Network Working Group
Internet Draft
Expiration Date: August 2001

Eric C. Rosen
Cisco Systems, Inc.
Yakov Rekhter
Juniper Networks, Inc.

Tony Bogovic
Ravichander Vaidyanathan
Telcordia Technologies

Stephen John Brannon
Monique Jeanne Morrow
Swisscom AG

Marco Carugi
France Telecom

Christopher J. Chase
Luyuan Fang
ATT

Ting Wo Chung
Bell Nexxia

Jeremy De Clercq
Alcatel

Eric Dean
Global One

Paul Hitchin
Adrian Smith
BT

Manoj Leelanivas
Juniper Networks, Inc.

Dave Marshall
Worldcom

Luca Martini
Level 3 Communications, LLC

Vijay Srinivasan
CoSine Communications

Alain Vedrenne
SITA EQUANT

February 2001

BGP/MPLS VPNs

draft-rosen-rfc2547bis-03.txt

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.

Abstract

This document describes a method by which a Service Provider may use an IP backbone to provide VPNs for its customers. MPLS is used for forwarding packets over the backbone, and BGP is used for distributing routes over the backbone. The primary goal of this method is to support the case in which a client obtains IP backbone services from a Service Provider or Service Providers with which it maintains contractual relationships. The client may be an enterprise, a group of enterprises which need an extranet, an Internet Service Provider, an application service provider, another VPN Service Provider which uses this same method to offer VPNs to clients of its own, etc. The method makes it very simple for the client to use the backbone services. It is also very scalable and flexible for the Service Provider, and allows the Service Provider to add value.

This document obsoletes RFC 2547.

Table of Contents

1. Introduction

1.1. Virtual Private Networks

Consider a set of "sites" which are attached to a common network which we may call the "backbone". Let's apply some policy to create a number of subsets of that set, and let's impose the following rule: two sites may have IP interconnectivity over that backbone only if at least one of these subsets contains them both.

The subsets we have created are "Virtual Private Networks" (VPNs). Two sites have IP connectivity over the common backbone only if there is some VPN which contains them both. Two sites which have no VPN in common have no connectivity over that backbone.

If all the sites in a VPN are owned by the same enterprise, the VPN is a corporate "intranet". If the various sites in a VPN are owned by different enterprises, the VPN is an "extranet". A site can be in more than one VPN; e.g., in an intranet and in several extranets. In general, when we use the term VPN we will not be distinguishing between intranets and extranets.

We wish to consider the case in which the backbone is owned and operated by one or more Service Providers (SPs). The owners of the sites are the "customers" of the SPs. The policies that determine whether a particular collection of sites is a VPN are the policies of the customers. Some customers will want the implementation of these policies to be entirely the responsibility of the SP. Other customers may want to implement these policies themselves, or to

share with the SP the responsibility for implementing these policies. In this document, we are primarily discussing mechanisms that may be used to implement these policies. The mechanisms we describe are general enough to allow these policies to be implemented either by the SP alone, or by a VPN customer together with the SP. Most of the discussion is focused on the former case, however.

The mechanisms discussed in this document allow the implementation of a wide range of policies. For example, within a given VPN, we can allow every site to have a direct route to every other site ("full mesh"), or we can restrict certain pairs of sites from having direct routes to each other ("partial mesh").

In this document, we are interested in the case where the common backbone offers an IP service. We are NOT focused on the case where the common backbone is part of the public Internet, but rather on the case where it is the backbone network of an SP or set of SPs with which the customer maintains contractual relationships. That is, the customer is explicitly purchasing VPN service from the SP, rather

than purchasing Internet access from it. (The customer may or may not be purchasing Internet access from the same SP as well.)

The customer itself may be a single enterprise, a set of enterprises needing an extranet, an Internet Service Provider, an application service provider, or even another SP which offers the same kind of VPN service to its own customers.

In the rest of this introduction, we specify some properties which VPNs should have. The remainder of this document outlines a VPN model which has all these properties.

1.2. Edge Devices

We suppose that at each site, there are one or more Customer Edge (CE) devices, each of which is "attached" via some sort of data link (e.g., PPP, ATM, ethernet, Frame Relay, GRE tunnel, etc.) to one or more Provider Edge (PE) routers. Routers in the Provider's network which do not attach to CE devices are known as "P routers".

If a particular site has a single host, that host may be the CE

device. If a particular site has a single subnet, the CE device may be a switch. In general, the CE device can be expected to be a router, which we call the CE router.

We will say that a PE router is attached to a particular VPN if it is attached to a CE device which is in that VPN. Similarly, we will say that a PE router is attached to a particular site if it is attached to a CE device which is in that site.

When the CE device is a router, it is a routing peer of the PE(s) to which it is attached, but it is NOT a routing peer of CE routers at other sites. Routers at different sites do not directly exchange routing information with each other; in fact, they do not even need to know of each other at all. As a consequence, the customer has no backbone or "virtual backbone" to manage, and does not have to deal with any inter-site routing issues. In other words, in the scheme described in this document, a VPN is NOT an "overlay" on top of the SP's network.

With respect to the management of the edge devices, clear administrative boundaries are maintained between the SP and its customers. Customers are not required to access the PE or P routers for management purposes, nor is the SP required to access the CE devices for management purposes.

1.3. Multiple Forwarding Tables in PEs

Each PE router maintains a number of separate forwarding tables. Every site to which the PE is attached must be mapped to one of those forwarding tables. When a packet is received from a particular site, the forwarding table associated with that site is consulted in order to determine how to route the packet. The forwarding table associated with a particular site S is populated ONLY with routes that lead to other sites which have at least one VPN in common with S. This prevents communication between sites which have no VPN in common.

A PE router is attached to a site by virtue of being the endpoint of an interface or "sub-interface" (e.g., PVC, VLAN, GRE tunnel, etc.) whose other endpoint is a CE device. If there are multiple

attachments between a site and a PE router, all the attachments may be mapped to the same forwarding table, or different attachments may be mapped to different forwarding tables. When a PE router receives a packet from a CE device, it knows the interface or sub-interface over which the packet arrived, and this determines the forwarding table used for processing that packet. The choice of forwarding table is NOT determined by the user content of the packet.

Different sites can be mapped to the same forwarding table, but ONLY if they have all their VPNs in common.

A PE router will also have a "default forwarding table," which is not associated with any particular VPN site or sites. The default forwarding table is used for handling traffic which is not VPN traffic, as well as for VPN traffic which is simply transiting this router (i.e., traffic which was not received over a sub-interface whose other endpoint is a CE device, and which is not being sent over a sub-interface whose other endpoint is a CE device.

1.4. VPNs with Overlapping Address Spaces

If two VPNs have no sites in common, then they may have overlapping address spaces. That is, a given address might be used in VPN V1 as the address of system S1, but in VPN V2 as the address of a completely different system S2. This is a common situation when the VPNs each use an RFC1918 private address space. (In fact, two VPNs which do have sites in common may have overlapping address spaces, as long as the overlapping part of the address space does not belong to any of the sites which the two VPNs have in common.)

The fact that sites in different VPNs are mapped to different forwarding tables makes it possible for different VPNs to have

overlapping address spaces, without creating any ambiguity.

1.5. VPNs with Different Routes to the Same System

Although a site may be in multiple VPNs, it is not necessarily the case that the route to a given system at that site should be the same in all the VPNs. Suppose, for example, we have an intranet

consisting of sites A, B, and C, and an extranet consisting of A, B, C, and the "foreign" site D. Suppose that at site A there is a server, and we want clients from B, C, or D to be able to use that server. Suppose also that at site B there is a firewall. We want all the traffic from site D to the server to pass through the firewall, so that traffic from the extranet can be access controlled. However, we don't want traffic from C to pass through the firewall on the way to the server, since this is intranet traffic.

