Network Working Group
Internet Draft
Expiration Date: May 2001

Kireeti Kompella
Manoj Leelanivas
Quaizar Vohra
Juniper Networks

Javier Achirica
Telefonica Data

Ronald Bonica
WorldCom

Chris Liljenstolpe
Cable & Wireless

Eduard Metz
KPN Dutch Telecom

Chandramouli Sargor
Vijay Srinivasan
CoSine Communications

# MPLS-based Layer 2 VPNs

draft-kompella-mpls-l2vpn-02.txt

## 1. Status of this Memo

## 2. Abstract

Virtual Private Networks (VPNs) based on Frame Relay or ATM circuits have been around a long time. While these VPNs work well, the costs of maintaining separate networks for Internet traffic and VPNs and the administrative burden of provisioning these VPNs have led Service Providers to look for alternative solutions. In this document, we present a VPN solution where from the customer's point of view, the VPN is based on Layer 2 circuits, but the Service Provider maintains and manages a single MPLS-based network for IP, MPLS IP VPNs, and Layer 2 VPNs.

## 3. Introduction

The first corporate networks were based on dedicated leased lines interconnecting the various offices of the corporation. Such networks offered connectivity and little else: they didn't scale well, they were expensive for the service providers (and hence for their customers), and provisioning them was a slow and arduous task.

The first Virtual Private Networks (VPNs) were based on Layer 2 circuits: X.25, Frame Relay and ATM (see [VPN]). Layer 2 VPNs were easier to provision, and virtual circuits allowed the service provider to share a common infrastructure for all the VPNs. These features were passed on to the customers in terms of cost savings. However, while Layer 2 VPNs were a significant step forward from dedicated lines, they still had their drawbacks. First, they tied the service provider VPN infrastructure to a single medium (e.g., ATM). This became even more of a burden if the Internet infrastructure was to share the same physical links. Second, the Internet infrastructure and the VPN infrastructure, even if they shared the same physical network, needed separate administration and maintenance. Third, while provisioning was much easier than for dedicated lines, it was still complex. This was especially evident in the effort to add a site to an existing VPN.

This document offers a solution that preserves the advantages of a Layer 2 VPN while allowing the Service Provider to maintain and manage a single (MPLS-based) network for IP, MPLS IP VPNs ([IPVPN]) and Layer 2 VPNs, and reducing the provisioning problem significantly. In particular, adding a site to an existing VPN in most cases requires configuring just the Provider Edge router connected to the new site.

The rest of this section discusses the relative merits of MPLS-based Layer 2 and Layer 3 VPNs. Section 4 describes the operation of an MPLS-based Layer 2 VPN. Sections 5 and 6 offer two alternative means

of signalling Layer 2 VPNs, one using LDP and the other using BGP.

### 3.1. Terminology

We assume that the reader is familiar with Multi-Protocol Label Switching (MPLS [MPLS]), the Label Distribution Protocol (LDP [LDP]) and the Border Gateway Protocol version 4 (BGP [BGP]).

The terminology we use follows. A "customer" is a customer of a Service Provider seeking to interconnect the various "sites" (independently connected networks) through the Service Provider's network, while maintaining privacy of communication and address space. The device in a customer site that connects to a Service Provider router is termed the CE (customer edge device); this device may be a router or a switch. The Service Provider router to which a CE connects is termed a PE. A router in the Service Provider's network which doesn't connect directly to any CE is termed P. These definitions follow those given in [IPVPN].

### 3.2. Advantages of Layer 2 VPNs

We define a Layer 2 VPN as one where a Service Provider provides a layer 2 network to the customer. As far as the customer is concerned, they have (say) Frame Relay circuits connecting the various sites; each CE is configured with a DLCI with which to talk to other CEs. Within the Service Provider's network, though, the layer 2 packets are transported within MPLS Label-Switched Paths (LSPs).

The Service Provider does not participate in the customer's layer 3 network, in particular, in the routing, resulting in several advantages to the SP as a whole and to PE routers in particular.

#### 3.2.1. Separation of Administrative Responsibilities

In a Layer 2 VPN, the Service Provider is responsible for Layer 2 connectivity; the customer is responsible for Layer 3 connectivity, which includes routing. If the customer says that host x in site A cannot reach host y in site B, the Service Provider need only demonstrate that site A is connected to site B. The details of how routes for host y reach host x are the customer's responsibility.

Another very important factor is that once a PE provides Layer 2 connectivity to its connected CE, its job is done. A misbehaving CE can at worst flap its interface. On the other hand, a misbehaving CE

in a Layer 3 VPN can flap its routes, leading to instability of the PE router or even the entire SP network. This means that the Service Provider must aggressively damp route flaps from a CE; this is common enough with external BGP peers, but in the case of VPNs, the scale of the problem is much larger; also, the CE-PE routing protocol may not be BGP, and thus not have BGP's flap damping control.

### 3.2.2. Migrating from Traditional Layer 2 VPNs

Since "traditional" Layer 2 VPNs (i.e., real Frame Relay circuits connecting sites) are indistinguishable from MPLS-based VPNs from the customer's point-of-view, migrating from one to the other raises few issues. With Layer 3 VPNs, special care has to be taken that routes within the traditional VPN are not preferred over the Layer 3 VPN routes (the so-called "backdoor routing" problem, whose solution requires protocol changes that are somewhat ad hoc).

### 3.2.3. Privacy of Routing

In a Layer 2 VPN, the privacy of customer routing is a natural fallout of the fact that the Service Provider does not participate in routing. The SP routers need not do anything special to keep customer routes separate from other customers or from the Internet; there is no need for per-VPN routing tables, and the additional complexity this imposes on PE routers.

