Network Working Group
Internet-Draft
Intended status: Informational
Expires: August 30, 2012

K. Kompella
Juniper Networks
B. Kothari
Cisco Systems
R. Cherukuri
Juniper Networks

February 27, 2012

Layer 2 Virtual Private Networks Using BGP for Auto-discovery and Signaling
draft-kompella-l2vpn-l2vpn-10.txt

Abstract

Layer 2 Virtual Private Networks (L2VPNs) based on Frame Relay or ATM circuits have been around a long time; more recently, Ethernet VPNs, including Virtual Private LAN Service, have become popular. Traditional L2VPNs often required a separate Service Provider infrastructure for each type, and yet another for the Internet and IP VPNs. In addition, L2VPN provisioning was cumbersome. This document presents a new approach to the problem of offering L2VPN services where the L2VPN customer's experience is virtually identical to that offered by traditional Layer 2 VPNs, but such that a Service Provider can maintain a single network for L2VPNs, IP VPNs and the Internet, as well as a common provisioning methodology for all services.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

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."

This Internet-Draft will expire on August 30, 2012.

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction 4
1.1. Terminology 6
1.2. Advantages of Layer 2 VPNs 6
1.2.1. Separation of Administrative Responsibilities 6
1.2.2. Migrating from Traditional Layer 2 VPNs 7
1.2.3. Privacy of Routing 7
1.2.4. Layer 3 Independence 7
1.2.5. PE Scaling 8
1.2.6. Ease of Configuration 8
1.3. Advantages of Layer 3 VPNs 9
1.3.1. Layer 2 Independence 9
1.3.2. SP Routing as Added Value 10
1.3.3. Class-of-Service 10
1.4. Multicast Routing 10
1.5. Conventions used in this document 11
2. Operation of a Layer 2 VPN 12
2.1. Network Topology 12
2.2. Configuration 13
2.2.1. CE Configuration 14
2.2.2. PE Configuration 15
2.2.3. Adding a New Site 15
2.2.4. Deleting a Site 15
2.2.5. Managing CE ID Mappings 16
2.2.6. Managing Label Blocks 16
2.3. Operations, Administration and Maintenance (OAM) 17
3. PE Information Exchange 18
3.1. Circuit Status Vector 20
3.2. Generalizing the VPN Topology 21
4. Layer 2 Interworking 22
5. Packet Transport 23
5.1. Layer 2 MTU 23
5.2. Layer 2 Frame Format 23
5.3. IP-only Layer 2 Interworking 24
6. Security Considerations 25
7. Acknowledgments 26
8. Contributors 27
9. IANA Considerations 28
10. References 29
10.1. Normative References 29
10.2. Informative References 29
Authors' Addresses 31

1. Introduction

The earliest Virtual Private Networks (VPNs) were based on Layer 2 circuits: X.25, Frame Relay and ATM (see [Kosiur]). More recently, multipoint VPNs based on Ethernet Virtual Local Area Networks (VLANs) and Virtual Private LAN Service (VPLS) ([RFC4761] and [RFC4762]) have become quite popular. In contrast, the VPNs described in this document are point-to-point, and usually called Virtual Private Wire Service (VPWS). All of these come under the classification of Layer 2 VPNs (L2VPNs), as the customer to Service Provider (SP) hand-off is at Layer 2.

There are at least two factors that adversely affected the cost of offering L2VPNs. The first is that the easiest way to offer a L2VPN of a given type of Layer 2 was over an infrastructure of the same type. This approach required that the Service Provider build a separate infrastructure for each Layer 2 encapsulation -- e.g., an ATM infrastructure for ATM VPNs, an Ethernet infrastructure for Ethernet VPNs, etc. In addition, a separate infrastructure was needed for the Internet and IP VPNs ([RFC4364], and possibly yet another for voice services. Going down this path meant a proliferation of networks.

