends back a VPN-Query-Response indicating against each VPN listed, whether it has been configured on the remote PE end or not.
- When the query response comes back to the local PE, each of the VPN contained in the response message go to a "Clean" state (from a previous Dirty state). A Clean state means that its VPN membership has been conveyed to the remote PE.
- All the VPNs in the query response, which are not configured on the remote system remain in the clean state, until the VPN membership is removed from the local PE.
- For each VPN in the query response which is configured on the remote PE, either a VPN session request is initiated or is passively awaited, depending upon the relative numerical relationship between the IP addresses of the two PEs.
- As stated earlier, it is assumed that the underlying transport does not guarantee any reliable delivery. Hence, every time a new query

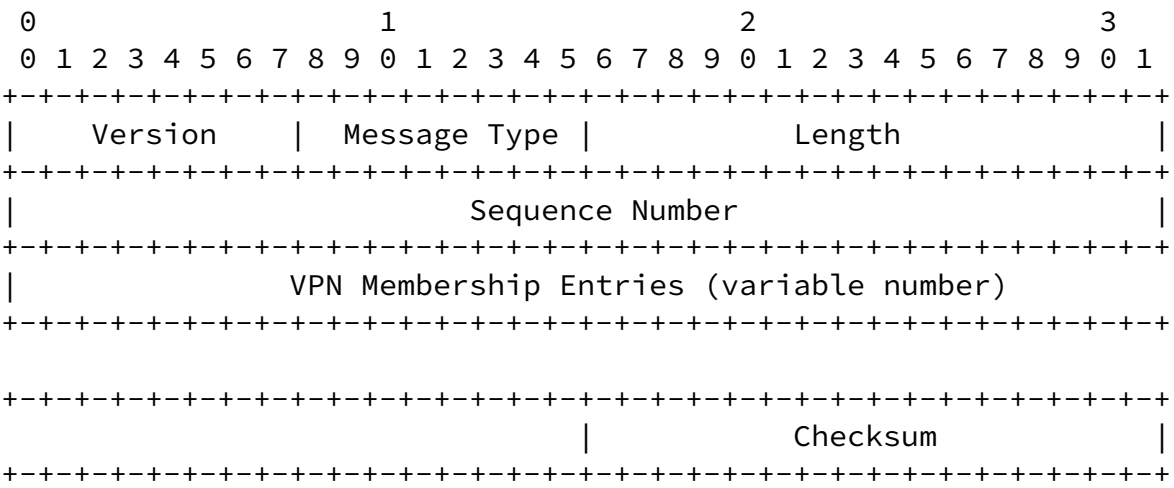
request is sent, a (4-byte) sequence number is incremented and is included in the PDU message.
- The receiver end copies the sequence number in the query request to the query response. The sender of the query request does not accept any query response which has a sequence number different from the one that was included in the last query request sent.
- A checksum is included to protect the integrity of the PDU.

At the end of initial run of VDP, both PEs know whether each of the VPN on the local side has a member VPN site on the remote PE as well. The idea is to have a mechanism where VPN memberships of a specific

PE need not be configured at all other PEs in the network.

5.2. Message format

The VDP message format is as shown below:



Version (one byte): Set to 1
Message Type (one byte):
0 for VDP Request
1 for VDP Response
Length (two bytes): length of the entire PDU beginning version through the checksum field
Sequence Number (four bytes): Starts from 0 and wraps over after reaching the maximum value. The sender of VDP Request MUST increment the number by one every time it sends a new request. The sender of VDP Response MUST copy the Sequence number from the Request it is responding to.
VPN Membership Entries (nine bytes per entry): There can be multiple entries in this field. Each entry consists of two parts:
- An 8-byte VPN ID
- A 1-byte value that shows whether the VPN is present or not: 1 denotes Present, 0 denotes Not Present. This field SHOULD be ignored on receiving the Request. The sender of Response MUST set it to proper value based on whether the VPN represented by the corresponding VPN ID is present or not.

Checksum (two bytes): This is an IP-style checksum over the VDP message beginning the Version field through the Checksum field.

