Network Working Group
Internet Draft
Intended status: Standard Track
Expires: May 19, 2026

K. Majumdar
Oracle
L. Dunbar
Futurewei
V.Kasiviswanathan
NextHop AI
A. Ramchandra
Google
A. Choudhary
Cisco

November 19, 2025

Multi-segment SD-WAN via Cloud DCs
draft-ietf-rtgwg-multisegment-sdwan-11

Abstract

This document describes a method for seamlessly interconnecting geographically separated SD-WAN segments via a Cloud Backbone without requiring Cloud Gateways (GWs) to decrypt and re-encrypt traffic. By encapsulating IPsec-encrypted payloads within GENEVE headers (RFC 8926), the approach enables Cloud GWs to forward encrypted traffic directly between distant Customer Premises Equipment (CPEs). This reduces processing overhead, improves scalability, and preserves the confidentiality of enterprise data while ensuring secure and efficient multi-segment SD-WAN. connectivity.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

This Internet-Draft will expire on May 19, 2026 .

Copyright Notice

Table of Contents

1. Introduction

Enterprises are increasingly turning to SD-WAN to connect on-premises CPEs with cloud services, as discussed in detail in [Net2Cloud]. Each SD-WAN segment typically connects a CPE to its nearest Cloud Gateway (GW). Some of this traffic terminates at the cloud services and must be decrypted by the Cloud GW. Other traffic is destined for remote CPEs located in different geographic regions and only require forwarding across a Cloud Backbone, without decryption.

Multi-segment SD-WAN refers to the architecture in which two or more SD-WAN segments are interconnected via a Cloud Backbone. This model enables traffic that originates in one SD-WAN segment to reach a distant CPE through transit Cloud GWs without decryption. It supports hybrid traffic handling: local cloud-bound traffic is decrypted by the Cloud GW, while CPE-to-CPE traffic is forwarded securely across the backbone.

Interconnecting these SD-WAN segments via a Cloud Backbone provides several key benefits:

a) Seamless connectivity - Enterprises can integrate geographically dispersed SD-WAN segments into a unified network without complex manual configurations.
b) Scalability - The Cloud Backbone's elasticity accommodates increased traffic demands without requiring extensive on-premises infrastructure.
c) Simplified operations - Centralized orchestration streamlines policy enforcement and network management across all segments.

The challenges and motivations for this architecture are further detailed in [Net2Cloud], which outlines issues enterprises face when interconnecting branch sites with dynamic workloads in third-party Cloud DCs, particularly when leveraging existing VPN infrastructure.

A key requirement in Cloud Backbone stitching SD-WAN segments is the ability to forward encrypted traffic across the Cloud Backbone without requiring decryption at Cloud GWs. Since IPsec Security Associations (SAs) are established end-to-end between CPEs, Cloud GWs cannot access the payload for routing. Introducing an additional IPsec tunnel layer between CPE and Cloud GW just for routing purposes is inefficient-it adds processing overhead, increases latency due to decryption and re-encryption, and imposes scalability limits due to cloud provider restrictions on IPsec capacity per GW instance.

This document defines a GENEVE-based method that avoids these inefficiencies. SD-WAN CPEs encapsulate IPsec-encrypted packets with GENEVE headers [RFC8926] that include Sub-TLVs to signal when traffic should transit the Cloud Backbone without decryption. This enables Cloud GWs to forward encrypted traffic efficiently to remote CPEs, without accessing the payload. The result is secure, low-latency, and scalable interconnection of geographically distributed SD-WAN segments using the Cloud Backbone.

2. Conventions used in this document

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 BCP14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

The following acronyms and terms are used in this document:

Cloud Backbone: The global, private network infrastructure operated by a cloud provider that interconnects its regions, zones, and points of presence.

Cloud DC: Off-Premises Data Center, managed by the third party, that hosts applications, services, and workload for different organizations or tenants.

CPE: Customer (Edge) Premises Equipment.

OnPrem: On Premises data centers and branch offices.

RR Route Reflector.

SA IPsec Security Association

SD-WAN An overlay connectivity service that optimizes transport of IP Packets over one or more Underlay Connectivity Services and determining forwarding behavior by applying Policies to them. [MEF-70.1]

VPN Virtual Private Network.

3. Use Cases

3.1. Multi-segment SD-WAN via a Single Cloud GW

Enterprise branches with established SD-WAN paths to a Cloud GW for accessing cloud services can also use the Cloud GW to interconnect with one another, as shown in Figure 1. Stitching SD-WAN segments through a Cloud Gateway provides a way to extend policy enforcement and traffic control across branches, particularly when direct branch-to-branch paths over the public internet are insufficient. This approach is beneficial for several reasons:

- The public internet between branches may suffer from limited bandwidth, unpredictable performance, and security risks.
- Centralized enforcement of enterprise security policies can be enabled through cloud-hosted services. Traffic destined to cloud-resident applications can be decrypted for full inspection (e.g., firewall, threat detection), while CPE-to-CPE traffic that remains IPsec-encrypted can still benefit from header- or flow-based functions-such as DDoS mitigation, rate limiting, anomaly detection, and SLA/usage analytics-especially when the same CPE also sends traffic terminating in the cloud.
- Cloud platforms often offer enhanced monitoring, proprietary threat detection tools, and analytics services that can inspect and respond to suspicious traffic crossing segments.

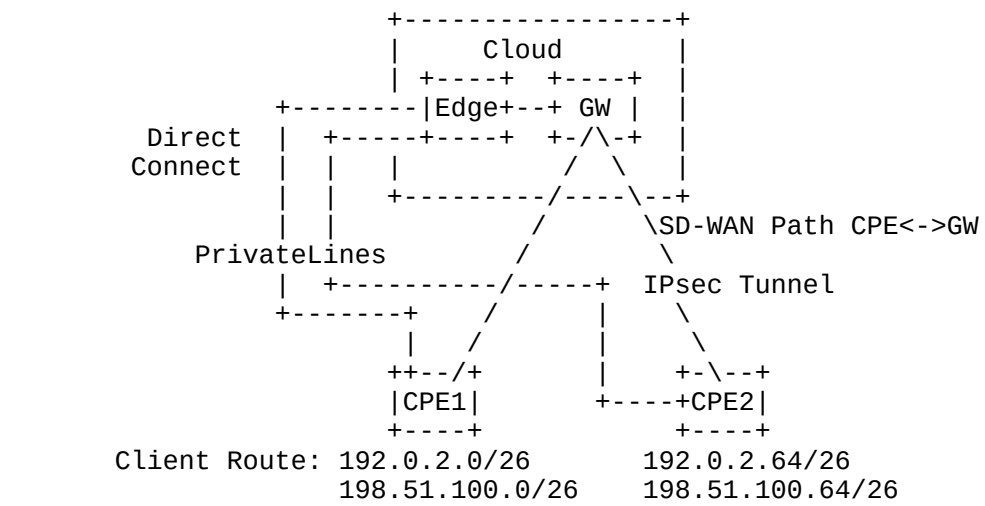


Figure 1 Multi-Segment SD-WAN stitching via a Cloud GW
Note: For clarity, each line in this figure represents connectivity that may consist of multiple parallel paths. Multiple paths are not shown to avoid excessive complexity in the illustration.

3.2. Multi-segment SD-WAN via Cloud Backbone

For geographically distant enterprise branches that have established SD-WAN paths to their respective Cloud GWs for accessing cloud services, the Cloud Backbone provides an efficient way to interconnect these branches, as shown in Figure 2. As outlined in the Introduction section, this approach enhances network integration, supports dynamic scaling, and simplifies overall management, making it well-suited for multi-segment SD-WAN deployments across different regions.