The other is that each of these networks had different provisioning methodologies. Furthermore, the provisioning of a L2VPN was fairly complex. It is important to distinguish between a single Layer 2 circuit, which connects two customer sites, and a Layer 2 VPN, which is a set of circuits that connect sites belonging to the same customer. The fact that two different circuits belonged to the same VPN was typically known only to the provisioning system, not to the switches offering the service; this complicated the setting up, and subsequently, the troubleshooting, of a L2VPN. Also, each switch offering the service had to be provisioned with the address of every other switch in the same VPN, requiring, in the case of full-mesh VPN connectivity, provisioning proportional to the square of the number of sites. This made full-mesh L2VPN connectivity prohibitively expensive for the SP, and thus in turn for customers. Finally, even setting up a individual circuit often required the provisioning of every switch along the path.

Of late, there has been much progress in network "convergence", whereby Layer 2 traffic, Internet traffic and IP VPN traffic can be carried over a single, consolidated network infrastructure based on IP/MPLS tunnels; this is made possible by techniques such as those described in [RFC4448], [RFC4618], [RFC4619], and [RFC4717] for Layer 2 traffic, and [RFC4364] for IP VPN traffic. This development goes a long way toward addressing the problem of network profileration. This document goes one step further and shows how a Service Provider

can offer Layer 2 VPNs using protocol and provisioning methodologies similar to that used for VPLS ([RFC4761]) and IP VPNs ([RFC4364]), thereby achieving a significant degree of operational convergence as well. In particular, all of these methodologies include the notion of a VPN identifier that serves to unify components of a given VPN, and the concept of auto-discovery, which simplifies the provisioning of dense VPN topologies (for example, a full mesh). In addition, similar techniques are used in all of the above-mentioned VPN technologies to offer inter-AS and inter-provider VPNs (i.e., VPNs whose sites are connected to multiple Autonomous Systems (ASs) or service providers).

Technically, the approach proposed here uses the concepts and solution and described in [RFC4761], which describes a method for VPLS, a particular form of a Layer 2 VPN. That document in turn borrowed much from [RFC4364]. This includes the use of BGP for auto-discovery and "demultiplexor" (see below) exchange, and the concepts of Route Distinguishers to make VPN advertisements unique, and Route Targets to control VPN topology. In addition, all three documents share the idea that routers not directly connected to VPN customers should carry no VPN state, restricting the provisioning of individual connections to just the edge devices. This is achieved by using tunnels to carry the data, with a demultiplexor that identifies individual VPN circuits. These tunnels could be based on 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 an MPLS label as the demultiplexor.

Layer 2 VPNs typically require that all sites in the VPN connect to the SP with the same Layer 2 encapsulation. To ease this restriction, this document proposes a limited form of Layer 2 interworking, by restricting the Layer 3 protocol to IP only (see Section 5).

It may be instructive to compare the approach in [RFC4447] and [RFC6074] (these are the IETF-approved technologies for the functions described in this document, albeit using two separate protocols) with the one described here. To comply with IETF standards, it is recommended that devices implementing the solution described in this document also implement the approach in [RFC4447] and [RFC6074].

The rest of this section discusses the relative merits of Layer 2 and Layer 3 VPNs. Section 3 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.

1.1. Terminology

The terminology used is from [RFC4761] and [RFC4364], and is briefly repeated here. A "customer" is a customer of a Service Provider seeking to interconnect their various "sites" (each an independent network) at Layer 2 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 (provider edge). A router in the Service Provider's network which doesn't connect directly to any CE is termed P ("provider" device). Every pair of PEs is connected by a "tunnel"; within a tunnel, VPN data is distinguished by a "demultiplexor", which in this document is an MPLS label.

Each CE within a VPN is assigned a 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, since a CE may be connected to multiple PEs (or multiply connected to a PE); in such a case, the CE would have a CE ID for each connection. 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. The number space for CE IDs is scoped to a given VPN.

In 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.

Within each physical connection from a CE to a PE, there may be multiple virtual circuits. These will be referred to as Attachment Circuits (ACs), following [RFC3985]. Similarly, the entity that connects two attachment circuits across the Service Provider network is called a pseudowire (PW).

