Network Working Group
Internet Draft
Intended status: Informational
Expires: July 29, 2023

L. Dunbar
J. Guichard
Futurewei
Ali Sajassi
Cisco
J. Drake
Juniper
B. Najem
Bell Canada
Ayan Barnerjee
D. Carrel
IPsec Research
January 29, 2023

BGP Usage for SDWAN Overlay Networks
draft-ietf-bess-bgp-sdwan-usage-07

Abstract

The document discusses the usage and applicability of BGP as the control plane for multiple SDWAN scenarios. The document aims to demonstrate how the BGP-based control plane is used for large-scale SDWAN overlay networks with little manual intervention.

SDWAN edge nodes are commonly interconnected by multiple types of underlay networks owned and managed by different network providers.

This Internet-Draft will expire on July 29, 2023.

Copyright Notice

Table of Contents

1. Introduction

SDWAN optimizes the transport of IP Packets over multiple underlay connectivity services. Here are some of the main characteristics of "SDWAN" networks:

- Augment of transport, which refers to utilizing paths over different underlay networks. There are often multiple parallel overlay paths between any two SDWAN edges; some are private networks over which traffic can traverse with or without encryption; others require encryption, e.g., over untrusted public networks.
- Direct Internet breakout from remote branch offices is allowed instead of all traffic hauled to Corporate HQ for centralized policy control.
- Some traffic can be forwarded based on their application identifiers instead of based on destination IP addresses by the

edge nodes placing the traffic onto specific overlay paths based on the application-specific policies.

- The traffic forwarding can also be based on specific performance criteria (e.g., packets delay, packet loss, jitter) to provide better application performance by choosing the underlay that meets or exceeds the specified policies.

[Net2Cloud-Problem] describes the network-related problems to connect enterprises' branch offices to dynamic workloads in different Cloud Data Centers (DC). SDWAN has been positioned as a flexible way to reach dynamic workloads in third-party Cloud DCs. However, scaling becomes a significant issue when hundreds or thousands of nodes need to be interconnected by SDWAN overlay networks.

This document describes using BGP as the control plane to scale large SDWAN overlay networks.

2. Conventions used in this document

Cloud DC: Third party data centers that host applications and workloads owned by different organizations or tenants.

Controller: Used interchangeably with SDWAN controller to manage SDWAN overlay path creation/deletion and monitor the path conditions between sites.

CPE: Customer Premise Equipment

CPE-Based VPN: Virtual Private Secure network formed among CPEs. This differentiates from more commonly used PE-based VPNs [RFC 4364].

Homogeneous Encrypted SDWAN: A SDWAN network in which all traffic to/from the SDWAN edges are carried by IPsec tunnels regardless of underlay networks. I.e., the client traffic is carried by IPsec tunnel even over MPLS private networks.

ISP: Internet Service Provider

NSP: Network Service Provider. NSP usually provides more advanced network services, such as MPLS VPN, private leased lines, or managed Secure WAN connections, often within a private, trusted domain. In contrast, an ISP usually provides plain Internet services over public untrusted domains.

PE: Provider Edge

SDWAN Edge Node: an edge node, which can be physical or virtual, maps the attached clients' traffic to the wide area network (WAN) overlay tunnels.

SDWAN: SD-WAN: Software Defined Wide Area Network. A connectivity service offered by a Service Provider that optimizes the transport of IP Packets over multiple underlay connectivity services by recognizing applications at Ingress and determining forwarding behavior by applying policies to them.

SDWAN IPsec SA: IPsec Security Association between two SDWAN ports or nodes.

SDWAN over Hybrid Networks: SDWAN over Hybrid Networks typically have edge nodes utilizing bandwidth resources from different types of underlay networks, some being private networks and others being public Internet.

WAN Port: A Port or Interface facing an ISP or Network Service Provider (NSP), with address allocated by the ISP or the NSP.

C-PE: SDWAN Edge node, which can be CPE for customer managed SDWAN, or PE for provider managed SDWAN services.

ZTP: Zero Touch Provisioning

3. Use Case Scenario Description and Requirements

This section describes some essential requirements for SDWAN networks and several SDWAN scenarios used by the subsequent sections to explain how the BGP control plane is applied.

3.1. Requirements

3.1.1. Supporting SDWAN Segmentation

"SDWAN Segmentation" is a frequently used term in SDWAN deployment, referring to policy-driven network partitioning. An SDWAN segment is a virtual private network (SDWAN VPN) consisting of a set of edge nodes interconnected by the tunnels, such as IPsec tunnels and MPLS VPN tunnels.

This document assumes that an SDWAN VPN configuration on a PE follows the same way as MPLS VPN, i.e., via VRFs. One SDWAN VPN can be mapped to one or multiple virtual topologies governed by the SDWAN controller's policies.

When using BGP for SDWAN, the Client Route UPDATE is the same as MPLS VPN. Route Target in the BGP Extended Community can be used to differentiate routes belonging to different SDWAN VPNs.

As SDWAN is an overlay network arching over multiple types of networks, MPLS L2VPN/L3VPN or pure L2 underlay can continue using the VPN ID, VN-ID, or VLAN in the data plane to differentiate packets belonging to different SDWAN VPNs. For packets carried by an IPsec tunnel, the IPsec's inner encapsulation header can have the SDWAN VPN Identifier to distinguish the packets belonging to different SDWAN VPNs.

3.1.2. Client Service Requirement

The Client interface of SDWAN edges can be IP or Ethernet-based.

For Ethernet-based client interfaces, SDWAN edge should support VLAN-based service interfaces (EVI100), VLAN bundle service interfaces (EVI200), or VLAN-Aware bundling service interfaces. EVPN service requirements apply to the Client traffic, as described in Section 3.1 of RFC8388.