5.3. Example

For example, suppose PE1 in Figure 1 has member VPNs A, B, C and D, and PE2 has member VPNs B, D, E and F. The following exchange will take place, assuming PE2 receives PE1's query before it sends its own query:

- PE1 sends a request with VPNs A, B, C and D in the VPN Membership Entries.
- PE2 sends a response with VPNs A, B, C and D in the VPN Membership Entries. The Present/Not Present field will reflect the membership status of the corresponding VPN at PE2. This means VPNs B and D will be marked Present, while VPNs A and C will be marked as Not Present. The sequence number will be the same as in the request received from PE1. Checksum is recomputed.
- PE2 sends a request for VPNs E and F to PE1.
- PE1 sends a response showing VPNs E and F Not Present.
- PE1 and PE2 have one ES session set up for each of the VPNs B and D.

VDP does not retransmit a PDU if no response is received. However, it periodically scans the database for "Dirty" entries, and sends a new VDP Request message if one or more such entries are found. These may be older entries which were never acknowledged through VDP Response by the peer PE, or they may newly configured VPNs since the last scan.

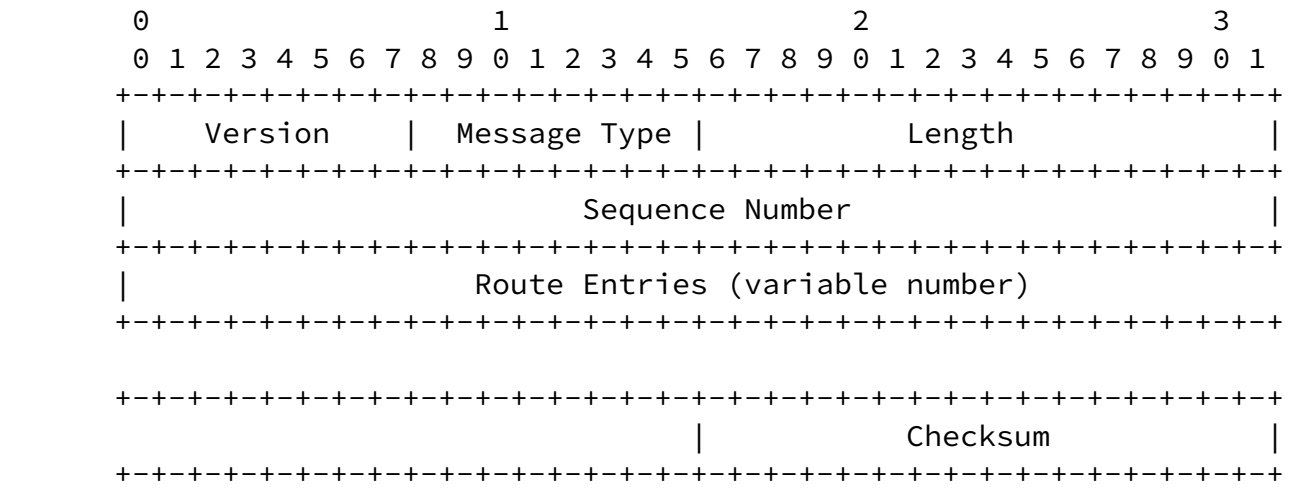
If a VPN membership is removed from a PE, the PE tears down the ES/L2TP sessions corresponding to the VPN from each PE which had that VPN as a member. Thus, there is no need for an explicit VDP message for informing VPN membership removal.

6. Route Distribution Protocol (RDP)

RDP is used to distribute subscriber addresses to other sites in the VPN. RDP is VPN specific, and therefore is carried in the session created for the specific VPN. In the case of ES over L2TP, the RDP for VPN A is sent to PE X via the L2TP tunnel set up with PE X inside the ES/L2TP session set up for VPN A. This is shown in Figure 1.

6.1. Message format

The RDP message format is as shown below:



Version (one byte): Set to 1
Message Type (one byte):
0 for Route Advertise
1 for Route Advertise Response
Length (two bytes): length of the entire PDU beginning version through checksum
Sequence Number (four bytes): Starts from 0 and wraps over after reaching the maximum value. The sender of Route Advertise message MUST increment the number by one every time it sends a new message. The sender of Route Advertise Response MUST copy the Sequence number from the Advertise message it is responding to.
Route Entries (five bytes per entry): Each route entry is a 3-tuple:
- A 32-bit prefix
- A 4-bit mask-length, and
- A 4-bit field that shows whether the route is being Added or Withdrawn. 0 denotes ### and 1 denotes ###

Checksum (two bytes): This is an IP-style checksum over the RDP message beginning the Version field through the Checksum field.