1.2. Advantages of Layer 2 VPNs

A Layer 2 VPN is one where a Service Provider provides Layer 2 connectivity to the customer. 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 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, but route flaps in the customer network have little effect on the SP network. 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. Thus, when offering a Layer 3 VPN, a SP should proactively protect itself from Layer 3 instability in the CE network.

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. Layer 3 VPNs, on the other hand, require a considerable re-design of the customer's Layer 3 routing architecture. Furthermore, 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 choose to perform Route Target Filtering, described in [RFC4684].

1.2.6. Ease of Configuration

Configuring traditional Layer 2 VPNs with dense topologies 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 (Data Link Connection Identifier) 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. Finally, each PVC requires state in every transit switch.

In our proposal, PVCs are tunnelled across the SP network. The tunnels used are provisioned independently 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 ([RFC4364] 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 media 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. In the case of Virtual Private LAN Service (a specific type of L2 VPN; see [RFC4761]), however, the CE-PE link need transport only one copy of a multicast packet.

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.

1.5. 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 [RFC2119].

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 media, and each CE has multiple DLCIs to its PE, each to connect to another CE in the VPN. If the Layer 2 media were ATM, then each CE would have multiple VPI/VCIs (Virtual Path Identifier/Virtual Channel Identifier) 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 media 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 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 4 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 provisioned; CE1 and CE3 have DLCIs 200 through 209 provisioned; CE4 has DLCIs 107, 209, 265, 301, 414, 555, 654, 777 and 888 provisioned; and CE5 has DLCIs 417-426.

