Network Working Group
Internet Draft
Expiration Date: August 2001

Jeremy De Clercq
Yves T'Joens
Olivier Paridaens
Alcatel

Chandru Sargor
Vijay Srinivasan
CoSine Communications

February 2001

BGP/IPsec VPN

<draft-declercq-bgp-ipsec-vpn-01.txt>

Status of this Memo

Abstract

This document describes a method by which a Service Provider may use

an IP backbone to provide VPNs for its customers. In this method IPsec tunnels are deployed through the backbone, and the forwarding of packets over the backbone relies on normal IP forwarding. BGP is

used for distributing (private) routes over the backbone.

This model is based on the model described in RFC 2547 [RFC2547], and the internet-draft that obsoletes it [RFC2547bis]. The main difference is that in this model IPsec is used as tunneling mechanism instead of MPLS.

The purpose of extending on the procedures defined in [RFC2547], is to offer an increased level of security by building upon the authentication and encryption services of IPsec, particularly for interdomain VPN operation, while it has the added benefit that no PE to PE MPLS backbone is required.

Note however that the model does not exclude the use of MPLS in segments of the backbone to improve on traffic engineering and/or QoS aspects.

Table of Contents

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY" and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

1.0 Introduction

Most of the definitions used in this introduction come from the model presented in [RFC2547bis]. For the sake of completeness, we introduce them in this document too.

1.1 Motivation

This document proposes a model to deploy VPNs that extends on the model presented in [RFC2547bis]. The purpose of extending the VPN procedures described in [RFC2547bis] is twofold:

a) The IPsec based VPN model does not require the presence of an MPLS-aware backbone. Thereby allowing a wider deployment of inter-provider backbone VPNs. Note that the usage of MPLS over certain segments of the backbone is not excluded, and could be used for traffic engineering and QoS purposes.

b) The IPsec based VPN model offers stronger security services than the "layer-2" security offered by the model described in [RFC2547bis]. Note that while "layer-2" security might be sufficient for intra-domain VPN operation, it might be short in providing security when building VPNs over multiple adjacent backbones, some of which might not even be VPN aware.

This document proposes an IPsec based VPN model that is easy to combine with [RFC2547bis] and that offers the additional advantages stated before.

Note further that the procedures as described in this document do not require any extensions to the IPsec framework, and as such can make use of an existing implementation base.

1.2 Virtual Private Networks

Like in [RFC2547bis], we consider a set of "sites" which are attached to a common network which we call the "backbone". On this topology we apply some policy to create a number of subsets of that set of sites, and we impose the following rule: two sites may have IP connectivity over that backbone only if at least one of these subsets contains them both.

The subsets that we have created are "Virtual Private Networks" (VPNs). Two sites have IP connectivity over the common backbone only

if there is some VPN that contains them both. Two sites which have no VPN in common, have no connectivity over that backbone.

We consider the case where the owner and operator of the "backbone" is a Service Provider (SP), and where the owners of the "sites" are the customers of that SP. In this document, we discuss mechanisms that may be used to implement the policies to determine whether a set of sites belong to a certain VPN. We don't focus on the question whether it is the SP or the customer that implements these policies, though this model allows for both approaches.

The model presented in this document allows for the deployment of a wide range of policies (leading to different communication topologies: full mesh, hub and spoke, ...).

1.3 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 to one or more Provider Edge (PE) routers. Routers in the Provider's network backbone which do not attach to CE devices are known as "P routers".

The CE device may be a single host, it may be a switch if the considered site is a single subnet, and in general, the CE device may be expected to be a router.

We will say that a PE router is 'attached to a VPN' if it is attached to a CE device that is in that VPN.

When the CE device is a router, it is a routing peer of the PE(s) it is attached to (by means of any routing protocol), but it is NOT a routing peer of the other CE routers in the other sites, even if they are in a common VPN. Routers at different sites do not directly exchange routing information with each other, they even don't have to know of each other's existence at all. Like in the BGP/MPLS VPN model [RFC2547bis], in this document, a VPN is not an "overlay" on top of the SP's network, and a VPN customer does not have to manage a "virtual backbone" in the SP's backbone.

1.4 Multiple Routing and Forwarding Instances in PEs

Every PE router maintains a number of separate forwarding tables. Every site to which the PE is attached must be mapped on one of these forwarding tables. In the scope of this document, we will use the term VRF (VPN Routing and Forwarding Instance) to describe the instances in the PE that do the forwarding and the routing in a

specific context and that contain a specific separate forwarding table.