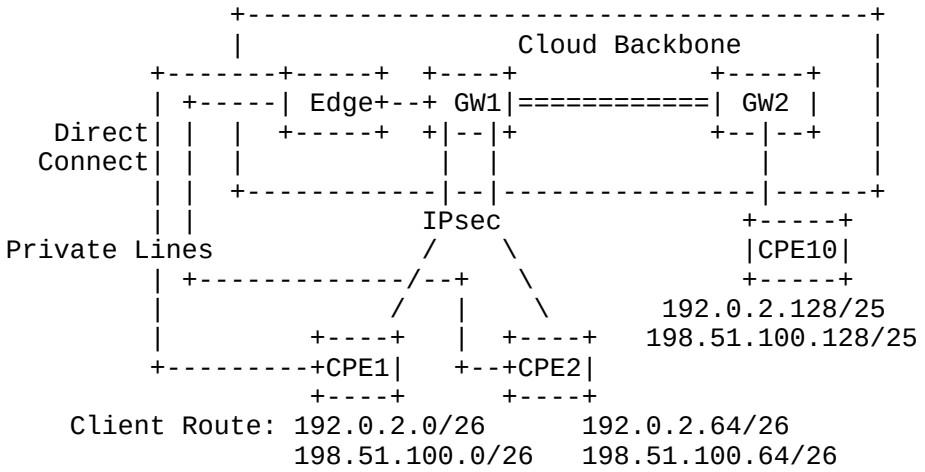


Figure 2 Multi-Segment SD-WAN Stitching via Cloud Backbone

3.3. Traffic Steering Challenges in Multi-Segment SD-WAN

Many well-established traffic engineering methods, such as SRv6 and MPLS-TE, effectively steer traffic through specific network nodes when the entire network operates under a single administrative domain.

However, in typical SD-WAN deployments, CPE-to-CPE traffic is carried as best-effort over the public Internet or other shared transport. Forwarding in the underlay is destination-based, and the on-premises CPEs cannot directly control the specific path that packets take. This limits the ability to enforce precise traffic engineering (TE) to reach destination CPEs.

This lack of predictable routing makes traffic steering between branch offices highly challenging. Unlike private MPLS networks or provider-controlled backbones, SD-WAN cannot inherently dictate the intermediate paths for branch-to-branch traffic. As a result, policies intended to optimize performance, enforce security, or ensure compliance can be difficult to implement.

To address this issue, this document describes a method where Cloud GWs explicitly interconnect SD-WAN segments, ensuring that branch-to-branch traffic is steered through the Cloud Backbone rather than taking unpredictable internet routes. This approach provides greater control over traffic flows, improving reliability, security, and policy enforcement.

Note: The mechanism described in this document does not alter the forwarding behavior of the underlay network. Traffic from the source CPE to the ingress Cloud GW and from the egress Cloud GW to the destination CPE continues to follow normal underlay forwarding. Since these are typically short hops, the more useful traffic engineering (TE) occurs across the longer-range Cloud Backbone. In this model, the overlay steering defined here enables predictable selection of ingress and egress Cloud GWs, while TE within the backbone is offloaded to the Cloud Backbone provider

4. Data Plane encoding for SD-WAN Transit

To enable Cloud GWs to distinguish between packets requiring decryption for internal cloud services and transit packets that should be forwarded to destination CPEs, proper packet marking is essential. Many encapsulation methods, such as VLAN tags, IP-in-IP, GRE, etc., can be used to steer traffic from a CPE to its nearest (or chosen) Cloud GW. However, GENEVE encapsulation [RFC8926] offers significant advantages, including flexible option Sub-TLVs that can signal routing and policy preferences, such as Restricted Regions, Exclude Regions, preferred egress Cloud GWs, and other service specific requirements. In addition, GENEVE Encapsulation [RFC8926] is widely supported by major Cloud Service Providers, which allows Cloud GWs to efficiently steer IPsec-encrypted packets between CPEs via Cloud Backbone without decryption, reducing processing overhead and improving performance while maintaining end-to-end encryption.

4.1. Multi-Segment SD-WAN Option Class

Geneve header format is specified in Section 3 of [RFC8926]. This document uses the GENEVE Option Class value 0x0163, which has been assigned by IANA to identify Multi-Segment SD-

WAN-specific Sub-TLVs encoded within the GENEVE header. This enables Cloud GWs to interpret and process SD-WAN transit packets efficiently without requiring decryption.

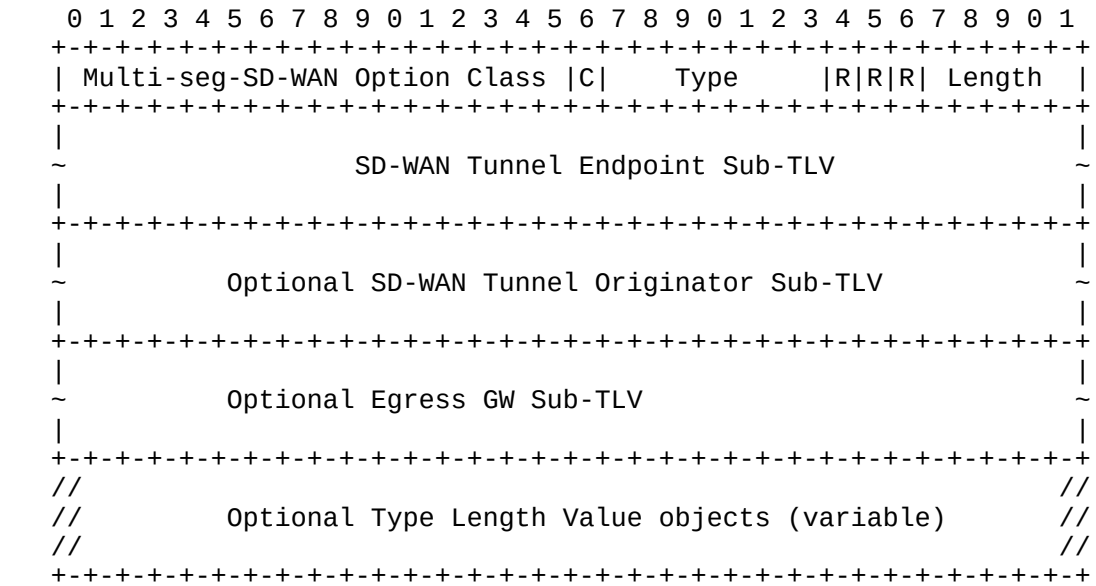


Figure 3 Multi Segment SD-WAN Option Class

- Multi-seg-SD-WAN Option Class: value 0x0163 (assigned by IANA).
- C-bit: Must be set to ensure that a receiving node drops the packet if it does not recognize the option, as per [RFC8926].
- Type (8 bits): Specifies the multi-segment SD-WAN forwarding model:
  Type = 1: Single-hop transit SD-WAN

  Type = 2: Multi-Hop transit SD-WAN with an explicitly specified egress Cloud GW (via Egress GW Sub-TLV).

  Type = 3: Multi-hop transit SD-WAN without an explicitly specified egress Cloud GW.

- Length (5 bits): Indicates the total length of the option fields in 4-byte units. If no options are present, this field is zero [RFC8926].

Note: the payload following the multi-seg-SD-WAN Option Class can be IPv4 or IPv6. The Protocol Type of the GENEVE header is set to 50, indicating the GENEVE payload carries IPsec ESP [RFC8926][IPsecOverGENEVE].

4.2. SD-WAN Tunnel Endpoint Sub-TLV

The SD-WAN Endpoint sub-TLV indicates the destination CPE, which is the endpoint of the IPsec Tunnel between branch CPEs. This Sub-TLV is used by the Cloud Backbone to determine the optimal egress Cloud GW for forwarding the encrypted traffic.