### 3.2.4. Layer 3 Independence

Since the Service Provider simply provides Layer 2 connectivity, the customer can run any Layer 3 protocols they choose. If the SP were participating in customer routing, it would be vital that the customer and SP both use the same layer 3 protocol(s) and routing protocols.

### 3.2.5. Multicast Routing

Supporting IP multicast over MPLS-based Layer 3 VPN is as yet undocumented.

In the Layer 2 VPN case, the CE routers run native multicast routing directly. The SP backbone just provides pipes to connect the CE routers; whether the CE routers run IP unicast or IP multicast or some other network protocol is irrelevant to the SP routers.

### 3.2.6. PE Scaling

In the Layer 2 VPN scheme described below, each PE transmits a single small chunk of information about every CE that the PE is connected to to every other PE. That means that each PE need only maintain a single chunk of information from each CE in each VPN, and keep a single "route" to every site in every VPN. This means that both the Forwarding Information Base and the Routing Information Base scale well with the number of sites and number of VPNs. Furthermore, the scaling properties are independent of the customer: the only germane quantity is the total number of VPN sites.

This is to be contrasted with Layer 3 VPNs, where each CE in a VPN may have an arbitrary number of routes that need to be carried by the SP. This leads to two issues. First, both the information stored at each PE and the number of routes installed by the PE for a CE in a VPN can be (in principle) unbounded, which means in practice that a PE must restrict itself to installing routes associated with the VPNs that it is currently a member of. Second, a CE can send a large number of routes to its PE, which means that the PE must protect itself against such a condition. Thus, the SP must enforce limits on the number of prefixes accepted from a CE; this in turn requires the PE router to offer such control.

The scaling issues of Layer 3 VPNs come into sharp focus at a BGP route reflector (RR). An RR cannot keep all the advertised routes in every VPN since the number of routes will be too large. The following solutions/extensions are needed to address this issue:

1) RRs could be partitioned so that each RR services a subset of VPNs so that no single RR has to carry all the routes. This method has the disadvantage that a PE changing its VPN membership could force a change in the RR configuration, and would require carefully constructing RR topologies.
2) An RR could use a preconfigured list of Route-Targets for its inbound route filtering. The RR may also need to install Outbound Route Filters [BGP-ORF] which contain the above list of Route-Targets on each of its peers so that they do not send unnecessary VPN routes. This method also requires significant extensions along with the fact that multiple RRs are needed to service different sets of VPNs.

### 3.2.7. Ease of Configuration

Configuring traditional Layer 2 VPNs was a burden primarily because of the O(n*n) nature of the task. If there are n CEs in a Frame Relay VPN, say full-mesh connected, n*(n-1)/2 DLCI PVCs must be provisioned across the SP network. At each CE, (n-1) DLCIs must be configured to reach each of the other CEs. Furthermore, when a new CE is added, n new DLCI PVCs must be provisioned; also, each existing CE must be updated with a new DLCI to reach the new CE.

In our proposal, the provisioning of "PVCs" across the SP network is handled by signalling protocols (LDP, RSVP-TE), reducing a large part of the provisioning burden. Furthermore, we assume that DLCIs at the CE edge are relatively cheap; and labels in the SP network are cheap. This allows the SP to "over-provision" VPNs: for example, allocate 50 CEs to a VPN when only 20 are needed. With this over-provisioning, adding a new CE to a VPN requires configuring just the new CE and its associated PE; existing CEs and their PEs need not be re-configured.

## 3.3. Advantages of Layer 3 VPNs

Layer 3 VPNs ([IPVPN] in particular) offer a good solution when the customer traffic is wholly IP, customer routing is reasonably simple, and the customer sites connect to the SP with a variety of Layer 2 technologies.

### 3.3.1. Layer 2 Independence

One major restriction in a Layer 2 VPN is that the Layer 2 medium with which the various sites of a single VPN connect to the SP must be uniform. On the other hand, the various sites of a Layer 3 VPN can connect to the SP with any supported media; for example, some sites may connect with Frame Relay circuits, and others with Ethernet.

A corollary to this is that the number of sites that can be in a Layer 2 VPN is determined by the number of Layer 2 circuits that the Layer 2 technology provides. For example, if the Layer 2 technology is Frame Relay with 2-octet DLCIs, a CE can connect to at most about a thousand other CEs in a VPN.

### 3.3.2. SP Routing as Added Value

Another problem with Layer 2 VPNs is that the CE router in a VPN must be able to deal with having N routing peers, where N is the number of sites in the VPN. This can be alleviated by manipulating the topology of the VPN. For example, a hub-and-spoke VPN architecture means that only one CE router (the hub) needs to deal with N neighbors. However, in a Layer 3 VPN, a CE router need only deal with one neighbor, the PE router. Thus, the SP can offer Layer 3 VPNs as a value-added service to its customers.

Moreover, with layer 2 VPNs it is up to a customer to build and operate the whole network. With Layer 3 VPNs, a customer is just responsible for building and operating routing within each site, which is likely to be much simpler than building and operating routing for the whole VPN. That, in turn, makes Layer 3 VPNs more suitable for customers who don't have sufficient routing expertise, again allowing the SP to provide added value.

### 3.3.3. Class-of-Service

Class-of-Service issues have been addressed for Layer 3 VPNs. Since the PE router has visibility into the network layer (IP), the PE router can take on the tasks of CoS classification and routing.

Class-of-Service issues for Layer 2 VPNs will be addressed in a future revision.

