Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: December 13, 2018

E. Rosen, Ed.
R. Bonica
Juniper Networks, Inc.

June 11, 2018

Augmenting RFC 4364 Technology to Provide Secure Layer L3VPNs over Public Infrastructure
draft-rosen-bess-secure-l3vpn-00

Abstract

The Layer 3 Virtual Private Network (VPN) technology described in RFC 4364 is focused on the scenario in which a network Service Provider (SP) maintains a secure backbone network and offers VPN service over that network to its customers. Customers access the SP's network by attaching "Customer Edge" (CE) routers to "Provider Edge" (PE) routers, and exchanging cleartext IP packets. PE routers generally serve multiple customers, and prevent unauthorized communication among customers. Customer data sent across the backbone (from one PE to another) is encapsulated in MPLS, using an MPLS label to associate a given packet with a given customer. The labeled packets are then sent across the backbone network in the clear, using MPLS transport. However, many customers want a VPN service that is secure enough to run over the public Internet, and which does not require them to send cleartext IP packets to a service provider. Often they want to connect directly to edge nodes of the public Internet, which does not provide MPLS support. Each customer may itself have multiple tenants who are not allowed to intercommunicate with each other freely. In this case, the customer many need to provide a VPN service for the tenants. This document describes a way in which this can be achieved using the technology of RFC 4364. The functionality assigned therein to a PE router can be placed instead in Customer Premises Equipment. This functionality can be augmented by transmitting MPLS packets through IPsec Security Associations. The BGP control plane sessions can also be protected by IPsec. This allows a customer to use RFC 4364 technology to provide VPN service to its internal departments, while sending only IPsec-protected packets to the Internet or other backbone network, and eliminating the need for MPLS transport in the backbone.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

1.1. Review of L3VPN Concepts and Terminology

In conventional Virtual Private Networks (L3VPNs) based on the technology of [RFC4364], a Service Provider (SP) maintains a secure private network (known as the "SP backbone"). An SP maintains a number of "Provider Edge" (PE) routers to which customers may attach. A customer router that attaches to a PE router is known as a "Customer Edge" (CE) router.

Multiple customers may connect to a single PE router. Within a given PE, each customer is associated with a routing context of its own (known as a Virtual Routing and Forwarding table, or VRF). A particular customer attaches to the PE via a set of one or more interfaces or Virtual LANs (VLANs) that are not shared with other customers. (In the subsequent text, the term "interface" will include VLANs and other "virtual" interfaces.) Each such interface is associated with a particular customer's VRF; thus such interfaces are known as "VRF interfaces". These are the PE's customer-facing interfaces. The VRF interfaces carry IP datagrams, either IPv4 or IPv6 or both.

A given customer's VRF is automatically populated with, and only with, routes that lead out the local VRF interfaces, and routes that lead to remote VRF interfaces of the same customer. Routes leading outside a customer's VPN are excluded from that customer's VRF unless explicitly allowed by policy. Thus two customers can attach to the same PE even if they are not allow to communicate with each other through that PE.

The PE at which a customer data packet enters the SP backbone network is known as the packet's "ingress PE". The PE at which a customer data packet leaves the SP backbone is known as the packet's "egress

PE". Generally, the ingress PE pushes two MPLS labels onto each data packet. The top label (sometimes known as the "transport label") directs the packet to its egress PE. The second label (sometimes known as the "VPN label") is used at the egress PE to associate a given customer's packets with that customer's VRF at the egress PE.

These labeled packets travel across the SP backbone "in the clear" (i.e., with no cryptographic protection to provide privacy, authentication, or integrity), as the SP backbone is presumed to be adequately secure.

The control plane protocol for this type of VPN is BGP. A given customer's routes are distributed among the PEs to which that customer attaches by means of a BGP address family known as "VPN-IP" (either VPN-IPv4 or VPN-IPv6). Distribution of these routes is controlled in such a way as to ensure that a given customer's routes, exported from one of that customer's VRFs, are imported only by other VRFs associated with the same customer.

