Network Working Group
Request for Comments: 2547
Category: Informational

E. Rosen
Y. Rekhter
Cisco Systems, Inc.

March 1999

BGP/MPLS VPNs

Status of this Memo

Copyright Notice

Abstract

This document describes a method by which a Service Provider with an IP backbone may provide VPNs (Virtual Private Networks) for its customers. MPLS (Multiprotocol Label Switching) is used for forwarding packets over the backbone, and BGP (Border Gateway Protocol) is used for distributing routes over the backbone. The primary goal of this method is to support the outsourcing of IP backbone services for enterprise networks. It does so in a manner which is simple for the enterprise, while still scalable and flexible for the Service Provider, and while allowing the Service Provider to add value. These techniques can also be used to provide a VPN which itself provides IP service to customers.

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 several extranets. We regard both intranets and extranets as VPNs. 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 particularly interested in the case where the common backbone offers an IP service. We are primarily concerned with the case in which an enterprise is outsourcing its backbone to a service provider, or perhaps to a set of service providers, with which it maintains contractual relationships. We are not focused on providing VPNs over the public Internet.

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. The VPN Model of this document appears to be an instance of the framework described in [4].

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.

If a particular site has a single host, that host may be the CE device. If a particular site has a single subnet, that 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 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 (except in the case where this is necessary for security purposes, see section 9). As a consequence, very large VPNs (i.e., VPNs with a very large number of sites) are easily supported, while the routing strategy for each individual site is greatly simplified.

It is important to maintain clear administrative boundaries between the SP and its customers (cf. [4]). The PE and P routers should be administered solely by the SP, and the SP's customers should not have any management access to it. The CE devices should be administered solely by the customer (unless the customer has contracted the management services out to the SP).

1.3. VPNs with Overlapping Address Spaces

We assume that any two non-intersecting VPNs (i.e., VPNs with no sites in common) may have overlapping address spaces; the same address may be reused, for different systems, in different VPNs. As long as a given endsystem has an address which is unique within the scope of the VPNs that it belongs to, the endsystem itself does not need to know anything about VPNs.

In this model, the VPN owners do not have a backbone to administer, not even a "virtual backbone". Nor do the SPs have to administer a separate backbone or "virtual backbone" 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.

1.4. 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.

1.5. Multiple Forwarding Tables in PEs

Each PE router needs to maintain 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, and it allows two VPNs with no site in common to use address spaces that overlap with each other.

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 to require therefore that the routing information about a particular VPN be present ONLY 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.

VPNs may span multiple service providers. We assume though that when the path between PE routers crosses a boundary between SP networks, it does so via a private peering arrangement, at which there exists mutual trust between the two providers. In particular, each provider must trust the other to pass it only correct routing information, and to pass it labeled (in the sense of MPLS [9]) packets only if those packets have been labeled by trusted sources. We also assume that it is possible for label switched paths to cross the boundary between service providers.

1.7. Security

A VPN model should, even without the use of cryptographic security measures, 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 should not be possible for systems in one VPN to gain access to systems in another VPN.

It should also be possible to deploy standard security procedures.

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 between 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; two geographic locations connected via a leased line, over which OSPF is running, will constitute a single site, because communication between the two locations does not involve the use of the backbone.

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.

3. Per-Site Forwarding Tables in the PEs

Each PE router maintains one or more "per-site 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 per-site forwarding table only if that packet has arrived directly from a site which is associated with that table.

How are the per-site forwarding tables 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 forwarding tables 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 forwarding tables, 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 forwarding table associated with that site can contain routes from the full set of VPNs of which the site is a member.

A PE generally maintains only one forwarding table per site, even if it is multiply connected to that site. Also, different sites can share the same forwarding table if they are meant to use exactly the same set of routes.

