PPVPN Working Group
Internet Draft
Expiration Date: May 2002

Heinrich Hummel,
Jochen Grimminger
Siemens AG

November 2001

Partially meshed base tunnels plus hierarchical mp2p tunnel sequence LSPs
draft-hummel-ppvpn-mp2p-tunnel-sequencing-00.txt

Status of this Memo

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 presented ID fits into ppvpn, ccamp and into idr of the Routing Area.

WHY IS IT TARGETED AT THIS WG(s)

The application in mind for which to setup hierarchical multipoint-to-point tunnel sequence LSPs, as described, is CE-based VPN as well as Network(PE)-based VPN.

The detailed C-Plane aspects (procedures, messages, TLVs) for setting up hierarchical mp2p tunnel sequence LSPs, i.e. for concatenating

some base tunnels to different tree-shaped tunnel sequences, would be a work item for ccamp.

The details w.r.t. distribution/discovery of all base tunnels to/by all targetted communities (VRFs) would be an extension of MP-BGP and subject for idr.

JUSTIFICATION

So far, any-CE-to-any-CE connectivity may either mean full mesh CE-CE-tunneling (in CE-based VPNs) or full mesh PE-PE tunneling (in Network(PE)-based VPNs). Accordingly, a CE-based VPN with 50 000 CEs (which is a stated requirement) would need 2,499,950 unidirectional CE-CE-tunnels; a network(PE)-based VPN with 1000 PEs would need 999,000 PE-PE uni-dir.tunnels. However, by using only a partial mesh, e.g. a chessboard mesh, n nodes may be fully interconnected using less than 4 * n unidir. base tunnels plus n tree-shaped tunnel sequence LSPs where the base tunnels are reused again and again for conveying traffic to each egress node. Even more, by installing several, differently routed tree shaped tunnel sequence LSPs rooted at the same egress node, such nice services like path protection, traffic balancing and QoS-/SLA-/traffic type-specific tunneling can easily be supported without needing any extra tunnel.

Abstract

In order to provide any-CE-to-any-CE connectivity it is proposed to deploy an O(n)-sized set of base tunnels which form a reasonable partial mesh (e.g. chessboard topology), and to use them and re-use them many times as elements of concatenated hierarchical multipoint-to-point tunnel sequence LSPs. The number of saved tunnels is of order n-square. Multiple,differently routed tunnel sequence LSPs but rooted at the same node may cater for features like traffic balancing, path protection, QoS-routing, etc., without needing any extra tunnel. It is outlined how to establish such an optimal VPN inter-site tunneling.

1 Introduction

In order to provide any-CE-to-any-CE connectivity it is proposed to deploy an O(n)-sized set of base tunnels which form a reasonable partial mesh (e.g. chessboard topology), and to use them and re-use them many times as elements of concatenated hierarchical multipoint-to-point tunnel sequence LSPs.

For a CE-based VPN with N CEs, N such mp2p hierarchical tunnel sequence LSPs would be required, whereby each such hierarchical LSP is composed of some CE-CE base tunnels and is egressing at a different CE. There may be even x (e.g. x=3 or 4) such hierarchical LSPs per egress-CE, which are differently routed, i.e. which are composed by different CE-CE base tunnels, as to support traffic balancing, path protection, QoS-specific routing, etc.

For a network(PE)-based VPN with n PEs, n such hierarchical mp2p tunnel sequence LSPs would be required, whereby each such hierarchical LSP is composed of some PE-PE base tunnels and is egressing at a different PE. There may be even x (e.g. x=3 or 4) such hierarchical LSPs per egress-PE, which are differently routed, i.e. which are composed by different PE-PE base tunnels, as to support traffic balancing, path protection, QoS-specific routing,etc. The so-called VC-Label,which selects the remote interface to the remote CE, would become the "third" label instead of the "second" label in the label stack.

The draft outlines how such optimal VPN inter-site tunneling may be established. Hereby, for reasons of simplicity, the draft is focussed on network(PE)-based VPNs using uni-directional LSPs (covering CE-based VPNs and bi-directional LSPs as well wouldn't take extra inspiration, but extra transpiration).

Extensions/modifications to LDP and RSVP-TE as well as to MP-BGP will be required.

2 Savings and advantages

Full mesh tunneling among n nodes requires n*(n-1) uni-directional tunnels.

