Network Working Group
Internet Draft
Expiration Date: October 2003

draft-kompella-ppvpn-l2vpn-03.txt

K. Kompella (Juniper)
M. Leelanivas (Juniper)
Q. Vohra (Juniper)
J. Achirica (Telefonica)
R. Bonica (WorldCom)
D. Cooper (Global Crossing)
C. Liljenstolpe (C & W)
E. Metz (KPN Dutch Telecom)
H. Ould-Brahim (Nortel)
C. Sargor (CoSine)
H. Shah (Tenor)
V. Srinivasan (CoSine)
Z. Zhang (Unisphere)

Layer 2 VPNs Over Tunnels

Status of this Memo

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

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.

Copyright Notice

Copyright (C) The Internet Society (2003). All Rights Reserved.

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 network for IP, IP VPNs, and Layer 2 VPNs.

0.1. ID Summary

SUMMARY

This ID describes an approach to provisioning Layer 2 VPNs in a Service Provider network. From the VPN customers' point of view, the VPNs look like the traditional Layer 2 connections (Frame Relay, ATM, ...); the benefits here are to the SP, whose job provisioning and managing the connections within their network is simplified.

RELATED DOCUMENTS

draft-rosen-ppvpn-l2-signaling-02.txt

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

Belongs in PPVPN.

WHY IS IT TARGETED AT THIS WG

This document describes a mechanism for Provider-Provisioned Layer 2 VPNs.

JUSTIFICATION

"Traditional" Layer 2 VPNs are very common, widely deployed and incontrovertably useful. The techniques described here show how the work that a provider must do within its network to provision Layer 2 VPNs can be made much simpler and more automated.

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 [KEYWORDS].

1. 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 network for IP, 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.

To ease the restriction that all sites within a single VPN connect via the same layer 2 technology, this document proposes a limited form of layer 2 interworking, restricted to IP only as the layer 3 protocol.

The solution we propose scales well because the amount of forwarding state maintained in the core routers of the Service Provider Network is independent of the number of layer 2 VPNs provisioned over the SP network. This is achieved by using tunnels to carry the data, with a "demultiplexing field" that identifies individual VCs. These tunnels

could be MPLS, GRE, or any other tunnel technology that offers a demultiplexing field; the signaling of these tunnels is outside the scope of this document. The specific approach taken here is to use a 32-bit demultiplexing field formatted as an MPLS label; other sizes and formats are clearly possible, and will be defined as needed.

This approach combines auto-discovery of VPN sites with the signalling of the demultiplexing fields for L2VPN PVCs. This is possible because the mechanism used for auto-discovery (BGP) is also capable of distributing Layer 2 information as well as the demultiplexing field.

The rest of this section discusses the relative merits of Layer 2 and Layer 3 VPNs. Section 4 describes the operation of a Layer 2 VPN. Section 5 describes IP-only layer 2 interworking. Section 6 describes how the L2 packets are transported across the SP network. Section 7 discusses BGP as a mechanism for auto-discovery and signalling of Layer 2 VPNs.

1.1. Terminology

We assume that the reader is familiar with Multi-Protocol Label Switching (MPLS [MPLS]) 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].

We also introduce three new terms:

VPN Label - the demultiplexing field which identifies an L2VPN PVC to the edge of the SP network, i.e., the PE.

Tunnel - a PE-to-PE tunnel that is used to carry multiple types of data. P routers in the SP core forward this data based on the tunnel header and not on the data within, thus limiting the Layer 2 state to the PE routers who host the Layer 2 circuit.

CE ID - a number that uniquely identifies a CE within an L2 VPN. More accurately, the CE ID identifies a physical connection from the

CE device to the PE. Say a CE connected to a PE over a DS-3 for Frame Relay access to a VPN; this DS-3 would need a CE ID. The CE would also have N DLCIs over this DS-3 to speak to N other sites in the VPN.

A CE may be connected to multiple PEs (or multiply connected to a PE), in which case it would have a CE ID for each connection. If these connections are in the same VPN, the CE IDs would have to be different. A CE may also be part of many L2 VPNs; it would need one (or more) CE ID(s) for each L2 VPN of which it is a member.

For the case of inter-Provider L2 VPNs, there needs to be some coordination of allocation of CE IDs. One solution is to allocate ranges for each SP. Other solutions may be forthcoming.

1.2. 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 tunnels, which could be MPLS Label-Switched Paths (LSPs) or GRE tunnels, as examples.