Suppose a packet is received by a PE router from a particular directly attached site, but the packet's destination address does not match any entry in the forwarding table associated with that site. If the SP is not providing Internet access for that site, then the packet is discarded as undeliverable. If the SP is providing Internet access for that site, then the PE's Internet forwarding table will be consulted. This means that in general, only one forwarding table per PE need ever contain routes from the Internet, even if Internet access is provided.

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 (a) the label at the top of the label stack was actually distributed by the backbone router to the non-backbone device, and (b) 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. These restrictions are necessary in order to prevent packets from entering a VPN where they do not belong.

The per-site forwarding tables in a PE are ONLY used for packets which arrive from a site which is directly attached to the PE. They are not used for routing packets which arrive from other routers that belong to the SP backbone. As a result, there may be 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).

3.1. Virtual Sites

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. The PE then needs to contain a separate forwarding table for each virtual site. 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, and this could be used by the PE, along with the interface over which the packet is received, to assign the packet to a particular virtual site.

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, our out different network interfaces.

These schemes do NOT require the CE to support MPLS. Section 8 contains a brief discussion of how the CE might support multiple virtual sites if it does support MPLS.

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).

A BGP speaker can only install and distribute one route to a given address prefix. Yet we allow each VPN to have its own address space, which means that the same address can be used in any number of VPNs, where in each VPN the address denotes a different system. It follows

that we need to allow BGP to install and distribute multiple routes to a single IP address prefix. Further, we must ensure that POLICY is used to determine which sites can be use which routes; given that several such routes are installed by BGP, only one such must appear in any particular per-site forwarding table.

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 [3] 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 4.2).

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 4.2.3) 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 IANA has assigned that ASN. 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.

If the Administrator subfield and the Assigned Number subfield of a VPN-IPv4 address are both set to all zeroes, the VPN-IPv4 address is considered to have exactly the same meaning as the corresponding globally unique IPv4 address. In particular, this VPN-IPv4 address and the corresponding globally unique IPv4 address will be considered comparable by BGP. In all other cases, a VPN-IPv4 address and its corresponding globally unique IPv4 address will be considered noncomparable by BGP.

A given per-site forwarding table will only have one VPN-IPv4 route for any given 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 to associate routes which lead to particular CE with a particular RD. The PE may be configured to associate all routes leading to the same CE with the same RD, or it may be configured to associate different routes with different RDs, even if they lead to the same CE.

4.2. Controlling Route Distribution

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

4.2.1. The Target VPN Attribute

Every per-site forwarding table is associated with one or more "Target VPN" attributes.

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

Any route associated with Target VPN T must be distributed to every PE router that has a forwarding table associated with Target VPN T. When such a route is received by a PE router, it is eligible to be installed in each of the PE's per-site forwarding tables that is associated with Target VPN T. (Whether it actually gets installed depends on the outcome of the BGP decision process.)

In essence, a Target VPN attribute identifies a set of sites. Associating a particular Target VPN attribute with a route allows that route to be placed in the per-site forwarding tables that are used for routing traffic which is received from the corresponding sites.

There is a set of Target VPNs that a PE router attaches to a route received from site S. And there is a set of Target VPNs that a PE router uses to determine whether a route received from another PE router could be placed in the forwarding table associated with site S. The two sets are distinct, and need not be the same.

The function performed by the Target VPN attribute is similar to that performed by the BGP Communities Attribute. However, the format of the latter is inadequate, since it allows only a two-byte numbering space. It would be fairly straightforward to extend the BGP Communities Attribute to provide a larger numbering space. It should also be possible to structure the format, similar to what we have described for RDs (see section 4.1), 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.

When a BGP speaker has received two routes to the same VPN-IPv4 prefix, it chooses one, according to the BGP rules for route preference.

Note that a route can only have one RD, but it can have multiple Target VPNs. In BGP, scalability is improved if one has a single route with multiple attributes, as opposed to multiple routes. One could eliminate the Target VPN attribute by creating more routes (i.e., using more RDs), but the scaling properties would be less favorable.