The absolute minimal number of tunnels to interconnect all nodes contiguously would be 2 * (n-1), whereby any two nodes, which are connected by some tunnel in one direction, are also connected by the inverse-directional tunnel. Hereby the tunnel topology has no meshes and looks like a tree. An algorithm for selecting those tunnels which form the tree with the smallest weight sum can be found in [1].

However, without any meshes it is impossible to use alternate routes for traffic balancing, path protection, etc. Therefore, let's assume a set of such pairs of mutual inverse base tunnels which form a reasonable partial mesh, e.g. a chessboard topology. Hereby n nodes are contiguously webbed together by less than 4 * n (uni-directional) tunnels.

Compared with full mesh we save at least n*(n-1) -4*n = n*(n-5) tunnels. Note that the establishment of each of these tunnels would involve P-routers and would consume one label at each P-router. Accordingly are the effects, if we can save so many "base" tunnels.

Also note, that from the statistical multiplex performance's point of view it is better to have only 4*n tunnels with more bandwidth than n*(n-1) tunnels with less bandwidth.

In order to provide any-to-any connectivity based on only the chessboard meshed set of base tunnels, we need to establish some tree-shaped tunnel sequence LSPs, which would convey all user traffic from any of n-1 ingress nodes to one specific egress node.

We may even want to have x such LSPs which eventually are differently routed (i.e. composed by different sets of base tunnels) but convey traffic to the same egress node. We may want this for reasons of traffic balancing, fast rerouting, QoS/SLA/traffic-type-specific tunneling.
Note that in a full mesh scenario you would need x * n * (n-1) tunnels to get the same benefits !

The immense saving with respect to base tunnels allows us to be less keen on PE-PE tunnel sharing: We can afford NOT to share some PE-PE base tunnel in case the respective customers are not allied. This opens the capability to do VPN-specific bandwidth policing. Otherwise you won't know which customer is to blame for his SLA-violation.

Multicast:
The fundamental goal of multicast is to avoid retransmitting the same user data multiple times over the same physical link. In the full mesh scenario the chance is big that the same data is transmitted multiple times (up to n-1 times) over the same physical link. This is much less likely, in case there are multiple transit base tunnels involved. Nevertheless, Multicast is for further study.

A particular base tunnel may be used many times, i.e. for transmitting user traffic to different egress nodes. A hierarchical label, let's call it the "Tunnel-Sequence-Label" will take care that the user traffic which exits some base tunnel will enter the next base tunnel on its way to some egress node. As mentioned above, we

need trees of tunnel sequences. As we will see, no P-router will ever become aware of the existence of any tunnel sequence LSP! This includes: A P-router will never have to use its label space for assigning some Tunnel-Sequence-Label.

3 Establishing the elementary base tunnels

Each PE which is involved in a particular VPN/Community needs to know the entire respective set of PE-PE-base-tunnels. This information may either be provided by network configuration or by MP-BGP advertisement.

Hereby, a particular VRF at a particular PE must EXACTLY know this set. Which tunnels are part of this set may depend on some policy:

a) Base tunnels dedicated to some customer:
PE-PE-base tunnel exclusively belongs to some specific VPN/Community. It may be appropriate to have several base tunnels from PE-x to PE-y, whereby some of them belong to VPN/Community A while others belong to VPN/Community B.

b) All base tunnels are shared by all customers:
Base tunnel from PE-x to PE-y participates (transitively) in a tree-shaped tunnel sequence LSP rooted in PE-z, though the customers with sites at PE-z have neither sites at PE-x nor at PE-y.

c) Some base tunnels are shared by all customers, some other base tunnels are dedicated to individual customers.

MP-BGP may advertise, by means of a new TLV inside the MP_REACH_NLRI, all characteristic data of each base tunnel to be built and may convey them to the right VRFs at the right PEs by means of some Route Target Community in the Extended Communities attribute - in accordance with the mentioned policy.

Characteristic data of a base tunnel comprises: tunnel endpoints, direction, usage (i.e for User and/or Control plane), bandwidth,FEC at its egress endpoint, "color".

The respective adjacent VRFs may initiate the establishment of these base tunnels, which includes assigning their LSP-IDs. As a result of a base tunnel establishment its ingress-PE must store its first hop interface and its first hop Tunnel-Label in such a way that these information can be retrieved based on the LSP-ID again.