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.

1.2.1. 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.

1.2.2. Migrating from Traditional Layer 2 VPNs

Since "traditional" Layer 2 VPNs (i.e., real Frame Relay circuits connecting sites) are indistinguishable from tunnel-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).

1.2.3. 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.

1.2.4. 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.

Note that IP-only layer 2 interworking doesn't have this benefit as it restricts the layer 3 to IP only.

1.2.5. 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 routes 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.
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.

1.2.6. 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, PVCs are tunnelled across the SP network. The tunnels used are provisioned independent of the L2VPNs, using signalling protocols (in case of MPLS, LDP or RSVP-TE can be used), or set up by configuration; and the number of tunnels is independent of the number of L2VPNs. This reduces a large part of the provisioning burden.

Furthermore, we assume that DLCIs at the CE edge are relatively cheap; and VPN 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. Note that if DLCIs at the CE edge are expensive, e.g. if these DLCIs are provisioned across a switched network, one could provision them as and when needed, at the expense of extra configuration. This need not still result in extra state in the SP network, i.e. an intelligent

implementation can allow overprovisioning of the pool of VPN labels.

1.3. 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.

1.3.1. 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.

This restriction of layer 2 VPN is alleviated by the IP-only layer 2 interworking proposed in this document. This comes at the cost of losing the layer 3 independence.

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.

1.3.2. 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.

As mentioned later, multicast routing and forwarding is another

value-added service that an SP can offer.

1.3.3. 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. This restriction on layer 2 VPNs is again eased in the case of IP-only layer 2 interworking, as the PE router has visibility into the network layer (IP).

1.4. Multicast Routing

There are two aspects to multicast routing that we will consider. On the protocol front, supporting IP multicast in a Layer 3 VPN requires PE routers to participate in the multicast routing instance of the customer, and thus keep some related state information.

In the Layer 2 VPN case, the CE routers run native multicast routing directly. The SP network just provides pipes to connect the CE routers; PEs are unaware whether the CEs run multicast or not, and thus do not have to participate in multicast protocols or keep multicast state information.

On the forwarding front, in a Layer 3 VPN, CE routers do not replicate multicast packets; thus, the CE-PE link carries only one copy of a multicast packet. Whether replication occurs at the ingress PE, or somewhere within the SP network depends on the sophistication of the Layer 3 VPN multicast solution. The simple solution where a PE replicates packets for each of its CEs may place considerable burden on the PE. More complex solutions may require VPN multicast state in the SP network, but may significantly reduce the traffic in the SP network by delaying packet replication until needed.

In a Layer 2 VPN, packet replication occurs at the CE. This has the advantage of distributing the burden of replication among the CEs rather than focusing it on the PE to which they are attached, and thus will scale better. However, the CE-PE link will need to carry the multiple copies of multicast packets.

Thus, just as in the case of unicast routing, the SP has the choice to offer a value-added service (multicast routing and forwarding) at some cost (multicast state and packet replication) using a Layer 3 VPN, or to keep it simple and use a Layer 2 VPN.

2. 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. In the case of IP-only layer 2 interworking, each CE could have a mix of one or more of the above layer 2 mediums to connect to other CEs.

2.1. 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. (Note that the signalling mechanism discussed in section 7 will allow a site to grow in terms of connectivity to other sites at a later point of time at the cost of additional signalling, i.e., over-provisioning is not a must but a recommendation).