So this means that every site is mapped to a particular VRF.

When a packet is received by a PE router from a specific site, the VRF associated with that site must be used to handle that packet. The forwarding table in a VRF associated with a particular site must be populated ONLY with routes that lead to sites that have at least a VPN in common with the considered site. This prevents communication between sites that are not in the same VPN.

The way this 'selective population' of routing tables is done, is explained further in this document.

The relationship between sites and VRFs is a one-to-one or a multi-to-one relationship. More than one site can be associated with the same VRF only if they have access to the same set of routes (if they have all their VPNs in common).

A PE router is "attached" to a site when it is the endpoint of an interface or "sub-interface" (PVC, VLAN, etc.) whose other endpoint is a CE device. It is the interface through which the PE received a packet that identifies the VRF to send the packet to.

1.5 VPNs with Overlapping Address Spaces

An important requirement for VPNs is that different VPNs must be able to use overlapping private address spaces. This model allows the usage of overlapping address spaces, for VPNs that do not have sites in common.

The fact that sites in different VPNs are mapped to different VRFs (thus to different routing and forwarding contexts) in the PEs, makes it possible for different VPNs to have overlapping address spaces.

The usage of the IPsec tunnel mode in the backbone network hides the private addresses in that backbone, so that also there all possible ambiguity disappears when using overlapping address spaces.

1.6 VPNs with Different Routes to the Same System

As it is stated in [RFC2547bis], the fact that routes are included independently in the different VRFs makes it possible to introduce (in different VRFs) different routes to the very same system, so that the route to a certain system is dependent of the VRF that handles the packet (i.e. dependent of the origin of the packet). This can be

used to create (more) complex communication topologies.

1.7 SP Backbone Routers

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

The model presented in this document does not impose any VPN knowledge on the P routers, nor does it request the use of e.g. MPLS in the backbone network. The only requirement for the P routers is that they are regular IP routers, and that they maintain /32 addresses for every PE participating in the BGP/IPsec VPN context.

The routing information about a particular VPN is only present in the PE routers that attach to that VPN.

1.8 Security

The model to deploy VPNs described in this document, offers two kinds of security measures. The first Security aspect is offered by the IPsec Security Protocols. The IP traffic sent over the backbone(s) is sent through IPsec tunnels, so that it can be encrypted and authenticated. This allows for the deployment of VPNs over untrusted, not participating backbones. This provides a PE to PE end-to-end security service.

In addition, 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.

1.9 Using IPsec in the backbone

In the model presented in this document, the IP security protocol [RFC2401] is used instead of MPLS to tunnel IP packets through the backbone of the network.

Because there is no concept of "label-stacking" in IPsec, the straightforward way of providing IPsec network-based VPNs would be to deploy a full mesh of Security Associations between the VRFs among the participating PEs. This would cause serious scalability problems and is therefor not applicable for large networks.

The scalability problems arise because:

a) every VRF in a PE needs to maintain Security Associations with every VRF from the peer PEs that are attached to the same VPN(s).

b) the creation of a new VRF in a certain PE requires the creation of

new Security Associations via IKE-exchanges with all the participating PEs.

The model presented in this document provides a way to deploy IPsec network-based VPNs in a scaleable manner. There is only a full mesh of SAs between participating PEs, not between VRFs. The selection of the correct VRF when a packet arrives at the end of the IPsec tunnel (the goal of the second label in the [RFC2547]-model) is based on the Security Parameter Index. This means that a method must be provided to link a pool of SPIs with a single Security Association instead of the usual one-to-one relationship between a SA and a SPI.

This model resolves this by introducing the concepts of an SPI-prefix and an SPI-label.

2.0 Sites and CEs

This document uses the same definitions and imposes the same global behaviour on the sites and the CEs as [RFC2547bis].

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 the use of the backbone.

A CE device is always regarded as being in a single site (though a site may consist of multiple "virtual sites", see later in this section). A site 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 for example) attach to multiple PEs, of the same or of different SPs.

While we use the site as the basic unit of interconnection, the architecture of [RFC2547bis] allows for a finer degree of granularity in the control of interconnectivity. These techniques are also applicable in this model. The customer itself may divide its site into different "virtual sites", each belonging to a different set of VPNs. The PE then needs to contain a separate VRF for each virtual site. The way this can be done is explained in [RFC2547bis].