This means that it needs to be possible to set up two routes to the server. One route, used by sites B and C, takes the traffic directly to site A. The second route, used by site D, takes the traffic instead to the firewall at site B. If the firewall allows the traffic to pass, it then appears to be traffic coming from site B, and follows the route to site A.

1.6. SP Backbone Routers

The SP's backbone consists of the PE routers, as well as other routers ("P routers") which do not attach to CE devices.

If every router in an SP's backbone had to maintain routing information for all the VPNs supported by the SP, this model would have severe scalability problems; the number of sites that could be supported would be limited by the amount of routing information that could be held in a single router. It is important therefore that the routing information about a particular VPN is only required to be present in those PE routers which attach to that VPN. In particular, the P routers should not need to have ANY per-VPN routing information whatsoever. (This condition may need to be relaxed somewhat when multicast routing is considered. This is not considered further in this paper.)

So just as the VPN owners do not have a backbone or "virtual backbone" to administer, the SPs themselves do not have a separate backbone or "virtual backbone" to administer for each VPN. Site-to-site routing in the backbone is optimal (within the constraints of the policies used to form the VPNs), and is not constrained in any way by an artificial "virtual topology" of tunnels.

VPNs may span multiple service providers. There are a number of

possible methods for implementing this, which shall be discussed later.

1.7. Security

VPNs of the sort being discussed here, even without making use of cryptographic security measures, are intended to provide a level of security equivalent to that obtainable when a level 2 backbone (e.g., Frame Relay) is used. That is, in the absence of misconfiguration or deliberate interconnection of different VPNs, it is not possible for systems in one VPN to gain access to systems in another VPN. This is discussed in more detail in section 13.

2. Sites and CEs

From the perspective of a particular backbone network, a set of IP systems constitutes a site if those systems have mutual IP interconnectivity, and communication among them occurs without use of the backbone. In general, a site will consist of a set of systems which are in geographic proximity. However, this is not universally true. If two geographic locations are connected via a leased line, over which OSPF is running, and if that line is the preferred way of communicating between the two locations, then the two locations can be regarded as a single site, even if each location has its own CE router. (This notion of "site" is topological, rather than geographical. If the leased line goes down, or otherwise ceases to be the preferred route, but the two geographic locations can continue to communicate by using the VPN backbone, then one site has become two.)

A CE device is always regarded as being in a single site (though as we shall see, a site may consist of multiple "virtual sites"). A site, however, may belong to multiple VPNs.

A PE router may attach to CE devices in any number of different sites, whether those CE devices are in the same or in different VPNs. A CE device may, for robustness, attach to multiple PE routers, of the same or of different service providers. If the CE device is a router, the PE router and the CE router will appear as router adjacencies to each other.

While the basic unit of interconnection is the site, the architecture described herein allows a finer degree of granularity in the control of interconnectivity. For example, certain systems at a site may be members of an intranet as well as members of one or more extranets,

while other systems at the same site may be restricted to being members of the intranet only.

In some cases, a particular site may be divided by the customer into several "virtual sites", perhaps by the use of VLANs. Each virtual site may be a member of a different set of VPNs. For example, if a CE supports VLANs, and wants each VLAN mapped to a separate VPN, the packets sent between CE and PE could be contained in the site's VLAN encapsulation. Then the VLAN tag could be used by the PE, along with the interface over which the packet is received, to assign the packet to a particular VPN.

Alternatively, one could divide the interface into multiple "sub-interfaces" (particularly if the interface is Frame Relay or ATM), and assign the packet to a VPN based on the sub-interface over which it arrives. Or one could simply use a different interface for each virtual site. In any case, only one CE router is ever needed per site, even if there are multiple virtual sites. Of course, a different CE router could be used for each virtual site, if that is desired.

Note that in all these cases, the mechanisms, as well as the policy, for controlling which traffic is in which VPN are in the hand of the customer.

If it is desired to have a particular host be in multiple virtual sites, then that host must determine, for each packet, which virtual site the packet is associated with. It can do this, e.g., by sending packets from different virtual sites on different VLANs, or out different network interfaces.

3. VRFs: Per-Site Forwarding Tables in the PEs

Each PE router maintains one or more "per-site forwarding tables." These are known as VRFs, or "VPN Routing and Forwarding" tables. Every site to which the PE router is attached is associated with one of these tables. A particular packet's IP destination address is looked up in a particular VRF only if that packet has arrived directly from a site which is associated with that table.

It would in fact be more precise to say that in the PE router,
  - sub-interfaces may be mapped to VRFs,
  - the mapping is many-to-one,
  - the VRF in which a packet's destination address is looked up is determined by the sub-interface over which it is received, and

- two sub-interfaces may not be mapped to the same VRF unless the same set of routes is meant to be available to packets received over either sub-interface.

A sub-interface which is mapped to a VRF may be referred to as a "VRF sub-interface".

How are the VRFs populated?

As an example, let PE1, PE2, and PE3 be three PE routers, and let CE1, CE2, and CE3 be three CE routers. Suppose that PE1 learns, from CE1, the routes which are reachable at CE1's site. If PE2 and PE3 are attached respectively to CE2 and CE3, and there is some VPN V containing CE1, CE2, and CE3, then PE1 uses BGP to distribute to PE2 and PE3 the routes which it has learned from CE1. PE2 and PE3 use these routes to populate the VRFs which they associate respectively with the sites of CE2 and CE3. Routes from sites which are not in VPN V do not appear in these VRFs, which means that packets from CE2 or CE3 cannot be sent to sites which are not in VPN V.

If a site is in multiple VPNs, the VRF associated with that site contains routes from the full set of VPNs of which the site is a member.

A PE generally associates only one VRF to each site, even if it is multiply connected to that site. However, different sites can share the same VRF if (and only if) they are meant to use exactly the same set of routes.

When a PE receives a packet from a directly attached site, it always looks up the packet's destination address in the VRF which is associated with that site. However, when a PE receives a packet which is destined to go to a particular directly attached site, it does not necessarily need to lookup the packet's destination address in the VRF (or anywhere else). The packet may already be carrying enough information (in the form of an MPLS label, see section 5) to determine the packet's outgoing sub-interface. That is, the packet's exit point from the backbone may be completely determined by the information in the VRF associated with its entry point to the

backbone.

This allows the backbone to support multiple different routes to the same system, where the route followed by a given packet is determined by the site from which the packet enters the backbone. E.g., one may have one route to a given system for packets from the extranet (where the route leads to a firewall), and a different route to the same system for packets from the intranet (including packets that have already passed through the firewall).

A PE router also contains a "default forwarding table", which is not a VRF. The default forwarding table is used for forwarding packets that arrive on sub-interfaces which are not associated with any VRF, and which are not destined to be sent on sub-interfaces associated with a VRF. The default forwarding table is populated in the normal way by the routing algorithm of the SP network; it does not contain routes from the VPNs.

4. VPN Route Distribution via BGP

PE routers use BGP to distribute VPN routes to each other (more accurately, to cause VPN routes to be distributed to each other).

We allow each VPN to have its own address space, which means that a given address may denote different systems in different VPNs. If two routes, to the same IP address prefix, are actually routes to different systems, it is important to ensure that BGP not treat them as comparable. Otherwise BGP might choose to install only one of them, making the other system unreachable. Further, we must ensure that POLICY is used to determine which packets get sent on which routes; given that several such routes are installed by BGP, only one such must appear in any particular VRF.

We meet these goals by the use of a new address family, as specified below.

4.1. The VPN-IPv4 Address Family

The BGP Multiprotocol Extensions [BGP-MP] allow BGP to carry routes from multiple "address families". We introduce the notion of the

"VPN-IPv4 address family". A VPN-IPv4 address is a 12-byte quantity, beginning with an 8-byte "Route Distinguisher (RD)" and ending with a 4-byte IPv4 address. If two VPNs use the same IPv4 address prefix, the PEs translate these into unique VPN-IPv4 address prefixes. This ensures that if the same address is used in two different VPNs, it is possible to install two completely different routes to that address, one for each VPN.