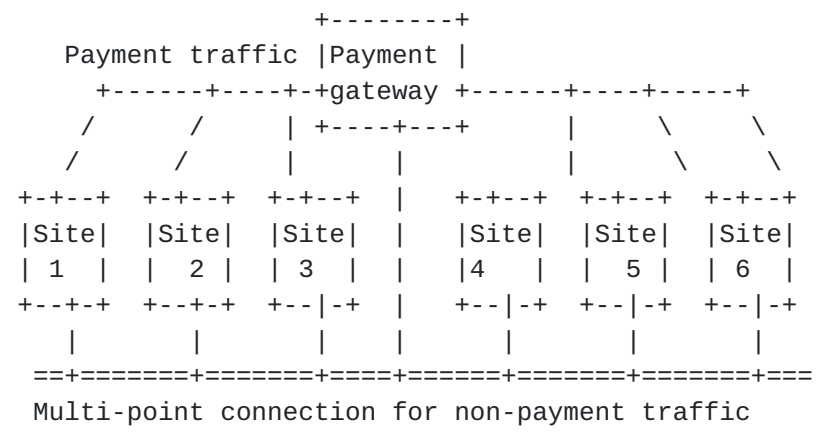
For IP-based client interfaces, L3VPN service requirements are applicable.

3.1.3. SDWAN Traffic Segmentation

SDWAN Traffic Segmentation enables the separation of the traffic based on the business and the security needs of different user groups and/or application requirements. Each user group and/or application may need different isolated topologies and/or policies to fulfill the business requirements.

For example, a retail business requires the point-of-sales (PoS) application to be on a different topology from other applications. The PoS application is routed only to the payment processing entity at a hub site; other applications can be routed to all other sites.

The traffic from the PoS application follows a Tree topology in the figure below, whereas other traffic can be multipoint-to-multipoint topology.



Another example is an enterprise that wants to isolate the traffic from different departments, with each department having its unique topology and policy. The HR department may need to access specific applications that are NOT accessible by the engineering department. Also, the contractors may have limited access to the enterprise resources.

3.1.4. Zero Touch Provisioning

SDWAN zero-tour provisioning (ZTP) allows devices to be configured and provisioned centrally. When an SDWAN edge is installed at a remote location, ZTP automates follow-up steps, including updates to the OS, software version, and configuration before client traffic being forwarded. The ZTP can bootstrap a remote SDWAN edge and establish a secure connection to the local SDWAN Controller, making

it convenient to add or delete an SDWAN edge node (virtual or physical). From the network control perspective, ZTP includes the following:

- Upon power-up, an SDWAN edge can establish the transport layer secure connection (such as TLS, SSL, etc.) to its controller, whose address can be burned or preconfigured on the device.

- The SDWAN Controller can designate a local network controller in the proximity of the SDWAN edge. Like the Route-Reflector (RR) for BGP-controlled SDWAN, the local network controller manages and monitors the communication policies for traffic to/from the edge node.

3.1.5. Constrained Propagation of SDWAN Edge Properties

One SDWAN edge node may only be authorized to communicate with a small number of other SDWAN edge nodes. Under this circumstance, the property of the SDWAN edge node cannot be propagated to other nodes that are not authorized to communicate. But a remote SDWAN edge node, upon powering up, might not have the right policies to know which peers are authorized to communicate. Therefore, SDWAN deployment needs to have a central point to distribute the properties of an SDWAN edge node to its authorized peers.

BGP is well suited for this purpose. RFC4684 has specified the procedure to constrain the distribution of BGP UPDATE to only a subset of nodes. Each edge node informs the Route-Reflector (RR) [RFC4456] on its interested SDWAN VPNs. The RR only propagates the BGP UPDATE for the relevant SDWAN VPNs to the edge.

The connection between an SDWAN edge and its RR can be over an insecure network. Therefore, an SDWAN edge must establish a secure transport layer connection (TLS, SSL, etc.) to its designated RR upon power-up. The BGP UPDATE messages need to be sent over the secure channel (TLS, SSL, etc.) to the RR.

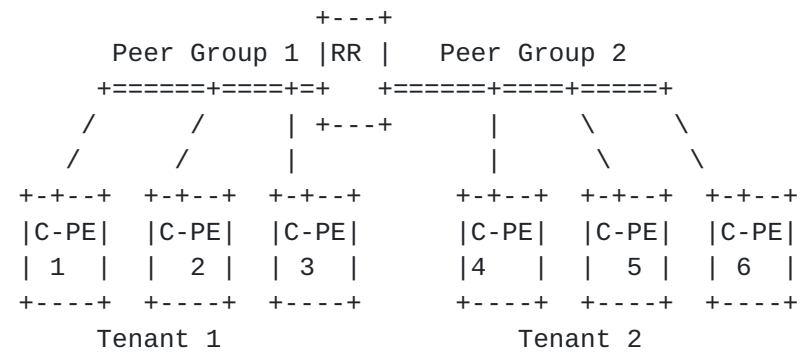


Figure 1: Peer Groups managed by RR

Tenant separation is achieved by the SDWAN VPN identifiers represented in the control plane and data plane, respectively.

3.2. Scenario #1: Homogeneous Encrypted SDWAN

Homogeneous Encrypted SDWAN refers to a type of SDWAN network with edge nodes encrypting all traffic over WAN to other edge nodes, regardless of whether the underlay is private or public. For lack of better terminology, we call this Homogeneous Encrypted SDWAN throughout this document.

Here are some typical scenarios for the using Homogeneous Encryption:

- A small branch office to connect to its HQ offices via the Internet. All sensitive traffic to/from this small branch office must be encrypted, usually achieved by IPsec Tunnels.

- A store in a shopping mall may need to securely connect to its applications in one or more Cloud DCs via the Internet. A common way of achieving this is to establish IPsec SAs to the Cloud DC gateway to carry the sensitive data to/from the store.

As described in [SECURE-EVPN], the granularity of the IPsec SAs for Homogeneous Encryption can be per site, per subnet, per tenant, or per address. Once the IPsec SA is established for a specific subnet/tenant/site, all traffic to/from the subnets/tenants/site is encrypted.