3.0 VPN Routing and Forwarding Instances

Each PE router maintains one or more "per-site-forwarding tables". As stated before, these are known as VRFs, or "VPN Routing and Forwarding" instances. Every site to which the PE router is attached

is associated with one of these tables. A particular packet's (private) 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.

In a PE router, the following rules apply:

- sub-interfaces may be mapped to VRFs

- the mapping between sub-interfaces (sites) to VRFs is many-to-one or one-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

- two sub-interfaces (sites) 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 (both sub-interfaces "are in the same set of VPNs").

The way by means of which VRFs are populated is explained further in this document.

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

[RFC2547bis] gives two basic methods for providing Internet access over an interface that is associated with a VRF: the VRF may contain a default route which leads to a firewall; or, if no entry in the VRF matches the destination address, the packet's destination address may be matched against the PE's Internet forwarding table.

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 look up the packet's destination address in the appropriate VRF (although in some cases it will need to). The packet may already be carrying enough information (in the form of a VPN-SPI, see section 4.4) to determine the packet's outgoing sub-interface.

4.0 The VPN-SPI

4.1 Introduction

The Security Architecture for the Internet Protocol [RFC2401] defines a Security Association (SA) as a unidirectional "connection" that

affords security services to the traffic carried by it. It is a relationship between two entities, represented by a set of information that can be considered a contract between the entities. A Security Association is identified by a triple consisting of a Security Parameter Index (SPI), an IP Destination Address, and a security protocol (AH or ESP). The SPI is used to differentiate between different Security Associations to the same IP Destination Address, using the same security protocol. The SPI is a pseudo-random 32-bit number for which no formal format has been defined.

4.2 The VPN-SPI used in the IKE Negotiation

This document presently defines two SPI-formats and interpretations to be used in the IKE negotiation phase: a VPN-SPI format (V-flag = 1), and a non-VPN-SPI format (V-flag = 0). The way to interpret the SPI in IKE is dependent on the value of the first SPI bit: the V-flag.

4.2.1 VPN-SPI format, V-flag = 1

It is assumed in this document that the peer PE that receives the packets through an IPsec tunnel identified by a certain SPI, has chosen this SPI associated with the considered tunnel during the IKE-negotiation.

Provider Edge Routers that use IKE to establish IPsec tunnels between them to be used in the BGP/IPsec VPN context, MUST interpret the negotiated SPIs according to the format defined in this document.

This model assumes that for the inbound ("inbound" is used to denote the process of handling packets coming out of the IPsec tunnel) IPsec SA selection, the following identifiers are used: the Destination IP address of the outer IP header, the Security Protocol (AH or ESP) in the security header of the IPsec packet, the SPI, and eventually the Source IP address of the outer IP header. Although the use of the Source IP Address in the outer IP header during the inbound SA selection process is not standardized, it is recommended to do so when using this model.

In the model described by this document, every PE should define at least one "SPI-prefix" per participating peer PE. If the source IP address in the outer IP header of an incoming IPsec packet can not be used to select the appropriate SA (next to the destination IP address, the security protocol and the SPI), then these SPI-prefixes must be different for every peer PE. But if the source IP address in the outer IP header of an incoming IPsec packet can be used to select the appropriate SA, then platform-wide SPI-prefixes can be assigned (this means the same SPI-prefix for every peer PE).

A reason to assign more than one SPI-prefix to a certain PE, could be that two PEs could maintain multiple SAs, because some VPN traffic needs stronger protection than other VPN traffic.

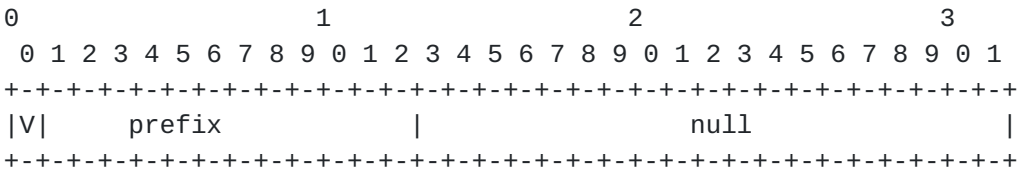
This means that every PE must do the following things (independently of the other PEs):

- if the Source IP address of the outer IP header can not be used by the inbound IPsec process in the PE for the SA selection process: associate at least one SPI-prefix with every peer PE. These SPI-prefixes must be unique within the context of a PE (each one identifies a peer PE).