When the base tunnels have been established MP-BGP shall advertise them once again, enhanced with their LSP-IDs.

As a result each VRF of some VPN/Community will learn the entire topology composed of all those established base tunnels it is supposed to know.

Prior to MP-BGP, network configuration has properly been done, if all PE nodes of a VPN/Community are contiguously interwebbed by base tunnels, and if any pair of nodes, which is connected by some U-plane base tunnel, is also connected by two (mutually inverse) C-Plane base tunnels.

4 Establishing the hierarch. Multipoint-to-point Tunnel Sequence LSP

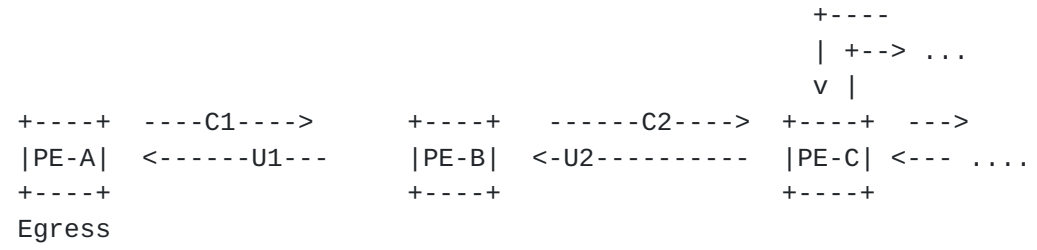
A particular egress-PE computes a Dijkstra-spanning tree of base tunnels which are all directed towards this egress-PE. Hereby QoS/SLA/traffic-type- specific constraints may have influenced the selection of the participating base tunnels. This tree of base tunnels will be the "U-plane tree".

For each of the participating tunnels there exists an inverse base tunnel. All these inverse tunnels form the "C-plane tree".

Using the unsolicit mode, the egress-PE sends out an explicitly routed LABEL MAPPING message along the p2mp C-plane tree. The LABEL MAPPING message will only be seen by the end nodes of the C-Plane base tunnels (which are PEs) and not by any P-router. In order to direct the message down the C-Plane tree, an extension to the existing Explicit Route TLV (ER-TLV) is needed:

New "("-TLVs and ")"-TLVs should be inserted between the ER-HOP-TLVs in accordance with the structure of the tree route. A simple algorithm at any transit PE will take care that the proper part of the received ER-TLV is forwarded, especially at the merging nodes of the U-Plane tree.

Furthermore, the ER HOP-TLV needs to be enhanced as well: It shall be able to carry two LSP-IDs (one for the next U-Plane base tunnel, and one for the respectively inverse C-Plane base tunnel).



Above figure shows some part of the p2mp C-plane tree (i.e. C1,C2) and some part of the mp2p U-plane tree (i.e. U1,U2), both rooted in

PE-A.

(Note: C-Plane base tunnels may eventually also be used as U-Plane base tunnels which participate in U-Plane trees rooted somewhere else. Also, any C-Plane base tunnel may participate in differently rooted C-plane trees. Any U-Plane base tunnel may participate in differently rooted U-Plane trees)

The U-plane tree is built by installing base tunnel bindings at the transit PEs, will say by "Multiprotocol Tunnel-Sequence-Label Switching". As a matter of fact, the U-plane tree is a hierarchical multipoint-to-point LSP, whose labels shall be called "Tunnel-Sequence-Label".

PE-A in above figure assigns an available Tunnel-Sequence-Label x (it is a label like any other Tunnel-Label) and writes it into the Label-TLV of the LABEL MAPPING message. Furthermore, PE-A provides an ER-TLV whose top-most ER HOP-TLV shall contain a) LSP-ID U1 and b) LSP-ID C1, followed by an ER HOP -TLV which contains a) LSP-ID U2 and b) LSP-ID C2.

When PE-B receives the LABEL MAPPING message, it assigns an own Tunnel-Sequence-Label y and creates an NHLFE which is retrievable based on y. The NHLFE entry shall contain the following information:

- PE-B's physical interface of base tunnel U1

- PE-B's Tunnel Label of base tunnel U1
  (PE-B must know the two information at this
  point in time, i.e. it must be able to retrieve them
  by means of LSP-ID U1 from the top-most ER HOP-TLV)

- Tunnel Sequence Label x.

