Network Working Group
Internet Draft
Document: draft-tsenevir-l2vpn-pmesh-00.txt
Category: Informational

Tissa Senevirathne
(Force10)

June, 2001

Use of Partial meshed tunnels to achieve forwarding behavior of full meshed tunnels.

Status of this Memo

This document is an Internet-Draft and is in full conformance with all provisions of Section 10 of RFC2026 [1].

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.

For potential updates to the above required-text see:
http://www.ietf.org/ietf/1id-guidelines.txt

Placement of this Memo in Sub-IP Area

RELATED DOCUMENTS:

See reference.

WHERE DOES IT FIT IN THE PICTURE OF THE SUB-IP WORK

The ID presented fits in to the PPVPN WG and/or CCAMP WG.

WHY IS IT TARGETED AT THIS WG(s)

Sub-IP and IP tunnels are becoming a popular method in carrying data transparently over the provider or the core network.

Use of such tunnels are key component of PPVPN infrastructure. On the other hand CCAMP WG charter includes defining common control and measurement plane. Hence optimal use of tunnels is an integral

part of the control infrastructure.

JUSTIFICATION

Increasing number of service providers are offering Ethernet services to the customer. In the core of the network IP or Sub-IP technologies are used. In general, Ethernet services provided to customers are VPN service having multi-points of service access (as opposed to point-to-point).

Requirement to use fully meshed networks seriously affects the scalability of Layer 2 NBVPN. The methods presented in this document facilitate service providers to offer scalable Layer 2 VPN solutions.

1. Abstract

This document presents methods to achieve proper forwarding of Broadcast, Multicast and Unknown traffic over a set of partial mesh tunnels. In addition, the methods presented in this document may be used to achieve loop free topology.

2. 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 RFC-2119 [2].

Table of Content

3. Introduction

Problem Definition

Layer 2 NBVPN services use tunneling methods such as IPSec or MPLS to carry customer's Layer 2 traffic over the provider's network. Layer 2 NBVPN deployment, in general, requires many-to-many connectivity. Unknown unicast, Broadcast and Multicast traffic are required to be forwarded to all end-points of the Layer 2 NBVPN. When there are more than two end-points in the Layer 2 NBVPN, to achieve the required Layer 2 behavior, a set of fully meshed tunnels are required. However requirement of such fully meshed tunnels seriously affects the scalability of Layer 2 NBVPN that require many-to-many connectivity.

In this document we provide methods to achieve required Layer 2 behavior using set of partially meshed tunnels.

In the first part of the document a typical deployment scenario is presented. In this section we also explain the proposed solution using an example.

In the second part required building blocks for the proposed solution is discussed.

Later in the document use of Spanning Tree and Shortest Path First Algorithms are discussed. Also discussed are the advantages, and disadvantages of using Spanning Tree Algorithms vs. Shortest Path First Algorithms for the purpose.

Scope of the Document

In this document word node and end-point is used to identify PE devices. The discussion in the document does not focus on devices

in the core of the network. The "tunnel" in this discussion refer to a logical connection between two PE devices.

4.0 Deployment scenario for Layer 2 NBVPN

Consider the scenario where there are (n)end points in the Layer 2 NBVPN. Now each PE device is required to maintain (n -1) tunnels to all other end points. As a result the network is required to maintain n*(n-1) set of tunnels.

Diagram below depicts partial mesh deployment of a 4 end-point Layer 2 NBVPN deployment. Typically, some end-points may be connected to all other end-points and some may only be connected to a sub set of end-points.

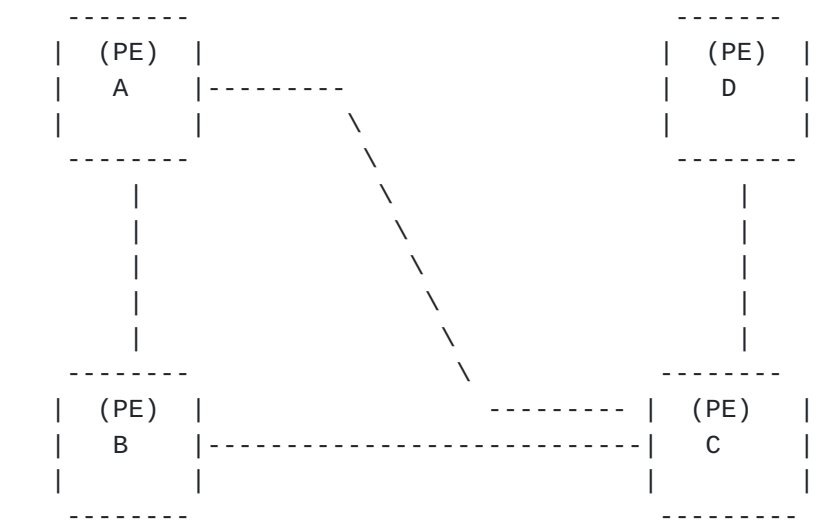


Fig: Partial Mesh connectivity

In theory, a partial mesh graph G may be represented by union of set of fully mesh graphs g and set of partial mesh graphs g'. In the above diagram graph G {A, B, C, D} contain fully mesh graph g {A,B,C} and partial mesh graph g' {C,D}.