- if the Source IP address of the outer IP header can be used by the inbound IPsec process in the PE: associate at least one SPI-prefix with every peer PE. These SPI-prefixes must not be unique.

In the regular IKE specifications, the SPI is defined as a pseudo-randomly generated number. This document imposes a formal format on the SPI used in the IKE negotiation under the conditions applicable in this section of the document. This allows the negotiating peers to interpret the SPIs as belonging to a BGP/IPsec VPN-environment, and to negotiate about SPI-pools instead of about single SPIs, so that multiple SPIs can be associated with a single Security Association.

The formal format for the VPN-SPI used in IKE-negotiations is the following:



V-flag:

This flag is used to differentiate IKE-negotiations about IPsec tunnels to be interpreted in the BGP/IPsec VPN context (V = 1) or in another context (V = 0). In the VPN-SPI context, this flag MUST be set to 1.

prefix:

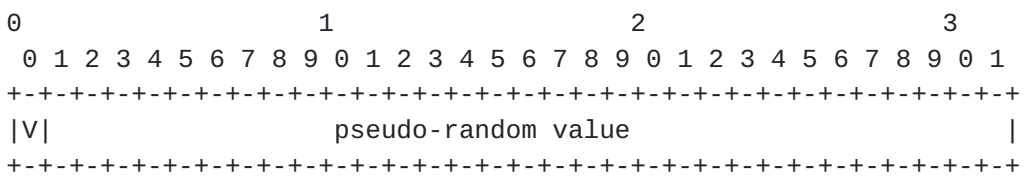
This 12-bit field contains the SPI prefix.

null:

These 19 bits must be set to 0.

The length of the VPN-SPI prefix is 12 bits because this allows to use a 20-bit MPLS label as the VPN-SPI label. The use of a smaller prefix-length could open the door for possible denial-of-service attacks.

4.2.2 non-VPN-SPI format, V-flag = 0



The SPI format used in the IKE negotiation in the non-VPN-SPI context does differ from the normal SPI format (a pseudo-random 32-bit number) on only one point: the first bit MUST be set to 0.

V-flag:

this flag is used to differentiate negotiations about IPsec tunnels to be interpreted in the BGP/IPsec VPN context (V = 1) from other contexts (V = 0). In the non-VPN-SPI context, this flag MUST be set to 0.

pseudo-random value:

this 31-bit field contains a pseudo-random number.

4.2.3 Use of the VPN-SPI during the IKE negotiation

During a normal IKE negotiation, two SAs are being set-up. The SPIs identifying these SAs are chosen by the receiving party.

Every PE participating in the BGP/IPsec VPN scenario MUST define a VPN-SPI prefix per peer PE.

The receiving PE (i.e. the PE that will receive the IPsec packets) must now form 32-bit VPN-SPIs as described in section 4.2.1 for the BGP/IPsec VPN SA negotiation via IKE with every peer PE. To create these VPN-SPIs, the PE sets the V-bit to 1 and inserts the correct SPI-prefix in the correct field. A particular VPN-SPI must be used in all the IKE-exchanges with the appropriate peer PE in the BGP/IPsec VPN context.

As a result of these IKE exchanges, every PE has at least 1 inbound SA with every other PE for the BGP/IPsec VPN context. These SAs are

identified by:

- the destination IP address

- the security protocol (AH or ESP)

- the V-bit of the VPN-SPI

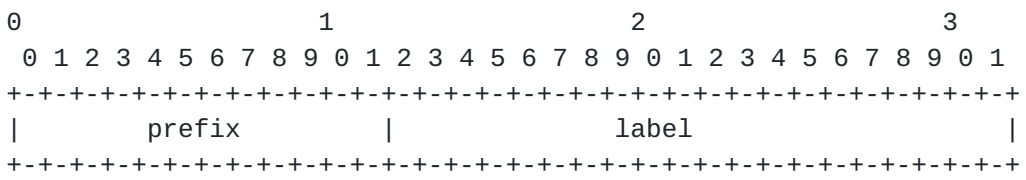
- the unique (i.e. different for every peer PE) SPI-prefix when the Source IP Addresses are not used

- the SPI-prefix and the Source IP address when the Source address may be used

4.3 The VPN-SPI in the IPsec processing

4.3.1 Format and interpretation of the VPN-SPI in the IPsec processing

The SPI that will finally been inserted in the security headers of the IPsec packets has the following format:



prefix:

This 12-bit field identifies the SA to associate an IPsec packet with.

label:

The length of the label field is the length of an ordinary MPLS label : 20 bits. This part of the VPN-SPI is not used to find the correct SA during inbound processing, but it is used to identify the correct outgoing interface or VRF after the IPsec processing in the egress PE (see later in this document).