## 4. Operation of a Layer 2 VPN

The following simple example of a customer with 4 sites connected to 3 PE routers in a Service Provider network will hopefully illustrate the various aspects of the operation of a Layer 2 VPN. For simplicity, we assume that a full-mesh topology is desired.

In what follows, Frame Relay serves as the Layer 2 medium, and each CE has multiple DLCIs to its PE, each to connect to another CE in the VPN. If the Layer 2 medium were ATM, then each CE would have multiple VPI/VCIs to connect to other CEs. For PPP and Cisco HDLC, each CE would have multiple physical interfaces to connect to other CEs.

### 4.1. Network Topology

Consider a Service Provider network with edge routers PE0, PE1, and PE2. Assume that PE0 and PE1 are IGP neighbors, and PE2 is more than one hop away from PE0.

Suppose that a customer C has 4 sites S0, S1, S2 and S3 that C wants to connect via the Service Provider's network using Frame Relay. Site S0 has CE0 and CE1 both connected to PE0. Site S1 has CE2 connected to PE0. Site S2 has CE3 connected to PE1 and CE4 connected to PE2. Site S3 has CE5 connected to PE2. (See the Figure 1 below.) Suppose further that C wants to "over-provision" each current site, in expectation that the number of sites will grow to at least 10 in the near future. However, CE4 is only provisioned with 9 DLCIs.

Suppose finally that CE0 and CE2 have DLCIs 100 through 109 free; CE1 and CE3 have DLCIs 200 through 209 free; CE4 has DLCIs 107, 209, 265, 301, 414, 555, 654, 777 and 888 free; and CE5 has DLCIs 417-426.

### 4.2. Configuration

The following sub-sections detail the configuration that is needed to provision the above VPN. For the purpose of exposition, we assume that the customer will connect to the SP with Frame Relay circuits, and that the customer's IGP of choice is OSPF.

While we focus primarily on the configuration that an SP has to do, we touch upon the configuration requirements of CEs as well. The main point of contact in CE-PE configuration is that both must agree on the DLCIs that will be used on the interface connecting them.

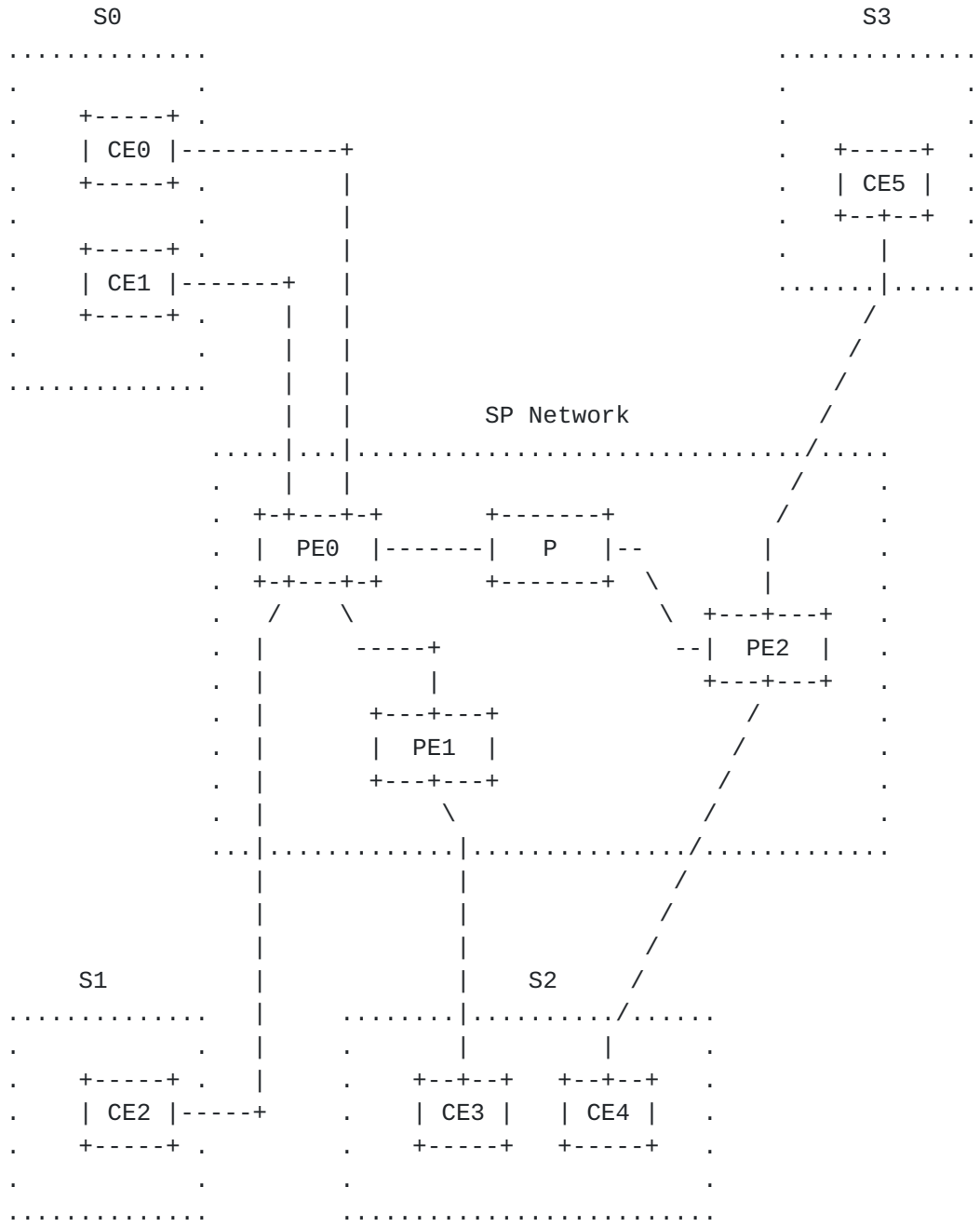
If the PE-CE connection is Frame Relay, it is recommended to run LMI between the PE and CE with the PE as DCE and the CE as DTE. For the case of ATM VCs, OAM cells may be used; for PPP and Cisco HDLC, keepalives may be used.