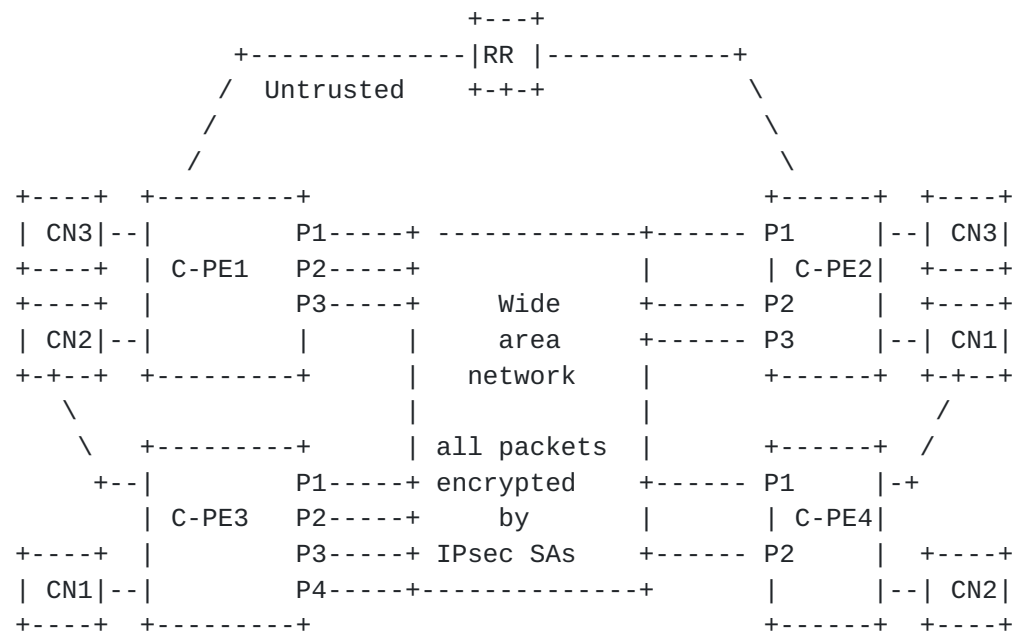


Figure 2: Homogeneous Encrypted SDWAN

One of the properties of Homogeneous Encryption is that the SDWAN Local Network Controller, e.g., RR in BGP-controlled SDWAN, might be connected to C-PEs via an untrusted public network, therefore, requiring a secure connection between RR and C-PEs (TLS, DTLS, etc.).

Homogeneous Encrypted SDWAN has some properties similar to the commonly deployed IPsec VPN, albeit the IPsec VPN is usually point-to-point among a small number of nodes and with heavy manual configuration for IPsec between nodes. In contrast, an SDWAN network can have many edge nodes and a central controller to manage the configurations on the edge nodes.

Existing Private VPNs (e.g., MPLS based) can use Homogeneous Encrypted SDWAN to extend over the public network to remote sites to which the VPN operator does not own or lease infrastructural connectivity, as described in [SECURE-EVPN] and [SECURE-L3VPN]

3.3. Scenario #2: Differential Encrypted SDWAN

The Differential Encrypted SDWAN refers to an SDWAN network in which traffic over the existing VPN is forwarded natively without encryption, and the traffic over the Public Internet is encrypted. Differential Encrypted SDWAN is over hybrid private VPN and public Internet underlays. Since IPsec requires additional processing power

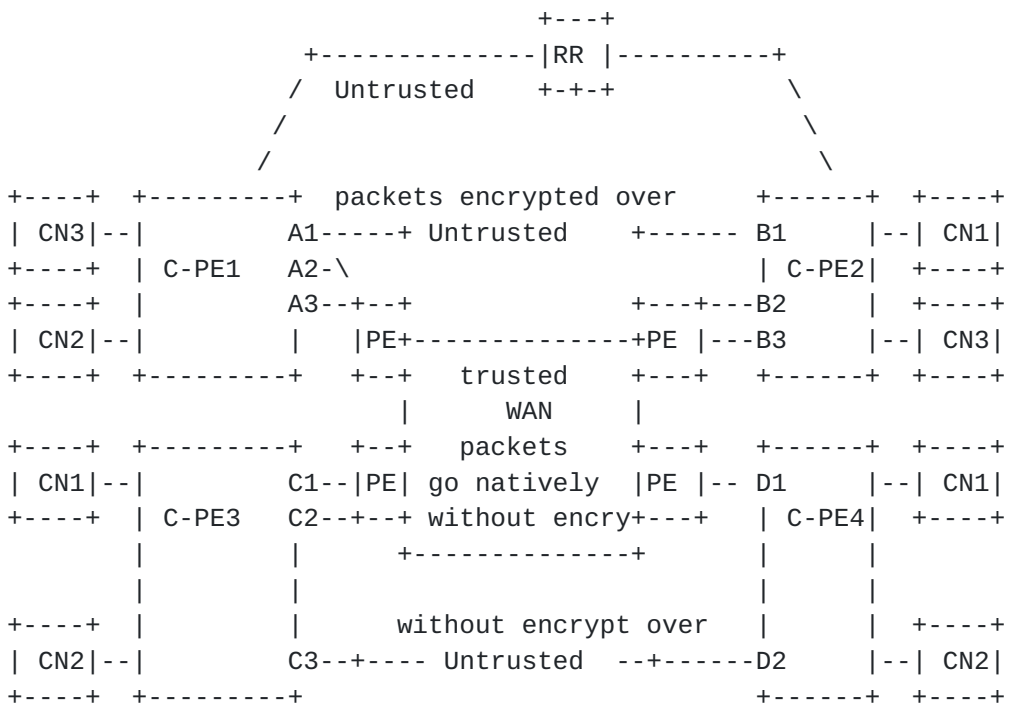
and the encrypted traffic over the Internet does not have the premium SLA commonly offered by Private VPNs, especially over a long distance, it is more desirable for traffic over a private VPN to be forwarded without encryption.

One C-PE might have the Internet-facing WAN ports managed by different ISPs/NSPs with the WAN ports' addresses assigned by the corresponding ISPs/NSPs. Clients might have policies to specify:

1) Some flows can only be forwarded over private VPNs.
2) Some flows can be forwarded over either private VPNs or the public Internet. The packets over the public Internet are encrypted.
3) Some flows, especially Internet-bound browsing ones, can be handed off to the Internet without any encryption.

Suppose a flow traversing multiple segments, such as A<->B<->C<->D, has Policy 2) above. The flow can cross different underlays in different segments, such as over Private underlay between A<->B without encryption or over the public Internet between B<->C protected by an IPsec SA.