How does a PE determine which Target VPN 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 Target VPN. Or the PE might be configured to associate certain routes leading to a particular site with one Target VPN, and certain with another. Or the CE router, when it distributes these routes to the PE (see section 6), might specify one or more Target VPNs 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 Target VPNs that, according to its own configuration, are not allowed, and/or to add in some Target VPNs that according to its own configuration are mandatory.

It might be more accurate, if less suggestive, to call this attribute the "Route Target" attribute instead of the "VPN Target" attribute. It really identifies only a set of sites which will be able to use the route, without prejudice to whether those sites constitute what might intuitively be called a VPN.

4.2.2. 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.

If two sites of VPN are in different Autonomous Systems (e.g., because they are connected to different SPs), then a PE router will need to use IBGP to redistribute 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 will then need to use EBGP to redistribute those routes to an ASBR in another AS. This allows one to connect different VPN sites to different Service Providers. However, 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.

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.

When a PE router distributes a VPN-IPv4 route via BGP, it uses its own address as the "BGP next hop". It also assigns and distributes an MPLS label. (Essentially, PE routers distribute not VPN-IPv4 routes, but Labeled VPN-IPv4 routes. Cf. [8]) When the PE processes a received packet that has this label at the top of the stack, the PE will pop the stack, and send the packet directly to the site from to which the route leads. This will usually mean that it just sends the packet to the CE router from which it learned the route. The label may also determine the data link encapsulation.

In most cases, the label assigned by a PE will cause the packet to be sent directly to a CE, and the PE which receives the labeled packet will not look up the packet’s destination address in any forwarding table. However, it is also possible for the PE to assign a label which implicitly identifies a particular forwarding table. In this case, the PE receiving a packet that label would look up the packet’s destination address in one of its forwarding tables. While this can

be very useful in certain circumstances, we do not consider it further in this paper.

Note that the MPLS label that is distributed in this way is only usable if there is a label switched path between the router that installs a route and the BGP next hop of that route. We do not make any assumption about the procedure used to set up that label switched path. It may be set up on a pre-established basis, or it may be set up when a route which would need it is installed. It may be a "best effort" route, or it may be a traffic engineered route. Between a particular PE router and its BGP next hop for a particular route there may be one LSP, 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.

All the usual techniques for using route reflectors [2] to improve scalability, e.g., route reflector hierarchies, are available. If route reflectors are used, there is no need to have any one route reflector know all the VPN-IPv4 routes for all the VPNs supported by the backbone. One can have separate route reflectors, which do not communicate with each other, each of which supports a subset of the total set of VPNs.

If a given PE router is not attached to any of the Target VPNs of a particular route, it should not receive that route; the other PE or route reflector which is distributing routes to it should apply outbound filtering to avoid sending it unnecessary routes. Of course, if a PE router receives a route via BGP, and that PE is not attached to any of the route’s target VPNs, the PE should apply inbound filtering to the route, neither installing nor redistributing it.

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

These distribution rules ensure that there is no one box 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 bound by the capacity of any single device, and therefore can increase virtually without bound.

4.2.3. The VPN of Origin Attribute

A VPN-IPv4 route may be optionally associated with a VPN of Origin attribute. This attribute uniquely identifies a set of sites, and identifies the corresponding route as having come from one of the sites in that set. Typical uses of this attribute might be to

identify the enterprise which owns the site where the route leads, or to identify the site's intranet. However, other uses are also possible. This attribute could be encoded as an extended BGP communities attribute.

In situations in which it is necessary to identify the source of a route, it is this attribute, not the RD, which must be used. This attribute may be used when "constructing" VPNs, as described below.

It might be more accurate, if less suggestive, to call this attribute the "Route Origin" attribute instead of the "VPN of Origin" attribute. It really identifies the route only has having come from one of a particular set of sites, without prejudice as to whether that particular set of sites really constitutes a VPN.

4.2.4. Building VPNs using Target and Origin Attributes

By setting up the Target VPN and VPN of Origin attributes properly, one can construct different kinds of VPNs.