1.2. Secured L3VPN

For security reasons, the L3VPN technology summarized in Section 1.1 is not generally used in the following scenarios:

o Some or all of the customer sites need to be reached over the public Internet, rather than over a secure SP backbone network.

o The customer does not want to expose any of his data in cleartext to the SP, even if the SP backbone network is secure.

o The customer does not want to expose any of his routing control information in cleartext to the SP, and/or wishes to hide his internal IP addressing structure from the SP.

In such situations, the customer needs to use cryptographic methods in order to ensure privacy, integrity, and authentication for the IP datagrams he sends over the backbone network; the cryptography must be applied before the datagrams are sent to the SP backbone network or Internet. (It is presumed of course that the customer's own sites and systems have been secured to his satisfaction; how that is achieved is outside the scope of this document.)

In these use cases, the customer may still want the benefits of the L3VPN service, e.g.:

- The customer may itself be providing a VPN service to multiple "tenants". E.g.,

    * The customer may be an enterprise or governmental agency that consists of multiple internal departments or organizations that are not allowed to communicate freely with each other, and that may even have independent IP address spaces. We will use the term "tenant" to refer to such a department or organization.

    * The customer may be a Data Center operator that is providing a virtual network to each of multiple Data Center tenants.

- In L3VPN, a CE router at one customer site does not have to be provisioned with the addresses of CE routers at other sites.

  Rather, these are auto-discovered via BGP. This sort of auto-discovery is just as valuable when the customer needs more security than is provided by conventional L3VPN. Auto-discovery also allows some or all of the CE routers to be mobile, changing their IP addresses from time to time.

It is possible to adapt the L3VPN technology to handle use cases where cryptographic methods must be applied before a packet is sent to an SP or to a backbone network . This document describes a way in which this may be done. We will refer to this adaptation as a "Secured L3VPN". Section 2 outlines the way this adaptation works. Subsequent sections of this document specify the necessary procedures in more detail.

Secured L3VPN makes use of IPsec technology. This document does not discuss the details of IPsec. A roadmap through the set of RFCs describing IPsec can be found in [RFC6071]. Of particular importance are [RFC4303] (IPsec Encapsulating Security Payload), [RFC7296] (Internet Key Exchange Protocol version 2), and [RFC8221] (Cryptographic Algorithm Implementation Requirements and Usage Guidance).

1.3. Terminology

In this document we shall use the following terminology:

SP

A network service provider (possibly an Internet service provider).

customer

An organization or other entity that obtains network service (private network service or Internet service) from an SP.

tenant

An organization or other entity that obtains VPN service from the customer. For example, if the customer is a governmental agency, its tenants might be the various departments of the agency. If the customer is an enterprise, its tenants might be the various organizations within the enterprise. If the customer is a Data Center provider, its tenants might be organizations to which it sells Data Center services.

C-PE router:

A router that performs the functions of an L3VPN PE router ([RFC4364]), but that is operated and managed not by a network service provider, but rather by a customer of the network service provider. The customer may use the C-PE to provide a Secured L3VPN service to one or more of its tenants.

Red interface:

A tenant-facing interface of a C-PE device, where the tenant in question is receiving Secured L3VPN service.

Black interface:

A C-PE interface that is not a red interface. This may be a network-facing interface, or a tenant-facing interface to a tenant that is not receiving any L3VPN service.

Red BGP session:

A BGP session, protected by IPsec, between two C-PEs, or between a C-PE and a BGP Route Reflector.

Black BGP session:

A BGP session other than a red BGP session.

Local red route:

A route whose next hop interface is a local red interface.

Remote red route:

A route received, as a VPN-IPv4 or VPN-IPv6 route, over a red BGP session.

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

2. Model of Operation

In a Secured L3VPN, the functions conventionally performed by an L3VPN PE router (as detailed in [RFC4364]) are instead performed by a router that is operated and managed by the customer, rather than by the SP. Since such a router is part of the customer's network, but has the functionality of an L3VPN PE router, we will refer to it as a

"C-PE router". The customer is responsible for ensuring that the C-PE itself is properly secured. The C-PE provides L3VPN functionality to the customer's tenants.