The RD does not by itself impose any semantics; it contains no information about the origin of the route or about the set of VPNs to which the route is to be distributed. The purpose of the RD is solely to allow one to create distinct routes to a common IPv4 address prefix. Other means are used to determine where to redistribute the route (see section 4.3).

The RD can also be used to create multiple different routes to the very same system. In section 3, we gave an example where the route to a particular server had to be different for intranet traffic than for extranet traffic. This can be achieved by creating two different VPN-IPv4 routes that have the same IPv4 part, but different RDs. This allows BGP to install multiple different routes to the same system, and allows policy to be used (see section 4.3.5) to decide which packets use which route.

The RDs are structured so that every service provider can administer its own "numbering space" (i.e., can make its own assignments of RDs), without conflicting with the RD assignments made by any other service provider. An RD consists of a two-byte type field, an administrator field, and an assigned number field. The value of the type field determines the lengths of the other two fields, as well as the semantics of the administrator field. The administrator field identifies an assigned number authority, and the assigned number field contains a number which has been assigned, by the identified authority, for a particular purpose. For example, one could have an RD whose administrator field contains an Autonomous System number (ASN), and whose (4-byte) number field contains a number assigned by the SP to whom that ASN belongs (having been assigned to that SP by the appropriate authority).

RDs are given this structure in order to ensure that an SP which provides VPN backbone service can always create a unique RD when it

needs to do so. However, the structuring provides no semantics. When BGP compares two such address prefixes, it ignores the structure entirely.

Note that VPN-IPv4 addresses and IPv4 addresses are always considered by BGP to be incomparable.

A VRF may have multiple VPN-IPv4 routes for a single IPv4 address prefix. When a packet's destination address is matched against a VPN-IPv4 route, only the IPv4 part is actually matched.

A PE needs to be configured such that routes which lead to particular CE become associated with a particular RD. The configuration may cause all routes leading to the same CE to be associated with the same RD, or it may be cause different routes to be associated with different RDs, even if they lead to the same CE.

4.2. Encoding of Route Distinguishers

As stated, a VPN-IPv4 address consists of an 8-byte Route Distinguisher followed by a 4-byte IPv4 address. The RDs are encoded as follows:

- Type Field: 2 bytes
- Value Field: 6 bytes

The interpretation of the Value field depends on the value of the Type field. At the present time, two values of the type field are defined: 0 and 1.

- Type 0: The Value field consists of two subfields:

    * Administrator subfield: 2 bytes
    * Assigned Number subfield: 4 bytes

The Administrator subfield must contain an Autonomous System number. If this ASN is from the public ASN space, it must have been assigned by the appropriate authority (use of ASN values from the private ASN space is strongly discouraged). The Assigned Number subfield contains a number from a numbering space which is administered by the enterprise to which the ASN has been assigned by an appropriate authority.

- Type 1: The Value field consists of two subfields:

  * Administrator subfield: 4 bytes
  * Assigned Number subfield: 2 bytes

  The Administrator subfield must contain an IP address. If this IP address is from the public IP address space, it must have been assigned by an appropriate authority (use of addresses from the private IP address space is strongly discouraged). The Assigned Number sub-field contains a number from a numbering space which is administered by the enterprise to which the IP address has been assigned.

4.3. Controlling Route Distribution

In this section, we discuss the way in which the distribution of the VPN-IPv4 routes is controlled.

4.3.1. The Route Target Attribute

Every VRF is associated with one or more "Route Target" attributes.

When a VPN-IPv4 route is created by a PE router, it is associated with one or more "Route Target" attributes. These are carried in BGP as attributes of the route.

Any route associated with Route Target T must be distributed to every PE router that has a VRF associated with Route Target T. When such a route is received by a PE router, it is eligible to be installed

those of the PE's VRFs which are associated with Route Target T. (Whether it actually gets installed depends on the outcome of the BGP decision process.)

A Route Target attribute can be thought of as identifying a set of sites. (Though it would be more precise to think of it as identifying a set of VRFs.) Associating a particular Route Target attribute with a route allows that route to be placed in the VRFs that are used for routing traffic which is received from the corresponding sites.

There is a set of Route Targets that a PE router attaches to a route received from site S; these may be called the "Export Targets". And there is a set of Route Targets that a PE router uses to determine whether a route received from another PE router could be placed in the VRF associated with site S; these may be called the "Import Targets". The two sets are distinct, and need not be the same. Note that a particular VPN-IPv4 route is only eligible for installation in a particular VRF if there is some Route Target which is both one of the route's Route Targets and one of the VRF's Import Targets.

The function performed by the Route Target attribute is similar to that performed by the BGP Communities Attribute. However, the format of the latter is inadequate for present purposes, since it allows only a two-byte numbering space. It is desirable to structure the format, similar to what we have described for RDs (see section 4.2), so that a type field defines the length of an administrator field, and the remainder of the attribute is a number from the specified administrator's numbering space. This can be done using BGP Extended Communities. The Route Targets discussed herein are encoded as BGP Extended Community Route Targets [BGP-EXTCOMM].

When a BGP speaker has received more than one route to the same VPN-IPv4 prefix, the BGP rules for route preference are used to choose which route are installed.

Note that a route can only have one RD, but it can have multiple

Route Targets. In BGP, scalability is improved if one has a single route with multiple attributes, as opposed to multiple routes. One could eliminate the Route Target attribute by creating more routes (i.e., using more RDs), but the scaling properties would be less

favorable.

How does a PE determine which Route Target attributes to associate with a given route? There are a number of different possible ways. The PE might be configured to associate all routes that lead to a particular site with a particular Route Target. Or the PE might be configured to associate certain routes leading to a particular site with one Route Target, and certain with another. Or the CE router, when it distributes these routes to the PE (see section 7), might specify one or more Route Targets for each route. The latter method shifts the control of the mechanisms used to implement the VPN policies from the SP to the customer. If this method is used, it may still be desirable to have the PE eliminate any Route Targets that, according to its own configuration, are not allowed, and/or to add in some Route Targets that according to its own configuration are mandatory.

4.3.2. Route Distribution Among PEs by BGP

If two sites of a VPN attach to PEs which are in the same Autonomous System, the PEs can distribute VPN-IPv4 routes to each other by means of an IBGP connection between them. Alternatively, each can have an IBGP connection to a route reflector.

When a PE router distributes a VPN-IPv4 route via BGP, it uses its own address as the "BGP next hop". This address is encoded as a VPN-IPv4 address with an RD of 0. ([BGP-MP] requires that the next hop address be in the same address family as the NLRI.) It also assigns and distributes an MPLS label. (Essentially, PE routers distribute not VPN-IPv4 routes, but Labeled VPN-IPv4 routes. Cf. [MPLS-BGP]). When the PE processes a received packet that has this label at the top of the stack, the PE will pop the stack, and process the packet appropriately.

The PE may distribute the exact set of routes that appears in the VRF, or it may perform summarization and distribute aggregates of those routes, or it may do some of one and some of the other.

Suppose that a PE has assigned label L to route R, and has distributed this label mapping via BGP. If R is an aggregate of a set of routes in the VRF, the PE will know that packets from the backbone which arrive with this label must have their destination addresses looked up in a VRF. When the PE looks up the label in its

Label Information Base, it learns which VRF must be used. On the other hand, if R is not an aggregate, then when the PE looks up the label, it learns the output sub-interface and the data link encapsulation header for the packet. In this case, no lookup in the VRF is done.

We would expect that the most common case would be the case where the route is NOT an aggregate. The case where it is an aggregate can be very useful though if the VRF contains a large number of host routes (e.g., as in dial-in), or if the VRF has an associated LAN interface (where there is a different outgoing layer 2 header for each system on the LAN, but a route is not distributed for each such system). However, we do not consider this further in this paper.