As shown in the figure below, C-PE-1 has two different types of interfaces (A1 to Internet and A2 & A3 to VPN). The C-PE's loopback address and the attached client addresses may or may not be visible to the ISPs/NSPs. The WAN ports' addresses can be allocated by the service providers or dynamically assigned (e.g., by DHCP).



CN: Client Network

Figure 3: SDWAN with Hybrid Underlays

Also, the connection between C-PEs and their Controller (RR) might be via the untrusted public network. It is necessary to encrypt the communication between RR and C-PEs, by TLS, DTLS, etc.

There could be multiple SDWAN edges (C-PEs) sharing common property, such as a geographic location. Some applications over SDWAN may need to traverse specific geographic areas for various reasons, such as to comply with regulatory rules, to utilize specific value-added services, or others.

Services may not be congruent, i.e., the packets from A-> B may traverse one underlay network, and the packets from B -> A may go over a different underlay.

3.4. Scenario #3: Private VPN PE based SDWAN

This scenario refers to the existing VPN (e.g., EVPN or IPVPN) being expanded by adding extra ports facing the untrusted Internet for PEs to offload low-priority traffic when the VPN paths are congested.

Throughout this document, this scenario is also called Internet Offload for Private VPN, or PE-based SDWAN.

Here are some differences from the Hybrid Underlay scenario (Section 3.3):

- For MPLS-based VPN, PEs would have MPLS as payload encapsulated within the IPsec tunnel egressing the Internet WAN ports, MPLS-in-IP/GRE-in-IPsec.
- The BGP RR is connected to PEs in the same way as VPN, i.e., via the trusted network.

The PE-based SDWAN can be used by VPN service providers to temporarily increase bandwidth between sites when not sure if the demand will sustain for an extended period or as a temporary solution before the permanent infrastructure is built or leased.

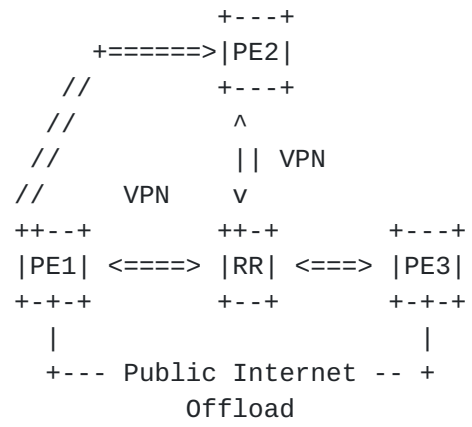


Figure 4: Additional Internet paths added to the VPN

4. Provisioning Model

4.1. Client Service Provisioning Model

Client service provisioning can follow the same approach as MPLS VRFs. A client VPN can establish the communication policy by specifying the Route Targets to be imported and exported. Alternatively, traditional Match and Action ACLs can identify the specific routes allowed or denied to or from the client VPN.

When an SDWAN edge node is dedicated to one client with one virtual network, all the prefixes attached to the client port(s) of the edge node can be considered in one VRF, and the RR can manage the policies for import/export of the VRF.

4.2. Policy Configuration

One of the characteristics of an SDWAN service is that packets can be forwarded over multiple types of underlays. Policies are needed to govern which underlay paths can carry an application flow, as described by Section 8 of MEF70.1. An Application Flow consists of packets that match specific criteria. For example, client-prefix-x can only be mapped to MPLS topology.

4.3. IPsec related parameters Provisioning

For the IPsec tunnel to be established, the SDWAN edge nodes need to support the common IPsec encryption algorithms (DES, 3DES, or AES), the hash algorithm (SHA or MD5), and the DH groups. Each SDWAN edge node can have the default supported values for those attributes or get the attributes from its controller to minimize the configuration. For BGP-controlled SDWAN, BGP UPDATE messages can propagate each node's IPsec-related attribute values for peers to choose the common values supported, which is traditionally done by IPsec IKEv2.

5. BGP Controlled SDWAN

5.1. BGP Walk Through for Homogeneous Encrypted SDWAN

For the BGP-controlled Homogeneous Encrypted SDWAN, a C-PE can advertise its attached client routes and the properties of the IPsec SA in one BGP UPDATE message.

In the Figure below, the BGP UPDATE message from C-PE2 to RR can have the client routes encoded in the MP-NLRI Path Attribute and the IPsec Tunnel associated information encoded in the Tunnel-Encap [RFC9012] Path Attributes as described in the [SECURE-EVPN].

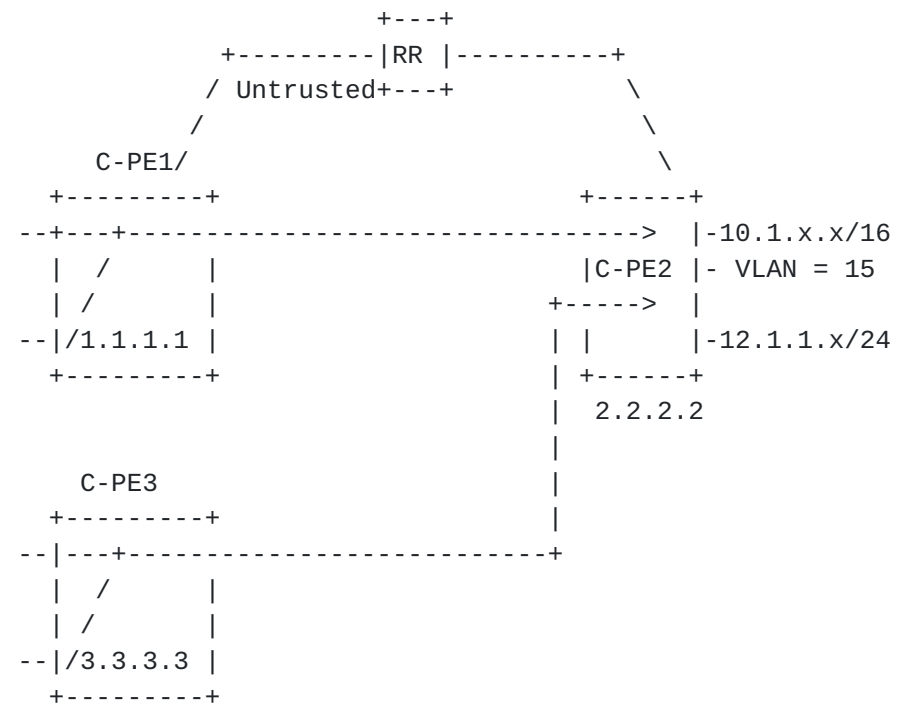


Figure 5: Homogeneous Encrypted SDWAN

Alternatively, the C-PE2 can use two separate BGP UPDATE messages to reduce the size of the BGP UPDATE messages, as IPsec SA tunnels have many attributes and IPsec SA keys periodical changes occur at different frequency than the clients' routes updates.