Note that every PE must take care not to assign SPI's for other IPsec contexts (than the BGP/IPsec VPN context) that might be identical to the combination of any distributed VPN-SPI prefix with any VPN-SPI label.

4.3.2 Outbound IPsec processing

As described in [RFC2547bis], an IP packet coming from a customer

site will be handled in a dedicated VPN Routing and Forwarding instance in the considered PE. The choice of the VRF is based on the packet's incoming sub-interface. In that VRF, a routing lookup will be done, based on the (private) destination address in the packet's IP header. As a result of that lookup, the information associated with a particular packet is: the next hop (PE) and the outgoing sub-interface to send the packet to, and a specific (SPI-)label to associate with that packet (assigned to that route by the next hop PE = BGP next hop). The mechanism by means of which the routes in the VRFs are associated with the correct SPI-labels is explained elsewhere in this document (section 5.0). If the outgoing sub-interface is associated with a VRF, then the next hop is a CE device attached to the same PE. The packet is then directly sent to that outgoing interface (although a new lookup in a VRF is also possible).

If the next hop PE is another PE, then the outgoing interface is an interface to the backbone, and the packet must be IPsec processed. Within a PE, the VRF that handles the considered packet, together with the 'next hop PE' found after the lookup (the 'selectors'), and eventually the SPI-label, uniquely identify a SA (= a certain security association for the BGP/IPsec VPN context between this PE and the next hop-PE).

The packet will now be processed according to that SA and the packet will be sent over the IPsec tunnel to the next hop PE, using the appropriate VPN-SPI (constructed using the SPI-label found after the lookup in the VRF and using the SPI-prefix associated with the considered SA) in the SPI-field of the security header of the IPsec packet.

4.3.3 Inbound IPsec processing

When a PE receives an IPsec packet from the core network that has it's own IP address in the destination IP address field of the outer IP header, the correct SA must be identified. This SA is identified by means of: the Destination IP address (it's own IP address) in the outer IP header, the security protocol (AH or ESP), the SPI (in fact only the prefix-part of the SPI can be enough to identify a BGP/IPsec VPN SA), and eventually the Source IP address of the outer IP header.

Once the correct SA is identified, the packet will be handled according to the IPsec inbound processing rules: decryption, authentication, etc. The result of this is a regular IP packet with -eventually- a private IP address in the IP header.

4.4 Use of the VPN-SPI after the inbound IPsec processing

Now, instead of routing this IP packet according to the global IP

Routing table of the PE (which is not possible because of the use of private addresses), the label-part of the VPN-SPI (that the considered PE distributed with the VPN addresses via BGP) found in the security header of the IPsec packet is used to direct the packet immediately to the correct customer-side interface or to the correct VRF for further processing in the right private address context. The label-part of the VPN-SPI has the same role as the first label in the BGP/MPLS VPN model [RFC2547bis].

4.5 Achievement

The introduction of the VPN-SPI does not affect the real IPsec processing. The SPI still identifies a SA. Also the IKE-mechanism is not changed radically: it is still two peers negotiating SAs, and assigning SPIs to them. The only difference is that some structure in these SPIs must be recognized. By introducing this structure to the SPI, and imposing the use of this structure on the participating PEs, two goals have been achieved:

- IKE is able to negotiate about SPI-pools, so that multiple SPIs can point to the same SA.

- the SPI (more precisely, a part of the SPI) can be used as a label to identify the correct VPN context to process certain packets in (a VRF), or the correct outgoing interface to send the packet to.

5.0 VPN Route Distribution via BGP

The distribution over the backbone of the (private) routes to the different sites participating in the BGP/IPsec VPN model, uses the same conceptual model as the BGP/MPLS VPN model [RFC2547bis]: PE routers use BGP to distribute VPN routes to each other.