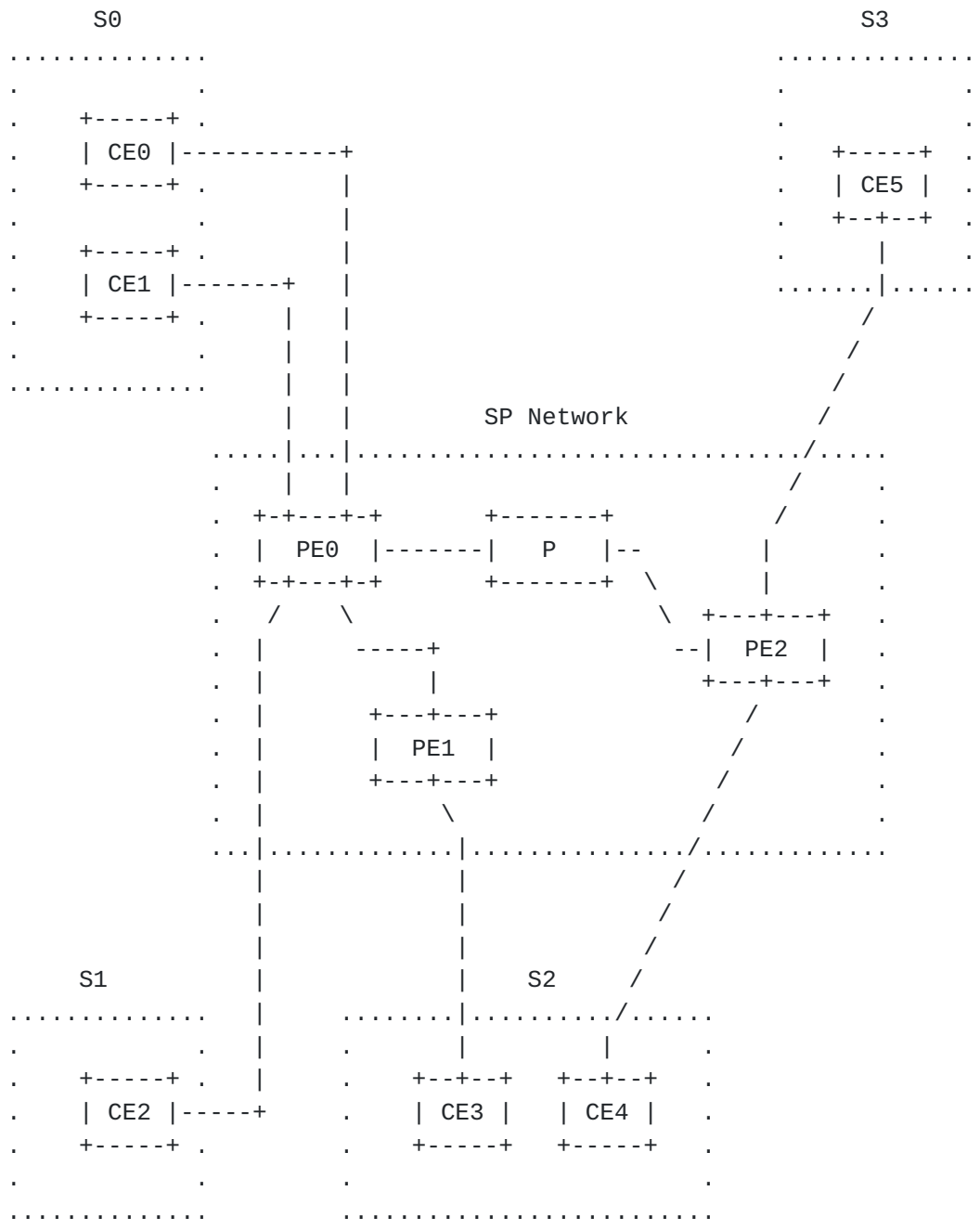


Figure 1: Example Network Topology

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.

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. For the case of ATM VCs, OAM cells may be used; for PPP and Cisco HDLC, keepalives may be used directly between CEs; however, in this case, PEs would not have visibility as to the liveness of customers circuits.

In case of IP-only Layer 2 interworking, if CE1, attached to PE0, connects to CE3, attached to PE1, via a L2VPN circuit, the Layer 2 media between CE1 and PE0 is independent of the Layer 2 media 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 4. 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. For the example, 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. 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. A CE's range must be at least the number of remote CEs that the CE will connect to in a given VPN; if the range exceeds this, then the CE is over-provisioned, in anticipation of growth of the VPN.

Each CE also "knows" which DLCI connects it to each other CE. The methodology followed in this example 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 just the methodology used in the example 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 associated with one or more Route Target communities [RFC4360] which serve to define the topology of the VPN. For each VPN, the PE must determine a Route Distinguisher (RD) to use; this may either be configured or chosen by the PE. RDs do not have to be unique across the VPN. For each CE attached to the PE in a given VPN, the PE must know the set of virtual circuits (DLCI, VCI/VPI or VLAN) connecting it to the CE, and a CE ID identifying the CE within the VPN. CE IDs must be unique in the context of a given VPN.

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. This usually requires adding a new virtual circuit between the PE and CE; in most cases, this configuration is limited to the PE in question.

The rest of the configuration is a local matter between the new CE and the PE to which it is attached. At this point, the PE can signal to other PEs that it has a new site in the VPN, by advertising a BGP Layer 2 route, and traffic connectivity will be set up.

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 between the PE and CE, and does not affect other PEs or CEs.

2.2.4. Deleting a Site

Deleting a site consists first of removing the CE ID of the site from the configuration of the PE to which the site is attached. The PE will then signal to other PEs that it no longer has access to that site, by withdrawing its previously advertised BGP Layer 2 route. Connectivity to the deleted site will cease.

The next steps are book-keeping: decommissioning the attachment circuit from the PE to the CE that corresponds to the site being removed, and noting that the CE ID is now free for future allocation. Note that each PE is now (further) over-provisioned; one may choose to actively "reap" CE IDs if desired.