PE-B forwards the LABEL MAPPING message to PE-C thru C-Plane base tunnel C2. For doing this, it locally looks up the physical interface as well as the first Tunnel-Label of C2 based on LSP-ID C2. Hereby it forwards the Tunnel-Sequence-Label y (inside the Label-TLV) and also the ER-TLV, however without the top-most ER-HOP-TLV. ,sp 1 Any transit PE, where the U-Plane tree is supposed to merge, may assign further Tunnel-Sequence-Labels (e.g. y2, y3,..) for each branch and map them to (the same) Label-Sequence-Label x analogously. Hereby a well deterministic algorithm must crack the received ER-TLV in compliance with the contained "("-TLVs and ")"-TLVs and determine all ER HOP-TLVs that contain LSP-IDs of adjacent U-Plane and C-Plane base tunnels.

Additionally, penultimate label popping may apply to Tunnel Sequence

Labels just like to plain Tunnel Labels.

So far the establishment of the hierarchical multipoint-to-point Tunnel Sequence LSP in LDP syntax. An analogous description in RSVP-TE syntax may be provided in a follow-up draft version.

5 Carrier's carrier network

The preceding sections have shown how Multiprotocol Tunnel-Sequence-Label Switching could be done as to save order-n-square many tunnels and provide even better services than in a simple full mesh scenario.

Additionally, we may consider network(PE)-based VPNs, which are provided by some "virtual service provider" whose network consists of a set of "Service Provider"-LSPs, provided by some "carrier's carrier".

Indeed, the technology outlined above, may somehow recursively be repeated in such a scenario. The VPN-related base tunnels may be tunnel sequence LSPs themselves, more precisely: The "VPN"-base tunnel is a (linear, not merging) hierarchical "Service Provider Tunnel Sequence" LSP.

Consequently, a user packet may have a label stack which contains:
- Service Provider Tunnel -Label
- Service Provider Tunnel-Sequence-Label
- VPN Tunnel-Sequence-Label
- VC-Label

The NHLFEs that perform the label binding for a "VPN"- U-Plane tree rooted at some VPN/Community-specific egress PE may contain:
- physical interface of the (first) SP-base tunnel of a particular VPN-base tunnel
- Tunnel-label of the (first) SP-base tunnel of a particular VPN-base tunnel
- Tunnel-Sequence Label which concatenates some SP-base tunnels to one VPN-base tunnel
- Tunnel-Sequence Label of the U-Plane tree rooted at some egress-PE of some specific VPN

The extensions of LDP, RSVP-TE and MP-BGP may be done such generalized that this scenario is covered as well.

6 Conclusion, outlook

By relatively small protocol extensions (LDP, RSVP-TE, MP-BGP) the notorious N-Square Problem would be eliminated once and forever: Signalling would catch up with what the Label Stack paradigma has promised since the beginning of MPLS.

Network(PE)-based VPNs as well as CE-based VPN may benefit immensely. But other applications/scenarios (typically overlay scenarios) may benefit as well. For instance, MPOA, the ATM Forum's Overlay models could reduce its n-square number of SVCs by introducing hierarchical multipoint-to-point SVCs as well.

7 Intellectual Property Considerations

This proposal in is full conformity with [RFC-2026].

Siemens may have patent rights on technology described in this document which employees of Siemens contribute for use in IETF standards discussions. In relation to any IETF standard incorporating any such technology, Siemens hereby agrees to license on fair, reasonable and non-discriminatory terms, based on reciprocity, any patent claims it owns covering such technology, to the extent such technology is essential to comply with such standard.

8 References

[1] H.Hummel (Siemens AG): Tree/Ring/Meshy VPN tunnel systems
draft-hummel-ppvpn-tunnel-systems-00.txt

8 Authors' Addresses

Heinrich Hummel
Siemens AG
Hofmannstrasse 51
81379 Munich, Germany
Tel: +49 89 722 32057
Email: heinrich.hummel@icn.siemens.de

Jochen Grimminger
Siemens AG
Otto-Hahn-Ring 6
81739 Munich, Germany
Tel.+49 89 636 417410
Email: Jochen.Grimminger@mchp.siemens.de

Full Copyright Statement