Network Working Group
Internet Draft
Intended status: Informational
Expires: May 4, 2020

L. Dunbar
Futurewei
Mehmet Toy
Verizon

November 4, 2019

Segment routing for SDWAN paths over hybrid networks
draft-dunbar-sr-sdwan-over-hybrid-networks-06

Abstract

This document describes a method for end-to-end (E2E) SDWAN paths to traverse specific list of underlay network segments, some of which can be private networks which include SR enabled segments, some of which can be the public IP networks that do not support SR, to achieve the desired optimal E2E quality.

The method described in this draft uses the principle of segment routing to enforce a SDWAN path's head-end selected route traversing through a list of specific nodes of multiple network segments without requiring the nodes in each network segment to have the intelligence (or maintaining states) of selecting next hop or next domain.

Status of this Memo

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

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79. This document may not be modified, and derivative works of it may not be created, except to publish it as an RFC and to translate it into languages other than English.

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.

Copyright Notice

Table of Contents

1. Introduction

This document describes a method to enforce a SDWAN path's head-end selected route traversing through a list of specific nodes of multiple network segments without requiring the nodes in each network segments to have the intelligence (or maintaining states) of selecting next hop or next segments. Those networks over which the SDWAN path traverse have at least one SR enabled network, and some network segments (especially the last mile access portion) being existing IP networks (such as existing IPv4, IPv6 or others).

Throughout this document, the term "Classic SDWAN" refers to a pair of CPEs in two locations aggregating N Service Providers' paths, such as MPLS Paths and public internet paths. Classic SDWAN is like what ONUG (Open Network User Group) has advocated, i.e. about pooling WAN bandwidth from multiple service providers to get better WAN bandwidth management, visibility & control.

[SR-SDWAN] describes using explicit routes within the SRv6 or SR-MPLS enabled networks to reach the desired quality for SDWAN paths over the SRv6 or SR-MPLS enabled networks respectively.

[BGP-SDWAN-Usage] describes three distinct SDWAN scenarios from an edge node's perspective:

1) Edge nodes interconnected by IPsec tunnels. All traffic among edge nodes are encrypted, i.e. over various IPsec tunnels.

2) Edge node has some ports connected to VPNs over which traffic can go natively without encryption and other ports to the public Internet over which traffic are encrypted; or

3) VPN PEs adding Internet facing WAN ports to offload low priority traffic when the VPN backbone paths/links are congested.

This document discusses using SR to steer traffic over combination of trusted network segments and untrusted segments, where traffic are forwarded natively over the trusted network segments and encrypted over the untrusted segments.

The goal is to place as large portion as possible of the SDWAN path over a provider VPN to achieve more optimal transport quality or steering the SDWAN path traversing specific ingress/egress PEs to reach optimal cost, quality, regulatory or other reasons.

2. Definition of terms

Cloud DC: Off-Premises Data Centers that usually host applications and workload owned by different organizations or tenants.

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

DMVPN: Dynamic Multipoint Virtual Private Network. DMVPN is a secure network that exchanges data between sites without needing to pass traffic through an organization's headquarter virtual private network (VPN) server or router.

Heterogeneous Cloud: applications & workloads split among Cloud DCs owned & managed by different Cloud Providers.

Hybrid Cloud: applications & workloads split between on-premises Data centers and Cloud DCs. In this document Hybrid Cloud also include heterogeneous cloud as well.

SDWAN: Software Defined Wide Area Network, "SDWAN" refers to the solutions of pooling WAN bandwidth from multiple underlay networks to get better WAN bandwidth management, visibility & control. When the underlay networks are private, traffic can traverse without additional encryption; when the underlay networks are public, such as the Internet, some traffic needs to be encrypted when traversing through (depending on user provided policies).

SP: Network Service Provider

SR: Segment Routing

SR Domain: A domain that supports Segment Routing

VPC: Virtual Private Cloud. A service offered by many Cloud DC operators to allocate a logically isolated cloud resources, including computing, networking and storage.