As described by Section 4 and 8 of [RFC9012], UPDATE U1 has its Nexthop to the node loopback address and is reclusively resolved to the IPsec SA tunnel detailed attributes advertised by the UPDATE U2 for the Node Loopback address.

Here are the details of the UPDATE messages:

- Suppose that a given packet P destined towards the client addresses attached to C-PE2 (e.g., prefix 10.1.x.x/16) can be carried by any IPsec tunnels terminated at C-PE2.
- The path along which P is to be forwarded is determined by BGP UPDATE U1.
- UPDATE U1 does not have a Tunnel Encapsulation attribute.
- UPDATE U1 can include the Encapsulation Extended Community with the option to have the Color Extended Community.
- The address of the next-hop of UPDATE U1 is router C-PE2.
- UPDATE U2 has a Tunnel Encapsulation attribute to describe the IPsec SA detailed attributes.

UPDATE U1:

- MP-NLRI Path Attribute:
  10.1.x.x/16
  12.1.1.x/24
- Nexthop: 2.2.2.2 (C-PE2)
- Encapsulation Extended Community: TYPE = IPsec

UPDATE U2:
- MP-NLRI Path Attribute:
  2.2.2.2 (C-PE2)
- Tunnel Encapsulation Path Attributes (as described in the [SECURE-EVPN]) for IPsec SA detailed attributes, including the WAN address to be used as the IP address of the IPsec encrypted packets.

If different client routes attached to C-PE2 need to be reached by separate IPsec tunnels, the Color-Extended-Community [RFC9012] is used to associate routes with the tunnels. See Section 8 of [RFC9012].

Suppose C-PE2 does not have a policy on the authorized peers for the specific client routes. Then, RR needs to check the client routes policies to constrain the BGP UPDATE messages propagation only to the remote authorized edge nodes.

5.2. BGP Walk Through for Differential Encrypted SDWAN

In this scenario, some client routes can be forwarded over any one of the tunnels terminating at the edge node. Some client routes can only be forwarded over specific tunnels (such as only MPLS VPN).

An edge node can use the Color Extended Community (Section 4 & 8 of [RFC9012]) in its BGP UPDATE message to associate the client routes with the specific tunnels.

For example, in Figure 5 above, suppose that Route 10.1.x.x/16 can be carried by either MPLS or IPsec and Route 12.1.1.x/24 can only be carried by MPLS; C-PE2 can use the following UPDATE messages:

UPDATE #1a for Route Route 10.1.x.x/16:

- MP-NLRI Path Attribute:
  10.1.x.x/16
  Nexthop: 2.2.2.2 (C-PE2)
- Encapsulation Extended Community: TYPE = SDWAN-Hybrid
- Color Extended Community: RED

UPDATE #1b for Route Route 12.1.1.x/24:
- MP-NLRI Path Attribute:
  12.1.1.x/24
  Nexthop: 2.2.2.2 (C-PE2)
- Encapsulation Extended Community: TYPE = MPLS-in-GRE
- Color Extended Community: YELLOW

UPDATE #2a: for IPsec tunnels terminated at the node:
- MP-NLRI Path Attribute:
  2.2.2.2 (C-PE2)
- Tunnel Encapsulation Path Attributes: TYPE=SDWAN-Hybrid
  Including the information about the WAN ports for receiving IPsec encrypted packets, the IPsec properties, etc.
- Color Extended Community: RED

UPDATE #2b: for MPLS-in-GRE terminated at the node:
- MP-NLRI Path Attribute:
  2.2.2.2 (C-PE2)
- Tunnel Encapsulation Path Attributes: TYPE=SDWAN-Hybrid
- Color Extended Community: YELLOW

SDWAN-Hybrid Tunnel Type is specified by [SDWAN-EDGE-Discovery].

5.3. BGP Walk Through for Application Flow-Based Segmentation

Suppose the application flow can be identified by the source or destination IP addresses. In that case, constraining the BGP UPDATE messages for the application only to the nodes that meet the criteria of the application flow can achieve the Application Flow-based Segmentation described in Section 3.1.2. In the Figure below, the following BGP Updates can be advertised to ensure that the Payment Application only communicates with the Payment Gateway:

BGP UPDATE #1a from C-PE2 to RR for the P2P topology that is only propagated to Payment GW node:

UPDATE #1a (only to the Payment GW node):

- MP-NLRI Path Attribute:
  - 30.1.1.x/24
  - Nexthop: 2.2.2.2
- Encapsulation extended community: TYPE = IPsec
- Color Extended Community: BLUE

BGP UPDATE #1b from C-PE2 to RR for the routes to be reached by C-PE1 and C-PE2:

- MP-NLRI Path Attribute:
  - 10.1.x.x
  - 12.4.x.x
  - Nexthop:2.2.2.2
- Encapsulation extended community: TYPE =IPsec
- Color Extended Community: RED

BGP UPDATE #2 describes the detailed IPsec attributes for IPsec tunnels terminated at C-PE2 2.2.2.2.