Each interface of a C-PE is either a "red interface" or a "black interface":

- A red interface is a tenant-facing interface that attaches to a tenant who is receiving Secured L3VPN service from the customer.

- A black interface is any interface that is not a red interface. Black interfaces may be network-facing interfaces (attached to an SP backbone), or may be tenant-facing interfaces attached to tenants that are not receiving any L3VPN from the customer. We assume in this document that a C-PE that provides the Secured L3VPN service to one or more tenants does not provide a conventional (unsecured) L3VPN service to any of the tenants.

A C-PE has one or more VRFs, one per tenant. Each VRF is associated with a distinct set of red interfaces, the ones that lead to the network(s), VLAN(s), or virtual network(s) that is (are) specific to the given tenant. Standard L3VPN techniques then prevent communication among the different tenants unless explicitly allowed by policy. In simpler scenarios, the customer may have sites with only a single tenant. The C-PEs at those sites require only a single VRF, and all the red interfaces will be associated with that VRF.

The black interfaces of a C-PE can attach to an access router of the public internet, or to a conventional L3VPN PE router belonging to an SP, or to any other router that provides IP connectivity among the customer's C-PE routers. (If a C-PE attaches to a conventional L3VPN PE router, then the C-PE appears to the conventional PE to be a CE router.)

As in any L3VPN, the VRFs are populated with a combination of local routes and remote routes:

- The local routes in a given VRF are those routes whose "next hop interface" is a local red interface associated with that VRF.

- The remote routes in a given VRF are those routes learned via BGP from the customer's other C-PEs. These routes may be learned directly via BGP sessions to those other C-PEs, or indirectly via one or more BGP Route Reflectors (RRs).

In this document, we will use the term "red routes" to refer to routes within a VRF. These are distinguished from the "black routes" that exist in a C-PE's global routing table.

A conventional PE router sends and receives MPLS packets over its network facing interfaces. A C-PE, on the other hand, sends "MPLS-in-IPsec" packets (see [RFC4023] and Section 5 of this document) over its black interfaces. Since an MPLS-in-IPsec packet is an IP datagram, there is no need for the backbone network to support MPLS transport. IPsec is used to provide privacy, integrity and authentication for the packets sent by the C-PE to the backbone network.

In a Secured L3VPN, protection of the control plane is just as important as is protection of the data plane. It is therefore necessary to ensure that the BGP messages used to disseminate the red routes also have privacy, integrity, and authentication. In order to ensure this, the BGP sessions used to disseminate information about red routes will be protected by IPsec. We will refer to such BGP sessions as "red BGP sessions". It is recommended to use IPsec Transport mode to protect these BGP sessions. This means that a C-PE MUST NOT send or receive VPN-IP routes over any BGP session that is not protected by IPsec. (A VPN-IP route is a route whose BGP Address Family (AFI) is 1 (IPv4) or 2 (IPv6) and whose Subsequent Address Family (SAFI) is 128, 129, or 5.)

Note that if RRs are used, the RRs must be as secure as the C-PEs. This likely means that they are managed by the customer and located at sites regarded by the customer as adequately secure.

Thus in a Secured L3VPN, red routes are propagated only among trusted systems, and always in red BGP sessions. The propagation of red routes on red BGP sessions is controlled by attaching Route Targets to those routes, as with any [RFC4364]-based technology.

As in any L3VPN, BGP uses the VPN-IPv4 and/or VPN-IPv6 address families when disseminating information about VPN routes from one VRF to another. Each such route carries an MPLS label that is to be pushed on the label stack of any tenant packet for which the address prefix in the route's NLRI is the best match to the packet's IP destination address.