Note that the use of BGP-distributed MPLS labels is only possible if there is a label switched path between the PE router that installs the BGP-distributed route and PE router which is the BGP next hop of that route. This label switched path may follow a "best effort" route, or it may follow a traffic engineered route. Between a particular PE router and its BGP next hop for a particular route there may be one label switched path, or there may be several, perhaps with different QoS characteristics. All that matters for the VPN architecture is that some label switched path between the router and its BGP next hop exists. However, to ensure interoperability among systems which implement this VPN architecture, all such systems must support LDP [MPLS-LDP].

A PE router, UNLESS it is a Route Reflector (see section 4.3.3) should not install a VPN-IPv4 route unless it has at least one VRF with an Import Target identical to one of the route's Route Target attributes. Inbound filtering should be used to cause such routes to be discarded. If a new Import Target is later added to one of the PE's VRFs (a "VPN Join" operation), it must then acquire the routes it may previously have discarded. This can be done using the refresh mechanism described in [BGP-RFSH]. The outbound route filtering mechanism of [BGP-ORF] can also be used to advantage to make the filtering more dynamic.

Similarly, if a particular Import Target is no longer present in any of a PE's VRFs (as a result of one or more "VPN Prune" operations), the PE may discard all routes which, as a result, no longer have any of the PE's VRF's Import Targets as one of their Route Target Attributes.

A router which is not attached to any VPN, and which is not a Route Reflector (i.e., a P router), never installs any VPN-IPv4 routes at all.

Note that VPN Join and Prune operations are non-disruptive, and do not require any BGP connections to be brought down, as long as the refresh mechanism of [BGP-RFSH] is used.

As a result of these distribution rules, no one PE ever needs to maintain all routes for all VPNs; this is an important scalability consideration.

4.3.3. Use of Route Reflectors

Rather than having a complete IBGP mesh among the PEs, it is advantageous to make use of BGP Route Reflectors [BGP-RR] to improve scalability. All the usual techniques for using route reflectors to improve scalability, e.g., route reflector hierarchies, are available.

Route reflectors are the only systems which need to have routing information for VPNs to which they are not directly attached. However, there is no need to have any one route reflector know all the VPN-IPv4 routes for all the VPNs supported by the backbone.

We outline below two different ways to partition the set of VPN-IPv4 routes among a set of route reflectors.

1. Each route reflector is preconfigured with a list of Route Targets. For redundancy, more than one route reflector may be preconfigured with the same list. A route reflector uses the preconfigured list of Route Targets to construct its inbound route filtering. The route reflector may use the techniques of [BGP-ORF] to install on each of its peers (regardless of whether the peer is another route reflector, or a PE) the set of "Outbound Route Filters" (ORFs) that contain the list of its preconfigured Route Targets. Note that route reflectors should accept ORFs from other route reflectors, which means that route reflectors should advertise the ORF capability to other route reflectors.

   A service provider may modify the list of preconfigured Route Targets on a route reflector. When this is done, the route

reflector modifies the ORFs it installs on all of its IBGP peers. To reduce the frequency of configuration changes on route reflectors, each route reflector may be preconfigured with a block of Route Targets. This way, when a new Route Target is needed for a new VPN, there is already one or more route reflectors that are (pre)configured with this Route Target.

Unless a given PE is a client of all route reflectors, when a new VPN is added to the PE ("VPN Join"), it will need to become a client of the route reflector(s) that maintain routes for that VPN. Likewise, deleting an existing VPN from the PE ("VPN Prune") may result in a situation where the PE no longer need to be a client of some route reflector(s). In either case, the Join or Prune operation is non-disruptive (as long as [BGP-RFSH] is used, and never requires a BGP connection to be brought down, only to be brought right back up.

(By "adding a new VPN to a PE", we really mean adding a new import Route Target to one of its VRFs, or adding a new VRF with an import Route Target not had by any of the PE's other VRFs.)

2. Another method is to have each PE be a client of some subset of the route reflectors. A route reflector is not preconfigured with the list of Route Targets, and does not perform inbound route filtering of routes received from its clients (PEs); rather it accepts all the routes received from all of its clients (PEs). The route reflector keeps track of the set of the Route Targets carried by all the routes it receives. When the route reflector receives from its client a route with a Route Target that is not in this set, this Route Target is immediately added to the set. On the other hand, when the route reflector no longer has any routes with a particular Route Target that is in the set, the route reflector should delay (by a few hours) the deletion of this Route Target from the set.

   The route reflector uses this set to form the inbound route filters that it applies to routes received from other route reflectors. The route reflector may also use ORFs to install the appropriate outbound route filtering on other route

reflectors. Just like with the first approach, a route reflector should accept ORFs from other route reflectors. To accomplish this, a route reflector advertises ORF capability to other route reflectors.

When the route reflector changes the set, it should immediately change its inbound route filtering. In addition, if the route reflector uses ORFs, then the ORFs have to be immediately changed to reflect the changes in the set. If the route reflector doesn't use ORFs, and a new Route Target is added to the set, the route reflector, after changing its inbound route filtering, must issue BGP Refresh to other route reflectors.

The delay of "a few hours" mentioned above allows a route reflector to hold onto routes with a given RT, even after it

loses the last of its clients which are interested in such routes. This protects against the need to reacquire all such routes if the clients' "disappearance" is only temporary.

With this procedure, VPN Join and Prune operations are also non-disruptive.

In both of these procedures, a PE router which attaches to a particular VPN "auto-discovers" the other PEs which attach to the same VPN. When a new PE router is added, or when an existing PE router attaches to a new VPN, no reconfiguration of other PE routers is needed.

Just as there is no one PE router that needs to know all the VPN-IPv4 routes that are supported over the backbone, these distribution rules ensure that there is no one RR which needs to know all the VPN-IPv4 routes that are supported over the backbone. As a result, the total number of such routes that can be supported over the backbone is not bounded by the capacity of any single device, and therefore can increase virtually without bound.

4.3.4. How VPN-IPv4 NLRI is Carried in BGP

The BGP Multiprotocol Extensions [BGP-MP] are used to encode the NLRI. If the AFI field is set to 1, and the SAFI field is set to

128, the NLRI is an MPLS-labeled VPN-IPv4 address. AFI 1 is used since the network layer protocol associated with the NLRI is still IP. Note that this VPN architecture does not require the capability to distribute unlabeled VPN-IPv4 addresses.

In order for two BGP speakers to exchange labeled VPN-IPv4 NLRI, they must use BGP Capabilities Negotiation to ensure that they both are capable of properly processing such NLRI. This is done as specified in [BGP-MP], by using capability code 1 (multiprotocol BGP), with an AFI of 1 and an SAFI of 128.

The labeled VPN-IPv4 NLRI itself is encoded as specified in [MPLS-BGP], where the prefix consists of an 8-byte RD followed by an IPv4 prefix.

4.3.5. Building VPNs using Route Targets

By setting up the Import Targets and Export Targets properly, one can construct different kinds of VPNs.

Suppose it is desired to create a a fully meshed closed user group,

i.e., a set of sites where each can send traffic directly to the other, but traffic cannot be sent to or received from other sites. Then each site is associated with a VRF, a single Route Target attribute is chosen, that Route Target is assigned to each VRF as both the Import Target and the Export Target, and that Route Target is not assigned to any other VRFs as either the Import Target or the Export Target.

Alternatively, suppose one desired, for whatever reason, to create a "hub and spoke" kind of VPN. This could be done by the use of two Route Target values, one meaning "Hub" and one meaning "Spoke". At the VRFs attached to the hub sites, "Hub" is the Export Target and "Spoke" is the Import Target. At the VRFs attached to the spoke site, "Hub" is the Import Target and "Spoke" is the Export Target.

Thus the methods for controlling the distribution of routing information among various sets of sites are very flexible, which in turn provides great flexibility in constructing VPNs.