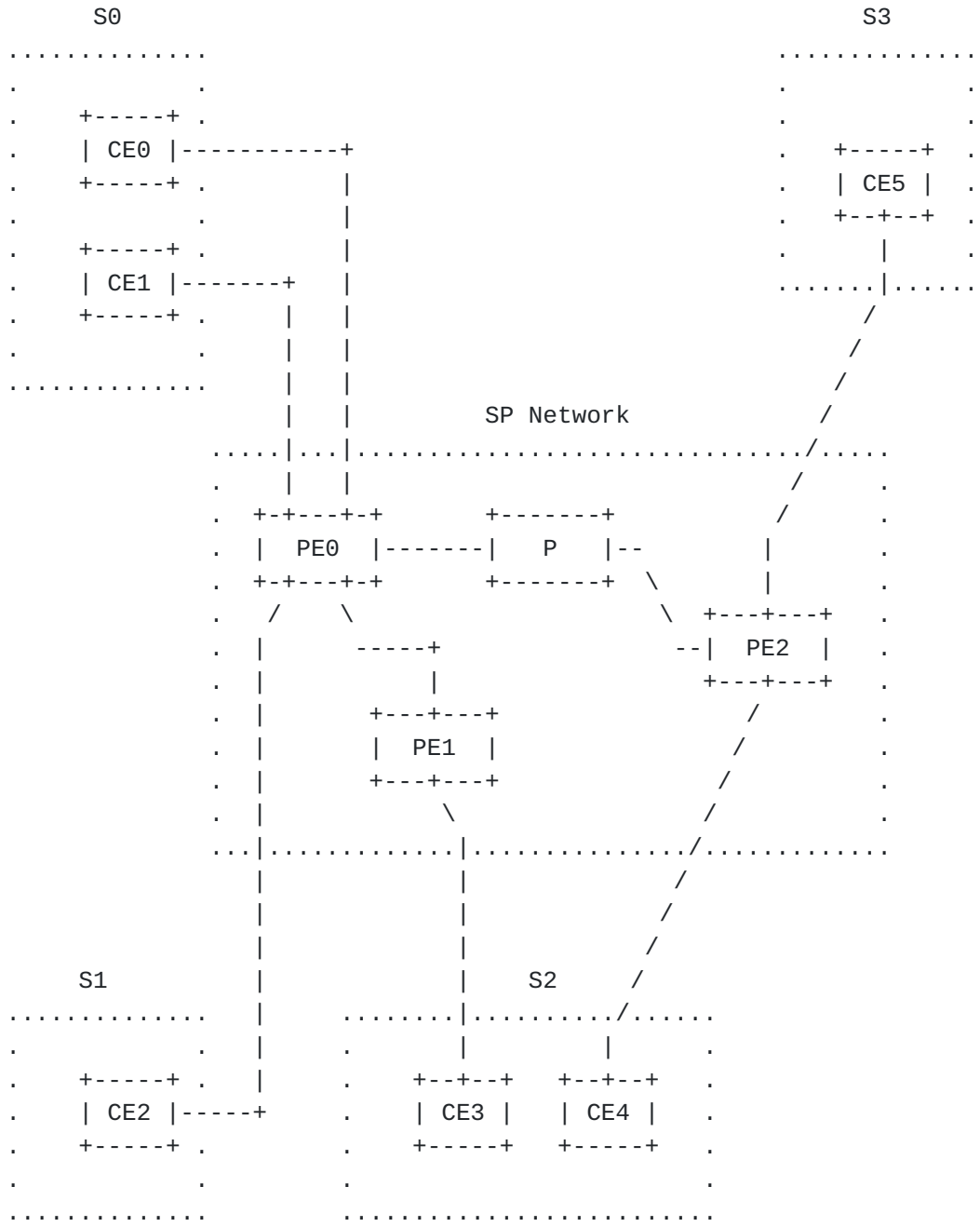
Suppose finally that CE0 and CE2 have DLCIs 100 through 109 free; CE1 and CE3 have DLCIs 200 through 209 free; CE4 has DLCIs 107, 209, 265, 301, 414, 555, 654, 777 and 888 free; and CE5 has DLCIs 417-426.

2.2. 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,

Figure 1: Example Network Topology



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. The PPP and cisco hdlc keepalives could be between local and remote CE if both CEs connect via the same layer 2 medium.

In case of IP-only layer 2 interworking, if CE1, attached to PE0, connects to CE3, attached to PE1, via an L2VPN circuit, the layer 2 medium between CE1 and PE0 is independent of the layer 2 medium between CE3 and PE1. Each side will run its own layer 2 specific link management protocol, e.g., LMI, LCP, etc. PE0 will inform PE1 about the status of its local circuit to CE1 via the circuit status vector TLV defined in section 7. Similarly PE1 will inform PE0 about the status of its local circuit to CE3.

2.2.1. 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 configured to communicate with its corresponding PE with the set of DLCIs given above; for example, CE0 is configured with DLCIs 100 through 109. OSPF is configured to run over each DLCI. In general, a CE is configured with a list of circuits, all with the same layer 2 encapsulation type, e.g., DLCIs, VCIs, physical PPP interface etc. (IP-only layer 2 interworking allows a mix of layer 2 encapsulation types). The size of this list/set determines the number of remote CEs a given CE can communicate with. Denote the size of this list/set as the CE's range.

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.