2.2.5. Managing CE ID Mappings

In the data plane, an attachment circuit, identified say by a DLCI, is mapped to a label via the control plane abstraction of a CE ID. At the egress PE, the label is mapped back to an attachment circuit via the same CE ID. It is up to the VPN administrator

to provision attachment circuits (e.g., DLCIs);

to allocate CE IDs; and

to keep a clear mapping of CE IDs to attachment circuits (and reflect this in PE configurations).

The PEs manage the mappings between attachment circuits and labels, i.e., the data plane mappings.

Note that in the N-to-one modes listed in Table 1, a single attachment circuit may correspond to several Layer 2 virtual circuits. Nevertheless, there is a one-to-one mapping between an attachment circuit and a CE ID (and thus a label).

2.2.6. Managing Label Blocks

Label blocks and label values are managed by the PEs. As sites get added and removed, labels are allocated and released. The easiest way to manage these are using fixed size label blocks rather than variable sized blocks, although the signaling described here supports either. If an implementation uses fixed size blocks, then allocating a label for a new site may requiring allocating a new block; similarly, freeing a label may require freeing a block.

If the implementation requires fixed size blocks, there is probably a default block size, but the implementation SHOULD allow the administrator to choose a size. Larger label block sizes mean more potential "wasted" labels, but less signaling overhead, a trade-off that the administrator might want to control.

Also, as sites get added and deleted, a PE may receive packets with a label that reflects a site that has been deleted locally but remote PEs haven't processed yet, or that reflects a new site added remotely but that hasn't been processed locally. In either of these cases,

the PE SHOULD silently discard the packet; it may choose to log the event once for each such label, but not for every such packet.

2.3. Operations, Administration and Maintenance (OAM)

Many Layer 2 mediums have OAM mechanisms. For example, the Point-to-Point Protocol (PPP) has Echo Request and Echo Reply messages; Frame Relay has the Local Management Interface. Among other things, OAM is used for trouble-shooting and as keepalives.

There are two ways to carry OAM information across Layer 2 VPNs. The first is convey OAM packets as any other Layer 2 packets across the VPN. This is the most general method, and maintains full Layer 2 transparency, and preserves all OAM information. The other method applies only to the link liveness aspect of OAM; it consists of transmitting the status of each attachment circuit across the control plane using the Circuit Status Vector (Section 3.1). This method is the only one applicable to Layer 2 Interworking VPNs (Section 4), since OAM packets are not IP frames, and thus cannot be transmitted across such Layer 2 VPNs.

3. PE Information Exchange

When a PE is configured with all the required information for a CE, it advertises to other PEs the fact that it is participating in a VPN via BGP messages, as per [RFC4761], section 3. BGP was chosen as the means for exchanging L2 VPN information for two reasons: it offers mechanisms for both auto-discovery and signaling, and allows for operational convergence, as explained in Section 1. A bonus for using BGP is a robust inter-AS solution for L2VPNs.

There are two modifications to the formating of messages. The first is that the set of Encaps Types carried in the L2-info extended community has been expanded to include those from Table 1. The value of the Encaps Type field identifies the Layer 2 encapsulation, e.g., ATM, Frame Relay etc.

Encaps Type	Description	Reference
0	Reserved	-
1	Frame Relay	RFC 4446
2	ATM AAL5 SDU VCC transport	RFC 4446
3	ATM transparent cell transport	RFC 4816
4	Ethernet (VLAN) Tagged Mode	RFC 4448
5	Ethernet Raw Mode	RFC 4448
6	Cisco HDLC	RFC 4618
7	PPP	RFC 4618
8	SONET/SDH Circuit Emulation Service	RFC 4842
9	ATM n-to-one VCC cell transport	RFC 4717
10	ATM n-to-one VPC cell transport	RFC 4717
11	IP Layer 2 Transport	RFC 3032
15	Frame Relay Port mode	RFC 4619
17	Structure-agnostic E1 over packet	RFC 4553