4.3.6. Route Distribution Among VRFs in a Single PE

It is possible to distribute routes from one VRF to another, even if both VRFs are in the same PE, even though in this case one cannot say that the route has been distributed by BGP. Nevertheless, the decision to distribute a particular route from one VRF to another within a single PE is the same decision that would be made if the VRFs were on different PEs. That is, it depends on the route target attribute which is assigned to the route (or would be assigned if the route were distributed by BGP), and the import target of the second VRF.

5. Forwarding Across the Backbone

If the intermediate routers in the backbone do not have any information about the routes to the VPNs, how are packets forwarded from one VPN site to another?

This is done by means of MPLS with a two-level label stack.

PE routers (and ASBRs which redistribute VPN-IPv4 addresses) need to insert /32 address prefixes for themselves into the IGP routing tables of the backbone. This enables MPLS, at each node in the backbone network, to assign a label corresponding to the route to each PE router. To ensure interoperability among different implementations, it is required to support LDP for setting up the label switched paths across the backbone. However, other methods of

setting up these label switched paths are also possible. (Some of these other methods may not require the presence of the /32 address prefixes in the IGP.)

When a PE receives a packet from a CE device, it chooses a particular VRF in which to look up the packet's destination address. This choice is based on the packet's incoming sub-interface. Assume that a match is found. As a result we learn a "next hop" and an "outgoing sub-interface".

If the packet's outgoing sub-interface is associated with a VRF, then the next hop is a CE device. The packet is sent directly to the CE

device. However, if the outgoing sub-interface and the incoming sub-interface are associated with different VRFs, and the route which best matches the destination address in the incoming sub-interface's VRF is an aggregate of several routes in the outgoing sub-interface's VRF, it may be necessary to look up the packet's destination address in the VRF of the outgoing interface as well.

If the packet's outgoing sub-interface is NOT associated with a VRF, then the packet must travel at least one hop through the backbone. The packet thus has a "BGP Next Hop", and the BGP Next Hop will have assigned a label for the route which best matches the packet's destination address. This label is pushed onto the packet's label stack, and becomes the bottom label. The packet will also have an "IGP Next Hop", which is the next hop along the IGP route to the BGP Next Hop. The IGP Next Hop will have assigned a label for the route which best matches the address of the BGP Next Hop. This label gets pushed on as the packet's top label. The packet is then forwarded to the IGP next hop. (Of course, if the BGP Next Hop and the IGP Next Hop are the same, and if penultimate hop popping is used, the packet may be sent with only the BGP-supplied label.)

MPLS will then carry the packet across the backbone. The egress PE router's treatment of the packet will depend on the label that was first pushed on by the ingress PE. In many cases, the PE will be able to determine, from this label, the sub-interface over which the packet should be transmitted (to a CE device), as well as the proper data link layer header for that interface. In other cases, the PE may only be able to determine that the packet's destination address needs to be looked up in a particular VRF before being forwarded to a CE device. Information in the MPLS header itself, and/or information associated with the label, may also be used to provide QoS on the interface to the CE. In any event, when the packet finally gets to a CE device, it will again be an ordinary unlabeled IP packet.

Note that it is the two-level labeling that makes it possible to keep all the VPN routes out of the P routers, and this in turn is crucial

to ensuring the scalability of the model. The backbone does not even need to have routes to the CEs, only to the PEs.

If it is necessary to carry VPN packets through a sequence of P routers which do not support MPLS, the top label (which represents a

route to the BGP next hop) could in theory be replaced with an "MPLS in IP (or in GRE or in IPsec, etc.)" encapsulation, where the IP destination address is the address of the BGP next hop. The use of such techniques is for further study.

6. Maintaining Proper Isolation of VPNs

To maintain proper isolation of one VPN from another, it is important that no router in the backbone accept a labeled packet from any adjacent non-backbone device unless the following two conditions hold:

1. the label at the top of the label stack was actually distributed by that backbone router to that non-backbone device, and

2. the backbone router can determine that use of that label will cause the packet to leave the backbone before any labels lower in the stack will be inspected, and before the IP header will be inspected.

The first condition ensure that any labeled packets received from non-backbone routers have a legitimate and properly assigned label at the top of the label stack. The second condition ensures that the backbone routers will never look below that top label. Of course, the simplest way to meet these two conditions is just to have the backbone devices refuse to accept labeled packets from non-backbone devices.

7. How PEs Learn Routes from CEs

The PE routers which attach to a particular VPN need to know, for each of that VPN's sites, which addresses in that VPN are at each site.

In the case where the CE device is a host or a switch, this set of addresses will generally be configured into the PE router attaching to that device. In the case where the CE device is a router, there are a number of possible ways that a PE router can obtain this set of addresses.

The PE translates these addresses into VPN-IPv4 addresses, using a configured RD. The PE then treats these VPN-IPv4 routes as input to BGP. Routes from a site are not leaked into the backbone's IGP.

Exactly which PE/CE route distribution techniques are possible depends on whether a particular CE is in a "transit VPN" or not. A "transit VPN" is one which contains a router that receives routes from a "third party" (i.e., from a router which is not in the VPN, but is not a PE router), and that redistributes those routes to a PE router. A VPN which is not a transit VPN is a "stub VPN". The vast majority of VPNs, including just about all corporate enterprise networks, would be expected to be "stubs" in this sense.

The possible PE/CE distribution techniques are:

1. Static routing (i.e., configuration) may be used. (This is likely to be useful only in stub VPNs.)

2. PE and CE routers may be RIP peers, and the CE may use RIP to tell the PE router the set of address prefixes which are reachable at the CE router's site. When RIP is configured in the CE, care must be taken to ensure that address prefixes from other sites (i.e., address prefixes learned by the CE router from the PE router) are never advertised to the PE. More precisely: if a PE router, say PE1, receives a VPN-IPv4 route R1, and as a result distributes an IPv4 route R2 to a CE, then R2 must not be distributed back from that CE's site to a PE router, say PE2, (where PE1 and PE2 may be the same router or different routers), unless PE2 maps R2 to a VPN-IPv4 route which is different than (i.e., contains a different RD than) R1.

3. The PE and CE routers may be OSPF peers. A PE router which is an OSPF peer of a CE router appears, to the CE router, to be an area 0 router. If a PE router is an OSPF peer of CE routers which are in distinct VPNs, the PE must of course be running multiple instances of OSPF.

   IPv4 routes which the PE learns from the CE via OSPF are redistributed into BGP as VPN-IPv4 routes. Extended community attributes are used to carry, along with the route, all the information needed to enable the route to be distributed to other CE routers in the VPN in the proper type of OSPF LSA. OSPF route tagging is used to ensure that routes received from the MPLS/BGP backbone are not sent back into the backbone.

   Specification of the complete set of procedures for the use of OSPF between PE and CE can be found in [VPN-OSPF].

4. The PE and CE routers may be BGP peers, and the CE router may use BGP (in particular, EBGP to tell the PE router the set of address prefixes which are at the CE router's site. (This technique can be used in stub VPNs or transit VPNs.)

   This technique has a number of advantages over the others:

   a) Unlike the IGP alternatives, this does not require the PE to run multiple routing algorithm instances in order to talk to multiple CEs

   b) BGP is explicitly designed for just this function: passing routing information between systems run by different administrations

   c) If the site contains "BGP backdoors", i.e., routers with BGP connections to routers other than PE routers, this procedure will work correctly in all circumstances. The other procedures may or may not work, depending on the precise circumstances.

   d) Use of BGP makes it easy for the CE to pass attributes of the routes to the PE. A complete specification of the set of attributes and their use is outside the scope of this document. However, some examples of the way this may be used are the following:

      - The CE may suggest a particular Route Target for each route, from among the Route Targets that the PE is authorized to attach to the route. The PE would then attach only the suggested Route Target, rather than the full set. This gives the CE administrator some dynamic control of the distribution of routes from the CE.