2.2.2. PE Configuration

Each PE is configured with the VPNs in which it participates. Each VPN is configured with a Route Target community [IPVPN] which uniquely identifies the VPN within the SP network. For each VPN, the PE has a list of CEs, which are members of that VPN. For each CE, the PE knows the CE ID, its range and which DLCIs to expect from the CE.

2.2.3. 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 by allocating additional local circuits 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 and the algorithm for mapping a CE-id to a DLCI which is used for connecting to the corresponding CE. 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.

2.3. 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 n labels, where n is the CE's initial range. Denote a contiguous set of labels by a label-block. Call the smallest label in this label-block the label-base and the number of labels in the label-block as label-range.

To allow a CE to grow its connectivity at a later point of time additional DLCIs might be added between the CE and its PE. To advertise the additional capacity of a CE without disrupting existing connectivity to the site, a new label-block is picked with k labels, where k is the the number of additional circuits. This process might be repeated several times as and when a CE's range needs growing.

The PE then advertises for this CE all its label-blocks. Each label-block is propagated in a separate BGP NLRI (see figure 3). 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.

So as to be able to distinguish between the multiple label-blocks of a given CE, notion of a block offset is introduced. The block offset identifies the position of a given label-block in the set of label blocks of a given CE. A remote site with CE ID m will connect to this CE using a label selected from one of the label blocks such that the following condition holds true for that label-block :

block offset <= m < block offset + label-range

If the PE-CE physical link goes down, or the CE configuration is removed, all its advertised label-blocks are withdrawn.

Note that an implementation can easily allow allocation of a label-block which is larger than the actual number of DLCIs provisioned. This allows DLCIs to be provisioned as and when needed without increasing the state in the network, at the cost of extra signalling and configuration.

2.3.1. PE Advertisement Processing

When a PE receives a Layer 2 VPN advertisement, it checks if the received Route Target community matches any VPN that it is a member of. If not, the PE may store the advertisement for future use, or may discard it. Since we use BGP as the auto-discovery and signalling protocol, a PE can use the BGP Route Refresh capability to learn all the discarded advertisements pertaining to a VPN at a later time, when the VPN is configured on the PE.

Otherwise, suppose the advertisement is from PE A for VPN X, CE m, and a label-block Lm. Add this label-block to the existing label-blocks for CE m in VPN X. For the purpose of further discussion we denote a label-block from CE m as Lm. Denote Lm's block offset as LOm, label-base as LBm, and label-range as LRm.

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. (Note that for IP-only layer 2 interworking a separate

encapsulation type is defined).

1) Say the configured CE ID is k, and the DLCI list is Dk[]. A label-block of k is denoted by Lk. Denote Lk's block offset as LOk, label-base as LBk, and label-range as LRk.
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) Search among all the label-blocks from m for one which satisfies LOm <= k < LOm + LRm. If none found, issue a warning : "Cannot communicate with CE m (PE A) of VPN X: outside range" and stop. Otherwise let Lm be the label-block found.
4) Search among all the label-blocks of k for one which satisfies LOk <= m < LOk + LRk. If none found, issue a warning : "Cannot communicate with CE m (PE A) of VPN X: outside range" and stop. Otherwise let Lk be the label-block found.
5) Look in the appropriate table to see which label-stack will get to PE A. This is the "tunnel" label-stack, Z.
6) The DLCI that CE-k will use to talk to CE-m is Dk[m]. Then "VPN" label for sending packets to CE-m is (LBm + k - LOm) if The "VPN" label on which to expect packets from CE-m is (LBk + m - LOk).
7) Install a "route" such that packets from CE-k with DLCI Dk[m] will be sent with tunnel label-stack Z, VPN label (LBm + k - LOm). Also, install a route such that packets received with label (LBk + m - LOk) will be mapped to DLCI Dk[m] and be sent to CE k.
8) Activate DLCI Dk[m] to the CE. This can be done using LMI.