Suppose it is desired to create a Closed User Group (CUG) which contains a particular set of sites. This can be done by creating a particular Target VPN attribute value to represent the CUG. This value then needs to be associated with the per-site forwarding tables for each site in the CUG, and it needs to be associated with every route learned from a site in the CUG. Any route which has this Target VPN attribute will need to be redistributed so that it reaches every PE router attached to one of the sites in the CUG.

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 Target Attribute values, one meaning "Hub" and one meaning "Spoke". Then routes from the spokes could be distributed to the hub, without causing routes from the hub to be distributed to the spokes.

Suppose one has a number of sites which are in an intranet and an extranet, as well as a number of sites which are in the intranet only. Then there may be both intranet and extranet routes which have a Target VPN identifying the entire set of sites. The sites which are to have intranet routes only can filter out all routes with the "wrong" VPN of Origin.

These two attributes allow great flexibility in allowing one to control the distribution of routing information among various sets of sites, which in turn provides great flexibility in constructing VPNs.

5. Forwarding Across the Backbone

If the intermediate routes 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. (Certain procedures for setting up label switched paths in the backbone may not require the presence of the /32 address prefixes.)

When a PE receives a packet from a CE device, it chooses a particular per-site forwarding table in which to look up the packet's destination address. Assume that a match is found.

If the packet is destined for a CE device attached to this same PE, the packet is sent directly to that CE device.

If the packet is not destined for a CE device attached to this same PE, the packet's "BGP Next Hop" is found, as well as the label which that BGP next hop assigned for the packet's destination address. This label is pushed onto the packet's label stack, and becomes the bottom label. Then the PE looks up the IGP route to the BGP Next Hop, and thus determines the IGP next hop, as well as the label assigned to the address of the BGP next hop by the IGP next hop. This label gets pushed on as the packet's top label, and the packet is then forwarded to the IGP next hop. (If the BGP next hop is the same as the IGP next hop, the second label may not need to be pushed on, however.)

At this point, MPLS will carry the packet across the backbone and into the appropriate CE device. That is, all forwarding decisions by P routers and PE routers are now made by means of MPLS, and the packet's IP header is not looked at again until the packet reaches the CE device. The final PE router will pop the last label from the MPLS label stack before sending the packet to the CE device, thus the CE device will just see an ordinary IP packet. (Though see section 8 for some discussion of the case where the CE desires to received labeled packets.)

When a packet enters the backbone from a particular site via a particular PE router, the packet's route is determined by the contents of the forwarding table which that PE router associated with that site. The forwarding tables of the PE router where the packet

leaves the backbone are not relevant. As a result, one may have multiple routes to the same system, where the particular route chosen for a particular packet is based on the site from which the packet enters the backbone.

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.

6. 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. In no case will routes from a site ever be 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. In this case, the site should be a single OSPF area, the CE should be an ABR in that area, and the PE should be an ABR which is not in that area. Also, the PE should report no router links other than those to the CEs which are at the same site. (This technique should be used only in stub VPNs.)

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.)

   From a purely technical perspective, this is by far the best technique:

   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. For example, the CE may suggest a particular Target for each route, from among the Target attributes that the PE is authorized to attach to the route.

   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).

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).

If a set of sites constitute a transit VPN, it is convenient to represent them as a BGP Confederation, so that the internal structure of the VPN is hidden from any router which is not within the VPN. In this case, each site in the VPN would need two BGP connections to the backbone, one which is internal to the confederation and one which is external to it. The usual intra-confederation procedures would have to be slightly modified in order to take account for the fact that the backbone and the sites may have different policies. The backbone is a member of the confederation on one of the connections, but is not a member on the other. These techniques may be useful if the customer for the VPN service is an ISP. This technique allows a customer that is an ISP to obtain VPN backbone service from one of its ISP peers.