In Secured L3VPNs, these routes MUST also carry a Tunnel Encapsulation attribute ([tunnel_encaps]) specifying an "MPLS-in-IPsec" tunnel type (see Section 5, as well as Sections 3 and 8.1 of [RFC4023]). This indicates that before a tenant's MPLS packet is sent to the backbone network, it must be encapsulated in IP and then sent on an IPsec Transport Mode Security Association (SA).

A C-PE may have unprotected (black) BGP sessions, e.g., to gather public Internet routes. However, the black BGP sessions MUST NOT be enabled for the VPN-IP AFI/SAFIs. This prevents any routes learned over the black BGP sessions from being imported into the VRFs.

As we shall see in Section 3.3, there are also some IP routes (as distinguished from VPN-IP routes) that MUST NOT be transmitted on black BGP sessions, and that MUST be ignored if received on black BGP sessions. These are known as the "red loopback routes".

The procedures of this document result in a network overlay whose control plane consists of red BGP sessions, and whose data plane consists of MPLS-in-IPsec Security Associations. This allows an SP's customer to provide Secured L3VPN service to its tenants.

When RRs are used, C-PEs "register" with the RRs by setting up BGP sessions to them, running the BGP sessions through IPsec SAs. The necessary authentication of the C-PE is provided during the course of setting up the IPsec SA. One C-PE learns of another other C-PE's presence when the RR propagates routes from the latter C-PE to the former.

The procedures specified in this document result in one MPLS-in-IPsec SA between a given pair of C-PEs. This one SA will carry the traffic of all the tenants that are attached to both C-PEs. That should provide adequate security, as the tenants' data is already exposed to the C-PEs. If for some reason it is desired to have a distinct SA for each tenant, a method of doing so is mentioned in Section 4.

3. How the C-PEs Advertise Red Routes

3.1. Red and Black C-PE Loopback Addresses

To support the Secured L3VPN control plane, each C-PE MUST have two loopback addresses. One of these will be known as its "red loopback", the other as its "black loopback".

The black loopbacks MUST be addresses that are globally routable. That is, they are public addresses. (Strictly speaking, the black loopback only needs to be routable in any network that might be used to carry traffic between two C-PEs. But we will assume that traffic between two C-PEs might need to traverse the public Internet.) Typically a C-PE's black loopback will be in the address space administered by the network service provider to which the C-PE attaches. The service provider may assign it dynamically, or it may be assigned statically and configured in the C-PE by the customer.

In addition to having a globally routable black loopback, a C-PE will of course have globally routable interface addresses for each of its black interfaces.

Interface addresses of the red interfaces SHOULD NOT be globally routable.

If the C-PE attaches to multiple service providers, the black loopback is likely to be be a provider-independent address. However, it MUST be routable in the backbone network of both providers, and most likely will need to be globally routable.

The C-PE may have one or more (black) BGP sessions with service provider peers, in which case it may advertise the black loopback; the next hop field of such an advertisement would be the interface address of the interface over which that BGP session runs.

Each C-PE of a given customer MUST be provisioned with a red loopback that is unique among the set of C-PEs of that customer. The red loopback MUST NOT be a routable address in the public Internet or in the backbone networks of any service provider to which any of the C-PEs is attached. If a C-PE has a (black) BGP session with a service provider peer, it MUST NOT advertise a route to its red loopback over that session. That is, any IP route to a red loopback is considered to be a red route, and MUST NOT be advertised or received on a black BGP session. These "red loopback routes" can thus be considered to be "red routes", even though they are IP rather than VPN-IP routes.

3.2. Setting Up Red BGP Sessions Between C-PEs and RRs

The customer is expected to have two or more BGP Route Reflectors (red RRs). The red RRs are presumed to be secure; making them so is the responsibility of the customer. As with the C-PEs, each red RR has a black loopback and a red loopback. If the RR is not also a C-PE, it will have only black interfaces, each of course with a globally routable interface address.

A customer's red RRs will form BGP sessions with that customer's C-PEs. These BGP sessions MUST be protected by IPsec. The use of IPsec transport mode is RECOMMENDED. If the RR's red loopback is an IPv4 address, it may be used as the RR's BGP Identifier (see [RFC4271]).