18	Structure-agnostic T1 (DS1) over packet	RFC 4553
19	VPLS	RFC 4761
20	Structure-agnostic T3 (DS3) over packet	RFC 4553
21	Nx64kbit/s Basic Service using Structure-aware	RFC 5086
25	Frame Relay DLCI	RFC 4619
40	Structure-agnostic E3 over packet	RFC 4553
41 (1)	Octet-aligned payload for Structure-agnostic DS1 circuits	RFC 4553
42 (2)	E1 Nx64kbit/s with CAS using Structure-aware	RFC 5086
43	DS1 (ESF) Nx64kbit/s with CAS using Structure-aware	RFC 5086
44	DS1 (SF) Nx64kbit/s with CAS using Structure-aware	RFC 5086

Table 1: Encaps Types

Note (1): Allocation of a separate code point for Encaps Type eliminates the need for TDM payload size.

Note (2): Having separate code points for Encaps Types 42-44 allows specifying the trunk framing (i.e, E1, T1 ESF or T1 SF) with CAS.

The second is the introduction of TLVs (Type-Length-Value triplets) in the VPLS NLRI (Network Layer Reachability Information). L2VPN TLVs can be added to extend the information carried in the NLRI, using the format shown in Figure 2. In L2VPN TLVs, Type is 1 octet, Length is 2 octets and represents the size of the Value field in bits.

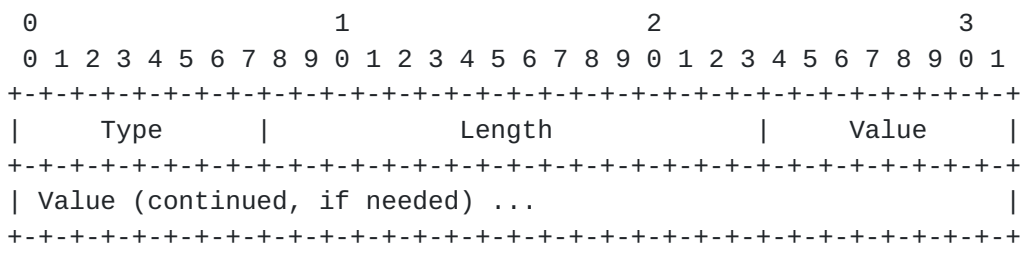


Figure 2: Format of TLVs

TLV Type	Description
1	Circuit Status Vector

Table 2: TLV Types

3.1. Circuit Status Vector

This sub-TLV carries the status of a L2VPN PVC between a pair of PEs. Note that a 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. The local status information for all PVCs corresponding to a label-block is advertised along with the NLRI for the label-block using the status vector TLV. The Type field of this TLV is 1. The Length field of the TLV specifies the length of the value field in bits. 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 corresponds to the PVC associated with the first label in the label block, and 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.

The Value field is padded to the nearest octet boundary.

If PE A receives a VPLS NLRI, 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.

3.2. 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.

4. 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.

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 remote attachment circuit (and thus remote 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 AC). 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.

5.1. Layer 2 MTU

This document requires that the Layer 2 MTU configured on all the access circuits connecting CEs to PEs in a 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.

5.2. 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 [RFC4448], [RFC4618], [RFC4619], and [RFC4717], except in the case of IP-only Layer 2 Interworking which is described next.

5.3. IP-only Layer 2 Interworking

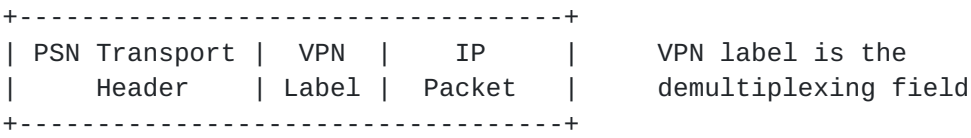


Figure 3: Format of IP-only Layer 2 interworking packet

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 3). 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.

6. Security Considerations

RFC 4761, on which this document is based, has a detailed discussion of security considerations. As in RFC 4761, the focus here is the privacy of customer VPN data (as opposed to confidentiality, integrity, or authentication of said data); to achieve the latter, one can use the methods suggested in RFC 4761. The techniques described in RFC 4761 for securing the control plane and protecting the forwarding path apply equally to L2 VPNs, as do the remarks regarding multi-AS operation. The mitigation strategies and the analogies with [RFC4364] also apply here.