(However, if a VPN customer is itself an ISP, and its CE routers support MPLS, a much simpler technique can be used, wherein the ISP is regarded as a stub VPN. See section 8.)

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 certain attributes to the route. There are three such attributes:

- Site of Origin

  This attribute uniquely identifies the site from which the PE router learned the route. All routes learned from a particular site must be assigned the same Site of Origin attribute, even if a site is multiply connected to a single PE, or is connected to multiple PEs. Distinct Site of Origin attributes must be used for distinct sites. This attribute could be encoded as an extended BGP communities attribute (section 4.2.1).

- VPN of Origin

  See section 4.2.1.

- Target VPN

  See section 4.2.1.

7. How CEs learn Routes from PEs

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

In general, a PE may distribute to a CE any route which the PE has placed in the forwarding table which it uses to route packets from that CE. There is one exception: 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.

8. What if the CE Supports MPLS?

In the case where the CE supports MPLS, AND is willing to import the complete set of routes from its VPNs, the PE can distribute to it a label for each such route. When the PE receives a packet from the CE with such a label, it (a) replaces that label with the corresponding label that it learned via BGP, and (b) pushes on a label corresponding to the BGP next hop for the corresponding route.

8.1. Virtual Sites

If the CE/PE route distribution is done via BGP, the CE can use MPLS to support multiple virtual sites. The CE may itself contain a separate forwarding table for each virtual site, which it populates as indicated by the VPN of Origin and Target VPN attributes of the routes it receives from the PE. If the CE receives the full set of routes from the PE, the PE will not need to do any address lookup at all on packets received from the CE. Alternatively, the PE may in some cases be able to distribute to the CE a single (labeled) default route for each VPN. Then when the PE receives a labeled packet from

the CE, it would know which forwarding table to look in; the label placed on the packet by the CE would identify only the virtual site from which the packet is coming.

8.2. Representing an ISP VPN as a Stub VPN

If a particular VPN is actually an ISP, but its CE routers support MPLS, then the VPN can actually be treated as a stub VPN. The CE and PE routers need only exchange routes which are internal to the VPN. The PE router would distribute to the CE router a label for each of these routes. Routers at different sites in the VPN can then become BGP peers. When the CE router looks up a packet's destination address, the routing lookup always resolves 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.

9. Security

Under the following conditions:

a) labeled packets are not accepted by backbone routers from untrusted or unreliable sources, 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

b) labeled VPN-IPv4 routes are not accepted from untrusted or unreliable sources,

the security provided by this architecture is virtually identical to that provided to VPNs by Frame Relay or ATM backbones.

It is worth noting that the use of MPLS makes it much simpler to provide this level of security than would be possible if one attempted to use some form of IP-within-IP tunneling in place of MPLS. It is a simple matter to refuse to accept a labeled packet unless the first of the above conditions applies to it. It is rather more difficult to configure the a router to refuse to accept an IP packet if that packet is an IP-within-IP tunnelled packet which is going to a "wrong" place.

The use of MPLS also allows a VPN to span multiple SPs without depending in any way on the inter-domain distribution of IPv4 routing information.

It is also possible for a VPN user to provide himself with enhanced security by making use of Tunnel Mode IPSEC [5]. This is discussed in the remainder of this section.

9.1. Point-to-Point Security Tunnels between CE Routers

A security-conscious VPN user might want to ensure that some or all of the packets which traverse the backbone are authenticated and/or encrypted. The standard way to obtain this functionality today would be to create a "security tunnel" between every pair of CE routers in a VPN, using IPSEC Tunnel Mode.

However, the procedures described so far do not enable the CE router transmitting a packet to determine the identify of the next CE router that the packet will traverse. Yet that information is required in order to use Tunnel Mode IPSEC. So we must extend those procedures to make this information available.

A way to do this is suggested in [6]. Every VPN-IPv4 route can have an attribute which identifies the next CE router that will be traversed if that route is followed. If this information is provided to all the CE routers in the VPN, standard IPSEC Tunnel Mode can be used.

