

Network Working Group
Internet Draft
Intended status: Standard Track
Expires: February 4, 2026

K. Majumdar
Oracle
L. Dunbar
Futurewei
V.Kasiviswanathan
Arista
A. Ramchandra
Microsoft
A. Choudhary
Aviatrix
August 4, 2025

Multi-segment SD-WAN via Cloud DCs
draft-ietf-rtgwg-multisegment-sdwan-05

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.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at <http://www.ietf.org/ietf/lid-abstracts.txt>

The list of Internet-Draft Shadow Directories can be accessed at <http://www.ietf.org/shadow.html>

This Internet-Draft will expire on Dec 4, 2020.

Copyright Notice

Copyright (c) 2025 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

| | |
|--|----|
| 1. Introduction..... | 3 |
| 2. Conventions used in this document..... | 5 |
| 3. Use Cases..... | 5 |
| 3.1. Multi-segment SD-WAN via a Single Cloud GW..... | 5 |
| 3.2. Multi-segment SD-WAN via Cloud Backbone..... | 7 |
| 3.3. Traffic Steering Challenges in Multi-Segment SD-WAN.. | 8 |
| 4. Data Plane encoding for SD-WAN Transit..... | 9 |
| 4.1. Multi-Segment SD-WAN Option Class..... | 9 |
| 4.2. SD-WAN Tunnel Endpoint Sub-TLV..... | 11 |
| 4.3. SD-WAN Tunnel Originator Sub-TLV..... | 12 |

| | |
|--|----|
| 4.4. Egress GW Sub-TLV..... | 13 |
| 4.5. Exclude Transit Sub-TLV..... | 14 |
| 5. Packet Header Processing..... | 15 |
| 6. Error Handling..... | 17 |
| 7. Control Plane considerations..... | 17 |
| 7.1. Control Plane for CPEs..... | 17 |
| 7.2. Control Plane between CPEs and Cloud GWs..... | 18 |
| 8. Observability Consideration..... | 18 |
| 9. Security Considerations..... | 19 |
| 9.1. Threat Analysis..... | 19 |
| 9.2. HMAC-based Integrity and Authentication..... | 20 |
| 9.3. AH based Integrity and Authentication..... | 22 |
| 10. Manageability Considerations..... | 23 |
| 11. IANA Considerations..... | 24 |
| 12. References..... | 24 |
| 12.1. Normative References..... | 24 |
| 12.2. Informative References..... | 25 |
| 13. Acknowledgments..... | 26 |
| Appendix A: Illustration of Packets through Cloud GWs..... | 26 |
| A.1 Single Hop Cloud GW..... | 26 |
| A.2 Multi-hop Transit GWs..... | 28 |
| Appendix B: Illustration from Private VPN to IPsec Tunnel... | 29 |

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 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 is possible through cloud-hosted security services (e.g., firewalls, DDoS protection), ensuring consistent treatment of traffic across sites.
- Cloud platforms often offer enhanced monitoring, proprietary threat detection tools, and analytics services that can inspect and respond to suspicious traffic crossing segments.

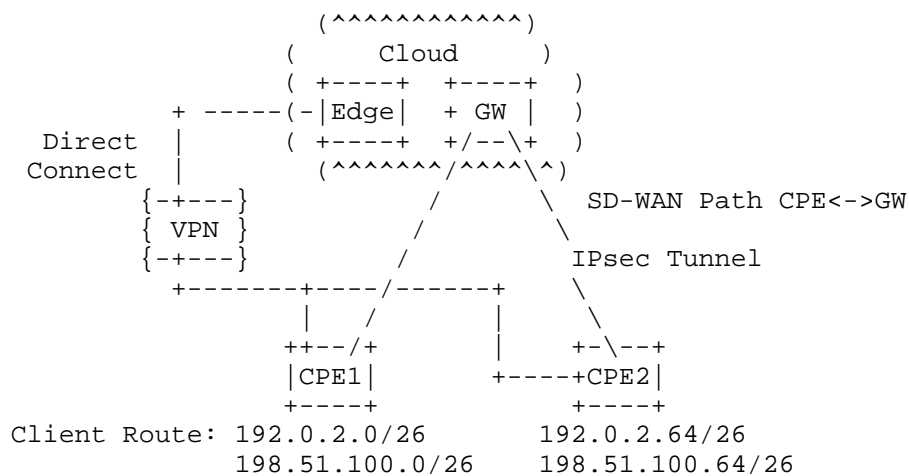


Figure 1 Multi-Segment SD-WAN stitching via a Cloud GW

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.