#### 4.2.1. CE Configuration

Each CE that belongs to a VPN is given a "CE ID". CE IDs must be unique in the context of a VPN. We assume that the CE ID for CE-k is k. Each CE is also configured with a maximum number of CEs that it can connect to; this is the CE's "range".

Each CE is configured to communicate with its corresponding PE with the set of DLCIs given above; for example, CE0 is configured with

Figure 1: Example Network Topology




DLCIs 100 through 109. OSPF is configured to run over each DLCI.

Each CE also "knows" which DLCI connects it to each other CE. A simple algorithm is to use the CE ID of the other CE as an index into the DLCI list this CE has (with zero-based indexing, i.e., 0 is the first index). For example, CE0 is connected to CE3 through its

fourth DLCI, 103; CE4 is connected to CE2 by the third DLCI in its list, namely 265. This is the methodology used in the examples below; the actual methodology used to pick the DLCI to be used is a local matter; the key factor is that CE-k may communicate with CE-m using a different DLCI from the DLCI that CE-m uses to communicate to CE-k, i.e., the SP network effectively acts as a giant Frame Relay switch. This is very important, as it decouples the DLCIs used at each CE site, making for much simpler provisioning.

#### 4.2.2. PE Configuration

Each PE is configured with the VPNs in which it participates. Each VPN has an VPN ID that is unique within the SP network. For each VPN, the PE has a list of CEs that are members of that VPN. For each CE, the PE knows the CE ID, which DLCIs to expect from the CE, and the CE's range.

#### 4.2.3. Adding a New Site

The first step in adding a new site to a VPN is to pick a new CE ID. If all current members of the VPN are over-provisioned, i.e., their range includes the new CE ID, adding the new site is a purely local task. Otherwise, the sites whose range doesn't include the new CE ID and wish to communicate directly with the new CE must have their ranges increased to incorporate the new CE ID.

The next step is ensuring that the new site has the required connectivity (see below). This may require tweaking the connectivity mechanism; however, in several common cases, the only configuration needed is local to the PE to which the CE is attached.

The rest of the configuration is a local matter between the new CE and the PE to which it is attached.

It bears repeating that the key to making additions easy is over-provisioning. However, what is being over-provisioned is the number of DLCIs/VCIs that connect the CE to the PE. This is a local matter, and generally is not an issue.

### 4.3. PE Information Exchange

When a PE is configured with all the needed information for a CE, it first of all chooses a contiguous set of labels with n labels, where n is the CE's range. Call the smallest label in this set the label-base. The PE then advertises (for this CE): its Router ID, the VPN

ID, the CE ID, the CE's range, and the label-base. This is the basic Layer 2 VPN advertisement. This same advertisement is sent to all other PEs. Note that PEs that may not be part of the VPN can receive and keep this information, in case at some future point, a CE connected to the PE joins the VPN.

If the PE-CE connection goes down, or the CE configuration is removed, the above advertisement is withdrawn.

#### 4.3.1. PE Advertisement Processing

When a PE receives a Layer 2 VPN advertisement, it checks if the VPN ID matches any VPN that it is a member of. If not, the PE just stores the advertisement for future use.

Otherwise, suppose the advertisement is from PE A for VPN X, CE m with range Rm and label base Lm. For each CE that the receiving PE B is connected to that is a member of VPN X, PE B does the following.

0) Look up the configuration information associated with the CE. If the encapsulation type for VPN X in the advertisement does not match the configured encapsulation type for VPN X, stop.
1) Say the configured CE ID is k, the range is Rk, and the DLCI list is Dk[]. Also, get the label base PE B allocated for this CE, say Lk.
2) Check if k = m. If so, issue an error: "CE ID k has been allocated to two CEs in VPN X (check CE at PE A)". Stop.
3) Check if k >= Rm, or m >= Rk. If so, issue a warning: "Cannot communicate with CE m (PE A) of VPN X: outside range". Stop.
4) Look in the appropriate table to see which label will get to PE A. This is the "outer" label, Z.
5) The DLCI that CE-k will use to talk to CE-m is Dk[m]. The "inner" label for sending packets to CE-m is (Lm + k). The "inner" label on which to expect packets from CE-m is (Lk + m).
6) Install a "route" such that packets from CE-k with DLCI Dk[m] will be sent with outer label Z, inner label (Lm + k). Also, install a route such that packets received with label (Lk + m) will be mapped to DLCI Dk[m] and be sent to CE-k.
7) Activate DLCI Dk[m] to the CE. This can be done using LMI.

If an advertisement is withdrawn, the appropriate DLCI must be de-activated, and the corresponding routes must be removed from the forwarding table.