Discussions in this document will focus on the forwarding at node A and C. Node A represents partial mesh node. Node C represents a fully meshed node. However, the discussion is equally applicable to

all other nodes in other graphs with any arbitrarily number of nodes. Only restriction is that graph is required to be connected (ie no disconnected nodes).

5.0 Objective and Methodology

Objective

Each node must be able to derive the mesh topology of the Layer 2 NBVPN domain. The derived topology by each node must be identical. As a result each node is capable of deriving nodes that have already received traffic from the upstream. Forwarding policies deduce set of outgoing tunnels for each incoming tunnel. Note: for fully meshed nodes the outgoing tunnel set is NULL.

Methodology

Each end-point advertise its preference (by some means) to receive traffic over a given tunnel T that is terminating/starting at the end-point. A tunnel T is represented using end-point addresses (id). The preference value assigned must be coordinated via global policies. As an example, lower numerical value represents higher preference.

We assume that there is control plane IP connectivity between all end-points. Intermediate devices in the IP plane do not modify the tunnel preference as they forward advertisements.

Each source node chooses set of optimal tunnels, using the preference information received. If there is no direct tunnel the best intermediate end-point is selected. The set of tunnels that a source end-point use to forward unknown, multicast and broadcast traffic is called Source Mesh and denoted by Ms{A} where A is the end-point id. Ms{A} = [A-B,A-C].

Each end-point also maintains set of tunnels for each incoming tunnel. These tunnels are called intermediate mesh and represented by receiving Tunnel and intermediate end-point address. The intermediate Mesh is denoted by Mi{k,t}; where j is the intermediate node and t is the tunnel that traffic is arriving. As an example intermediate mesh for tunnel A-C at end-point C is Mi{C, A-C}. Mi{C,A-C} = [D].

The above source and intermediate graphs(mesh) are derived using a Minimum Spanning Tree or a Shortest Path First Algorithm. Based on the methods presented in discussion the intermediate mesh derived

by an end-point is a subset of source mesh of the corresponding source end-point. As an example Mi{C, A-C} is a subset of Ms{A}. Thus guaranteeing loop free forwarding.

[3] has specified the requirement to maintain a separate Virtual Forwarding Instance (VFI) by each PE device for each Layer 2 NBVPN domain. [3] also specify the requirement for each VFI to contain a flooding scope. Flooding scope of VFI represents tunnels and local ports that any unknown, broadcast or multicast packets should be forwarded. We propose to use multiple flooding scopes; flooding scope for locally originating traffic is called source flooding and denoted by Fs{i} where i is the end-point identifier. At end-point A source flooding scope is denoted by Fs{A}. Flooding scope of non locally originating traffic is called intermediate flooding scope and denoted by Fi{j-t} where j is the intermediate endpoint id and t is the receiving tunnel id. As an example at end-point C there are three intermediate flooding scopes: Fi{C,A-C}, Fi{C, B-C}, Fi{C, D-C}

As a result; in theory each end-point for each Layer 2 NBVPN domain has a single source flooding scope and multiple intermediate flooding scopes (each for each tunnel).

Although exact implementation details of multiple flooding scopes are beyond the scope of this document we would like to present a simple method to implement multiple flooding scopes. A CAM (Content Addressable Memory) lookup with ingress tunnel (port) id and Layer 2 NBVPN domain Id may be used to obtain the appropriate flooding scope. Similarly, if MPLS is used as the tunneling method; incoming Label may be used to derive the corresponding flooding scope.

6.0 Building Blocks

The methods presented in this document can be broadly classified in to four major blocks. These building blocks collectively specify the implementation of Layer 2 NBVPN using partial meshed tunneling topologies.

1. Global policy for reachability preference. It is important that all end-points use the same set of polices. Uses of such policies assure proper forwarding behavior. Reachability preference policies are used to derive the source and intermediate mesh.

2. Advertisement protocol for advertisement of reachability preferences. The protocol used for advertisement of preferences MAY be Link State. Use of such protocol guarantees faster convergence. OSPF Opaque LSA can be easily adapted for the purpose.