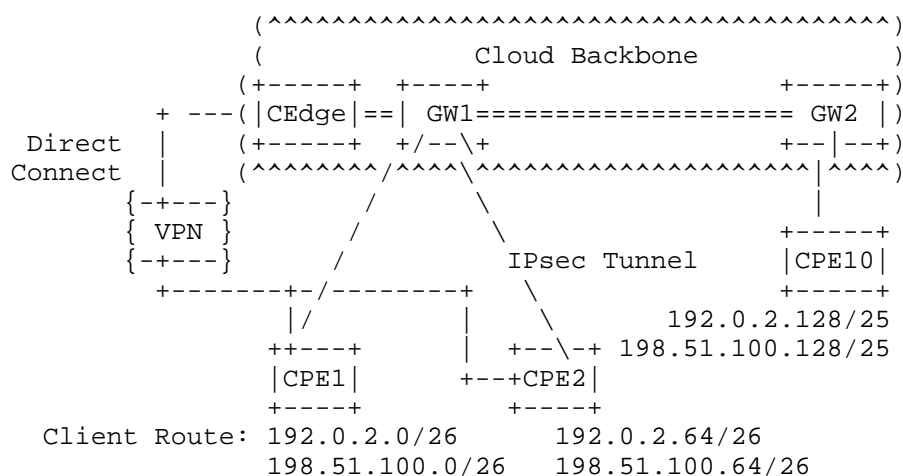


Figure 2 Multi-Segment SD-WAN Stitching via Cloud Backbone

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 SD-WAN deployments where on-premises CPEs connect to Cloud GWs over the public internet, traffic forwarding is typically destination-based, limiting the ability to enforce precise traffic steering policies. Even when branch CPEs have dedicated SD-WAN paths to Cloud GWs, the paths between branches may still be routed over the public internet via any available route, bypassing the Cloud Backbone entirely.

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 proposes 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.

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. Since GENEVE Encapsulation [RFC8926] is widely supported by major Cloud Service Providers, it is chosen as the encapsulation method. This allows Cloud GWs to efficiently steer IPsec-encrypted packets between CPEs 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.

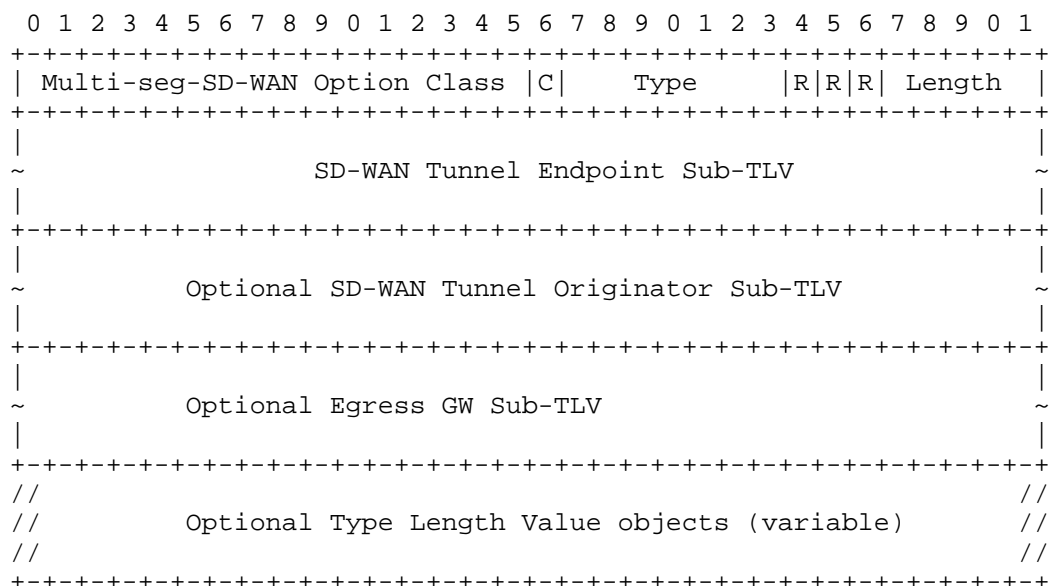


Figure 3 Multi Segment SD-WAN Option Class

- 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 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.