For example, in an SD-WAN deployment where CPE1 establishes an IPsec SA with CPE2 (as shown in Figure 1), this Sub-TLV within the GENEVE header contains CPE2's IP address, ensuring that encrypted traffic is correctly routed to the terminating CPE of the IPsec tunnel while enabling the Cloud Backbone to steer the packet to the most suitable egress Cloud GW.

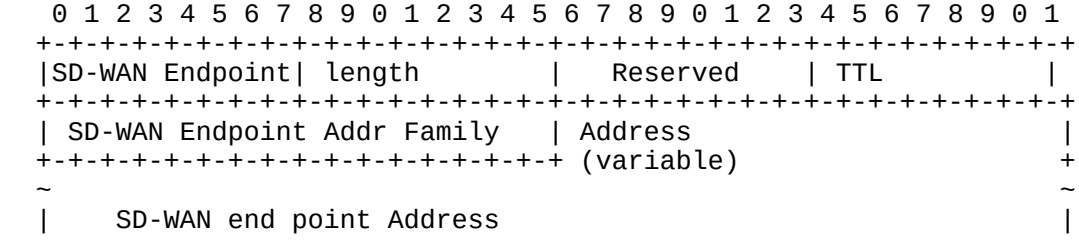


Figure 4 SD-WAN Endpoint Sub-TLV

- SD-WAN Endpoint (8 bits): Identifies the SD-WAN Tunnel Endpoint Sub-TLV with a Type value of 1.

- Length (8 bits): Specifies the total length of the value field in 4-byte units.

- TTL (Time to Live): This field is set by the originating CPE to indicate the maximum number of logical transit nodes or regions, those that are visible to the CPEs, that a packet is permitted to traverse across the Cloud Backbone. Only transit nodes or regions that are externally visible (i.e., known to or tracked by the CPEs) MUST decrement the TTL by one. Internal cloud forwarding elements that are opaque to the CPEs MUST NOT modify the TTL. If the TTL reaches zero, the packet MUST be dropped, and an alert MAY be generated. This mechanism allows enterprises to constrain the path scope of their packets, enforce traversal policies, and detect anomalies (e.g., excessive transit hops).

- SD-WAN Dst Addr Family (16 bits): Identifies the address family of the destination endpoint. Values follow the Address Family Numbers registry. For example, a value of 1 indicates an IPv4 address and a value of 2 indicates an IPv6 address.

4.3. SD-WAN Tunnel Originator Sub-TLV

The SD-WAN Tunnel Originator Sub-TLV is an optional Sub-TLV within the multi-seg-SD-WAN Option Class to indicate the originating CPE of the IPsec Tunnel.

For example, in an SD-WAN deployment where CPE1 establishes an IPsec SA with CPE2 (as shown in Figure 1), this Sub-TLV within the GENEVE header carries CPE1’s address, allowing transit nodes and Cloud GWs to recognize the source of the encrypted traffic.

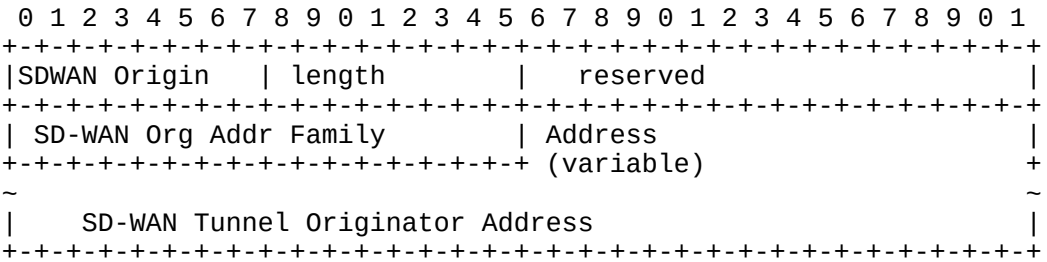


Figure 5 SD-WAN Tunnel Originator Sub-TLV

- SDWAN Origin (8 bits): Identifies the SDWAN Tunnel Originator Sub-TLV with a Type value of 2.
- Length (8 bits): Specifies the total length of the value field in 4-byte units, excluding the first 4 bytes, which include the SD-WAN Origin (1 byte), Length (1 byte), and Reserved (2 bytes) fields.
- Reserved (16 bits): Reserved for future. Must set to 0. Ignored by recipients.
- SD-WAN Org Addr Family (16 bits): Identifies the family address of the originator. A value of 1 indicates an IPv4 address and a value of 2 indicates an IPv6 address.

This Sub-TLV allows Cloud GWs and transit nodes to identify the packet's source, allowing them to apply source specific policies for forwarding. These policies may include traffic engineering rules specific to the originating CPE, security enforcement tailored to the source, or path selection constraints based on the origin.

4.4. Egress GW Sub-TLV

In a multi-segment SD-WAN deployment over the Cloud Backbone, the originating CPE can use the Egress GW Sub-TLV to explicitly specify the egress Cloud GW responsible for forwarding traffic to the destination CPE. This ensures predictable routing behavior and enables policy-driven packet delivery across the Cloud Backbone.

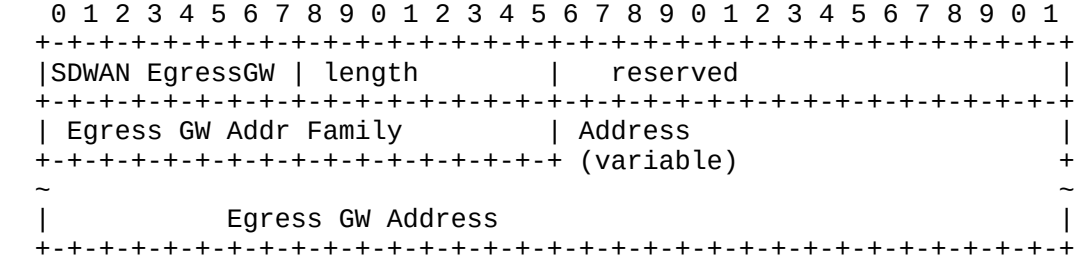


Figure 6 SD-WAN Egress GW Sub-TLV

- SDWAN EgressGW (8 bits): Identifies Egress GW Sub-TLV with a Type value of 3.

- Length (8 bits): Specifies the total length of the value field in 4-byte units, excluding the first 4 bytes, which include the SD-WAN Origin Sub-TLV Type (1 byte), Length (1 byte), and Reserved (2 bytes) fields.
- Reserved (16 bits): Reserved for future. Must set to 0. Ignored by recipients.
- Egress GW Addr Family: Identifies the family address of the Egress GW. A value of 1 indicates an IPv4 address and a value of 2 indicates an IPv6 address.

The Egress GW Sub-TLV allows the originating CPE to specify the Egress Cloud GW responsible for forwarding traffic to the destination CPE. This Egress GW address can be either preconfigured or dynamically discovered through a control plane protocol exchange with the destination CPE. By explicitly defining the egress GW, this Sub-TLV ensures predictable traffic steering, reducing reliance on destination-based routing and optimizing packet delivery across the Cloud Backbone. The details of the control plane protocol used for GW discovery are beyond the scope of this document.