6.2. RDP Operation

RDP maintains a "dirty" list of routes for each VRF. The list is maintained per peer PE, and tells the local PE that the status of that specific route has not been confirmed by the peer PE.

A route is added in the dirty list and put in Dirty/Add state when the route was recently added. The route is sent to the peer PE in an RDP Advertise message with the add/withdraw value set to Add. When the peer PE acknowledges the message after processing the routes in the Advertise message, it includes the route in its Route Entries list. The local PE at this point removes the route from the dirty list.

Similarly, when a route is deleted from the VRF for the VPN, it is added to the dirty list and put in Dirty/Withdraw state. It retained in the list till the peer PE acknowledges the Route Advertise message that announced the withdrawal of the route.

When a PE receives an RDP Advertise message, it identifies the corresponding VRF by the ES tunnel/session. This ensures that the routes belonging to a specific VPN are injected only into the VRF

corresponding to that VPN. This is essential for an L3 VPN, since it allows SP's customers to have overlapping private address space without causing any confusion in the core.

Once a VRF is identified by the receiving PE, the routes are added or deleted based on the Add/Withdraw field. When the PE is done processing the Route Advertise message, it sends the packet back to the PE which sent the Route Advertise message. This serves as an acknowledgement of Route Advertise. The Message Type field is set to Route Advertise Response, and the checksum is recomputed.

The PE receiving the Route Advertise Response compares the sequence number of the Response message with the last Route Advertise message sent to the peer PE. If the sequence numbers do not match, the Route Advertise Response is silently discarded. If the sequence numbers match, the receiving PE finds all the routes listed in the Route Advertise Response and removes them from the dirty list.

7. Data plane operation

When L3 VPN data is received from a CE, the VRF is chosen based on the interface. The Destination IP address in the VRF tells the PE the peer PE, as well as the ES tunnel/session corresponding to the peer PE and the VPN. The customer data packet is encapsulated in ES (and the lower transport protocol such as L2TP) and sent to the peer PE.

The peer PE identifies the outbound interface based on the ES tunnel/ session information in the packet from the sending PE. The ES encapsulation is removed and the packet is sent out on the outbound interface.

8. Interface with ES

VDP, RDP and data plane traffic is encapsulated in ES [ES]. If ES runs over L2TP as shown in [ES], all the sessions inside each tunnel between the PEs will need to negotiate ES as the L2TP service type, as defined in [L2TPES].

8. Future Work

Modify ES header to accommodate multiple types of traffic inside ES. Assigning unique VPN-ID for inter-SP VPNs.

9. Security Considerations

All the underlying ES Security considerations remain, though no new ones are introduced.

10. IANA Considerations

None at present.

11. Intellectual Property Considerations

Amber 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 Amber Networks, Amber intends to disclose those patents and license them on reasonable and non-discriminatory terms.

12. Acknowledgments

Many thanks to Himansu Sahu, Danny McPherson, Stanley Fong and Indira Mitchell for their help in reviewing this draft.

13. References

[RFC2547bis] Rosen, et. al., "BGP/MPLS VPNs", Work in Progress, February 2001

[GRE_IP_MPLS] Rekhter, et. al., "Use of PE-PE GRE or IP in RFC2547 VPNs", Work in Progress, June 2001

[ES] Vasavada, N., "ESP: Encapsulation Services Protocol", Work in Progress, July 2001

[L2TPES] Vasavada N., "Encapsulation Services Protocol Service Type for L2TP", draft-vasavada-l2tpext-es-svctype-00.txt, Work in Progress, July 2001

14. Author's Address

Nishit Vasavada
Amber Networks, Inc.
48664 Milmont Drive
Fremont, CA 94538
Phone: +1 510.687.5200
Email: nishit@ambernetworks.com