We allow each VPN to have its own address space, which means that a given address may denote different physical systems in different VPNs (concept of 'private addresses'). If two routes, to the same IP address prefix, are actually routes leading to separate systems, we must make sure that BGP treats them as two different routes. Otherwise BGP might choose to install only one of them, making the other system unreachable. Further, we must make sure that policy is used to determine which packets get sent on which routes. Given that several routes are installed by BGP, only one of them may appear in a particular VRF.

These goals are met by the use of the VPN-IPv4 address family and by the use of the Route Target attribute.

5.1 The VPN-IPv4 Address Family

The BGP Multiprotocol extensions [BGP-MP] allow BGP to carry routes from multiple "address families". [RFC2547] introduces the notion of "VPN-IPv4 address family". The model described in this document uses the same address family (but does not use the "labeled"-version of it). A VPN-IPv4 address is a 12-octet quantity, beginning with an 8-octet "Route Distinguisher" (RD), and ending with a 4-octet IPv4 address. If two VPNs use the same IPv4 address prefix for a different system, the PEs transform these into two different VPN-IPv4 address prefixes (using two different RDs).

For the possible other uses of the RD and the structure and encoding of the RD, we refer to [RFC2547bis].

5.2 Controlling Route Distribution

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

5.2.1 The Route Target Attribute

The BGP/MPLS VPN model [RFC2547bis] introduces the concept of "Route Target" attributes. These BGP attributes are encoded as BGP Extended Community Route Targets [BGP-EXTCOMM].

Every VRF in a PE is associated with one or more Route Target attributes (= the "import" Route Target attributes). Every site attached to a PE is also associated with one or more Route Target attributes (= the "export" Route Target attributes). These Export Route Target attributes may be associated per route, per site or per VRF (thus possibly associated with more than one site, as more than one site may be served by the same VRF). The two sets of Route Target attributes need not be identical, they are distinct.

When a PE learns a customer-route from one of his attached CEs, the PE creates a VPN-IPv4 route to distribute with BGP. The PE then associates one or more "Route Target" attributes with that route (the "export" Route Targets associated with the considered site). These Route Targets are carried in BGP as attributes of the route.

Any route associated with a certain Route Target T must be distributed to every PE router that has at least one VRF associated with Route Target T (an "import" Route Target of that VRF). When such a route (carrying a Route Target attribute T) is received by a PE router, it is eligible to be installed only in those VRFs that are associated with an "import" Route Target T. Whether the route actually gets installed is dependent on the outcome of the BGP decision process.

A Route Target Attribute can be thought of as identifying a set of sites or a set of VRFs. It is used to filter the appropriate routes into the correct VRFs.

Several methods can be used to associate routes with Route Target attributes: a PE can be configured to associate every route coming from a certain site with a set of Route Targets; or to associate some routes with one set of Route Targets and some routes with another set; or alternatively, the control could be shifted to the CE: the CE could specify every route it advertises to the PE with one or more Route Targets.

5.2.2 Route Distribution among PEs by BGP

If two sites of a VPN attach to PEs that are in the same Autonomous System, these PEs can distribute VPN-IPv4 routes to each other by means of an IBGP connection between them. The way it is done for PEs that do not belong to the same Autonomous System is explained later in this document (section 10).

Like in the model described in [RFC2547bis], the use of route reflectors [BGP-RR] is strongly recommended in order to scale the number of BGP connections. The model described here allows for the use of all the route reflector techniques to improve scalability. As it is explained in [RFC2547bis], the set of VPN-IPv4 routes may be partitioned among a set of route reflectors.

When a PE router distributes a VPN-IPv4 route via BGP, it uses its own address as the "BGP next hop". As BGP must use only one kind of Address Family ([BGP-MP]), this address is encoded as a VPN-IPv4 address with a RD of 0. In addition, the PE assigns an appropriate (SPI-)label Attribute (and a set of Route Target Attributes, see later) to the route and distributes it. This (SPI-)label identifies the destination within the PE (a certain customer-interface or a certain VRF) of the packets coming from the backbone and following the considered route.

When the PE processes a packet received from the core, it goes through the following steps:

- identify the correct SA by means of the SPI, the destination address and the security protocol.

- process the packet according to the IPsec process defined by the SA.

- (if the SPI is a VPN-SPI) use the label-part of the VPN-SPI in the security header of the IPsec packet (more precisely compare it to the

SPI-labels distributed via BGP in the SPI-label attribute) to direct the packet to the correct outgoing interface or to a certain VRF for further processing.