When a C-PE device comes up, it attempts to set up an IPsec-protected BGP session with the red RRs. This requires first setting up an IPsec SA with each red RR, and then using IPsec Transport Mode to protect the BGP session.

If the C-PE's red loopback is an IPv4 address, the C-PE's BGP Identifier (see [RFC4271]) may be the red loopback.

The endpoint addresses of the IPsec SA are the black loopbacks of the endpoint systems.

Therefore, in order to initiate a BGP session to a red RR, a C-PE must be provisioned to know a publicly routable address (i.e., the black loopback) of the RR. A C-PE must also be provisioned with whatever additional information is needed in order to set up an IPsec SA with each of the red RRs. Each C-PE will attempt to continuously maintain live BGP sessions (protected by IPsec) with each red RR. Note that the source and destination IP address fields of the IP datagrams carrying the IPsec-encapsulated BGP messages will be publicly routable adresses.

In some scenarios, it may be desirable to provision each red RR with the publicly routable address and pre-shared secret of every C-PE. This makes it easy for the C-PEs to authenticate themselves to the RR, but requires each RR to be reprovisioned every time a new C-PE is added to the network.

In other scenarios, it may be considered desirable to allow the RRs to auto-discover the C-PEs, without the need for any per-C-PE pre-provisioning of the RRs. In this case, a certificate-based authentication method [reference?] can be used when setting up the IPsec SAs that carry the BGP sessions.

In either type of scenario, the C-PE SHOULD NOT be assumed to have a fixed black loopback address or fixed black interface addresses; rather, it SHOULD be assumed that a C-PE might be a mobile device whose globally routable addresses change from time to time.

If a customer's C-PEs support multiple VPNs (for multiple tenants), that customer's red RRs will receive and disseminate the VPN-IP routes of all those VPNs.

Note that according to the above procedures, the C-PEs will only have red BGP sessions to the red RRs; the C-PEs will not have BGP sessions to each other. Thus it is not necessary for the C-PEs to know of each other in advance. Of course, if a particular customer deems it desirable for the C-PEs to have red BGP sessions to each other, each C-PE can be provisioned with a publicly routable address of each other C-PE, along with any additional information needed to set up an IPsec SA to each other C-PE.

It is RECOMMENDED that, for the purpose of setting up the red BGP sessions, all the RRs and C-PEs be considered to be in the same Autonomous System (AS). Then the red BGP sessions will all be IBGP sessions, and the next hop field of a red route will not be modified as the route is propagated. Note that if an implementation allows a given router to be attached to two different ASes, this does not require that all the C-PEs and red RRs attach to the Internet via the same AS. The "red overlay" may appear to be within a single AS, but the "black underlay" need not be within a single AS.

If it is necessary to use an EBGP session between a C-PE and an RR, the RR SHOULD have a configured policy to leave the next hop unchanged when propagating red VPN-IP routes on an EBGP session. See Section 3.4.

In some scenarios, the C-PEs may set up red BGP sessions to Autonomous System Border Routers (ASBRs), rather than to RRs. This is transparent to the C-PE.

3.3. Routes Transmited by the C-PE on Red BGP Sessions

A C-PE MUST transmit its local VPN-IP routes on the red BGP sessions, and only on the red BGP sessions. The next hop of each local VPN-IP route MUST be set to the red loopback of the C-PE. The choice to transmit a particular VPN-IP route on a particular session may of course be influenced by the route's Route Targets.

A C-PE MUST NOT transmit its local VPN-IP routes on black BGP sessions. VPN-IP routes MUST NOT be accepted from black BGP sessions.

In all other respects, the handling of VPN-IP routes is done by normal L3VPN procedures.