      - Additional types of Extended Community attributes may be defined, where the intention is to have those attributes passed transparently (i.e., without being changed by the PE routers) from CE to CE. This would allow CE administrators to implement additional route filtering, beyond that which is done by the PEs. This additional filtering would not require

coordination with the SP.

On the other hand, using BGP is likely to be something new for the CE administrators, except in the case where the customer itself is already an Internet Service Provider (ISP), or where the CE devices are managed by the SP.

If a site is not in a transit VPN, note that it need not have a unique Autonomous System Number (ASN). Every CE whose site which is not in a transit VPN can use the same ASN. This can be chosen from the private ASN space, and it will be stripped out by the PE. Routing loops are prevented by use of the Site of Origin Attribute (see below).

What if a set of sites constitute a transit VPN? This will generally be the case only if the VPN is itself an ISP's network, where the ISP is itself buying backbone services from another SP. The latter SP may be called a "Carrier's Carrier". In this case, the best way to provide the VPN is to have the CE routers support MPLS, and to use the technique described in section 9.

When we do not need to distinguish among the different ways in which a PE can be informed of the address prefixes which exist at a given site, we will simply say that the PE has "learned" the routes from that site.

Before a PE can redistribute a VPN-IPv4 route learned from a site, it must assign a Route Target attribute (see section 4.3.1) to the route, and it may assign a Site of Origin attribute to the route.

The Site of Origin attribute, if used, is encoded as a Route Origin Extended Community [BGP-EXTCOMM]. The purpose of this attribute is to uniquely identify the set of routes learned from a particular site. This attribute is needed in some cases to ensure that a route learned from a particular site via a particular PE/CE connection is not distributed back to the site through a different PE/CE connection. It is particularly useful if BGP is being used as the PE/CE protocol, but different sites have not been assigned distinct ASNs.

8. How CEs learn Routes from PEs

In this section, we assume that the CE device is a router.

If the PE places a particular route in the VRF it uses to route packets received from a particular CE, then in general, the PE may distribute that route to the CE. Of course the PE may distribute that route to the CE only if this is permitted by the rules of the PE/CE protocol. (For example, if a particular PE/CE protocol has "split horizon", certain routes in the VRF cannot be redistributed back to the CE.) We add one more restriction on the distribution of routes from PE to CE: if a route's Site of Origin attribute

identifies a particular site, that route must never be redistributed to any CE at that site.

In most cases, however, it will be sufficient for the PE to simply distribute the default route to the CE. (In some cases, it may even be sufficient for the CE to be configured with a default route pointing to the PE.) This will generally work at any site which does not itself need to distribute the default route to other sites. (E.g., if one site in a corporate VPN has the corporation's access to the Internet, that site might need to have default distributed to the other site, but one could not distribute default to that site itself.)

Whatever procedure is used to distribute routes from CE to PE will also be used to distribute routes from PE to CE.

9. Carriers' Carriers

Sometimes a VPN may actually be the network of an ISP, with its own peering and routing policies. Sometimes a VPN may be the network of an SP which is offering VPN services in turn to its own customers. VPNs like these can also obtain backbone service from another SP, the "carrier's carrier", using essentially the same methods described in this document. In particular:

- The CE routers should distribute to the PE routers ONLY those routes which are internal to the VPN. This allows the VPN to be

handled as a stub VPN.

- The CE routers should support MPLS, in that they should be able to receive labels from the PE routers, and send labeled packets to the PE routers. They do not need to distribute labels of their own though.

- The PE routers should distribute, to the CE routers, labels for the routes they distribute to the CE routers.

- Routers at the different sites should establish BGP connections among themselves for the purpose of exchanging external routes (i.e., routes which lead outside of the VPN).

- All the external routes must be known to the CE routers.

Then when a CE router looks up a packet's destination address, the routing lookup will resolve to an internal address, usually the address of the packet's BGP next hop. The CE labels the packet appropriately and sends the packet to the PE.

In the above procedure, the CE routers are the only routers in the VPN which need to support MPLS. If, on the other hand, all the routers at a particular VPN site support MPLS, then it is no longer required that the CE routers know all the external routes. All that is required is that the external routes be known to whatever routers are responsible for putting the label stack on a hitherto unlabeled packet, and that there be label switched path that leads from those routers to their BGP peers at other sites. In this case, for each internal route that a CE router distributes to a PE router, it must also distribute a label.

10. Inter-Provider Backbones

What if two sites of a VPN are connected to different Autonomous Systems (e.g., because the sites are connected to different SPs)? The PE routers attached to that VPN will then not be able to maintain IBGP connections with each other, or with a common route reflector. Rather, there needs to be some way to use EBGP to distribute VPN-IPv4 addresses.

There are a number of different ways of handling this case, which we present in order of increasing scalability.

a) VRF-to-VRF connections at the AS border routers.

In this procedure, a PE router in one AS attaches directly to a PE router in another. The two PE routers will be attached by multiple sub-interfaces, at least one for each of the VPNs whose routes need to be passed from AS to AS. Each PE will treat the other as if it were a CE router. That is, the PEs associate each such sub-interface with a VRF, and use EBGP to distribute unlabeled IPv4 addresses to each other.

This is a procedure that "just works", and that does not require MPLS at the border between ASes. However, it does not scale as well as the other procedures discussed below.

b) EBGP redistribution of labeled VPN-IPv4 routes from AS to neighboring AS.

In this procedure, the PE routers use IBGP to redistribute labeled VPN-IPv4 routes either to an Autonomous System Border Router (ASBR), or to a route reflector of which an ASBR is a client. The ASBR then uses EBGP to redistribute those labeled VPN-IPv4 routes to an ASBR in another AS, which in turn distributes them to the PE routers in that AS, or perhaps to another ASBR which in turn distributes them ...

When using this procedure, VPN-IPv4 routes should only be accepted on EBGP connections at private peering points, as part of a trusted arrangement between SPs. VPN-IPv4 routes should neither be distributed to nor accepted from the public Internet, or from any BGP peers which are not trusted. An ASBR should never accept a labeled packet from an EBGP peer unless it has actually distributed the top label to that peer.

If there are many VPNs having sites attached to different Autonomous Systems, there does not need to be a single ASBR between those two ASes which holds all the routes for all the VPNs; there can be multiple ASBRs, each of which holds only the routes for a particular subset of the VPNs.

This procedure requires that there be a label switched path leading from a packet's ingress PE to its egress PE. Hence the appropriate trust relationships must exist between and among the set of ASes along the path. Also, there must be agreement among the set of SPs as to which border routers need to receive routes with which Route Targets.

c) Multihop EBGP redistribution of labeled VPN-IPv4 routes between source and destination ASes, with EBGP redistribution of labeled IPv4 routes from AS to neighboring AS.

In this procedure, VPN-IPv4 routes are neither maintained nor distributed by the ASBRs. An ASBR must maintain labeled IPv4 /32 routes to the PE routers within its AS. It uses EBGP to distribute these routes to other ASes. ASBRs in any transit ASes will also have to use EBGP to pass along the labeled /32 routes. This results in the creation of a label switched path from the ingress PE router to the egress PE router. Now PE routers in different ASes can establish multi-hop EBGP connections to each other, and can exchange VPN-IPv4 routes over those connections.

If the /32 routes for the PE routers are made known to the P routers of each AS, everything works normally. If the /32 routes for the PE routers are NOT made known to the P routers (other than the ASBRs), then this procedure requires a packet's ingress PE to put a three label stack on it. The bottom label is assigned by the egress PE, corresponding to the packet's destination address in a particular VRF. The middle label is assigned by the ASBR, corresponding to the /32 route to the egress PE. The top label is assigned by the ingress PE's IGP Next Hop, corresponding to the /32 route to the ASBR.