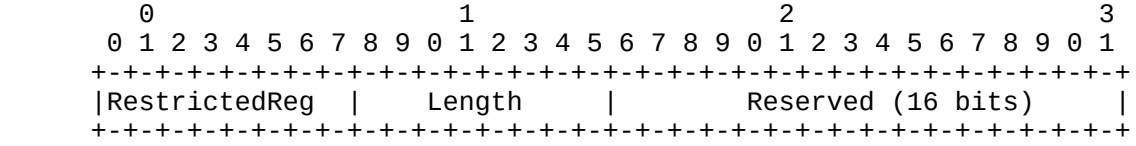
4.5. Restricted Regions Sub-TLV

Some enterprises may require that traffic across the Cloud Backbone is strictly confined to a specific set of regions. This Sub-TLV allows the ingress SD-WAN CPE to express such restrictions as part of the encapsulation metadata.

Traffic MUST be discarded if the Ingress Gateway, the Egress Gateway, or any transit node belongs to a region not listed in this Sub-TLV.

This restriction is commonly used to enforce regulatory, security, or latency-based geographic constraints, where data must remain confined to specified regions.

Format of the Restricted Regions Sub-TLV:



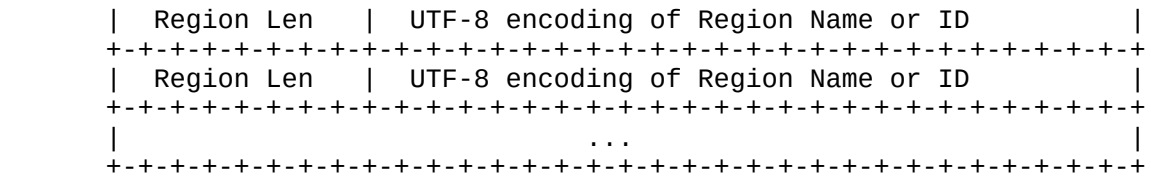


Figure 7 Restricted Regions Sub-TLV

- RestrictedReg (8 bits): Identifies the Restricted Regions Sub-TLV with a Type value of 4.
- Length (8 bits): Total length of the Value field (everything after the Type and Length fields), in octets.
- Reserved (16 bits): Reserved for future use. MUST be set to zero and ignored on receipt.
- Region Len (8 bits per region entry): Length of the UTF-8 encoding of the Region Name or identifier, in octets.
- UTF-8 encoding of the Region Name (e.g., "us-west", "eu-central") or numeric identifier.

Multiple regions MAY be present, each starting with its own Region Len field.

Processing notes:

- Receiving Cloud Gateway MUST check whether it and all intermediate transit regions are included in the listed regions.
- If any component of the path falls outside the listed regions, the packet MUST be discarded.
- Region interpretation is based on prior agreement between the enterprise and the Cloud Backbone provider (e.g., standard region names, operator-specific definitions, or standardized Region IDs).

Note:
It is beyond the scope of this document to specify how the Cloud Backbone enforces this restriction. Mechanisms for identifying region boundaries, enforcing region-based constraints, and generating alerts or alarm notifications when traffic violates region restrictions are subject to implementation decisions and based on prior agreement between the Cloud Backbone provider and the enterprise.

4.6. Exclude Transit Sub-TLV

Exclude Transit Sub-TLV is an optional field used to specify a list of Cloud Availability Regions, Zones, or Notes that must be avoided when forwarding packets across the Cloud Backbone. This can be used for:

- Regulatory compliance, ensuring traffic does not traverse restricted or non-compliant regions.
- Risk mitigation, preventing traffic from passing through regions with known security, performance, or geopolitical concerns.

Multiple region entries MAY be specified in a single Sub-TLV. Each region is identified by a variable length UTF-8 encoded name or numeric ID, preceded by a length field. This Sub-TLV expresses explicit exclusions and supports both soft and hard enforcement.

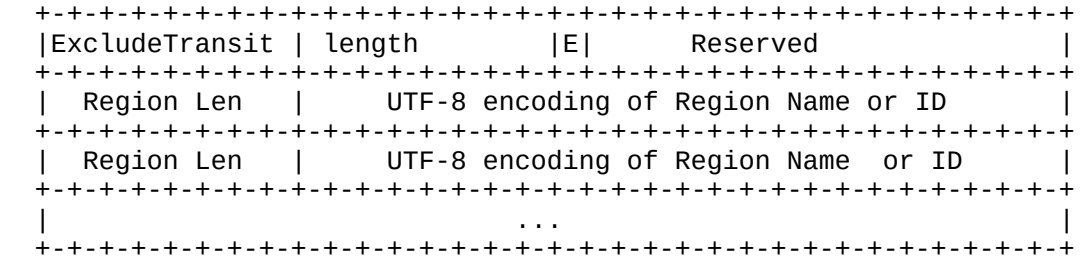


Figure 8 Exclude Transit Sub-TLV

- ExcludeTransit (8 bits): identifies the Exclude Transit Sub-TLV with a Type value of 5.
- Length (8 bits): Total length of the Value field in octets (everything after the first 2 bytes).
- E-bit (1 bit) - Exclusion severity indicator:
  - o 0: Soft exclusion - the listed region is undesirable; avoid when feasible.
  - o 1: Hard exclusion - the region MUST be avoided; if unavoidable, generate an alert or alarm.

- Reserved (15 bits): MUST be set to zero on transmission and ignored on receipt.
- Region Len (8 bits per region entry): Length of the UTF-8 encoding of the Region Name or identifier, in octets.
- UTF-8 encoding of the Region Name (e.g., "us-west", "eu-central") or numeric identifier.

Multiple region entries MAY be listed, each beginning with a Region Len byte.

Processing Notes:

The E-bit determines how strictly the exclusions are enforced. A value of 1 (hard exclusion) mandates the Cloud Backbone to drop the packet or raise an alert if the excluded region is traversed. A value of 0 allows best-effort avoidance without enforcement or notification. The meaning and granularity of region identifiers MUST be agreed upon between the enterprise and the Cloud Backbone provider (e.g., standardized names, or operator-defined zones). It is beyond the scope of this document to define how enforcement or alerting is implemented. These are subject to operator policies and implementation specifics.

5. Packet Header Processing

The procedures described in this section apply only to packets that carry the SD-WAN Option Class in the GENEVE header. Packets without this option are processed using default forwarding behavior.

As illustrated in Figure 1, when Cloud GW receives a GENEVE-encapsulated packet (i.e. Dst Port = 6081 (GENEVE); MultiSeg-SDWAN Option Class; inner IP header’s Protocol Type = 50 (ESP)), it processes the packet as follows:

Processing at the Ingress Cloud GW:

- Authenticate the packet using a preconfigured authentication method.
- Check if the Egress GW Sub-TLV is present:

    - If the Egress GW Sub-TLV exists, the Cloud Backbone uses it to identify the Egress Cloud GW.
    - If the Egress GW Sub-TLV is not present, the Cloud Backbone determines the optimal egress Cloud GW based on the destination CPE address.
  - Change the destination address in the outer IP header of the GENEVE packet to the address determined by the Cloud Backbone. This address is intended to reach the Egress Cloud GW identified by the Egress GW Sub-TLV (if present), or the optimal egress GW selected based on the destination CPE address.
  - Forward the packet to the egress Cloud GW.

To prevent unauthorized access, Cloud GW SHOULD drop any packets containing unrecognized source addresses or invalid values in the GENEVE Sub-TLVs, ensuring that only registered entities can utilize Cloud services.

Processing at the Egress Cloud GW:

- Decapsulate the GENEVE header to extract the IPsec-encrypted payload.
- Validate that the SD-WAN Tunnel Endpoint Sub-TLV corresponds to a registered destination CPE.
- Ensure the source Cloud GW is an authorized forwarding node to prevent unauthorized traffic injection.
- Forward the IPsec-encrypted payload to the destination CPE, preserving the end-to-end encryption.
- Drop any packet that lacks a valid destination CPE or originates from an untrusted source.