UPDATE #2a: for all IPsec SAs terminated at the node:
- MP-NLRI Path Attribute:
  2.2.2.2 (C-PE2)
- Tunnel Encapsulation Path Attributes: TYPE=IPsec (for all IPsec SAs)
- Color Extended Community: RED

UPDATE #2b: for the IPsec SA to the Payment Gateway:
- MP-NLRI Path Attribute:
  2.2.2.2 (C-PE2)
- Tunnel Encapsulation Path Attributes: TYPE=IPsec (for the IPsec SA to Payment GW).
- Color Extended Community: Blue

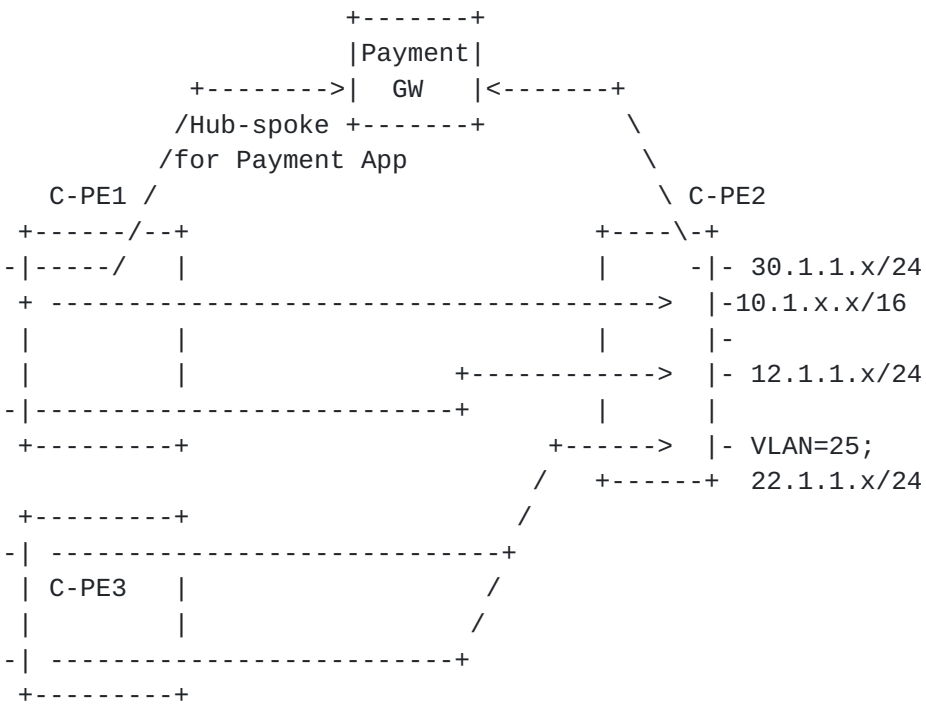


Figure 6: Application Based SDWAN Segmentation

5.4. Benefit of Using Recursive Next Hop Resolution

Using the Recursive Next Hop Resolution described in Section 8 of [RFC9012], the clients' routes UPDATE messages become very compact, and any changes of the underlay network tunnels attributes can be advertised without client route update. This method is handy when the underlay tunnels are IPsec based, which requires periodic message exchange for the pairwise re-keying process.

5.5. Why BGP as Control Plane for SDWAN?

For an SDWAN network with a small number of nodes, the traditional hub & spoke model utilizing NHRP or DSVPN/DMVPN protocol had worked reasonably well. DSVPN/DMVPN has a hub node (or controller) managing the edge nodes, including local & public addresses and tunnel identifiers mapping. However, for a sizeable SDWAN network, say more than 100 nodes with different underlays, the traditional approach becomes very messy, complex, and error prone.

Here are some of the compelling reasons for using BGP:

- Simplified peer authentication process:

  With a secure management channel established between an edge node and RR, RR can perform peer authentication on behalf of the edge node. Not only RR has policies on peer communication, but RR also has the built-in capability to constrain the propagation of the UPDATE messages to the authorized edge nodes [RFC4684].

- Scalable IPsec tunnel management

  When multiple IPsec tunnels are established between two pairwise edge nodes, BGP Tunnel Attribute Update can associate multiple IPsec tunnels with the client routes. In traditional IPsec VPN, separate routing protocols must run in parallel in each IPsec Tunnel if the client routes can be load shared among the IPsec tunnels.

- Simplified IPsec tunnel traffic selection configurations

  The IPsec tunnel's traffic selector or admission control can be inherently realized by specifying importing/exporting the Route Targets representing the SDWAN VPNs.

6. SDWAN Forwarding Model

This section describes how client traffic is forwarded in BGP Controlled SDWAN for the use cases described in Section 3.

The procedures described in Section 6 of RFC8388 are applicable for the SDWAN client traffic. Like the BGP-based VPN/EVPN client routes UPDATE message, Route Target can distinguish routes from different clients.

6.1. Forwarding Model for Homogeneous Encrypted SDWAN

6.1.1. Network and Service Startup Procedures

A single IPsec security association (SA) protects data in one direction. Under the Homogeneous Encrypted SDWAN Scenario, two SAs must be present to secure traffic in both directions between two C-PE nodes, two client ports, or two prefixes. Using Figure 2 of Section 3.2 as an example, for client CN2 attached to C-PE1, C-PE3, and C-PE4 to have full-mesh connection, six one-directional IPsec SAs must be established: C-PE1 <-> C-PE3; C-PE1 <-> C-PE4; C-PE3 <-> C-PE4.

SDWAN services to clients can be IP-based or Ethernet-based. An SDWAN edge can learn client routes from the client-facing ports via

OSPF, RIP, BGP, or static configuration for its IP-based services. For Layer-2 SDWAN services, the relevant EVPN parameters, such as the ESI (Ethernet Segment Identifier), EVI, and CE-VID to EVI mapping, can be configured similarly to EVPN described in RFC8388.

Instead of running IGP within each IPsec tunnel as done by the traditional IPsec VPN, BGP UPDATE messages propagate the client routes attached to SDWAN edge nodes.

In addition, BGP-RR (SDWAN Controller) facilitates the IPsec SA establishment and rekey management described in [SECURE-EVPN]. The Controller manages how clients' routes are associated with individual IPsec SA. Therefore, it is no longer necessary to manually configure the IPsec tunnel's endpoint addresses on each SDWAN edge node and set up policies for the allowed client prefixes.

6.1.2. Packet Walk-Through

For an IPsec SA terminated at a C-PE node, multiple client routes can be multiplexed in the IPsec SA (or tunnel). Traffic to the client prefixes is encapsulated in an inner tunnel, such as GRE or VxLAN, carried by the IPsec SA ESP tunnel. Different client traffic can be differentiated by a unique value in the inner encapsulation key or ID field.