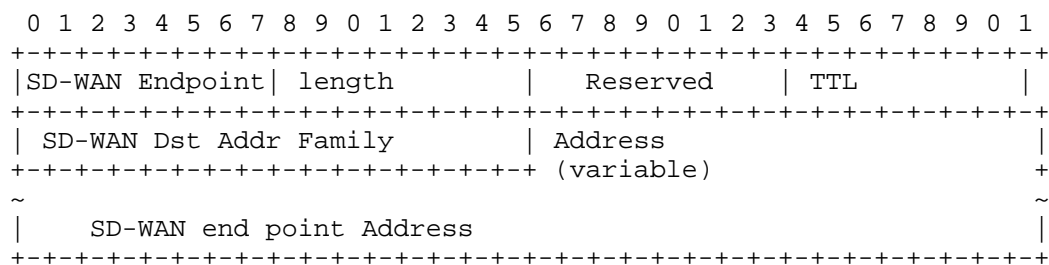


Figure 4 SD-WAN Endpoint Sub-TLV

- 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).

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.

```

0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|SDWAN Origin | length      | reserved      |
+-----+-----+-----+-----+-----+-----+-----+-----+
| SD-WAN Org  | Addr Family | Address      |
+-----+-----+-----+-----+-----+-----+-----+-----+
~                                                    ~
|   SD-WAN Tunnel Originator Address   |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

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.

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.

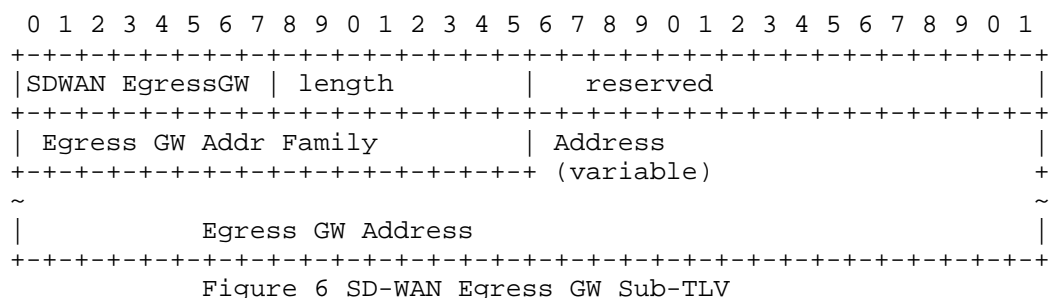


Figure 6 SD-WAN Egress GW Sub-TLV

- SDWAN EgressGW (8 bits): Identifies the 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.

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. Exclude Transit Sub-TLV

Exclude-Transit Sub-TLV is an optional field that explicitly specifies a list of Cloud Availability Regions or Zones that must be avoided when forwarding packets through 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 Exclude-Transit Sub-TLVs can be included within a single GENEVE header to specify multiple excluded transit nodes or regions. However, these Sub-TLVs form an unordered set, meaning there is no defined sequence in which exclusions are applied.

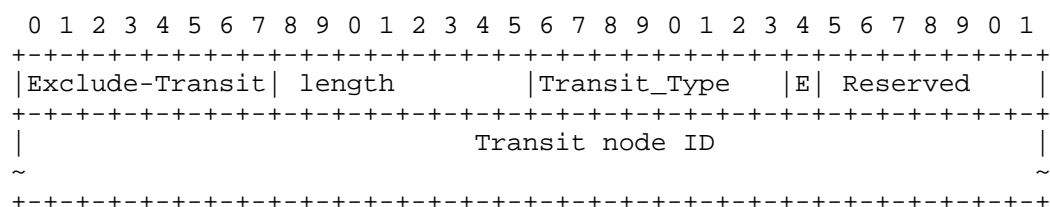


Figure 8 Exclude-Transit Sub-TLV

- Exclude -Transit (8 bits): identifies the Exclude - Transit Sub-TLV with a Type value of 5.

- Length (8 bits): Specifies the total length of the value field in 4-byte units, excluding the first 4 bytes, which include the Exclude-Transit Sub-TLV Type (1 byte), Length (1 byte), Transit_Type (1 byte), E-bit (1 bit), and Reserved bits (7 bits).
- Transit_type (8 bits): Defines how the transit node is identified:
 - Transit_type = 1: The Transit Node ID is represented as a numeric identifier, such as a Cloud Availability Region or Zone number assigned by the cloud provider.
 - Transit_type = 2: The Transit Node ID is represented as an IP address.
- E-bit (1 bit) - Exclusion Requirement:
 - o 0: The specified transit node is undesirable, and the Cloud Backbone will make a best-effort attempt to route traffic around it.
 - o 1: The specified transit node must be avoided. If the node cannot be bypassed, an alert or alarm must be generated to notify the enterprise via an out-of-band mechanism. The specifics of these alerts are beyond the scope of this document.

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 with 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:

- o If the Egress GW Sub-TLV exists, the Cloud Backbone uses it to identify the egress Cloud GW.
- o 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.