### 4.3.2. Example of PE Advertisment Processing

Consider the example network of Figure 1. Let the VPN connecting S0, S1, S2 and S3 has a VPN id of 1. Suppose PE2 receives an advertisement from PE0 for VPN 1, CE ID 0 with CE range R0 = 10 and label base L0 = 1000. Since PE2 is connected to CE4 which is also in VPN 1, PE2 does the following:

0) Look up the configuration information associated with CE4. The advertised encapsulation type matches the configured encapsulation type (both are Frame Relay), so proceed.
1) CE4's range R4 is 9, its DLCI list D4[] is [ 107, 209, 265, 301, 414, 555, 654, 777, 888], and its label base L4 is 4000.
2) CE0 and CE4 have ids 0 and 4 respectively, so step 2 of 4.3.1 is skipped.
3) Since CE4's id is less than R0, and CE0's id is less than R4, step 3 of 4.3.1 is skipped.
4) Look in the appropriate table on PE2 to see which label will get to PE0. Let the label be 10001.
5) The DLCI that CE4 will use to talk to CE0 is D4[0], i.e., 107. The inner label for sending packets to CE0 is (L0 + 4), i.e 1004. The inner label on which to expect packets from CE0 is (L4 + 0), i.e., 4000.
6) Install a "route" such that packets from CE4 with DLCI 107 will be sent with outer label 10001, inner label 1004. Also, install a route such that packets received with label 4000 will be mapped to DLCI 107 and be sent to CE4.
7) Activate DLCI 107 to CE4.

Since CE5 is also attached to PE2, PE2 needs to do processing similar to the above for CE5.

Similarly, when PE0 receives an advertisment from PE2 for VPN1, CE4, with range R4 = 9, and label base L4 = 4000. PE0 processes the advertisment for CE0 (and CE1, which is also in VPN 1).

0) Look up the configuration information associated with CE0. The advertised encapsulation type matches the configured encapsulation type (both are Frame Relay), so proceed.
1) CE0's range, R0, is 9, its DLCI list D0[] is [100 - 109], and its label base L0 is 1000.
2) CE0 and CE4 have ids 0 and 4 respectively, so step 2 of 4.3.1 is skipped.
3) Since CE4's id is less than R0, and CE0's id is less than R4, step 3 of 4.3.1 is skipped.
4) Let the outer label to reach PE2 be 9999.
5) The DLCI which CE0 will use to talk to CE4 is D0[4], i.e., 104. The inner label for sending packets to CE4 is (L4 + 0), i.e.

4000. The inner label on which to expect packets from CE4 is (L0 + 4), i.e., 1004.

6) Install a "route" such that packets from CE0 with DLCI 104 will be sent with outer label 9999, inner label 4000. Also, install a route that packets received with label 1004 will be mapped to DLCI 104 and be sent to CE0.
7) Activate DLCI 104 to CE0.

Note that the inner label of 4000, computed by PE0, for sending packets from CE0 to CE4 is the same as what PE2 computed as the incoming label for receiving packets originated at CE0 and destined to CE4. Similarly, the inner label of 1004, computed by PE0, for receiving packets from CE4 to CE0 is same as what PE2 computed as the outgoing label for sending packets originated at CE4 and destined to CE0.

### 4.3.3. Generalizing the VPN Topology

In the above, we assumed for simplicity that the VPN was a full mesh. To allow for more general VPN topologies when using LDP for signalling, we introduce the notion of node colors, and the "spoke" attribute; together, these constitute a node's "connectivity". A node (CE) in a VPN can be colored with one or more colors. Furthermore, a node may be a hub or a spoke. Two nodes are connected iff they share a color in common, and they are not both spokes.

To incorporate connectivity into the processing of advertisements, add step 3' to the above:

3') If CE k and CE m are not connected, stop.

This notion of connectivity does not allow arbitrary topologies to be built; however, it is a compromise of generality and efficiency.

A more general mechanism based on BGP extended communities can also be used; naturally, this mechanism can only be used when signalling VPNs with BGP. See below for details.

## 5. Packet Transport

When a packet arrives at a PE from a CE in a Layer 2 VPN, the layer 2 address of the packet identifies to which other CE the packet is destined. The procedure outlined above installs a route that maps the layer 2 address to an outer label (which identifies the PE to which the destination CE is attached) and an inner label (which identifies the destination CE). If the destination PE is one hop away from the source PE, and Penultimate Hop Popping (PHP) is used, there is no outer label. If the destination PE is the same as the source PE, no labels are needed.

The packet may then be modified (depending on the layer 2 encapsulation) and then sent to the destination PE with the appropriate number of labels.

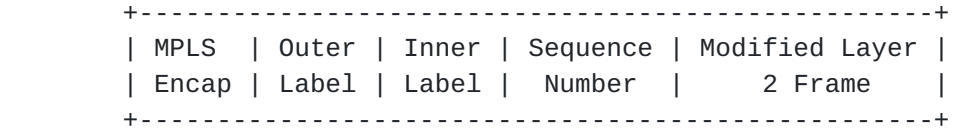
If the destination PE is the same as the source, the packet "arrives" with no labels. Otherwise, the packet arrives with one label (if PHP is used) or two labels, in which case the outer label is discarded; the remaining (inner) label is used to determine which CE is the destination CE. The packet is restored to a fully-formed layer 2 packet, and then sent to the CE.

The MTU on the Layer 2 access links MUST be chosen such that the size of the L2 frames plus the L2VPN header does not exceed the MTU of the MPLS network. Layer 2 frames that exceed the MPLS MTU after encapsulation MUST be dropped.

### 5.1. Layer 2 Frame Format

For each VPN encapsulation type (see section 5.1.3), we describe below the format of the frame as it is transported in the MPLS LSP.