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

2.3.2. Example of PE Advertisement Processing

Consider the example network of Figure 1. Let S0, S1, S2 and S3 belong to the same VPN, say VPN1. Suppose PE2 receives an advertisement from PE0 for VPN1, CE ID 0 and a label block L0 with block offset LO0 = 0, label-range LR0 = 10 and label base LB0 = 1000. Since PE2 is connected to CE4 which is also in VPN1, 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 is configured with DLCI list D4[] is [107, 209, 265, 301, 414, 555, 654, 777, 888]. A label-block L4 is allocated to CE4 with block offset LO4 = 0, label-range LR4 = 9 and

a label-base LB4 = 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 falls in the label-block L0 from CE0, i.e. LO0 <= 4 < LO0 + LR0, L0 is the label-block selected in step 3 of 4.3.1
4) Since CE0's id falls in the label-block L4 of CE4, i.e. LO4 <= 0 < LO4 + LR4, L4 is the label-block selected in step 4 of 4.3.1
5) Look in the appropriate table on PE2 to see which tunnel label-stack will get to PE0. Let the label-stack be a single label, 10001.
6) The DLCI that CE4 will use to talk to CE0 is D4[0], i.e., 107. The VPN label for sending packets to CE0 is (LB0 + 4 - LO0), i.e 1004. The VPN label on which to expect packets from CE0 is (LB4 + 0 - LO4), i.e., 4000.
7) Install a "route" such that packets from CE4 with DLCI 107 will be sent with top label 10001, VPN label 1004. Also, install a route such that packets received with label 4000 will be mapped to DLCI 107 and be sent to CE4.
8) 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 advertisement from PE2 for VPN1, CE4, with and a label block L4 with block offset LO4 = 0, label-range LR4 = 9 and label base LB4 = 4000. PE0 processes the advertisement for CE0 (and CE1, which is also in VPN1).

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 is configured with a DLCI list D0[] is [100 - 109], Label-block L0 is allocated to CE0 with block offset LO0 = 0, label-range LR0 = 10 and a label-base LB0 = 1000 (which was advertised to PE2)
2) CE0 and CE4 have ids 0 and 4 respectively, so step 2 of 4.3.1 is skipped.
3) Since CE0's id falls in the label-block L4 of CE4, i.e. LO4 <= 0 < LO4 + LR4, L4 is the label-block selected in step 4 of 4.3.1
4) Since CE4's id falls in the label-block L0 from CE0, i.e. LO0 <= 4 < LO0 + LR0, L0 is the label-block selected in step 3 of 4.3.1
5) Let the tunnel label-stack to reach PE2 be a single label, 9999.
6) The DLCI which CE0 will use to talk to CE4 is D0[4], i.e., 104.

The VPN label for sending packets to CE4 is (LB4 + 0 - LO4), i.e., 4000. The VPN label on which to expect packets from CE4 is (LB0 + 4 - LO4), i.e., 1004.

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

Note that the VPN 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 VPN 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.

2.3.3. 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, a mechanism based on filtering on BGP extended communities can be used (see section 7).

3. Layer 2 Interworking

As defined so far in this document, all CE-PE connections for a given Layer 2 VPN must use the same layer 2 encapsulation, e.g., they must all be Frame Relay. This is often a burdensome restriction. One answer is to use an existing Layer 2 interworking mechanism, for example, Frame Relay-ATM interworking.

In this document, we take a different approach: we postulate that the network layer is IP, and base Layer 2 interworking on that. Thus, one can choose between pure Layer 2 VPNs, with a stringent Layer 2 restriction but with Layer 3 independence, or a Layer 2 interworking VPNs, where there is no restriction on Layer 2, but Layer 3 must be IP. Of course, a PE may choose to implement Frame Relay-ATM interworking. For example, an ATM Layer 2 VPN could have some CEs connect via Frame Relay links, if their PE could translate Frame Relay to ATM transparent to the rest of the VPN. This would be private to the CE-PE connection, and such a course is outside the scope of this document.

For Layer 2 interworking as defined here, when an IP packet arrives at a PE, its Layer 2 address is noted, then all Layer 2 overhead is stripped, leaving just the IP packet. Then, a VPN label is added, and the packet is encapsulated in the PE-PE tunnel (as required by the tunnel technology). Finally, the packet is forwarded. Note that

the forwarding decision is made on the basis of the Layer 2 information, not the IP header. At the egress, the VPN label determines to which CE the packet must be sent, and over which virtual circuit; from this, the egress PE can also determine the Layer 2 encapsulation to place on the packet once the VPN label is stripped.

An added benefit of restricting interworking to IP only as the layer 3 technology is that the provider's network can provide IP Diffserv or any other IP based QOS mechanism to the L2VPN customer. The ingress PE can set up IP/TCP/UDP based classifiers to do DiffServ marking, and other functions like policing and shaping on the L2 circuits of the VPN customer. Note the division of labor: the CE determines the destination CE, and encodes that in the Layer 2 address. The ingress PE thus determines the egress PE and VPN label based on the Layer 2 address supplied by the CE, but the ingress PE can choose the tunnel to reach the egress PE (in the case that there are different tunnels for each CoS/DiffServ code point), or the CoS bits to place in the tunnel (in the case where a single tunnel carries multiple CoS/DiffServ code points) based on its own classification of the packet.