RFC 4761 perhaps should have discussed Denial of Service attacks based on the fact that VPLS PEs have to learn MAC addresses and replicate packets (for flooding and multicast). However, those considerations don't apply here, as neither of those actions are required of PEs implementing the procedures in this document.

7. 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.

8. Contributors

The following contributed to this document.

Manoj Leelanivas, Juniper Networks
Quaizar Vohra, Juniper Networks
Javier Achirica, Consultant
Ronald Bonica, Juniper Networks
Dave Cooper, Global Crossing
Chris Liljenstolpe, Telstra
Eduard Metz, KPN Dutch Telecom
Hamid Ould-Brahim, Nortel
Chandramouli Sargor
Himanshu Shah, Ciena
Vijay Srinivasan
Zhaohui Zhang, Juniper Networks

9. IANA Considerations

IANA is requested to create two new registries: the first is for the one-octet Encaps Type field of the L2-info extended community. The name of the registry is "BGP L2 Encapsulation Types"; the values already allocated are in Table 1 of Section 3. The allocation policy for new entries up to and including value 127 is "Expert Review". The allocation policy for values 128 through 251 is "First Come First Served". The values from 252 through 255 are for "Experimental Use".

The second registry is for the one-octet Type field of the TLVs of the VPLS NLRI. The name of the registry is "BGP L2 TLV Types"; the sole allocated value is in Table 2 of Section 3. The allocation policy for new entries up to and including value 127 is "Expert Review". The allocation policy for values 128 through 251 is "First Come First Served". The values from 252 through 255 are for "Experimental Use".

10. References

10.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, February 2006.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

[RFC4446] Martini, L., "IANA Allocations for Pseudowire Edge to Edge Emulation (PWE3)", BCP 116, RFC 4446, April 2006.

[RFC4448] Martini, L., Rosen, E., El-Aawar, N., and G. Heron, "Encapsulation Methods for Transport of Ethernet over MPLS Networks", RFC 4448, April 2006.

[RFC4618] Martini, L., Rosen, E., Heron, G., and A. Malis, "Encapsulation Methods for Transport of PPP/High-Level Data Link Control (HDLC) over MPLS Networks", RFC 4618, September 2006.

[RFC4619] Martini, L., Kawa, C., and A. Malis, "Encapsulation Methods for Transport of Frame Relay over Multiprotocol Label Switching (MPLS) Networks", RFC 4619, September 2006.

[RFC4717] Martini, L., Jayakumar, J., Bocci, M., El-Aawar, N., Brayley, J., and G. Koleyni, "Encapsulation Methods for Transport of Asynchronous Transfer Mode (ATM) over MPLS Networks", RFC 4717, December 2006.

[RFC4761] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.

10.2. Informative References

[RFC3985] Bryant, S. and P. Pate, "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, March 2005.

[RFC4447] Martini, L., Rosen, E., El-Aawar, N., Smith, T., and G. Heron, "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", RFC 4447, April 2006.

[RFC4684] Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk, R., Patel, K., and J. Guichard, "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, November 2006.

[RFC4762] Lasserre, M. and V. Kompella, "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.

[RFC6074] Rosen, E., Davie, B., Radoaca, V., and W. Luo, "Provisioning, Auto-Discovery, and Signaling in Layer 2 Virtual Private Networks (L2VPNs)", RFC 6074, January 2011.

[Kosiur] Kosiur, D., "Building and Managing Virtual Private Networks", November 1996.

Authors' Addresses

Kireeti Kompella
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: kireeti@juniper.net

Bhupesh Kothari
Cisco Systems
3750 Cisco Way
San Jose, CA 95134
USA

Email: bhupesh@cisco.com

Rao Cherukuri
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: cherukuri@juniper.net