By enforcing these processing steps at both the ingress and egress Cloud GWs, the system ensures secure, efficient, and policy-compliant forwarding of SD-WAN traffic across the Cloud Backbone.

6. Error Handling

To ensure secure and efficient traffic forwarding through the Cloud Backbone, Cloud GW SHOULD enforce the following error handling measures:

- Drop packets with unregistered or invalid source/destination addresses to prevent unauthorized access.
- Reject packets originating from unpaid or unregistered CPEs to enforce service subscription policies.
- Validate the SD-WAN Endpoint Sub-TLV and drop packets if the destination CPE is unauthorized, unreachable, or mismatched.
- Discard malformed packets with incorrect GENEVE headers, invalid Sub-TLV formats, or authentication failures.
- Drop packets with expired TTL values to prevent routing loops and log repeated occurrences.
- Reject misrouted packets if the Cloud Backbone cannot determine an optimal egress Cloud GW or if the specified egress GW is unreachable.
- Enforce rate limits on excessive traffic from a single source to prevent congestion and abuse.
- Verify compliance with transit node policies (e.g., ensuring mandatory transit nodes are included and excluded nodes are avoided).
- Mitigate replay attacks by tracking sequence numbers and rejecting duplicate packets.

By implementing these error handling mechanisms, Cloud GWs ensure network stability, security, and efficient resource utilization while preventing misconfigurations, abuse, and performance degradation.

7. Control Plane considerations

7.1. Control Plane for CPEs

The control plane enables SD-WAN CPEs to discover their network attributes, establish connectivity, and exchange routing information. In an SD-WAN deployment, on-premises

CPEs and virtual CPEs (vCPEs) in Cloud DCs may be managed under a common iBGP administrative domain, facilitating route propagation and policy enforcement.

Mechanisms such as BGP-based SD-WAN Edge Discovery [SD-WAN-Edge-Discovery] allow CPEs to dynamically discover each other's properties, improving automation and reducing manual configurations. Additionally, IPsec SAs parameters between CPEs and Cloud GWs can be exchanged through the iBGP control plane using a RR to simplify security policy management.

The iBGP control plane is used to exchange reachability and policy information among CPEs through Route Reflectors; it does not carry IPsec Security Association (SA) parameters, which are established separately via IKEv2 or out-of-band management systems.

Further details on the control plane between CPEs and Cloud Gateways (CGs) are described in Section 7.2.

7.2. Control Plane between CPEs and Cloud GWs

There are typically eBGP sessions between a CPE and a Cloud GW for exchanging routing information related to services that terminate within the cloud. This allows the CPE to learn routes to cloud-hosted resources and enables the Cloud GW to learn routes to the CPE's on-premises networks. This control-plane relationship is separate from the CPE-to-CPE encrypted traffic that transits the Cloud Backbone, which remains end-to-end encrypted and is not decrypted at the Cloud GWs.

When the connection between a CPE and a Cloud GW traverses a public or otherwise untrusted network, an IPsec tunnel may also be established to secure that traffic. In such cases, the IPsec Security Association (SA) parameters between the CPE and its corresponding Cloud GW are established out-of-band (e.g., via management or automation systems) or negotiated dynamically using IKEv2.

Control plane mechanisms must ensure that Cloud GWs can identify and authenticate SD-WAN CPEs, validate SD-WAN metadata, and apply appropriate routing policies based on dynamic network conditions. This ensures that route exchanges are trustworthy, policy-compliant, and adaptive to changing operational requirements.

8. Observability Consideration

Observability considerations encompass monitoring, analysis, and reporting mechanisms to gain insights into the behavior and performance of the multi-segment SD-WAN infrastructure. Key observability aspects include:

- Performance Metrics:
  Monitor and collect performance metrics related to link utilization, latency, and packet loss across the SD-WAN segments and Cloud DC backbone. This data provides insights into the overall health and efficiency of the network. IP Flow Information Export (IPFIX) [RFC7011] is one of the standardized methods to expose traffic flow over the network.

- Global Network Topology Visualization:
  Utilize visualization tools to depict the global network topology, showcasing the interconnections and traffic flows between different SD-WAN segments and Cloud DCs.

- Control Plane Monitoring:
  Monitor the control plane for both CPEs and the communication between CPEs and Cloud GWs. This includes tracking route discovery, path selection, and any changes in network state to ensure proper functioning of the SD-WAN control plane.

- Security Event Logging:
  The security event logging is to capture and analyze security-related events, including threat detection, authentication failures, and any unauthorized access attempts. Syslog [RFC5424] is a valuable tool for security monitoring and auditing.

These considerations contribute to the overall success of the multi-segment SD-WAN deployment connecting edge devices via a Cloud DC backbone.

9. Security Considerations

9.1. Threat Analysis

The GENEVE header used for steering is not encrypted, making it susceptible to man-in-the-middle (MitM) attacks between CPEs and Cloud GWs.

Key risks include:

a) Eavesdropping: Attackers can learn branch and Cloud GW locations, though payload remains protected by IPsec.
b) Header Manipulation: Altered Sub-TLVs may cause misrouting or packet drops.
c) Bandwidth Theft: A malicious or misconfigured CPE could spoof SD-WAN metadata to use Cloud Backbone resources without authorization.

Mitigation above risks requires authenticating and validating SD-WAN metadata to ensure it originates from authorized CPEs.

9.2. HMAC-based Integrity and Authentication

HMAC (Hash-based Message Authentication Code), a widely used cryptographic technique for ensuring both data integrity and authentication, can be used to ensure the integrity and authenticity of the GENEVE header between SD-WAN CPEs and Cloud GWs, protecting it from tampering. HMAC combines a shared secret key with a hash function to produce a fixed-size authentication value, which is appended to the packet. The receiver computes the HMAC over the received header and compares it with the transmitted value; a match confirms that the header has not been altered, provided the key remains secret.

This mechanism is scoped to communication between SD-WAN CPEs and Cloud GWs, with the shared key provisioned through a secure channel. For CPE-to-CPE traffic that only transits the Cloud Backbone, the HMAC key can be derived from the existing IPsec SAs established between each CPE and the Cloud GW, using a standard Key Derivation Functions (KDFs)[RFC5869]. This avoids the need for a new peer-to-peer IKEv2 exchange between CPEs, and because the key distribution occurs over the existing IPsec-protected CPE-GW channels, NAT traversal does not pose an issue. [lightweight-authenticate] describes a simplified method for applying HMAC to selected packets.

To reduce packet overhead, truncated HMAC values of 4 or 8 bytes are RECOMMENDED instead of full-length outputs (e.g., 32 bytes for HMAC-SHA-256). This is acceptable because the IPsec tunnel already protects the payload, and the HMAC only secures the steering metadata.

The HMAC value is carried in the HMAC-Auth-Val Sub-TLV in the GENEVE header:

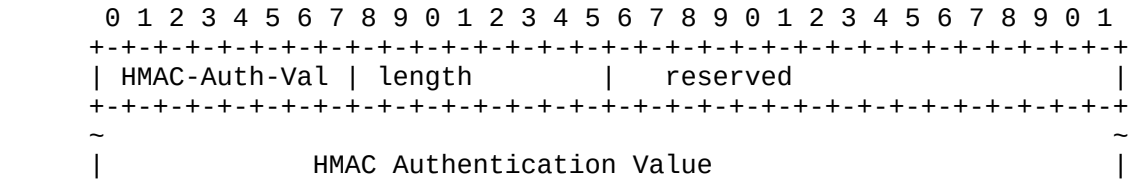


Figure 9 Multi Segment SD-WAN HMAC Sub-TLV

- HMAC-Auth-Val (8 bits): HMAC Authentication Value Sub-TLV Type = 6 (Assigned by this document).
- Length (8 bits): Total length of the value field, which is the length of the HMAC Authentication Value in bytes plus 2 reserved bytes. It is 6 bytes by default for a 4-byte HMAC. In deployments with higher security requirements, an 8-byte HMAC (total of 10 bytes) is RECOMMENDED.
- The HMAC Authentication Value (4 bytes or 8 bytes): Computed over the entire GENEVE header (excluding this Sub-TLV) using a pre-configured algorithm such as HMAC-SHA-256 and the shared key.