Each C-PE MUST also transmit the following IP (IPv4 or IPv6) route on the red BGP sessions. We refer to this route as the C-PE's "red loopback route":

- The address prefix field of the route's Network Layer Reachability Information (NLRI) contains the C-PE's red loopback as a host address.
- The Next Hop of the route is the C-PE's black loopback.
- The route carries a Tunnel Encapsulation Attribute [tunnel_encaps] with the the following parameters:
  - Tunnel Type = "MPLS-in-IPsec" (see Section 5.)
  - Remote Endpoint = the C-PE's black loopback

  - A "Security Handle" (see Section 6). This provides any information needed by another C-PE to set up an MPLS-in-IPsec Security Association with the advertising C-PE.

A C-PE MUST NOT transmit, on any black BGP session, an IP route whose NLRI contains its red loopback.

A given C-PE's red loopback route must be propagated to all other the C-PEs belonging to the same customer. Therefore, such routes SHOULD NOT carry Route Targets.

3.4. Propagating Red Routes

A route that is received over a red BGP session may need to be propagated to other red BGP sessions. A route that is received over a red BGP session MUST NOT be propagated over a black BGP session. Similarly, a route that is received over a black BGP session MUST NOT be propagated over a red BGP session.

When a route is propagated from one red BGP session to another, its next hop SHOULD be left unchanged. As specified in Section 4, this will ensure that a data packet sent on the path advertised by that route are sent on an IPsec SA between its ingress C-PE and its egress C-PE. Changing the next hop will change the IPsec SA endpoint.

This may be useful in certain deployments. For instance, the path from an ingress C-PE to an egress C-PE may traverse several ASBRS. If these ASBRs are secure, it may be desirable to set up a sequence of IPsec SAs, (e.g., C-PE1--ASBR1, ASBR1--ASBR2, ASBR2--C-PE2) instead of using a single IPsec SA between C-PE1 and C-PE2. If this is not the intention, the red BGP sessions MUST leave the next hop unchanged, even if those sessions are EBGP sessions.

In all other respects, propagation of red routes is governed by the normal procedures for propagating routes. If the route carries one or more Route Targets, these may affect its propagation. However, note that propagation of a route between a red BGP session and a black BGP session MUST NOT be done, irrespective of the Route Targets.

4. VPN-IP Routes and Recursive Route Resolution

Suppose a C-PE, say C-PE1, receives a packet, say packet P, on one of its local red interfaces. Suppose that packet P is addressed to a system that is reachable via one of the red interfaces at another C-PE, say C-PE2. C-PE1 looks up packet P's destination address in the VRF associated with P's incoming interface. The matching route will be a "Labeled VPN-IP route" [RFC4364] originated by C-PE2, and

disseminated to C-PE1 over a red BGP session. Per Section 3.3, the next hop of that route will be C-PE2's red loopback.

The labeled VPN-IP route matched by packet P's destination address will contain an MPLS label, the "VPN label". C-PE1 pushes the VPN label onto packet P's MPLS label stack. Then C-PE1 needs to determine how to transmit the resulting MPLS packet to the next hop of the VPN-IP route. The next hop of the labeled VPN-IP route will be the red loopback address C-PE2. So C-PE1 looks for the route to that red loopback address. This will be the red loopback route (i.e., the red IP route, see Section 3.3) originated by C-PE2.

C-PE2's red loopback will then be recursively resolved by means of C-PE2's red loopback route. By virtue of the Tunnel Encapsulation attribute carried by that route, C-PE1 will realize that to send packet P, it must set up an MPLS-in-IPsec SA (see Section 5, as well as Sections 3 and 8.1 of [RFC4023]) with C-PE2. Per the route's Tunnel Encapsulation attribute, the remote endpoint of this IPsec SA will be C-PE2's black loopback, and the Security Handle in the Tunnel Encapsulation attribute will carry any other information needed to set up the Security Association.

Note that the remote endpoint of the IPsec SA is determined by the Tunnel Encapsulation attribute of the red loopback route, rather than by the next hop field of that route. This ensures that the SA is made to the proper endpoint, even if the next hop field of the red loopback route was modified while the route was propagated.