Figure 2: Format of a Layer 2 Packet Carried in MPLS

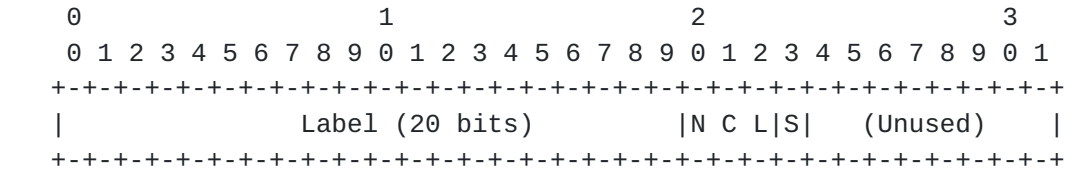



The "Outer Label" is used to transport the packet to the PE that is attached to the destination CE.

The "Inner Label" is used by the destination PE to distinguish which CE to send the packet to, and what layer 2 address to use (if

applicable). The Inner Label may also carry "non-address" information in its experimental bits. The label itself is 20 bits; the stack bit is as defined in [ENCAP]. The experimental bits are named N (Notification), C (Control) and L (Loss) as in the following figure. Note that the inner label is only used for forwarding from the destination PE to the destination CE, not within the MPLS network.

Figure 3: Format of the Inner Label




The "Sequence Number" is an optional two octet unsigned number that wraps back to zero that is used to ensure in-sequence delivery of L2 frames. The sequence number field is only included if its use is indicated via VPN signalling. A Layer 2 'connection' between two specific CEs is characterized within the MPLS network by the PEs to which the CEs are attached and a specific Inner Label in each direction. For each such Layer 2 connection, the sequence number field is set to zero for the first packet transmitted and incremented by 1 for each subsequent packet sent on the same Layer 2 connection. When an out-of-sequence packet arrives at the receiver, it MAY be buffered for future delivery or discarded.

The modification to the Layer 2 frame depends on the Layer 2 type. The following sections describe the modification for each protocol type, and other per-protocol information.

## 5.2. Frame Relay

A Frame Relay frame has the following format:

<DLCI><UI><proto><layer 3 packet>

For transport over an MPLS LSP, the <DLCI> octets are removed. The rest of the frame is transported as is.

At the destination PE, a new DLCI is added, and the fully-formed Frame Relay frame sent to the CE.

A DLCI contains "non-address" bits, namely, Forward and Backward Explicit Congestion Notification (FECN and BECN), the

Command/Response (C/R) bit and the Discard Eligible (DE) bit. The ingress LSR MAY set the experimental bits as follows: copy BECN to the N bit; copy the C/R bit to the C bit; and copy DE to the L bit. Otherwise, the ingress SHOULD set the experimental bits to 0. The egress LSR MAY in turn copy the N bit to BECN of the outgoing DLCI, the C bit to C/R and the L bit to DE.

Note that this is orthogonal to preferential treatment of the layer 2 frame in the MPLS network. If there are two LSPs (L-LSPs) to the destination PE, one for normal traffic and another for out-of-spec traffic, the ingress LSR MAY choose which LSP to use (i.e., which outer label) based on the DE bit. If there is one LSP (E-LSP), but an experimental bit is used to denote out-of-spec traffic, the ingress LSR MAY set this experimental bit based on the DE bit.

### 5.3. ATM AAL/5

For ATM AAL/5 VPNs, the AAL/5 PDU is transported without indication of the VPI/VCI. At the receiving PE, the AAL/5 PDU is fragmented, a cell header with the correct VPI/VCI added to each cell, and the cells sent to the CE.

If any of the cells that constitute the AAL/5 PDU have the CLP bit set, the ingress LSR MAY set the L bit. If the L bit is set in the inner label at the destination PE, this PE MAY set the CLP bit in each cell when fragmenting the AAL/5 PDU.

Again, the ingress PE may give preferential treatment to the ATM PDU based on whether any cell had the CLP bit set or all cells had their CLP bits clear.

### 5.4. ATM Cells

For ATM Cell VPNs, ATM cells (including the 5 octet header) are transported. At the receiving PE, the cells are sent to the CE.

The experimental bits of the inner label SHOULD be set to zero at the ingress and ignored by the destination PE.

### 5.5. PPP, Cisco HDLC, Ethernet

For PPP, Cisco HDLC and unswitched Ethernet VLANs VPNs, the Layer 2 frame is transported whole, without any modification. The Layer 2 frame does not include HLDC flags or Ethernet preamble, nor CRCs; we assume that bit/byte stuffing has been undone. At the receiving PE,

the frame is sent to the CE.

The experimental bits of the inner label SHOULD be set to zero at the ingress and ignored by the destination PE.

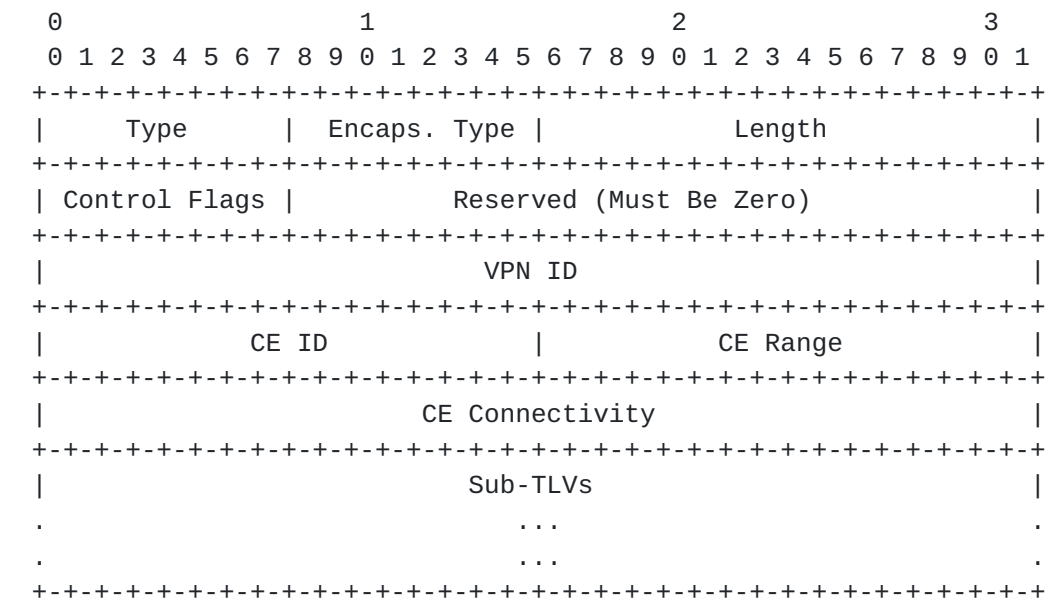
## 6. Signalling MPLS-Based Layer 2 VPNs