The advantages of using HMAC are:

- Data Integrity: Protects steering metadata from modification.
- Efficiency: Truncated values minimize overhead while retaining strong protection.
- Resistance to Tampering: Even truncated, HMAC values resist message tampering and replay attacks.
- Flexibility: Compatible with various hash functions like SHA-256 or SHA-512.
- Widely Supported: Mature and broadly implemented across platforms.

While truncated HMACs reduce collision resistance compared to full-length values, this is an acceptable tradeoff because the payload is encrypted by IPsec SAs, the HMAC covers only steering metadata, and attackers must possess the shared key to generate valid values.

9.3. AH based Integrity and Authentication

Some deployments may require stronger or more comprehensive integrity protection than a truncated HMAC, such as when mandated by security policy, regulatory compliance, or risk management practices. In these cases, an additional integrity layer can be applied using Authentication Header (AH) [RFC4301] or ESP-NULL [RFC2410] [RFC6071] on top of the existing IPsec encryption between CPEs.

AH and ESP-NULL provide cryptographic integrity for the entire IP packet, not just the GENEVE metadata. All approaches (including the HMAC) require cryptographic keys. The operational difference is that AH/ESP-NULL require dedicated IPsec SAs and IKE state between each Cloud GW and CPE, increasing per-peer state and processing. By contrast, the HMAC Sub-TLV (Type = 6, defined in this document) can use controller-distributed symmetric keys (e.g., per-tenant or per-CPE) without establishing additional IPsec SAs between Cloud GWs and CPEs.

NAT Considerations: AH is not compatible with NAT traversal because it authenticates the outer IP header, and any address change will cause verification to fail. ESP-NULL avoids this issue but still incurs additional per-packet processing.

10. Manageability Considerations

In multi-segment SD-WAN deployments where the Cloud GW and CPEs belong to different administrative domains, manageability must address the challenges of secure, interoperable, and policy-compliant operation across organizational boundaries, consistent with the service framework defined in MEF 70.1 [MEF70.1]. Key considerations include:

- Cross-Domain Authentication and Authorization:
  Ensure that CPEs connecting to the Cloud GW are authenticated using mutually agreed methods, and that authorization policies are enforced to prevent unauthorized use of Cloud Backbone resources.

- Metadata Validation and Policy Enforcement:
  Cloud GWs must validate SD-WAN metadata (e.g., GENEVE Sub-TLVs) against the registered information for each

CPE. This prevents spoofing, misrouting, and cross-tenant traffic leakage.

- Operational Coordination and Fault Handling:
  Define inter-organization procedures for troubleshooting and incident response. This should include point-of-contact directories, escalation processes, and shared logging formats for event correlation.

- Coordination of Configuration Changes:
  Coordinate configuration changes-such as policy updates, region restrictions, or authentication parameters-so that both the Cloud GW and CPEs apply them consistently, avoiding mismatches that disrupt traffic.

- Policy Automation Using I2NSF Principles ([RFC8192]):
  Where feasible, leverage I2NSF concepts to automate policy configuration, exchange, and enforcement between domains, reducing manual coordination and improving operational consistency.

11. IANA Considerations

IANA has assigned a new GENEVE Option Class from the IETF Review range as shown below:

Option Class	Description	Assignee/Contact	Reference
0x0163	Multi Segment SD-WAN	IETF	[this document]

IANA has assigned GENEVE Option Class value 0x0163 for identifying Multi-Segment SD-WAN. No further Option Class assignments are requested in this document.

IANA is requested to create the following new registry under the "Multi Segment SD-WAN GENEVE Option Class (0x0163):

Registry: Multi Segment SD-WAN Sub-TLVs
Assignment Policy: IETF Review
Reference: [this document]

Sub-TLV Type	Description	Reference

0	Reserved	
1	SD-WAN Endpoint	[Section 4.2]
2	SD-WAN Originator	[Section 4.3]
3	SD-WAN Egress GW	[Section 4.4]
4	Restricted Region	[Section 4.5]
5	Exclude Transit	[Section 4.6]
6	Multi SD-WAN-HMAC	[Section 9.2]
5-254	Unassigned	
255	Reserved	

12. References

12.1. Normative References

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

[RFC2403] C. Madson, R. Glenn, "The Use of HMAC-MD5-96 within ESP and AH", RFC2403, Nov. 1998.

[RFC2404] C. Madson, R. Glenn, "The Use of HMAC-SHA-1-96 within ESP and AH", RFC2404, Nov. 1998.

[RFC4301] S. Kent and K. Seo, "Security Architecture for the Internet Protocol", RFC4301, Dec. 2005.

[RFC5424] R. Gerhards, "The Syslog Protocol", RFC5424, March 2009.

[RFC7011] B. Claise, B. Trammell, and P. Aitken, "Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of Flow Information", RFC7011, Sept 2013.

[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>.

[RFC8926] J. Gross, et al, "Geneve: Generic Network Virtualization Encapsulation", RFC8926, Nov 2020.

12.2. Informative References

[IPsecOverGENEVE] S. Boutros, et al, "IPsec over GENEVE Encapsulation", draft-boutros-nvo3-ipsec-over-geneve-01, work-in-progress, Jan, 2018.

[RFC2410] R. Glenn and S. Kent, "The NULL encryption Algorithm and Its Use with IPsec", RFC2310, Nov. 1998.

[RFC5869] H. Krawczyk and P. Eronen, "HMAC-based Extract-and-Expand Key Derivation Function (HKDF)", RFC5869, May 2010.

[RFC6071] S. Frankel and S. Krishnan, "IP Security (IPsec) and Internet Key Exchange (IKE) Document Roadmap", Feb. 2011.

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

[MEF-70.1] MEF 70.1 SD-WAN Service Attributes and Service Framework. Nov. 2021.

[Net2Cloud] L. Dunbar and A. Malis, "Dynamic Networks to Hybrid Cloud DCs Problem Statement", draft-ietf-rtgwg-net2cloud-problem-statement-42, Jan, 2025.

[SD-WAN-Edge-Discovery] L. Dunbar, et al, "BGP UPDATE for SD-WAN Edge Discovery", draft-ietf-idr-sdwan-edge-discovery-25, July. 2025.

[lightweight-authenticate] L. Dunbar, K. Majumdar, S. Fluhrer, "Lightweight Authentication Methods for IP Header", draft-dunbar-ipsecme-lightweight-authenticate-01, July 2025.

13. Acknowledgments

Acknowledgements to Adrian Farrel, Joel Halpern, Donald Eastlake, Stephen Farrell, Ajeet Gill for their extensive review and suggestions.

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

Appendix A: Illustration of Packets through Cloud GWs

This section illustrates Cloud GWs connecting traffic flow carried by the IPsec tunnels.

A.1 Single Hop Cloud GW

Assuming that all CPEs are under one administrative control (e.g., iBGP).

Using Figure 1 as an example:

- There is a bidirectional IPsec tunnel between CPE1 and Cloud GW; with IPsec SA1 for the traffic from the CPE1 to the Cloud-GW; and IPsec SA2 for the traffic from the Cloud-GW to the CPE1.
- There is a bidirectional IPsec tunnel between CPE2 and Cloud GW; with IPsec SA3 for the traffic from the CPE2 to the Cloud-GW; and IPsec SA4 for the traffic from the Cloud-GW to the CPE2.
- All the CPEs are under one iBGP administrative domain, with a Route Reflector (RR) as their controller. The CPEs notify their peers of their corresponding Cloud GW addresses (which is out of the scope of this document).