To improve scalability, one can have the multi-hop EBGP

connections exist only between a route reflector in one AS and a route reflector in another. (However, when the route reflectors distribute routes over this connection, they do not modify the BGP next hop attribute of the routes.) The actual PE routers would then only have IBGP connections to the route reflectors in their own AS.

This procedure is very similar to the "Carrier's Carrier" procedures described in section 9. Like the previous procedure, it requires that there be a label switched path leading from a packet's ingress PE to its egress PE.

11. Accessing the Internet from a VPN

Many VPN sites will need to be able to access the public Internet, as well as to access other VPN sites. The following describes some of the alternative ways of doing this.

1. In some VPNs, one or more of the sites will obtain Internet Access by means of an "Internet gateway" (perhaps a firewall) attached to a non-VRF interface to an ISP. The ISP may or may not be the same organization as the SP which is providing the VPN service. Traffic to/from the Internet gateway would then be routed according to the PE router's default forwarding table.

   In this case, the sites which have Internet Access may be distributing a default route to their PEs, which in turn redistribute it to other PEs and hence into other sites of the VPN. This provides Internet Access for all of the VPN's sites.

   In order to properly handle traffic from the Internet, the ISP must distribute, to the Internet, routes leading to addresses that are within the VPN. This is completely independent of any of the route distribution procedures described in this document. The internal structure of the VPN will in general not be visible from the Internet; such routes would simply lead to the non-VRF interface that attaches to the VPN's Internet gateway.

   In this model, there is no exchange of routes between a PE router's default forwarding table and any of its VRFs. VPN route distribution procedures and Internet route distribution procedures are completely independent.

   Note that although some sites of the VPN use a VRF interface to communicate with the Internet, ultimately all packets to/from

the Internet traverse a non-VRF interface before leaving/entering the VPN, so we refer to this as "non-VRF Internet Access".

Note that the PE router to which the non-VRF interface attaches does not necessarily need to maintain all the Internet routes in its default forwarding table. The default forwarding table could have as few as one route, "default", which leads to another router (probably an adjacent one) which has the Internet routes. A variation of this scheme is to tunnel packets received over the non-VRF interface from the PE router to another router, where this other router maintains the full set of Internet routes.

2. Some VPNs may obtain Internet access via a VRF interface ("VRF Internet Access"). If a packet is received by a PE over a VRF interface, and if the packet's destination address does not match any route in the VRF, then it may be matched against the PE's default forwarding table. If a match is made there, the packet can be forwarded natively through the backbone to the Internet, instead of being forwarded by MPLS.

   In order for traffic to flow natively in the opposite direction (from Internet to VRF interface), some of the routes from the VRF must be exported to the Internet forwarding table. Needless to say, any such routes must correspond to globally unique addresses.

   In this scheme, the default forwarding table might have the full set of Internet routes, or it might have a little as a single default route leading to another router which does have the full set of Internet routes in its default forwarding table.

3. Suppose the PE has the capability to store "non-VPN routes" in a VRF. If a packet's destination address matches a "non-VPN route", then the packet is transmitted natively, rather than being transmitted via MPLS. If the VRF contains a non-VPN default route, all packets for the public Internet will match it, and be forwarded natively to the default route's next hop. At that next hop, the packets' destination addresses will be looked up in the default forwarding table, and may match more specific routes.

   This technique would only be available if none of the CE routers is distributing a default route.

4. It is also possible to obtain Internet access via a VRF interface by having the VRF contain the Internet routes. Compared with model 2, this eliminates the second lookup, but it has the disadvantage of requiring the Internet routes to be replicated in each such VRF.

   If this technique is used, the SP may want to make its interface to the Internet be a VRF interface, and to use the techniques of section 4 to distribute Internet routes, as VPN-IPv4 routes, to other VRFs.

It should be clearly understood that by default, there is no exchange of routes between a VRF and the default forwarding table. This is done ONLY upon agreement between a customer and a SP, and only if it suits the customer's policies.

12. Management VPNs

This specification does not require that the sub-interface connecting a PE router and a CE router be a "numbered" interface. If it is a numbered interface, this specification allows the addresses assigned to the interface to come from either the address space of the VPN or the address space of the SP.

If a CE router is being managed by the Service Provider, then the Service Provider will likely have a network management system which needs to be able to communicate with the CE router. In this case, the addresses assigned to the sub-interface connecting the CE and PE routers should come from the SP's address space, and should be unique within that space. The network management system should itself connect to a PE router (more precisely, be at a site which connects to a PE router) via a VRF interface. The address of the network management system will be exported to all VRFs which are associated with interfaces to CE routers that are managed by the SP. The addresses of the CE routers will be exported to the VRF associated with the Network Management system, but not to any other VRFs.

This allows communication between CE and Network Management system, but does not allow any undesired communication to or among the CE routers.

One way to ensure that the proper route import/exports are done is to

use two Route Targets, call them T1 and T2. If a particular VRF interface attaches to a CE router that is managed by the SP, then that VRF is configured to:

- import routes that have T1 attached to them, and

- attach T2 to addresses assigned to each end of its VRF interfaces.

If a particular VRF interface attaches to the SP's Network Management system, then that VRF is configured to attach T1 to the address of that system, and to import routes that have T2 attached to them.

13. Security

13.1. Data Plane

By security in the "data plane", we mean protection against the following possibilities:

- Packets from within a VPN travel to a site outside the VPN, other than in a manner consistent with the policies of the VPN.

- Packets from outside a VPN enter one of the VPN's sites, other than in a manner consistent with the policies of the VPN.

Under the following conditions:

1. a backbone router does not accept labeled packets over a particular data link, unless it is known that that data link attaches only to trusted systems, or unless it is known that such packets will leave the backbone before the IP header or any labels lower in the stack will be inspected, and

2. labeled VPN-IPv4 routes are not accepted from untrusted or unreliable routing peers,

3. no successful attacks have been mounted on the control plane,

the data plane security provided by this architecture is virtually identical to that provided to VPNs by Frame Relay or ATM backbones. If the devices under the control of the SP are properly configured, data will not enter or leave a VPN unless authorized to do so.

Condition 1 above can be stated more precisely. One should discard a labeled packet received from a particular neighbor unless one of the following two conditions holds:

- the packet's top label has a label value which the receiving system has distributed to that neighbor, or

- the packet's top label has a label value which the receiving system has distributed to a system beyond that neighbor (i.e., when it is known that the path from the system to which the label was distributed to the receiving system may be via that neighbor).

Condition 2 above is of most interest in the case of inter-provider VPNs (see section 10). For inter-provider VPNs constructed according to scheme b) of section 10, condition 2 is easily checked. (The issue of security when scheme c) of section 10 is used is for further study.)

It is worth noting that the use of MPLS makes it much simpler to provide data plane security than might be possible if one attempted to use some form of IP tunneling in place of the MPLS outer label. It is a simple matter to have one's border routers refuse to accept a labeled packet unless the first of the above conditions applies to it. It is rather more difficult to configure a router to refuse to accept an IP packet if that packet is an IP tunnelled packet whose destination address is that of a PE router; certainly this is not impossible to do, but it has both management and performance implications.

Note that if the PE routers support any "MPLS in IP" or "MPLS in GRE" or similar encapsulations, security is compromised unless either any such packets are filtered at the borders, or else some acceptable

means of authentication (e.g., IPsec authentication) is carried in the packet itself.

In the case where a number of CE routers attach to a PE router via a LAN interface, to ensure proper security, one of the following conditions must hold:

1. All the CE routers on the LAN belong to the same VPN, or

2. A trusted and secured LAN switch divides the LAN into multiple VLANs, with each VLAN containing only systems of a single VPN; in this case the switch will attach the appropriate VLAN tag to any packet before forwarding it to the PE router.

Cryptographic privacy is not provided by this architecture, nor by Frame Relay or ATM VPNs. These architectures are all compatible with the use of cryptography on a CE-CE basis, if that is desired.