3. Tunneling methods. Tunneling protocols such as IPSec or MPLS can be used to implement required tunnels.

4. Tree Algorithm. A Tree generation algorithm that is capable of building loop free graphs that use minimum cost concepts MUST be used. In this discussion we propose to use either a Minimum Spanning Tree Algorithm or Shortest Path First algorithm. Prim Algorithm [4] that is used in 802.1w [5] specification may be used for Minimum Spanning Tree. Dijkstra algorithm that is used in OSPF [6] may be used for Shortest Path First Algorithm.

7.0 Application of Minimum Spanning Tree Algorithm

When using ST there is a single tree for the entire Layer 2 NBVPN, rooted at some node. In order to identify the root node for Layer 2 NBPVN, the participating nodes advertise the node priority. In addition the nodes are also required advertise the reachability preference for each tunnel that originate at the node.

edge representation = {node-id, nexthop-node-id}

node-id = IP address of the node.

node-preference = [integer]

Semantics of node-preference is a global policy. As an example numerically lower numbers may represent higher preference.

7.1 Calculation of Minimum Spanning Tree

Step 1: Select the node with the highest preference as the root node. In the event of a tie use the node-id as the tie barker.

Step 2: Derive the Spanning Tree for the Layer 2 NBVPN using a minimum spanning tree algorithm. Here we propose to use Prim [4] Algorithm for the purpose.

Step 3. Let the derived Spanning Tree is T.

7.2 Calculation of Source Mesh

Step 1. Select source node v.

Step 2. Remove all the edges in the graph T except the edges that are directly connected to the node v. Let say this is graph T'.

Step 3. Graph T' is the source Mesh Ms for node v.

Step 4. The Graph T' represent the set of active tunnels to forward traffic (unknown, broadcast and multicast) originating from the node v.

7.3 Calculation of Intermediate Mesh

Let v is the local node.

Let T' is the source Mesh derived for the Local node v.

Let g is set of all nodes that have a directed edge with the local node v.

for each node u in g

Step 1: Remove corresponding edge in T'.

Step 2: The resultant graph T" is (intermediate mesh) Mi{v, u-v}. Broadcast, unknown and multicast traffic arriving on tunnel {u-v} MUST be forwarded on T".

Repeat the above process for each tunnel starting/terminating at node v.

8.0 Application of Shortest Path First (SPF)algorithm:

We propose to use Dijkstra algorithm for the purpose of calculating Source Mesh and Intermediate Mesh. Nodes (endpoints), edges (tunnels) are represented as below.

edge representation = {node-id, nexthop-node-id}

node-id = IP address of the node.

Preference = [integer]; preference to receive traffic over a tunnel. Deduction of preference is a global policy that all nodes agrees.

Treat each Layer 2 NBVPN domain as a single graph. In analogy to OSPF, that is single area.

8.1 Calculation of Source Mesh

Step 1. Calculate the SPF tree T for the source node v.

Step 2. Remove all the edges in the graph T except the edges that are directly connected to the node v. Let say this is graph T'.

Step 3. Graph T' is the source Mesh Ms for node v.

Step 4. The Graph T' represent the set of active tunnels to forward traffic (unknown, broadcast and multicast) originating from the node v.

8.2 Calculation of Intermediate Mesh

Let v is the local node.

Let g is set of all nodes that have a directed edge with the local node v.

for each node u in g

Step 1: calculate the SPF tree T for node u such that node u is member of g.

Step 2: Traverse from u to each node in T. Remove nodes that does not require traversing via local node v. Let the resultant graph T'.

Step 3: select local node v. Remove all the edges in T' that do not have a direct edge with v. Remove edge {u-v}. Let T'' is the resultant graph.

Step 4: The resultant graph T'' is the set of active tunnels for traffic arriving on tunnel {u-v}. Broadcast, unknown and multicast traffic arriving on tunnel {u-v} MUST be forwarded on T''.

Repeat the above process for each tunnel starting/terminating at node v.

9.0 Comparison between SPF and STP algorithms

When using Spanning Tree algorithm, there is a single Spanning Tree for the given Layer 2 NBVPN domain (graph). The Tree is rooted at the root node that was selected based on some criteria. As a result, path taken by some nodes to reach some other nodes may not be optimal.