4. 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 a tunnel (which identifies the PE to which the destination CE is attached) and a VPN label (which identifies the destination CE). If the egress PE is the same as the ingress PE, no tunnel or VPN label is needed.

The packet may then be modified (depending on the layer 2 encapsulation). In case of IP-only layer 2 interworking, the layer 2 header is completely stripped off till the IP header. Then, a VPN label and tunnel encapsulation are added as specified by the route described above, and the packet is sent to the egress PE.

If the egress PE is the same as the ingress, the packet "arrives" with no labels. Otherwise, the packet arrives with the VPN label, which 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.

4.1. Layer 2 MTU

This document requires that the Layer 2 MTU configured on all the access circuits connecting CEs to PEs in an L2VPN be the same. This can be ensured by passing the configured layer 2 MTU in the Layer2-info extended community when advertising L2VPN label-blocks. On receiving L2VPN label-block from remote PEs in a VPN, the MTU value carried in the layer2-info extendend community should be compared against the configured value for the VPN. If they don't match, then the label-block should be ignored.

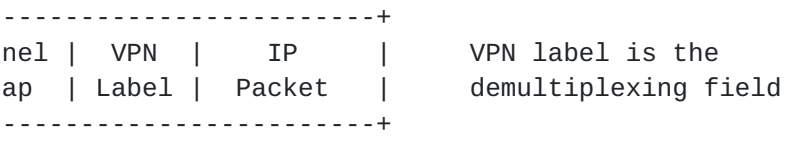
The MTU on the Layer 2 access links MUST be chosen such that the size of the L2 frames plus the L2VPN header does not exceed the MTU of the SP network. Layer 2 frames that exceed the MTU after encapsulation MUST be dropped. For the case of IP-only layer 2 interworking the IP MTU on the layer 2 access link must be chosen such that the size of the IP packet and the L2VPN header does not exceed the MTU of the SP network.

4.2. Layer 2 Frame Format

The modification to the Layer 2 frame depends on the Layer 2 type. This document requires that the encapsulation methods used in transporting of layer 2 frames over tunnels be the same as described in [L2-ENCAP], except in the case of IP-only Layer 2 Interworking which is described in section 6.2.

4.3. IP-only Layer 2 Interworking

Figure 2: Format of IP-only layer 2 interworking packet



At the ingress PE, an L2 frame's L2 header is completely stripped off and is carried over as an IP packet within the SP network (Figure 2). The forwarding decision is still based on the L2 address of the incoming L2 frame. At the egress PE, the IP packet is encapsulated back in an L2 frame and transported over to the destination CE. The forwarding decision at the egress PE is based on the VPN label as before. The L2 technology between egress PE and CE is independent of the L2 technology between ingress PE and CE.

5. Auto-discovery and Signalling of Layer 2 VPNs

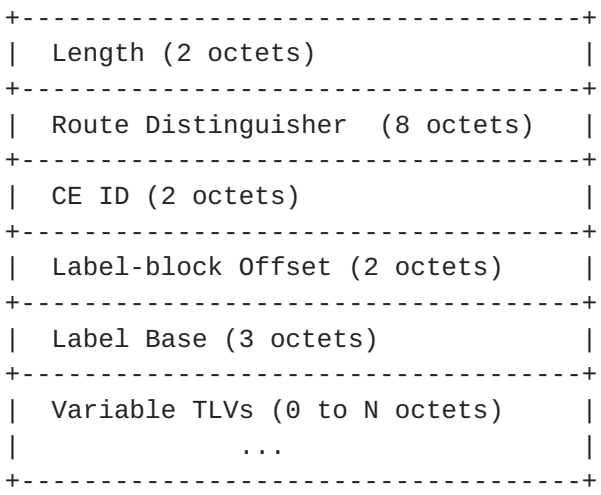
BGP version 4 ([BGP]) is used as the auto-discovery and signalling protocol for Layer 2 VPNs described in this document.

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 label-block information. One or more NLRIs will be carried in the above-mentioned BGP attributes. L2VPN NLRIs MUST be accompanied by one or more extended communities. This document proposes the reuse of ROUTE TARGET extended community defined in [EXT-COMM]. Its usage is exactly the same as in the case of [INETVPN].