7.2. Control Plane between CPEs and Cloud GWs

In multi-segment SD-WAN deployments, the control plane between CPEs and Cloud GWs must support dynamic routing updates and secure tunnel establishment.

- eBGP sessions are commonly used between enterprise CPEs and Cloud GWs to facilitate dynamic route exchange.
- If an IPsec tunnel is required between a Cloud GW and a vCPE, the IKEv2 key exchange must be performed to establish the tunnel securely.
- 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.

By leveraging these control plane considerations, enterprises can ensure efficient route discovery, security enforcement, and seamless SD-WAN interconnectivity across the Cloud Backbone.

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

As shown in Figure 3, the information carried by the GENEVE Header is not encrypted, which is susceptible to Man-in-the-Middle (MitM) attacks. An attacker can intercept and potentially alter the information in the GENEVE header between the branch CPEs and the Cloud GWs without the enterprise and the Cloud provider's knowledge or consent. Here is the threat analysis of the MitM attacks between CPEs and Cloud GWs:

- a) Eavesdropping: Attackers can get knowledge of the enterprise's branch locations and their respective contracted Cloud GWs. As the payload between the CPEs is encrypted, attackers can't get any data exchanged between CPEs. This threat is no different from direct IPsec SAs between two CPEs.

- b) Data Manipulation: Attackers alter the content (Sub-TLVs) in the GENEVE header. As packets with unrecognized source addresses or invalid values in the Sub-TLVs of the GENEVE header are dropped by Cloud GWs, there might be a higher packet drop rate between the CPEs. Packet drops are not a new problem. The transport layer, such as TCP or QUIC, can handle packet drop well.

- c) Potential stealing of Cloud Backbone bandwidth:
A threat actor such as a malicious or misconfigured C-PE, could inject SD-WAN metadata intended for another tenant, attempting to redirect traffic through Cloud Backbone paths it is not authorized to use. For example, a legitimate Cloud subscriber pays for Cloud Backbone transport between CPE-A and CPE-B. An attacker, who controls two locations (Node-A and Node-B), might construct a packet with CPE-A's address as the source and include CPE-B in the SD-WAN Endpoint Sub-TLV. The packet, when sent to the ingress Cloud GW, would appear to be a legitimate flow between CPE-A and CPE-B. After exiting the egress Cloud GW, the attacker could rewrite the outer addresses to resume communication between Node-A and Node-B, effectively tunneling traffic via the Cloud Backbone without authorization or payment.