IMPORTANT: A VPN-IP route MUST NOT be recursively resolved by an IP route that was received over a black BGP session. If a VPN-IP route's next hop resolves to a route received over a black BGP session, the existence of the latter route MUST be regarded as the result of an attempt to spoof the location of the egress C-PE. That is, the latter route MUST be considered to be a spoofed route. Note that it may not be possible to detect this spoofing attack until the attempt is made to recursively resolve the VPN-IP route. Implementors should take special care to ensure that their implementations are not vulnerable to this sort of spoofing attack. IF an implementation cannot detect this sort of attack during the recursive route resolution process, then the C-PE MUST NOT have any black BGP sessions.

When packet P is transmitted, it is transmitted through an MPLS-in-IPsec SA. Thus the only information that appears in the clear is the IP header needed to get the packet across the network. The IP source and destination addresses of that packet will be the black loopbacks of C-PE1 and C-PE2 respectively. The red loopback addresses do not appear in the packets at all, and no part of the payload packet

(neither the VPN label nor the IP datagram following the VPN label) appears in the clear.

The MPLS-in-IPsec SA between C-PE1 and C-PE2 may be initiated by

C-PE1 as soon as it receives a red loopback route originated by C-PE2. Alternatively, the initiation of the setup of the Security Association may be delayed until the SA is actually needed for transmitting packets.

These procedures will result in a single IPsec SA between a pair of C-PEs, with the data of multiple tenants carried on that single SA. If for some reason it is considered preferable to have an SA per tenant, the following procedures can be used:

- On each C-PE, provision a distinct red loopback for each tenant.

- Each C-PE will originate a red loopback route for each red loopback.

- Each red loopback route will have its own Tunnel Encapsulation attribute. The respective Security Handle sub-TLVs (if present) MUST be distinct.

5. MPLS-in-IPsec

Packets traveling from one C-PE to another travel through "MPLS-in-IPsec" tunnels. To transmit an MPLS packet through an MPLS-in-IPsec tunnel, one does the following:

- Encapsulate the MPLS packet in IP, as specified in Section 3 of [RFC4023].

- Use an IPsec transport mode Security Association to send the MPLS-in-IP packet from one C-PE to the other. This is specified in Section 8.1 of [RFC4023].

The result of encapsulating MPLS in IP and then transmitting the MPLS-in-IP packet on an IPsec transport mode Security Association is known as an MPLS-in-IPsec packet.

On the wire, an MPLS-in-IPsec packet consists of a cleartext IP header followed by a payload. The IP source and destination addresses of an MPLS-in-IPsec packet will be the black loopbacks of the source and destination C-PEs. The payload will be an MPLS packet. If the IPsec Security Association is providing privacy, authentication, and integrity, the payload is protected from inspection or alteration.

When the packet arrives at the destination C-PE, any necessary decryption is done, and packet appears to be an MPLS-in-IP packet addressed to the black address of the destination C-PE. The IP encapsulation is removed, yielding an MPLS packet. Per the usual L3VPN procedures, the label at the top of the MPLS label stack will be used to govern the further disposition of the packet. However, if a packet received over a black interface was not received though an IPsec SA, the packet MUST NOT be sent out any VRF interface.

MPLS-in-IP packets received in the clear (i.e., not received over an IPsec SA) MUST be discarded.

Note that this section is not intended to describe an implementation strategy.

6. Security Handle

This document defines a new BGP Tunnel Encapsulation attribute sub-TLV, the "Security Handle". This sub-TLV has a one-octet length field. It is intended for use in the Tunnel Encapsulation attribute carried by the red loopback routes. Its use is deployment specific.

As an example, in some deployments, this sub-TLV might be used to carry the IPsec Security Parameters Index (SPI). When setting up an SA to the originator of a particular Tunnel Encapsulation attribute, the SPI would be used as part of the SA setup procedure.

In deployments where the C-PEs auto-discover each other through RRs, and authenticate via certificate-based mechanisms, the Security Handle may not be needed at all. If a given deployment does not make use of the Security Handle, the sub-TLV SHOULD be omitted from the Tunnel Encapsulation attribue.