If the CE and PE are BGP peers, it is natural to present this information as a BGP attribute.

Each CE that is to use IPSEC should also be configured with a set of address prefixes, such that it is prohibited from sending insecure traffic to any of those addresses. This prevents the CE from sending insecure traffic if, for some reason, it fails to obtain the necessary information.

When MPLS is used to carry packets between the two endpoints of an IPSEC tunnel, the IPSEC outer header does not really perform any function. It might be beneficial to develop a form of IPSEC tunnel mode which allows the outer header to be omitted when MPLS is used.

9.2. Multi-Party Security Associations

Instead of setting up a security tunnel between each pair of CE routers, it may be advantageous to set up a single, multiparty security association. In such a security association, all the CE routers which are in a particular VPN would share the same security parameters (.e.g., same secret, same algorithm, etc.). Then the ingress CE wouldn’t have to know which CE is the next one to receive the data, it would only have to know which VPN the data is going to. A CE which is in multiple VPNs could use different security parameters for each one, thus protecting, e.g., intranet packets from being exposed to the extranet.

With such a scheme, standard Tunnel Mode IPSEC could not be used, because there is no way to fill in the IP destination address field of the "outer header". However, when MPLS is used for forwarding, there is no real need for this outer header anyway; the PE router can use MPLS to get a packet to a tunnel endpoint without even knowing the IP address of that endpoint; it only needs to see the IP destination address of the "inner header".

A significant advantage of a scheme like this is that it makes routing changes (in particular, a change of egress CE for a particular address prefix) transparent to the security mechanism. This could be particularly important in the case of multi-provider VPNs, where the need to distribute information about such routing changes simply to support the security mechanisms could result in scalability issues.

Another advantage is that it eliminates the need for the outer IP header, since the MPLS encapsulation performs its role.

10. 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 [10], or, where ATM is used as the backbone, through the use of ATM QoS capabilities. The traffic engineering work discussed in [1] is also directly applicable to MPLS/BGP VPNs. Traffic engineering could even be used to establish LSPs 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 [7] may be useful. An SP may apply either intserv or diffserv capabilities to a particular VPN, as appropriate.

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

Route reflectors and ASBRs can be partitioned among VPNs so that each partition carries routes for only a subset of the VPNs provided by the Service Provider. Thus no single Route Reflector or ASBR is required to maintain routes for all the VPNs.

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.

12. 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.

13. Security Considerations

Security issues are discussed throughout this memo.

14. Acknowledgments

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

15. Authors' Addresses

Eric C. Rosen
Cisco Systems, Inc.
250 Apollo Drive
Chelmsford, MA, 01824

EMail: erosen@cisco.com

Yakov Rekhter
Cisco Systems, Inc.
170 Tasman Drive
San Jose, CA, 95134

EMail: yakov@cisco.com

16. References

[1] Awduche, Berger, Gan, Li, Swallow, and Srinavasan, "Extensions to RSVP for LSP Tunnels", Work in Progress.

[2] Bates, T. and R. Chandrasekaran, "BGP Route Reflection: An alternative to full mesh IBGP", RFC 1966, June 1996.

[3] Bates, T., Chandra, R., Katz, D. and Y. Rekhter, "Multiprotocol Extensions for BGP4", RFC 2283, February 1998.

[4] Gleeson, Heinanen, and Armitage, "A Framework for IP Based Virtual Private Networks", Work in Progress.

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

[6] Li, "CPE based VPNs using MPLS", October 1998, Work in Progress.

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

[8] Rekhter and Rosen, "Carrying Label Information in BGP4", Work in Progress.

[9] Rosen, Viswanathan, and Callon, "Multiprotocol Label Switching Architecture", Work in Progress.

[10] Rosen, Rekhter, Tappan, Farinacci, Fedorkow, Li, and Conta, "MPLS Label Stack Encoding", Work in Progress.

17. Full Copyright Statement