To prevent such misuse, it is necessary to authenticate and verify the origin and legitimacy of the SD-WAN metadata to ensure it is associated with an authorized and registered CPE endpoint. This validation protects against both spoofing and cross-tenant policy violations. While it is not necessary for Cloud GWs to decrypt or re-encrypt payloads, they must enforce metadata integrity using authentication mechanisms such as those defined in [RFC2403] and [RFC2404]. These mechanisms rely on cryptographic hashing algorithms (e.g., SHA2 224/256/384/512) as part of a Hashed Message Authentication Code (HMAC) to validate packet authenticity.

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 data between CPEs and Cloud GWs to verify that GENEVE header has not been tampered with.

The basic idea behind HMAC is to combine a secret key and a hash function to produce a fixed-size authentication code for the GENEVE header between CPEs and Cloud GW. This authentication code is then sent along with the data itself. When Cloud GW and the destination CPEs receive the data and the authentication code, they can independently compute the HMAC using the same key and hash function. If the computed HMAC matches the received authentication code, it indicates that the data has not been altered, as long as the secret key remains confidential.

To minimize packet size overhead, especially as the HMAC value is appended to each packet, this document recommends using 4-byte or 8-byte truncated HMAC values, instead of full-length HMACs (e.g., 32 bytes for HMAC-SHA-256). This tradeoff is acceptable in this context because the actual user payload remains protected by IPsec encryption.

The HMAC authentication code can be carried by an HMAC Sub-TLV in the GENEVE Header, as specified below:

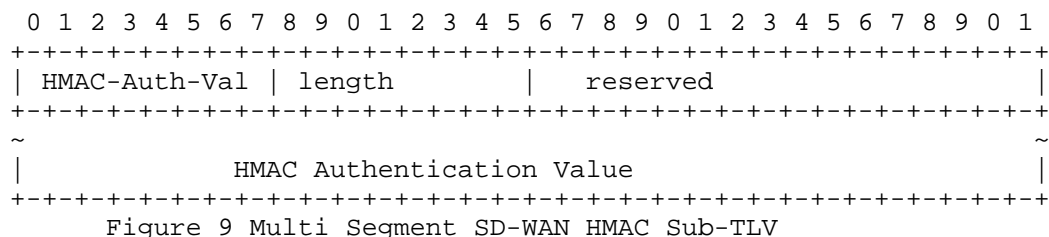


Figure 9 Multi Segment SD-WAN HMAC Sub-TLV

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): is computed over the entire GENEVE encapsulation header,

excluding the HMAC-auth-Val Sub-TLV, based on a pre-configured algorithm such as HMAC-SHA-256 and a shared key.

The advantages of using HMAC are:

- Data Integrity: HMAC ensures integrity and authenticity of steering metadata using a shared key, with efficient truncated values (4 or 8 bytes) for low-overhead protection.
- Efficiency: Truncated HMACs are computationally efficient and reduce packet overhead.
- Resistance to Tampering: Even truncated, HMAC values resist message tampering and replay attacks.
- Flexibility: HMAC works with various hash functions like SHA-256 or SHA-512.
- Widely Supported: HMAC is well-established and supported across platforms.

While truncated HMACs reduce collision resistance compared to full-length values, this tradeoff is acceptable due to the following mitigating factors:

- The payload remains encrypted by IPsec SAs.
- The HMAC protects only the steering metadata in the GENEVE header.
- Attackers still require knowledge of the shared key to forge valid packets.