There are two alternative means of signalling the MPLS-based Layer 2 VPNs described in this document: using LDP ([LDP]) or using BGP version 4 ([BGP]).

In LDP, VPN CE information and its associated label base are carried in a Label Mapping message, distributed in the downstream unsolicited mode described in [LDP]. A new FEC element is defined below to carry all the information corresponding to a VPN CE, except from the label base. The label base is carried in the Label TLV following the FEC TLV. If a FEC element in a FEC TLV encodes Layer 2 VPN information, it MUST be the only FEC element in the FEC TLV.

The Layer 2 VPN FEC element is depicted in Figure 4 below.

Figure 4: L2 VPN FEC Element



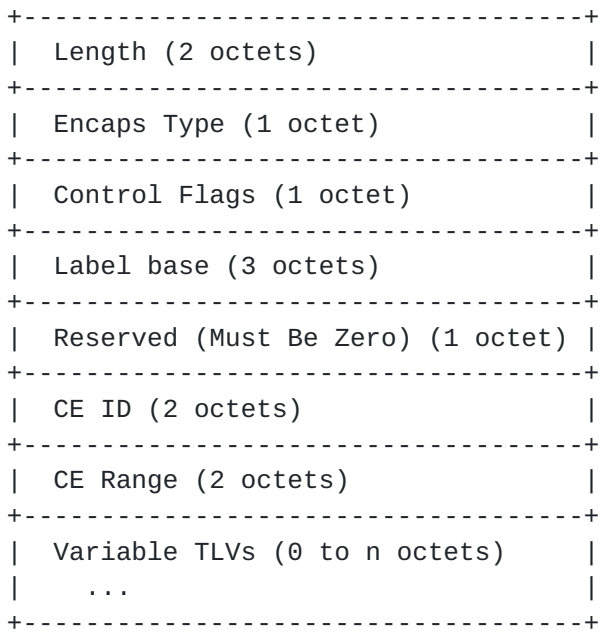

In BGP, the Multiprotocol Extensions [BGP-MP] are used to carry L2-VPN signalling information. [BGP-MP] defines the format of two BGP attributes (MP_REACH_NLRI and MP_UNREACH_NLRI) that can be used to announce and withdraw the announcement of reachability

information. We introduce a new address family identifier (AFI) for L2-VPN [to be assigned by IANA], a new subsequent address family identifier (SAFI) [to be assigned by IANA], and also a new NLRI format for carrying the individual L2-VPN CE information. This NLRI will be carried in the above-mentioned BGP attributes. This NLRI MUST be accompanied by one or more extended communities. The extended community type is "Layer 2 VPN" (to be assigned by IANA); and the format is <VPN-ID>:<connectivity>, where <VPN-ID> is 4 octets in length, and <connectivity> is two octets. All extended communities accompanying one or more Layer 2 VPN NLRIs MUST have the same <VPN-ID>.

PEs receiving VPN information may filter advertisements based on the extended communities, thus controlling CE-to-CE connectivity.

The format of the Layer 2 VPN NLRI is as shown in Figure 5 below.

Figure 5: BGP NLRI for L2 VPN Information




## 6.1. Signalled Information

### 6.1.1. Type (LDP only)

The Type is L2-VPN (to be decided by IETF Consensus Action).

### 6.1.2. Length

In LDP, the Length is the entire length of the L2 VPN FEC element, including the fixed header and all the sub-TLVs.

In BGP, the Length field indicates the length in octets of the L2-VPN address prefix.

### 6.1.3. Encapsulation Type

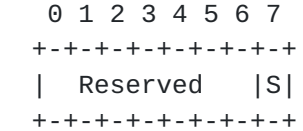
Identifies the layer 2 encapsulation, e.g., ATM, Frame Relay etc. The following encapsulation types are defined:

| Value | Encapsulation |
|---|---|
| 0 | Reserved |
| 1 | ATM PDUs (AAL/5) |
| 2 | ATM Cells |
| 3 | Frame Relay |
| 4 | PPP |
| 5 | Cisco-HDLC |
| 6 | Ethernet VLAN (unswitched) |
| 7 | MPLS |

### 6.1.4. Control Flags

This is a bit vector, defined as in the following Figure.

Figure 6: Control Flags Bit Vector




The following bit is defined; the rest MUST be set to zero.

| Name | Bit | Meaning |
|---|---|---|
| S | 0 | Sequenced delivery of frames is required |

#### 6.1.5. Label base (BGP only)

The label-base which is to be used for determining the inner label for forwarding packets to the CE identified by CE ID. (Note: LDP carries the label-base in the Label TLV following the FEC TLV.)