The use of the BGP refresh mechanism [BGP-RFSH] and the outbound route filtering mechanism [BGP-ORF] is strongly recommended to assure maximum scalability of the model.

In the BGP-point of view, the model described in this document has the same advantages as the model described in [RFC2547bis]: a PE router should not install VPN-IPv4 routes belonging to VPNs it is not attached to; a router which is not attached to any VPN (a P router), never installs any VPN-IPv4 routes at all. This also means that there is no box that needs to know all the VPN-IPv4 routes that are supported over the backbone.

5.2.3 How VPN-IPv4 NLRI is Carried by BGP

VPN-IPv4 NLRI is carried by BGP in the same way as in [RFC2547bis]. Labelled addresses are used.

5.2.4 Building VPNs using Route Targets

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

[RFC2547bis] gives two examples: a fully meshed closed user group (i.e. a set of sites where each can send traffic directly to the other, but where no communication is possible with other non-member sites), and a hub and spoke model (i.e. a communication topology where all the traffic between sites (the "spoke sites") must go through a central site (the "hub site").

To form a fully meshed closed user group for example, a single Route Target is needed. That Route Target is assigned to the VRFs associated with the participating sites, as both the Import and the Export Route Target.

The method for controlling the distribution of routing information among various sets of sites are very flexible. This provides great flexibility in constructing VPNs.

6.0 Forwarding Across the Backbone

As PE to PE IPsec tunnels are deployed across the backbone, the forwarding in the backbone is based on regular IP forwarding, using the destination addresses in the outer IP-headers. These outer IP

headers contain no VPN information. The addresses included in the outer IP headers are PE global IP addresses. This means that no VPN awareness is needed in the backbone at all, and that all the forwarding relies on regular IP, so no non-IP tunneling mechanisms are needed (such as MPLS). The only requirement is that PE routers need to insert /32 address prefixes for themselves (the IPsec tunnel endpoints) into the IGP routing tables of the backbone.

7.0 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.

If the CE device is a switch or a host, the set of addresses will generally be configured into the PE router. If the CE is a user dialing in, it will usually receive a temporary IP address from the PE. In the case where the CE is a router, there are a number of possible ways that a PE router can obtain this set of addresses.

The PE translates these private IPv4 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.

We can imagine a lot of route distribution techniques from CE to PE. However, the distinction must be made between CEs in a "transit VPN" (= a VPN that contains a router that receives routes from another, non-PE router that is not in the VPN) and CEs in a "stub VPN" (= a VPN without "third party" routing exchanges).

The possible PE/CE distribution techniques are: static routing, RIP, OSPF, EBGP, etc.

[RFC2547bis] gives a more detailed overview of the possible CE/PE route distribution scenario's.

Once the CE-routes are learned by the PE, it distributes the resulting VPN-IPv4 routes via BGP. These routes are associated with the following attributes: an SPI-label attribute, one or more Route Target attributes and eventually a Site of Origin attribute.

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.

8.0 How CEs Learn Routes from PEs

In this section, it is assumed that the CE device is a router.

If the PE places a particular route in the VRF associated with a certain site, 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. Note that whatever procedure is used to distribute routes from CE to PE will also be used to distribute routes from PE to CE.

One more restriction is added 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 in that site.

Note also that in most cases, it will be sufficient for the PE to simply distribute the default route to the CE.

9.0 Inter-Provider Backbones

A usual requirement for VPNs is that VPNs must be able to span across multiple backbones. This allows sites that are connected to different SPs to 'be in the same VPN'. This requirement introduces two main issues:

- the PE routers participating in the VPN topology are not able to establish IBGP connections with each other or with a common route reflector

- the security aspect becomes an important issue

The latter issue is covered by the use of the PE-PE end to end IPsec tunnels. For the first issue, [RFC2547bis] discusses 3 different solutions.

The first two solutions ('VRF-to-VRF connections at the AS border routers' and 'EBGP redistribution of labeled VPN IPv4 routes from AS to neighboring AS') are not applicable in the model presented in this document, because this model requires PE-PE end-to-end IPsec tunnels.

The third solution is perfectly applicable: multihop EBGP redistribution of VPN-IPv4 routes (associated with VPN-SPI attributes) between source and destination ASs. The participating PEs still need to set up IPsec tunnels with each other (whether they are in the same AS or not) via IKE. The /32 routes to all the participating PEs must be known in all the participating ASs (PE and

P routers) to allow for normal end-to-end IP forwarding.