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 3 below. One or more such NLRIs can be carried in a single MP_REACH_NLRI or MP_REACH_NLRI attribute. An L2VPN NLRI is uniquely identified by the RD, CE ID and the Label-block Offset. So an L2VPN NLRI carried in MP_UNREACH_NLRI attribute must contain only these 3 fields other than the length field.

Figure 3: BGP NLRI for L2 VPN Information



5.1. L2VPN NLRI Format

5.1.1. Length

The Length field indicates the length in octets of the L2-VPN address information.

5.1.2. Route Distinguisher

Has the same meaning as in [IPVPN].

5.1.3. CE ID

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

5.1.4. Label-Block Offset

A 16 bit number which identifies the position of a label-block within a set of label-blocks of a given CE. This enables a remote CE to select a label block when picking the VPN label for sending traffic destined to the CE this label-block corresponds to, such that :

label-block offset <= remote CE id.

5.1.5. Label base

The label-base which is to be used for determining the VPN label for forwarding packets to the CE identified by CE ID

5.1.6. Sub-TLVs

New sub-TLVs can be introduced as needed.

L2VPN TLVs can be added to extend the information carried in the L2 VPN NLRI. In L2VPN TLVs, type is 1 octet, length is 2 octets and represents the size of the value field in bits.

5.1.7. Circuit Status Vector

A new sub-TLV is introduced to carry the status of an L2VPN PVC between a pair of PEs. This sub-TLV is a mandatory part of MP_REACH_NLRI.

Note that an L2VPN PVC is bidirectional, composed of two simplex connection going in opposite directions. A simplex connection consists of the 3 segments: 1) the local access circuit between the source CE and the ingress PE, 2) the tunnel LSP between the ingress and egress PEs, and 3) the access circuit between the egress PE and

the destination CE.

To monitor the status of a PVC, a PE needs to monitor the status of both simplex connections. Since it knows that status of its access circuit, and the status of the tunnel towards the remote PE, it can inform the remote PE of these two. Similarly, the remote PE can inform the status of its access circuit to its local CE and the status of the tunnel to the first PE. Combining the local and the remote information, a PE can determine the status of a PVC.

The basic unit of advertisement in L2VPN for a given CE is a label-block. Each label within a label-block corresponds to a PVC on the CE. So its natural to advertise the local status information for all PVCs corresponding to a label-block along with the label-block's NLRI. This is done by introducing the circuit status vector TLV. The value field of this TLV is a bit-vector, each bit of which indicates the status of the PVC associated with the corresponding label in the label-block. Bit value 0 indicates that the local circuit and the tunnel LSP to the remote PE is up, while a value of 1 indicates that either or both of them are down.

PE A, while selecting a label from a label-block (advertised by PE B, for remote CE m, and VPN X) for one of its local CE n (in VPN X) can also determine the status of the corresponding PVC (between CE n and CE m) by looking at the appropriate bit in the circuit status vector.

Type field for the circuit status vector TLV is TBD.

The length field of the TLV specifies the length of the value field in bits. The value field is padded to the nearest octet boundary.

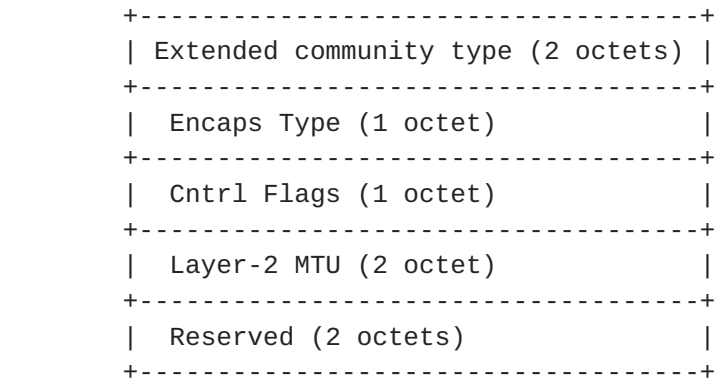
Note that the length field corresponds to the number of labels in the label-block, i.e., the label-block range. Label-block range enables a CE to select a label block (among several label-blocks advertised by a CE) when picking the VPN label for sending traffic destined to the CE this label-block corresponds to, such that :

received label-block offset <= local CE id < received label-block range.