When CPE1 (192.0.2.0/26) and CPE2 (192.0.2.64/26) need to communicate with each other, CPE1 and CPE2 establish a bidirectional IPsec Tunnel, with SA5 for the traffic from

CPE1 to CPE2 and SA6 for the traffic from CPE2 to CPE1. Assume the IPsec ESP Tunnel Mode is used. A packet from 192.0.2.1 to 192.0.2.65 has the following outer header:

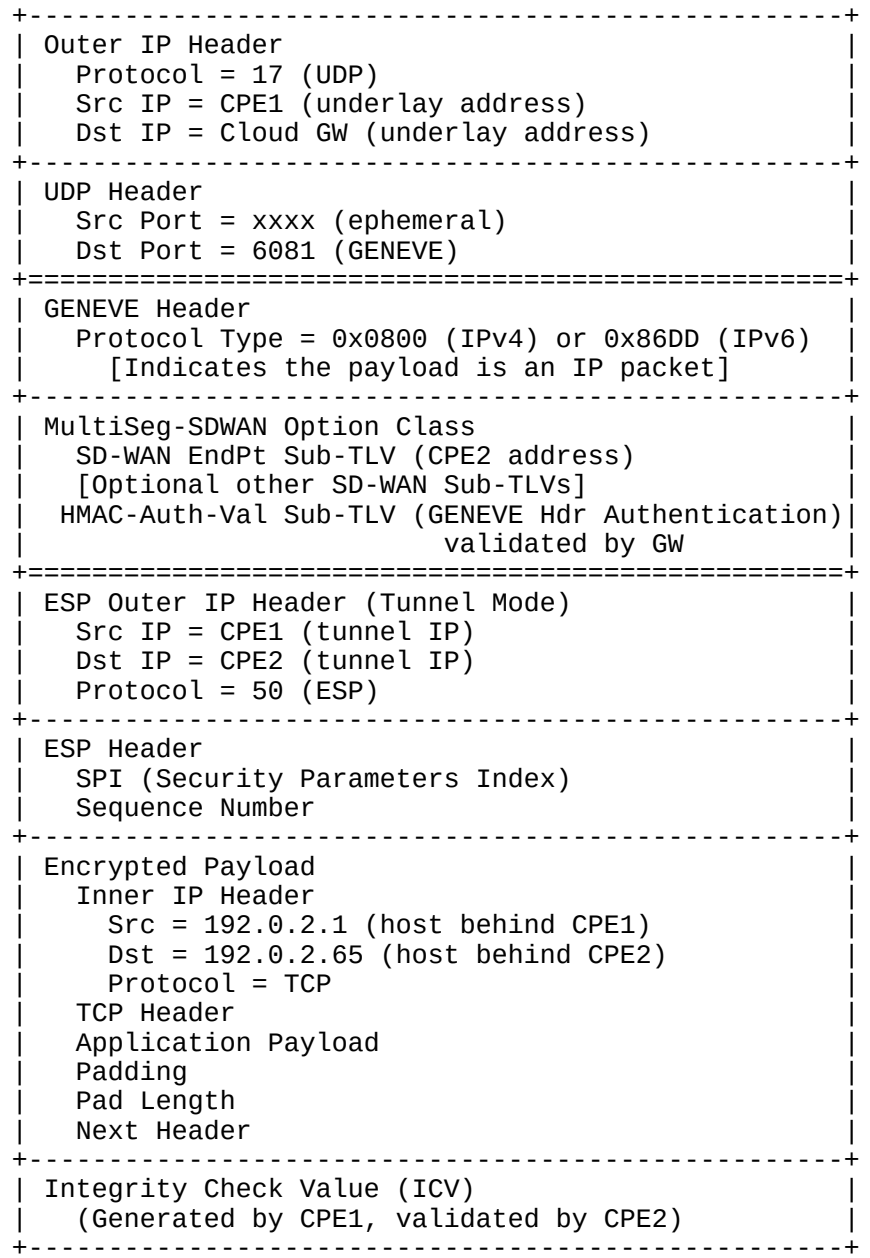


Figure 10 Packet header illustration to Cloud GWs

A.2 Multi-hop Transit GWs

Traffic to/from geographic apart CPEs can cross multiple Cloud DCs via Cloud backbone.

The on-premises CPEs are under one administrative control (e.g., iBGP).

Using Figure 2 as an example:

- There is a bidirectional IPsec tunnel between CPE1 and the Cloud GW1; with IPsec SA1 for the traffic from the CPE1 to the Cloud-GW1; and IPsec SA2 for the traffic from the Cloud-GW1 to the CPE1.
- There is a bidirectional IPsec tunnel between CPE10 and the Cloud GW2; with IPsec SA3 for the traffic from the CPE10 to the Cloud-GW2; and IPsec SA4 for the traffic from the Cloud-GW2 to the CPE10.
- All the CPEs are under one iBGP administrative domain, with a Route Reflector (RR) as their controller. CPEs notify their peers of their corresponding Cloud GW addresses.

When CPE1(192.0.2.0/26) and CPE10(192.0.2.128/25) need to communicate with each other, CPE1 and CPE10 establish a bidirectional IPsec Tunnel, with SA5 for the traffic from CPE1 to CPE10 and SA6 for the traffic from CPE10 to CPE1. Assume the IPsec ESP Tunnel Mode is used, a packet from 192.0.2.1 to 192.0.2.129 has the following outer header:

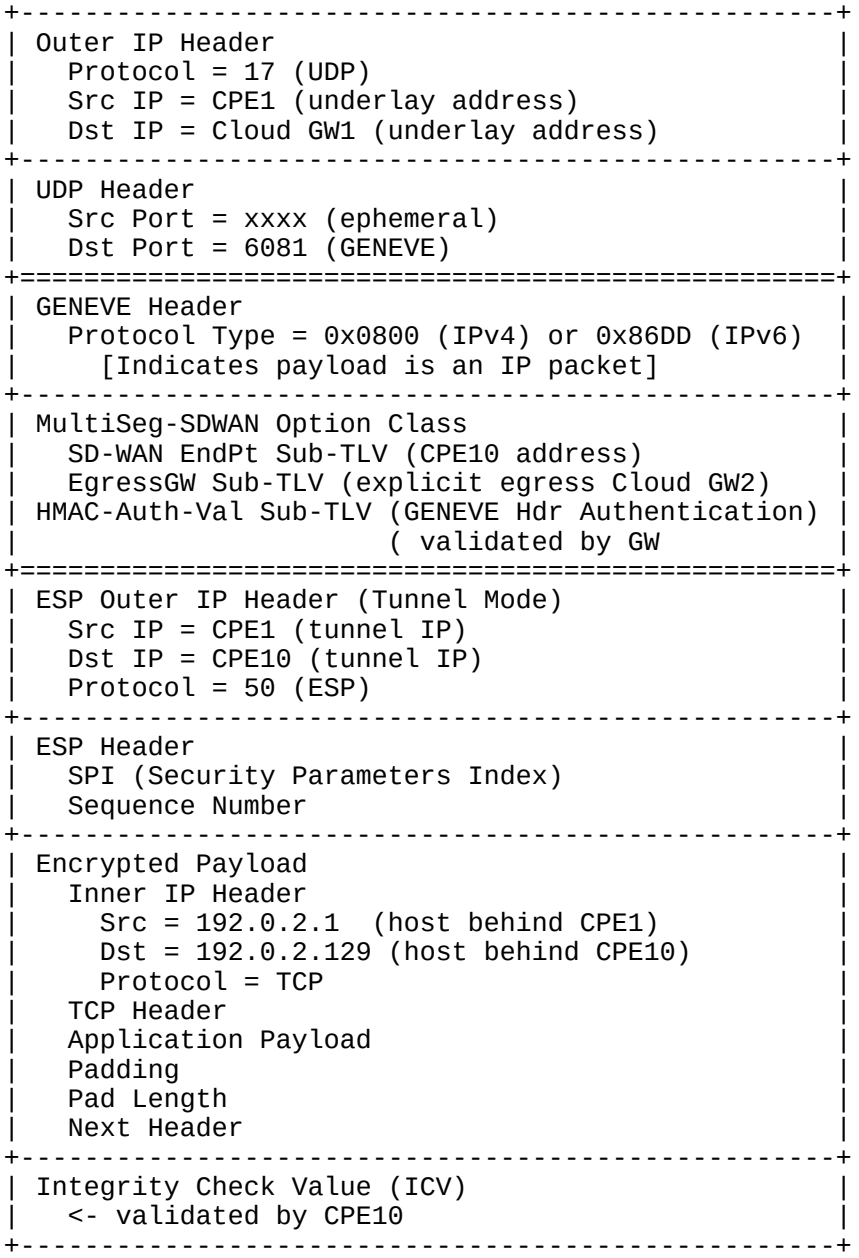


Figure 11: Packet header to Cloud GWs

Appendix B: Illustration from Private VPN to IPsec Tunnel

This section illustrates a Cloud GW connecting client traffic from a branch CPE via a Private VPN to another CPE via an IPsec tunnel.

Using Figure 1 as an example:

- CPE1 sends traffic via a Private VPN (Direct Connect to the Cloud Edge) to the Cloud GW. The traffic is not encrypted.
- There is a bidirectional IPsec tunnel between CPE2 and the Cloud GW; with IPsec SA1 for the traffic from the CPE2 to the Cloud-GW; and IPsec SA2 for the traffic from the Cloud-GW to the CPE2.
- All the CPEs are under one iBGP administrative domain, with a Route Reflector (RR) as their controller. CPEs notify their peers of their corresponding Cloud GW addresses.

Assume the IPsec ESP Tunnel Mode is used for the IPsec SA between Cloud GW and CPE2. For a packet from 192.0.2.1 to 192.0.2.129, the following header is added by CPE1 sending over the Private VPN:

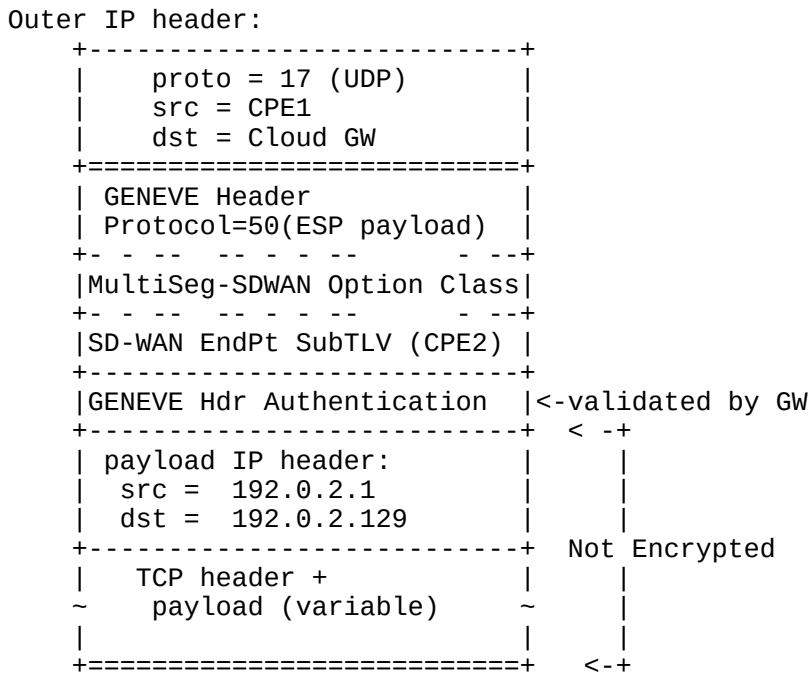


Figure 12 Illustration of packet through VPN

Upon receiving the GENEVE encapsulated packet with the "Multi-Segment-SD-WAN" option, the Cloud GW extracts the destination CPE from the GENEVE header and encrypts the packet with the IPsec SA2 to forward to the destination (i.e., CPE2). The GENEVE Header is carried to the CPE2.

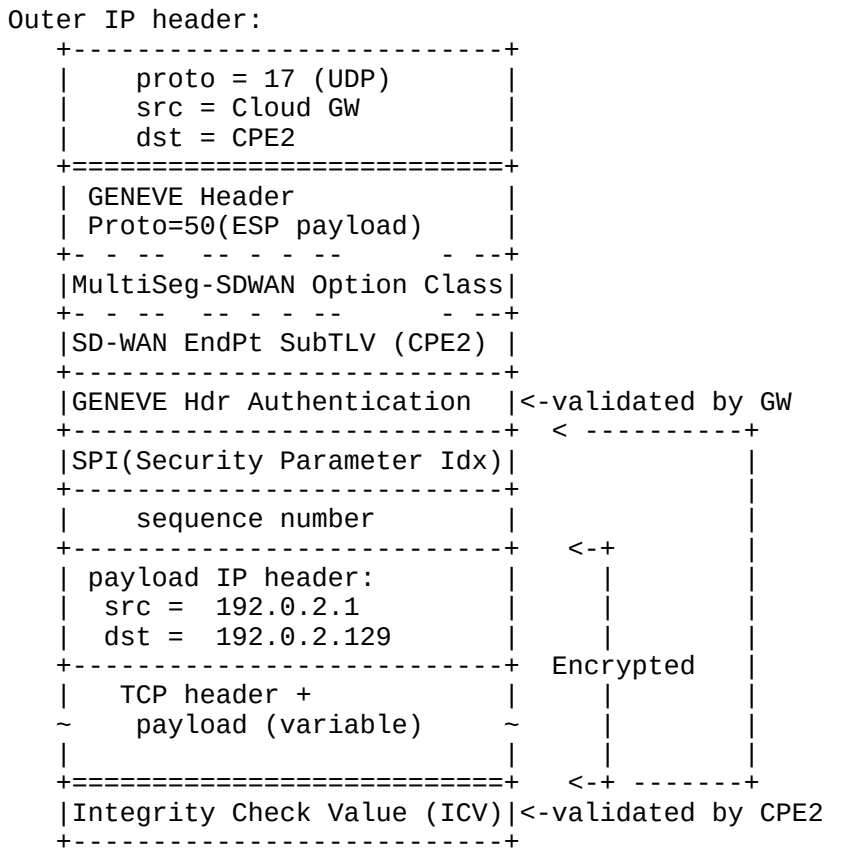


Figure 13 Illustration of packet from the Egress Cloud GW

Authors' Addresses

Kausik Majumdar
Oracle
Email: kausik.majumdar@oracle.com

Linda Dunbar
Futurewei
Email: ldunbar@futurewei.com

Venkit Kasiviswanathan
Nexthop AI
Email: venkit@nexthop.ai

Ashok Ramchandra
Google
Email: archiashok@gmail.com

Aseem Choudhary
Cisco
Email: asechoud@cisco.com

Contributors' Addresses

Ajeet Pal Singh Gill
Microsoft Azure
Email: ajeetgill@microsoft.com