For unicast packets forwarding:

The C-PE node address (or loopback address) acts as the Next Hop address for the prefixes attached to the C-PE nodes.

C-PE Node-based IPsec tunnel is inherently protected when the C-PE has multiple WAN ports to different underlay paths. As shown in Figure 2, when one of the underlay paths fails, the IPsec traffic can be forwarded to or received from a different physical port.

When a C-PE receives a packet from its client port, the packet is encapsulated inside the IPsec SA, whose destination address matches the Next Hop address of the packet's destination and forwarded to the target C-PE.

When a C-PE receives an IPsec encrypted packet from its WAN ports, it decrypts the packet and forwards the inner packet to the client port based on the inner packet's destination address.

For multicast packets forwarding:

IPsec was created to be a security protocol between two and only two devices, so multicast service using IPsec is problematic. An

IPsec peer encrypts a packet so that only one other IPsec peer can successfully perform the de-encryption. A straight way to forward a multicast packet for the Homogeneous Encrypted SDWAN is to encapsulate the multicast packet in separate unicast IPsec SA tunnels. More optimized forwarding multicast packets for the Homogeneous Encrypted SDWAN is out of the scope of this document.

6.2. Forwarding Model for Hybrid Underlay SDWAN

In this scenario, as shown in Figure 3 of Section 3.3, traffic forwarded over the trusted VPN paths can be native (i.e., unencrypted). The traffic forwarded over untrusted networks need to be protected by IPsec SA.

6.2.1. Network and Service Startup Procedures

Infrastructure setup: The proper MPLS infrastructure must be set up among the edge nodes, i.e., the C-PE1/C-PE2/C-PE3/C-PE4 of Figure 3. The IPsec SA between WAN ports or nodes must be set up as well. IPsec SA related attributes on edge nodes can be distributed by BGP UPDATE messages as described in Section 5.

There could be policies governing how flows can be forwarded, as specified by MEF70.1. For example, "Private-only" indicates that the flows can only traverse the MPLS VPN underlay paths.

6.2.2. Packet Walk-Through

For unicast packets forwarding:

Upon receiving a packet from a client port, if the packet belongs to a flow that can only be forwarded over the MPLS VPN, the forwarding processing is the same as the MPLS VPN. Otherwise, the C-PE node can make the local decision in choosing the least cost path, including the prior established MPLS paths and IPsec Tunnels, to forward the packet. Packets forwarded over the trusted MPLS VPN can be native without any additional encryption, while the packets sent over the untrusted networks need to be encrypted by IPsec SA.

For a C-PE with multiple WAN ports provided by different ISPs, separate IPsec SAs can be established for the different WAN ports. In this case, the C-PE have multiple IPsec tunnels in addition to the MPLS path to choose from to forward the packets from the client ports.

If the IPsec SA is chosen, the packet is encapsulated by the IPsec inner packet header and encrypted by the IPsec SA before forwarding to the WAN.

For packets received from a MPLS path, processing is the same as MPLS VPN.

For IPsec SA encrypted packets received from the WAN ports, the packets are decrypted, and the inner payload is decapsulated and forward per the forwarding table of the C-PE. For all packets from the Internet-facing WAN ports, the additional anti-DDoS mechanism has to be enabled to prevent potential attacks from the Internet-facing ports. Control Plane should not learn routes from the Internet-facing WAN ports.

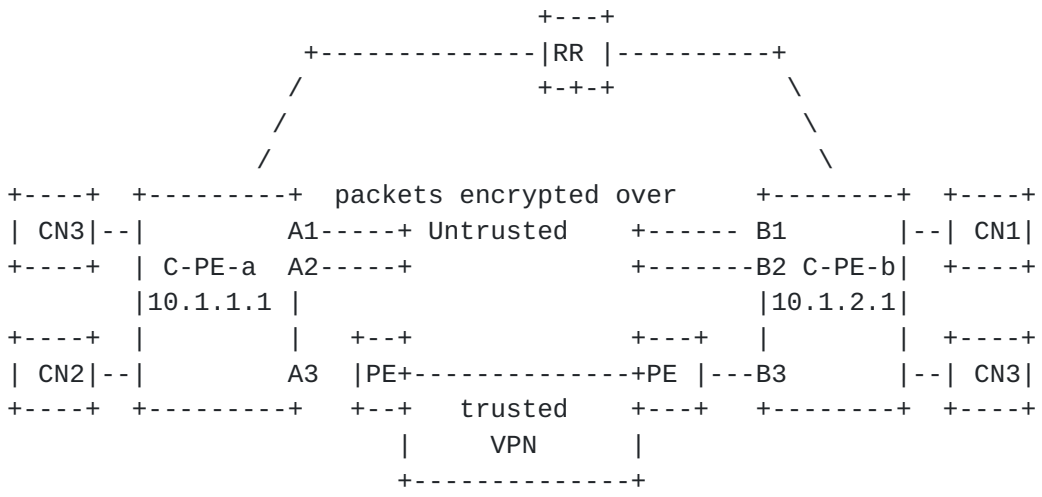


Figure 8: SDWAN with Hybrid Underlays

For multicast packets forwarding:

For multicast traffic, MPLS multicast [RFC6513, RFC6514, or RFC7988] can be used to forward multicast traffic.

If IPsec tunnels are chosen for a multicast packet, the packet is encapsulated and encrypted by multiple separate IPsec tunnels to the desired destinations.

6.3. Forwarding Model for PE based SDWAN

6.3.1. Network and Service Startup Procedures

In this scenario, all PEs have secure interfaces facing the clients and facing the MPLS backbone, with some PEs having extra connections

by untrusted public Internet. The public Internet paths are for offloading low priority traffic when the MPLS paths get congested. The PEs are already connected to their RRs, and the configurations for the clients and policies are already established.

6.3.2. Packet Walk-Through

For PEs to offload some MPLS packets to the Internet path, each MPLS packet is wrapped by an outer IP header as MPLS-in-IP or MPLS-in-GRE [RFC4023]. The outer IP address can be an interface address or the PE's loopback address.

When IPsec Tunnel mode is used to protect an MPLS-in-IP packet, the entire MPLS-in-IP packet is placed after the IPsec tunnel header.