In summary, HMAC-based integrity and authentication offer effective protection for SD-WAN steering information without compromising the performance or increasing packet size significantly.

9.3. AH based Integrity and Authentication

For enterprises or Cloud providers worrying about secret HMAC keys being compromised, they can add another layer of AH encryption [RFC4301] or ESP-NUL [RFC2410] [RFC6071] on top of the IPsec encryption between the two CPEs. Both AH and ESP with NUL encryption require pairwise IPsec key management between Cloud GWs and the CPEs, which introduces additional processing overhead both endpoints. In addition, the AH

encrypted packets can't traverse NAT because of outer IP address changes.

10. Manageability Considerations

The following manageability considerations are crucial for the successful deployment and ongoing operation of the proposed strategies outlined in this document:

- Centralized Orchestration:
A centralized orchestration system is needed to manage and authenticate multiple SD-WAN segments through the Cloud GWs.
- Policy-based Configuration:
Utilize policy-driven configurations to streamline the deployment of SD-WAN segments and their connectivity options. This approach allows for efficient management of network policies, ensuring consistent and coherent behavior across diverse deployment scenarios. [RFC8192] can be used to automate the security policy configurations.
- Real-time Monitoring and Analytics:
Integrate robust monitoring and analytics tools to provide real-time visibility into the performance and health of SD-WAN segments. This includes monitoring bandwidth utilization, latency, packet loss, and other key performance indicators to promptly identify and address any issues.
- Automated Alerting and Reporting:
Implement automated alerting mechanisms to promptly notify network administrators of potential issues or anomalies within the SD-WAN infrastructure. Additionally, generate regular reports to facilitate performance analysis, capacity planning, and compliance monitoring.

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 | Include Transit | [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.
- [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.

13. Acknowledgments

Acknowledgements to Adrian Farrel, Joel Halpern, Donald Eastlake, Stephen Farrell 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 192.0.2.0/26 and 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:

Outer IP header:

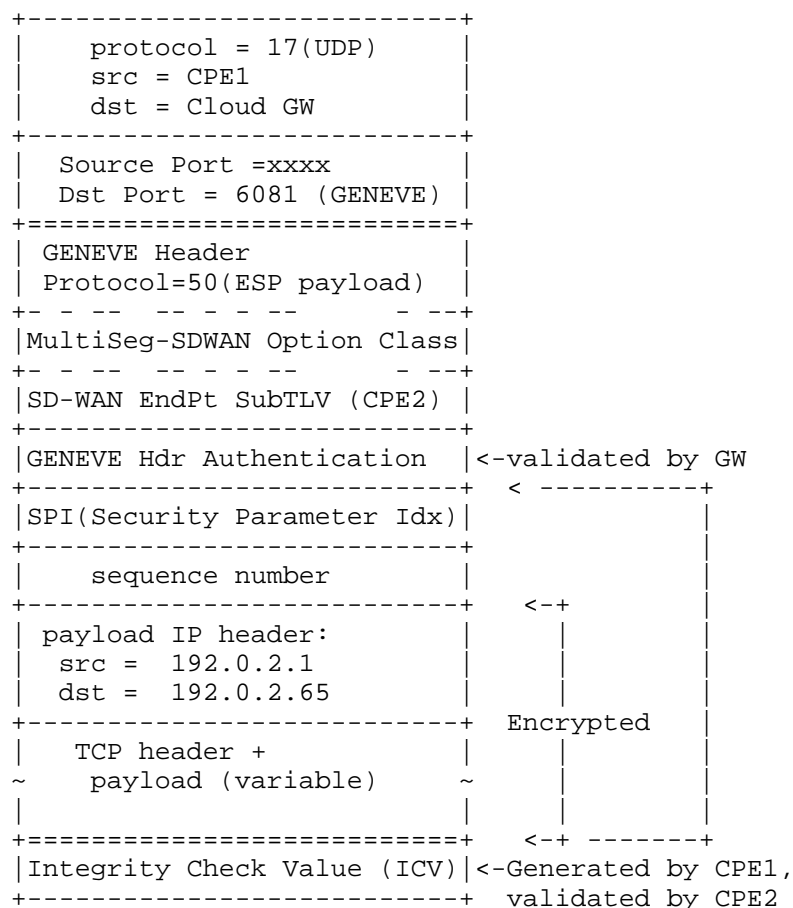


Figure 10 Packet header illustration to Cloud GWs

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 192.0.2.0/26 and 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:

Outer IP header:

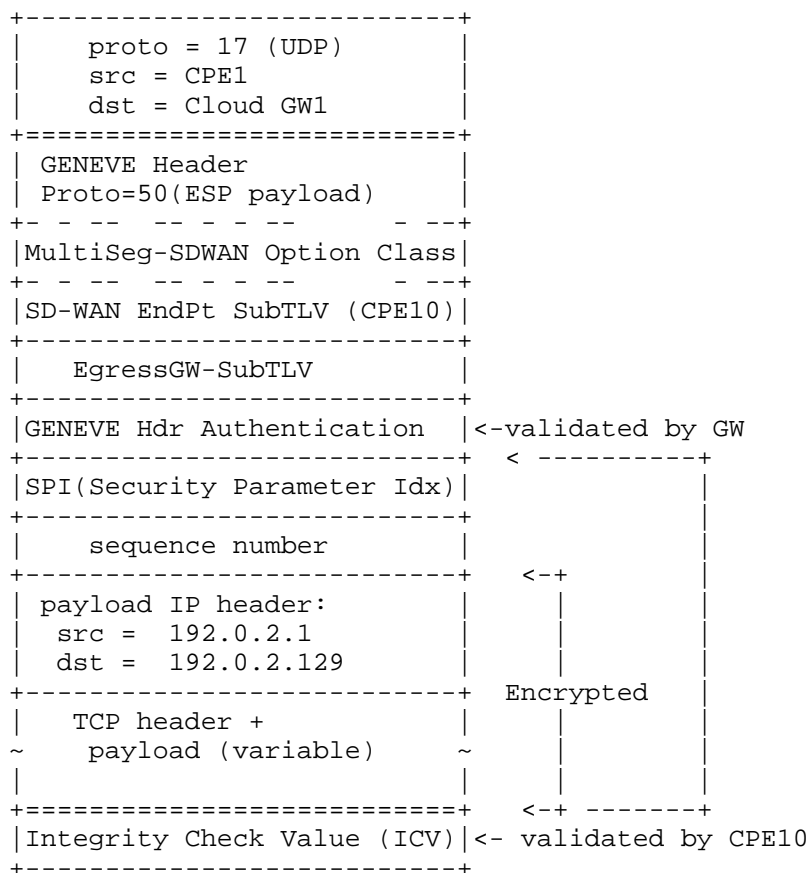


Figure 11 GENEVE header encapsulated IPsec packet

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:

Outer IP header:

```

+-----+
|      proto = 17 (UDP)      |
|      src = CPE1           |
|      dst = Cloud GW       |
+-----+
| GENEVE Header             |
| Protocol=50(ESP payload)  |
+-----+
| MultiSeg-SDWAN Option Class |
+-----+
| SD-WAN EndPt SubTLV (CPE2) |
+-----+
| GENEVE Hdr Authentication | <-validated by GW
+-----+ <--+
| payload IP header:        |
|   src = 192.0.2.1         |
|   dst = 192.0.2.129       |
+-----+
| TCP header +              | Not Encrypted
| ~   payload (variable)   ~ |
|                           |
+-----+ <--+

```

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.

Outer IP header:

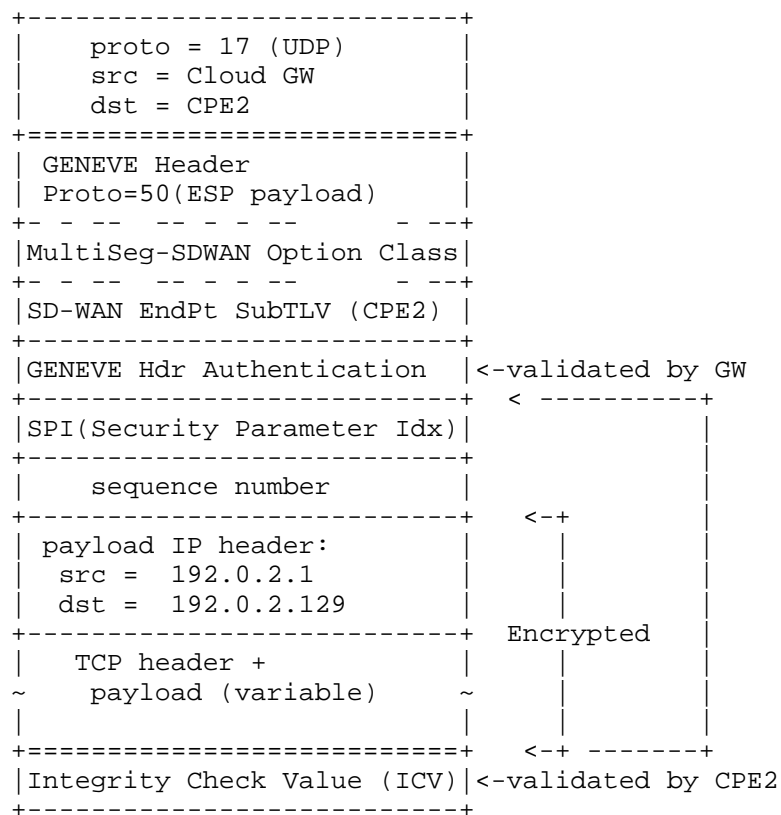


Figure 13 Illustration of packet from the Egress Cloud GW

Authors' Addresses

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

Linda Dunbar
Futurewei
Email: ldunbar@futurewei.com

Venkit Kasiviswanathan
Arista
Email: venkit@arista.com

Ashok Ramchandra
Microsoft
Email: aramchandra@microsoft.com

Internet-Draft

Multi-segment SD-WAN

Aseem Choudhary

Aviatrix

Email: achoudhary@aviatrix.com

Contributors' Addresses