#### 6.1.6. VPN ID (LDP only)

A 32 bit number which uniquely identifies a VPN in a provider's domain.

#### 6.1.7. CE ID

A 16 bit number which uniquely identifies a CE in a VPN.

#### 6.1.8. CE Range

A 16 bit number which describes the range of CE IDs to which the advertised CE is willing to connect. In particular, a PE receiving an L2 VPN TLV MUST NOT use a label greater than or equal to

<label-base> + <CE range>

when sending traffic for this VPN to the advertising PE.

#### 6.1.9. CE Connectivity (LDP only)

A 32-bit number encoding connectivity. If the leftmost bit is 1, the CE is a spoke. The remaining 31 bits encode the CE colors (bit i = 1 means the CE has color i).

#### 6.1.10. Sub-TLVs

New sub-TLVs can be introduced as needed.

In LDP, the TLV encoding mechanism described in [LDP] must be used.

In BGP, TLVs (type takes 1 octet) can be added to extend the information carried in the L2 VPN address prefix.

A TLV (type = 1) will be used for carrying VLAN IDs if the encapsulation is VLAN.

### 6.2. BGP L2 VPN capability

The BGP Multiprotocol capability extension [BGP-CAP] is used to indicate that the BGP speaker wants to negotiate L2 VPN capability with its peers. The capability code is 1, the capability length is 4, and the AFI and SAFI values will be set to the L2 VPN AFI and L2 VPN SAFI (discussed in section 5) respectively.

### 6.3. Advantages of Using BGP

PE routers in an SP network typically run BGP v4. This means that SPs are familiar with using BGP, and have already configured BGP on their PEs, so configuring and using BGP to signal Layer 2 VPNs is not much of an additional burden to the SP operators. This is especially the case when the protocol of choice for signalling MPLS LSPs across the SP network is RSVP (perhaps for its Traffic Engineering properties); in this case, the SP may find using LDP to signal Layer 2 VPN information undesirable.

Another advantage of using BGP is that with BPG it is easier to build inter-provider VPNs. Mechanisms for this will be described in a future version.

## 7. Acknowledgments

The authors would like to thank Dennis Ferguson, Der-Hwa Gan, Dave Katz, Nischal Sheth, John Stewart, and Paul Traina for the enlightening discussions that helped shape the ideas presented here, and Ross Callon for his valuable comments.

The idea of using extended communities for more general connectivity of a Layer 2 VPN was a contribution by Yakov Rekhter, who also gave many useful comments on the text; many thanks to him.

## 8. Security Considerations

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

## 9. IANA Considerations

(To be filled in in a later revision.)

## 10. References

[BGP] Rekhter, Y., and Li, T., "A Border Gateway Protocol 4 (BGP-4)", RFC 1771, March 1995.

[BGP-CAP] Chandra, R., and Scudder, J., "Capabilities Advertisement with BGP-4", RFC 2842, May 2000.

[BGP-MP] Bates, T., Rekhter, Y., Chandra, R., and Katz, D., "Multiprotocol Extensions for BGP-4", RFC 2858, June 2000

[BGP-ORF] Chen, E., and Rekhter, Y., "Cooperative Route Filtering Capability for BGP-4", March 2000 (work in progress).

[BGP-RFSH] Chen, E., "Route Refresh Capability for BGP-4", draft-ietf-idr-bgp-route-refresh-01.txt, March 2000, (work in progress).

[ENCAP] Rosen, E., Rekhter, Y., Tappan, D., Fedorkow, G., Farinacci, D., Li, T., and Conta, A., "MPLS Label Stack Encoding", draft-ietf-mpls-label-encaps-08.txt (work in progress)

[IPVPN] Rosen, E., and Rekhter, Y., "BGP/MPLS VPNs", RFC 2547, March 1999.

[LDP] Andersson, L., Doolan, P., Feldman, N., Fredette, A., and Thomas, B., "LDP Specification", draft-ietf-mpls-ldp-11.txt, August 2000 (work in progress).

[MPLS] Callon, R., Doolan, P., Feldman, N., Fredette, A., Swallow, G., and Viswanathan, A., "A Framework for Multiprotocol Label Switching", draft-ietf-mpls-framework-05.txt, September 1999 (work in progress).

[VPN] Kosiur, Dave, "Building and Managing Virtual Private Networks", Wiley Computer Publishing, 1998.

## 11. Intellectual Property Considerations

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

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

## 12. Full Copyright Statement

## 13. Author Information

Kireeti Kompella
Juniper Networks
1194 N. Mathilda Ave
Sunnyvale, CA 94089
kireeti@juniper.net

Manoj Leelanivas
Juniper Networks
1194 N. Mathilda Ave
Sunnyvale, CA 94089
manoj@juniper.net

Quaizar Vohra
Juniper Networks
1194 N. Mathilda Ave
Sunnyvale, CA 94089
qv@juniper.net

Javier Achirica
Telefonica Data
javier.achirica@telefonica-data.com

Ronald P. Bonica
WorldCom
22001 Loudoun County Pkwy
Ashburn, Virginia, 20147
rbonica@mci.net

Chris Liljenstolpe
Cable & Wireless
11700 Plaza America Drive
Reston, VA 20190
chris@cw.net

Eduard Metz
KPN Royal Dutch Telecom
St. Paulusstraat 4
2264 XZ Leidschendam
The Netherlands
e.t.metz@kpn.com

Chandramouli Sargor
CoSine Communications
1200 Bridge Parkway
Redwood City, CA 94065

csargor@cosinecom.com

Vijay Srinivasan
CoSine Communications
1200 Bridge Parkway
Redwood City, CA 94065
vijay@cosinecom.com