When IPsec transport mode is used to protect the MPLS packet, the MPLS-in-IP packet's IP header becomes the outer IP header of the IPsec packet, followed by an IPsec header, and then followed by the MPLS label stack. The IPsec header must set the payload type to MPLS by using the IP protocol number specified in section 3 of RFC4023.

If IPsec transport mode is applied to an MPLS-in-GRE packet, the GRE header follows the IPsec header.

The IPsec SA's endpoints should not be the client-facing interface addresses unless the traffic to/from those clients always goes through the IPsec SA even when the MPLS backbone has enough capacity to transport the traffic.

When the PEs' Internet-facing ports are behind the NAT [RFC3715], an outer UDP field can be added outside the encrypted payload [RFC3948]. Three UDP ports must be open on the PEs: UDP port 4500 (used for NAT traversal), UDP port 500 (used for IKE), and IP protocol 50 (ESP). IPsec IKE (Internet Key Exchange) between the two PEs would be over the NAT [RFC3947] as well.

Upon receiving a packet from a client port, the forwarding processing is the same as the MPLS VPN. If the MPLS backbone path to the destination is deemed congested, the IPsec tunnel towards the target PEs is used to encrypt the MPLS-in-IP packet.

Upon receiving a packet from the Internet-facing WAN port, the packet is decrypted, and the inner MPLS payload is extracted to be sent to the MPLS VPN engine.

Same as Scenario #2, the additional anti-DDoS mechanism must be enabled to prevent potential attacks from the Internet-facing port.

Control Plane should not learn routes from the Internet-facing WAN ports.

7. Manageability Considerations

BGP-controlled SDWAN utilizes the BGP RR to facilitate the routes and underlay properties distribution among the authorized edge nodes. With RR having the preconfigured policies about the authorized peers, the peer-wise authentications of the IPsec IKE (Internet Key Exchange) are significantly simplified.

8. Security Considerations

Adding an Internet-facing WAN port to a C-PE can introduce the following security risks:

1) Potential DDoS attacks from the Internet-facing ports.

2) Potential risk of provider VPN network being injected with illegal traffic from the Internet-facing WAN ports.

Therefore, the additional anti-DDoS mechanism must be enabled on all Internet-facing ports to prevent potential attacks from those ports. Control Plane should not learn any routes from the Internet-facing WAN ports.

9. IANA Considerations

No Action is needed.

10. References

10.1. Normative References

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

[RFC4364] E. rosen, Y. Rekhter, "BGP/MPLS IP Virtual Private networks (VPNs)", Feb 2006.

[RFC7296] C. Kaufman, et al, "Internet Key Exchange Protocol Version 2 (IKEv2)", Oct 2014.

[RFC7432] A. Sajassi, et al, "BGP MPLS-Based Ethernet VPN", Feb 2015.

[RFC8365] A. Sajassi, et al, "A network Virtualization Overlay Solution Using Ethernet VPN (EVPN)", March 2018.

[RFC9012] K.Patel, et al "The BGP Tunnel Encapsulation Attribute", RFC9012, April 2021.

10.2. Informative References

[RFC8192] S. Hares, et al, "Interface to Network Security Functions (I2NSF) Problem Statement and Use Cases", July 2017

[RFC5521] P. Mohapatra, E. Rosen, "The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", April 2009.

[RFC8388] J. Rabadan, et al, "Usage and Applicability of BGP MPLS-Based Ethernet VPN", May 2018.

[Net2Cloud-Gap] L. Dunbar, A. Malis, C. Jacquenet, "Gap Analysis of Interconnecting Underlay with Cloud Overlay", draft-dm-net2cloud-gap-analysis-02, work in progress, Oct. 2018.

[SDWAN-EDGE-Discovery] L. Dunbar, S. Hares, R. Raszuk, K. Majumdar, "BGP UPDATE for SDWAN Edge Discovery", draft-ietf-idr-sdwan-edge-discovery-05, Aug 2022.

[VPN-over-Internet] E. Rosen, "Provide Secure Layer L3VPNs over Public Infrastructure", draft-rosen-bess-secure-l3vpn-00, work-in-progress, July 2018

[DMVPN] Dynamic Multi-point VPN: https://www.cisco.com/c/en/us/products/security/dynamic-multipoint-vpn-dmvpn/index.html

[DSVPN] Dynamic Smart VPN: http://forum.huawei.com/enterprise/en/thread-390771-1-1.html

[SECURE-EVPN] A. Sajassi, et al, "Secure EVPN", draft-sajassi-bess-secure-evpn-01, Work-in-progress, March 2019.

[SECURE-L3VPN] E. Rosen, R. Bonica, "Secure Layer L3VPN over Public Infrastructure", draft-rosen-bess-secure-l3vpn-00, Work-in-progress, June 2018.

[ITU-T-X1036] ITU-T Recommendation X.1036, "Framework for creation, storage, distribution and enforcement of policies for network security", Nov 2007.

[Net2Cloud-Problem] L. Dunbar and A. Malis, "Seamless Interconnect Underlay to Cloud Overlay Problem Statement", draft-rtgwg-net2cloud-problem-statement-12, March 2022

[Net2Cloud-gap] L. Dunbar, A. Malis, and C. Jacquenet, "Gap Analysis of Interconnecting Underlay with Cloud Overlay", draft-rtgwg-net2cloud-gap-analysis-07, work-in-progress, July 2020.

11. Acknowledgments

Acknowledgements to Adrian Farrel, Joel Halpern, John Scudder, Darren Dukes, Andy Malis, Donald Eastlake, and Victo Sheng for their review and contributions.

This document was prepared using 2-Word-v2.0.template.dot.

Authors' Addresses

Linda Dunbar
Futurewei
Email: ldunbar@futurewei.com

James Guichard
Futurewei
Email: james.n.guichard@futurewei.com

Ali Sajassi
Cisco
Email: sajassi@cisco.com

John Drake
Juniper
Email: jdrake@juniper.net

Basil Najem
Bell Canada
Email: basil.najem@bell.ca

David Carrel
IPsec Research
Email: carrel@ipsec.org

Ayan Banerjee
Cisco
Email: ayabaner@cisco.com