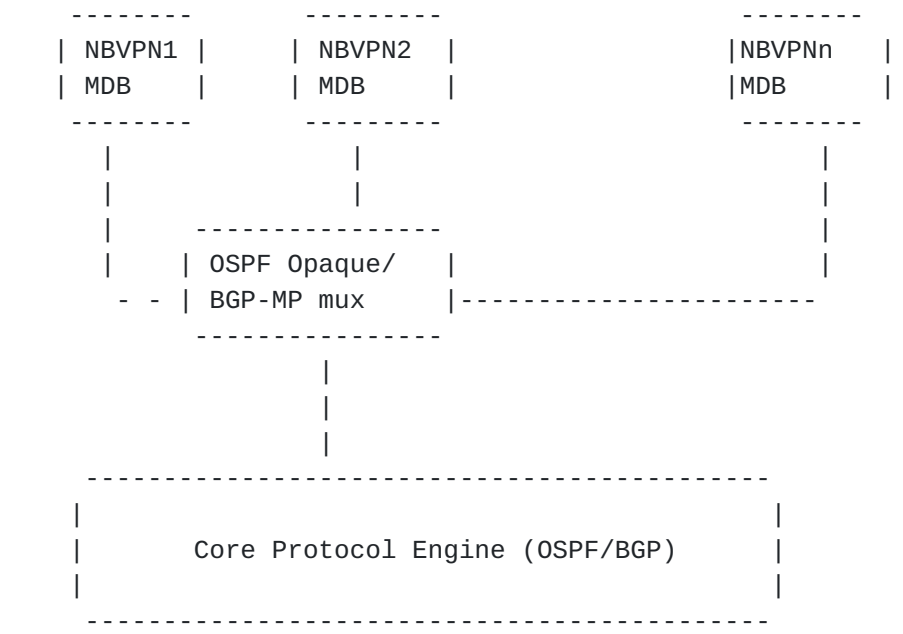
When using Shortest Path First Algorithms, there is a separate Shortest Path Tree for each node. As a result path taken by traffic originating at the node is always assured to take the best path.

However, SPF requires each node to derive SPF trees for each node. On the other hand Spanning Tree algorithm requires deriving only a single tree. All intermediate meshes and Source mesh can be derived from the Spanning Tree (there is only one tree for the network/VPN domain). Hence Spanning Tree requires less iteration of the algorithm.

Dijkstra is a very popular algorithm used to derive Shortest Path Trees. Dijkstra has computational complexity of O(n^2) ; where n is number of nodes.

Prim's algorithm is a popular algorithm used for Spanning Tree. Prim's Algorithm has computational complexity of O(n^2) ; where n is number of nodes. Kruskal's [7] algorithm is a variation of Prim's. Kruskal's Algorithm has a computational complexity of O(E*logE) where E is number of edges and E << N^2.

10.0 Interaction between various components



NBVPN(i)MDB - Represent reachability information received from other end-points. SPF or STP calculation for source Mesh and intermediate Mesh for the NBVPN domain is performed in this context.

OSPF Opaque/BGP-MP Mux - This module performs multiplexing of tunnel reachability information received to the correct NBVPN instance.

Core Protocol Engine - This module represent the protocol implementation.

11.0 Other Applications:

The methods presented in this document may be easily applicable to any other applications that require optimum path selection via transit node. Optical Lambda switching is such application that may use the methods presented in this document.

Internet Exchange Points (IEP) are a new evolving concept. IEP use a shared network fabric to provide multi-party peering for customer

sites. Methods presented in this discussion can be easily adapted to provide multi-party peering using partially meshed networks.

12.0 Security Considerations

A security analysis of the methods presented in this discussion has not yet been performed.

13.0 References

1 Bradner, S., "The Internet Standards Process -- Revision 3", BCP 9, RFC 2026, October 1996.

2 Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997

3 Senevirathne, T., and et.al, "Requirements for Layer 2 Network Based VPN", Work in Progress, May 2001.

4 Gross, J., and Yellen, J, "Graph Theory and its Applications", CRC Press, 1998.

5 IEEE/ISO , "Amendment 2-Rapid Reconfiguration", IEEE802.1w, March 26, 2001.

6 Moy, J., OSPF Version 2, RFC 1583, March 1994.

7 Aho, A.V., and et.al., "Data Structures and Algorithms", Addison-Wesley 1983.

14.0 Acknowledgments

Several people provided suggestions and comments and volunteered to review this draft. The suggestions and feedback received helped this document to evolve to a draft. Waldemar Augustyn and Andrew Smith provided valuable suggestions and comments.

15.0 Author's Addresses

Tissa Senevirathne
Force10 Networks
1440 McCarthy Blvd
Milipitas, CA 95035
Phone: 408-965-5103
Email: tissa@force10networks.com

Full Copyright Statement

"Copyright (C) The Internet Society (2001). All Rights Reserved. This document and translations of it may be copied and furnished to others, and derivative works that comment on or otherwise explain it or assist in its implmentation may be prepared, copied, published and distributed, in whole or in part, without restriction of any kind, provided that the above copyright notice and this paragraph are included on all such copies and derivative works. However, this document itself may not be modified in any way, such as by removing the copyright notice or references to the Internet Society or other Internet organizations, except as needed for the purpose of developing Internet standards in which case the procedures for copyrights defined in the Internet Standards process must be followed, or as required to translate it into