5.2. Layer2-Info Extended Community

This document introduces a new extended community, Layer2-Info, to allow carrying layer 2 specific information in a VPN. This extended community MUST be carried as part of path attribute in all BGP update messages carrying L2VPN NLRIs. The encoding of this community is shown in figure 4.

Figure 4: layer2-info extended community



5.2.1. Extended Community Type

TBD.

5.2.2. Encapsulation Type

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

Value	Encapsulation
0	Reserved
1	Frame Relay
2	ATM AAL5 VCC transport
3	ATM transparent cell transport
4	Ethernet VLAN
5	Ethernet
6	Cisco-HDLC
7	PPP
8	CEM [8]
9	ATM VCC cell transport
10	ATM VPC cell transport
11	MPLS
12	VPLS
64	IP-interworking

5.2.3. Control Flags

This is a bit vector, defined as in Figure 5.

The following bits are defined; the MBZ bits MUST be set to zero.

Name	Meaning

Figure 5: Control Flags Bit Vector

```
 0 1 2 3 4 5 6 7
+-+-+-+-+-+-+-+-+
|  MBZ  |Q|F|C|S|      (MBZ = MUST Be Zero)
+-+-+-+-+-+-+-+-+
```

C If set to 1(0), Control word is (not) required when encapsulating Layer 2 frames [L2-ENCAP].

S If set to 1(0), Sequenced delivery of frames is (not) required.

The Q and F flags are reserved for other use.

5.2.4. Layer-2 MTU

Specifies the layer-2 specific MTU of all the circuits in all the label-blocks advertised with this extended community. This allows for checking of the layer 2 MTU being same for all the circuits across all the sites in a VPN.

5.3. 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 seccion 7) respectively.

5.4. 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.

Another advantage of using BGP is that with BGP it is easier to build inter-provider VPNs. Mechanisms for this are similar as that described in [IPVPN]. Option a) and b) described there could be adapted with slight modification for the l2vpn case but have adverse scaling issue in the l2vpn context. So we recommend using option C) which in l2vpn context would require an ASBR to maintain labeled IPv4 /32 routes to PEs within its AS and use EBGP to distribute these routes to other ASes. This results in creation of an LSP from a PE in one AS to another PE in another AS. Now these PEs can run multihop EBGP to exchange L2VPN information. The L2VPN traffic will be tunnelled thru the inter-AS LSP established between PEs as described

above.

6. Acknowledgments

The authors would like to thank Chaitanya Kodeboyina 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.

7. Security Considerations

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

8. IANA Considerations

(To be filled in in a later revision.)

9. Normative 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", RFC2918, September 2000.

[EXT-COMM] Ramachandra, S., Tappan, D.,Rekhtar, Y., "BGP Extended Communities Attribute" (work in progress).

[L2-ENCAP] Martini, et. al., "Encapsulation Methods for Transport of Layer 2 Frames Over MPLS", November 2001 (work in progress).

10. Informative References

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

[IPVPN-MCAST] Rosen, et. al., "Multicast in MPLS/BGP VPNs", November 2000 (work in progress).

[MPLS] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.

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

Authors' Addresses

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

Dave Cooper
Global Crossing

960 Hamlin Court
Sunnyvale, CA 94089
email: dcooper@gblx.net

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

Eduard Metz
KPNQwest
Scorpius 60
2130 GE Hoofddorp, The Netherlands
email: eduard.metz@kpnqwest.com

Chandramouli Sargor
CoSine Communications
1200 Bridge Parkway
Redwood City, CA 94065
csargor@cosinecom.com

Himanshu Shah
Tenor Networks
100 Nanog Park
Acton, MA 01720
hshah@tenornetworks.com

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

Hamid Ould-Brahim
Nortel Networks
P O Box 3511 Station C
Ottawa ON K1Y 4H7 Canada
Phone: +1 (613) 765 3418
Email: hbrahim@nortelnetworks.com

Zhaohui Zhang
Juniper Networks
10 Technology Park Drive
Westford, MA 01886
zzhang@unispherenetworks.com

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.

Full Copyright Statement

Copyright (C) The Internet Society (2003). 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 implementation 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 languages other than English.

The limited permissions granted above are perpetual and will not be revoked by the Internet Society or its successors or assigns.

This document and the information contained herein is provided on an "AS IS" basis and THE INTERNET SOCIETY AND THE INTERNET ENGINEERING TASK FORCE DISCLAIMS ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.