Now PE routers in different ASs can establish multi-hop EBGP connections to each other, and can exchange VPN-IPv4 routes over these connections.

To improve scalability, one can have multi-hop EBGP connections only between a route reflector in one AS and an other route reflector in an other AS. Care must then be taken that these route reflectors do not change the BGP next hop attribute of the routes).

10.0 Use of an MPLS backbone

The model presented in this document does not in any way preclude the existence of an MPLS core network. An MPLS network can carry IPsec packets as easily as it can carry IP packets. This means that if an MPLS network is present in the backbone of the ISP network, all the extra functionalities that MPLS offers (Traffic Engineering, QoS, etc.) can still be used in the core of this BGP/IPsec VPN architecture.

11.0 Security

The model described in this document offers a higher level of security than [RFC2547bis]. The IPsec security mechanisms protect the IP packets when traversing the backbone(s). This is an ingress PE to egress PE end-to-end IPsec protection. Especially in the case when non-participating ('transit') SPs are traversed, this is an important requirement.

In addition, by introducing the VPN-SPI concept and formats, this architecture in itself provides a security level that is virtually identical to a 'layer-2' mechanism in the scope of an individual PE: the PE can decide to accept or reject an IP packet based on the SPI included in the security header of the IPsec packet, and the directing of the packets within the PE relies on the label-part of the VPN-SPI. If no misconfiguration occurs, the traffic from one VPN is perfectly shielded from the traffic in another VPN within the same PE.

12.0 Scalability

The model proposed in this document uses IPsec (tunnel mode) to tunnel the (private address) IPv4 packets through the shared backbone(s). As the model requires only one IPsec tunnel between every two PEs (and not a full mesh between sites or between VRFs), the solution remains scalable for large topologies.

All the scalability considerations that apply for [RFC2547bis] also apply for the model described in this document.

P routers (which are neither PE routers nor Route Reflectors) do not maintain any VPN routes. They only need to maintain global IP routes to all the participating PEs.

PE routers maintain VPN routes only for the VPNs they are attached to.

Route Reflectors can be partitioned among VPNs so that each partition carries only routes for a subset of the VPNs supported by the Service Provider.

These remarks apply also for the inter-provider VPNs, if multi-hop EBGP is used.

As a result, no single component within the SP network has to maintain all the routes for all the VPNs. This means that the support of an increasing number of VPNs is not limited by the capacity of an individual component.

13.0 References

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

[RFC2547] Rosen, E. and Rekhter, Y., "BGP/MPLS VPN", RFC 2547, March 1999.

[RFC2547bis] Rosen E., Rekhter Y., et al., "BGP/MPLS VPN", Work in Progress.

[RFC2401] Kent, S. and Atkinson R., "Security Architecture for the Internet Protocol", RFC 2401, November 1998.

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

[BGP-EXTCOMM] Ramachandra, S. and Tappan, D., "BGP Extended Communities Attribute", Work in Progress.

[BGP-MP] Bates, T., et al., "Multiprotocol Extensions for BGP4", RFC 2283, February 1998.

[BGP-RFSH] Chen, E., "Route Refresh Capability for BGP4", Work in Progress.

[BGP-ORF] Chen, E. and Rekhter, Y., "Cooperative Route Filtering Capability for BGP-4", Work in Progress.

14.0 Acknowledgements

The model presented in this document is based on a lot of ideas presented in [RFC2547bis]. We would like therefor to thank all the authors of "BGP/MPLS VPN" for their work that has been the basis for the ideas presented in this document. We also would like to thank Peter De Schrijver for his contribution to this draft.

15.0 Authors' Addresses

Jeremy De Clercq
Alcatel
Francis Wellesplein 1
2018 Antwerpen, Belgium
Phone: +32 3 240 4752
Email: jeremy.de_clercq@alcatel.be

Yves T'joens
Alcatel
Francis Wellesplein 1
2018 Antwerpen, Belgium
Phone: +32 3 240 7890
Email: yves.tjoens@alcatel.be

Olivier Paridaens
Alcatel
Francis Wellesplein 1
2018 Antwerpen, Belgium
Phone: +32 3 240 9320
Email: olivier.paridaens@alcatel.be

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

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