7. Data Plane Security Procedures

If a C-PE receives data over one of its local red interfaces, it may forward the data out another of its local red interfaces, as long as those two interfaces are associated with the same VRF, or if there is policy allowing communication ("extranet") between those two interfaces.

However, data received by a C-PE over one of its red interfaces MUST NOT be forwarded out a black interface, unless that data is being sent over the black interface through an IPsec SA.

Similarly, data received by a C-PE over one of its black interfaces MUST NOT be forwarded out a red interface unless the data arrived through an IPsec SA.

Typically an IPsec implementation has procedures to prevent unauthorized red-to-black or black-to-red forwarding. However, the conventional procedures are based on filtering of IP addresses, and hence do not apply directly if MPLS-in-IPsec is used. Implementors should take care to ensure that unauthorized red-to-black or black-to-red forwarding is prohibited.

8. Acknowledgments

The authors wish to thank John Scudder for his ideas and contributions to this work.

9. IANA Considerations

IANA is requested to create a new entry in the "BGP Tunnel Encapsulation Attribute Sub-TLVs" registry, "Security Handle". This sub-TLV is defined in Section 6 to have a one-octet length field. Thus it needs to be assigned a codepoint in the range 0-127 inclusive.

10. Implementation Challenges

This document specifies an architecture for Secured L3VPNs, but a successful implementation faces a number of challenges.

This document specifies a recursive route resolution process that makes use of the Tunnel Encapsulation attribute. This is a new feature.

This document specifies that during recurisve route resolution, resolution of a red route via a route received over a black BGP session must be prohibited. This is a new feature that may present challenges.

Ultimately, success will require a highly scalable IPsec implementation, that can set up SAs dynamically based on information disseminated by BGP. This presents a number of implementation challenges.

11. Security Considerations

12. References

12.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC4023] Worster, T., Rekhter, Y., and E. Rosen, Ed., "Encapsulating MPLS in IP or Generic Routing Encapsulation (GRE)", RFC 4023, DOI 10.17487/RFC4023, March 2005, <https://www.rfc-editor.org/info/rfc4023>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <https://www.rfc-editor.org/info/rfc4271>.

[RFC4303] Kent, S., "IP Encapsulating Security Payload (ESP)", RFC 4303, DOI 10.17487/RFC4303, December 2005, <https://www.rfc-editor.org/info/rfc4303>.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <https://www.rfc-editor.org/info/rfc4364>.

[RFC7296] Kaufman, C., Hoffman, P., Nir, Y., Eronen, P., and T. Kivinen, "Internet Key Exchange Protocol Version 2 (IKEv2)", STD 79, RFC 7296, DOI 10.17487/RFC7296, October 2014, <https://www.rfc-editor.org/info/rfc7296>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

[RFC8221] Wouters, P., Migault, D., Mattsson, J., Nir, Y., and T.

Kivinen, "Cryptographic Algorithm Implementation Requirements and Usage Guidance for Encapsulating Security Payload (ESP) and Authentication Header (AH)", RFC 8221, DOI 10.17487/RFC8221, October 2017, <https://www.rfc-editor.org/info/rfc8221>.

[tunnel_encaps]
Rosen, E., Patel, K., and G. Van de Velde, "The BGP Tunnel Encapsulation Attribute VPN", internet-draft draft-ietf-idr-tunnel-encaps-09, February 2018.

12.2. Informational References

[RFC6071] Frankel, S. and S. Krishnan, "IP Security (IPsec) and Internet Key Exchange (IKE) Document Roadmap", RFC 6071, DOI 10.17487/RFC6071, February 2011, <https://www.rfc-editor.org/info/rfc6071>.

Authors' Addresses

Eric C. Rosen (editor)
Juniper Networks, Inc.
10 Technology Park Drive
Westford, Massachusetts 01886
United States

Email: erosen@juniper.net

Ron Bonica
Juniper Networks, Inc.
2251 Corporate Park Drive
Herndon, Virginia 20171
United States

Email: rbonica@juniper.net