The use of cryptography on a PE-PE basis is for further study.

13.2. Control Plane

The data plane security of the previous section depends on the security of the control plane. To ensure security, neither BGP nor LDP connections should be made with untrusted peers. The TCP/IP MD5 authentication option should be used with both these protocols. The routing protocol within the SP's network should also be secured in a similar manner.

13.3. Security of P and PE devices

If the physical security of these devices is compromised, data plane security may also be compromised.

The usual steps should be take to ensure that IP traffic from the public Internet cannot be used to modify the configuration of these devices, or to mount Denial of Service attacks on them.

14. Quality of Service

Although not the focus of this paper, Quality of Service is a key component of any VPN service. In MPLS/BGP VPNs, existing L3 QoS capabilities can be applied to labeled packets through the use of the "experimental" bits in the shim header [MPLS-ENCAPS], or, where ATM is used as the backbone, through the use of ATM QoS capabilities. The traffic engineering work discussed in [MPLS-RSVP] is also directly applicable to MPLS/BGP VPNs. Traffic engineering could even be used to establish label switched paths with particular QoS characteristics between particular pairs of sites, if that is desirable. Where an MPLS/BGP VPN spans multiple SPs, the architecture described in [PASTE] may be useful. An SP may apply either intserv or diffserv capabilities to a particular VPN, as appropriate.

15. Scalability

We have discussed scalability issues throughout this paper. In this section, we briefly summarize the main characteristics of our model with respect to scalability.

The Service Provider backbone network consists of (a) PE routers, (b) BGP Route Reflectors, (c) P routers (which are neither PE routers nor Route Reflectors), and, in the case of multi-provider VPNs, (d) ASBRs.

P routers do not maintain any VPN routes. In order to properly forward VPN traffic, the P routers need only maintain routes to the PE routers and the ASBRs. The use of two levels of labeling is what makes it possible to keep the VPN routes out of the P routers.

A PE router maintains VPN routes, but only for those VPNs to which it is directly attached.

Route reflectors can be partitioned among VPNs so that each partition carries routes for only a subset of the VPNs supported by the Service Provider. Thus no single route reflector is required to maintain routes for all VPNs.

For inter-provider VPNs, if the ASBRs maintain and distribute VPN-IPv4 routes, then the ASBRs can be partitioned among VPNs in a

similar manner, with the result that no single ASBR is required to maintain routes for all the inter-provider VPNs. If multi-hop EBGP is used, then the ASBRs need not maintain and distribute VPN-IPv4 routes at all.

As a result, no single component within the Service Provider network has to maintain all the routes for all the VPNs. So the total capacity of the network to support increasing numbers of VPNs is not limited by the capacity of any individual component.

16. Intellectual Property Considerations

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

17. Acknowledgments

Significant contributions to this work have been made by Ravi Chandra, Dan Tappan and Bob Thomas.

We also wish to thank Shantam Biswas for his review and contributions.

18. Authors' Addresses

Eric C. Rosen
Cisco Systems, Inc.
250 Apollo Drive
Chelmsford, MA, 01824
E-mail: erosen@cisco.com

Yakov Rekhter
Juniper Networks
1194 N. Mathilda Avenue
Sunnyvale, CA 94089
E-mail: yakov@juniper.net

Tony Bogovic
Telcordia Technologies
445 South Street, Room 1A264B
Morristown, NJ 07960
E-mail: tjb@research.telcordia.com

Stephen John Brannon
Swisscom AG
Postfach 1570
CH-8301
Glattzentrum (Zuerich), Switzerland
E-mail: stephen.brannon@swisscom.com

Marco Carugi
France Telecom / CNET Research Centre
IP networks and services
CNET/DAC/NTR
Technopole Anticipa
2, av. P. Marzin
22307 Lannion
E-mail: marco.carugi@cnet.francetelecom.fr

Christopher J. Chase
AT&T
200 Laurel Ave
Middletown, NJ 07748
USA
E-mail: chase@att.com

Ting Wo Chung
Bell Nexxia
181 Bay Street
Suite 350
Toronto, Ontario
M5J2T3
E-mail: ting_wo.chung@bellnexxia.com

Eric Dean
Global One
12490 Sunrise Valley Dr.
Reston, VA 20170 USA
E-mail: edean@gip.net

Jeremy De Clercq
Alcatel Network Strategy Group
Francis Wellesplein 1
2018 Antwerp, Belgium
E-mail: jeremy.de_clercq@alcatel.be

Luyuan Fang
AT&T
IP Backbone Architecture
200 Laurel Ave.
Middletown, NJ 07748
E-mail: luyuanfang@att.com

Paul Hitchin
BT
BT Adastral Park
Martlesham Heath,
Ipswich IP5 3RE
UK
E-mail: paul.hitchen@bt.com

Manoj Leelanivas
Juniper Networks, Inc.
385 Ravendale Drive
Mountain View, CA 94043 USA
E-mail: manoj@juniper.net

Dave Marshall
Worldcom
901 International Parkway
Richardson, Texas 75081
E-mail: dave.marshall@wcom.com

Luca Martini
Level 3 Communications, LLC.
1025 Eldorado Blvd.
Broomfield, CO, 80021
E-mail: luca@level3.net

Monique Jeanne Morrow
Swisscom AG
Postfach 1570
CH-8301
Glattzentrum (Zuerich), Switzerland
E-mail: monique.morrow@swisscom.com

Ravichander Vaidyanathan
Telcordia Technologies
445 South Street, Room 1C258B
Morristown, NJ 07960
E-mail: vravi@research.telcordia.com

Adrian Smith
BT
BT Adastral Park
Martlesham Heath,
Ipswich IP5 3RE
UK
E-mail: adrian.ca.smith@bt.com

Vijay Srinivasan
1200 Bridge Parkway
Redwood City, CA 94065
E-mail: vsriniva@cosinecom.com

Alain Vedrenne
SITA EQUANT
3100 Cumberland Blvd, Suite 200
Atlanta, GA, 30339 USA
Email:Alain.Vedrenne@sita.int
Alain.Vedrenne@equant.com

19. References

[BGP-MP] Bates, Chandra, Katz, and Rekhter, "Multiprotocol Extensions for BGP4", June 2000, RFC 2858

[BGP-EXTCOMM] Ramachandra, Tappan, "BGP Extended Communities Attribute", January 2001, work in progress

[BGP-ORF] Chen, Rekhter, "Cooperative Route Filtering Capability for BGP-4", November 2000, work in progress

[BGP-RFSH] Chen, "Route Refresh Capability for BGP-4", March 2000, RFC 2918

[BGP-RR] Bates and Chandrasekaran, "BGP Route Reflection: An alternative to full mesh IBGP", RFC 2796, April 2000

[IPSEC] Kent and Atkinson, "Security Architecture for the Internet Protocol", November 1998, RFC 2401

[MPLS-ARCH] Rosen, Viswanathan, and Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001

[MPLS-BGP] Rekhter and Rosen, "Carrying Label Information in BGP4", January 2001, work in progress

[MPLS-LDP] Andersson, Doolan, Feldman, Fredette, Thomas, "LDP Specification", RFC 3036, January 2001

[MPLS-ENCAPS] Rosen, Rekhter, Tappan, Farinacci, Fedorkow, Li, and Conta, "MPLS Label Stack Encoding", RFC 3032, January 2001

[MPLS-RSVP] Awduche, Berger, Gan, Li, Srinavasan, Swallow, "Extensions to RSVP for LSP Tunnels", August 2000, work in progress

[PASTE] Li and Rekhter, "A Provider Architecture for Differentiated Services and Traffic Engineering (PASTE)", RFC 2430, October 1998.

[VPN-OSPF] Rosen and Psenak, "OSPF as the PE/CE Protocol in BGP/MPLS VPNs", February 2001, work in progress

20. Full Copyright Statement