3. Key Use Cases

3.1. SDWAN Path over LTE network and SR Domain

MEF Cloud Service Architecture [MEF-Cloud] describes a use case of network operators needing to use SDWAN over LTE for the last mile access that they do not have physical infrastructure, as shown below:

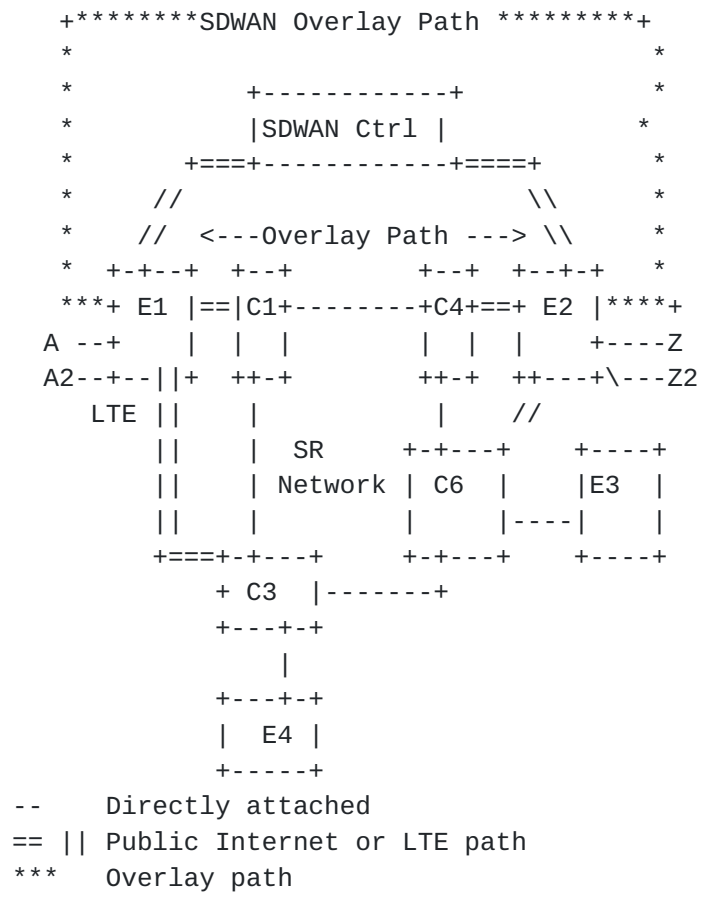


Figure 1: SDWAN end points are attached to VPN via LTE

3.2. SDWAN as Last Mile for Cloud DCs Access

Digital Transformation is propelling more and more enterprises to move their workloads/Apps to cloud DCs that are geographically close to their end users to improve end-to-end latency & overall user experience, or to comply with local data protection regulations. Conversely, workloads/Apps in those Cloud DCs can be easily shutdown when their end users' geographic base changes.

Because of the ephemeral property of the selected Cloud DCs, an enterprise or its network service provider may not have the direct links to the Cloud DCs that are optimal for hosting the enterprise's specific workloads/Apps. Under those circumstances, SDWAN is a very

flexible choice to interconnect the enterprise on-premises data centers & branch offices to its desired Cloud DCs.

However, SDWAN paths over public internet can have unpredictable performance, especially over long distances and cross state/country boundaries. Therefore, it is highly desirable to place as much as possible the portion of SDWAN paths over service provider VPN (e.g. enterprise's existing VPN) that have guaranteed SLA and to minimize the distance/segments over public internet.

Under this scenario, one or both of the SDWAN end points may not directly attached to the PEs of a SR Domain.

3.3. How & Why SR is useful for those use cases

Let us assume that the SDWAN Controller is capable of computing optimal paths between two end-points (e.g. E1<->E2 in the Figure 2), either by communicating with the SR Domain controller/management-system, or by other methods which is out of the scope of this document.

When a SR domain has multiple PEs with ports facing the external networks (such as the public internet or LTE termination), SDWAN paths can traverse the SR domain via different ingress/egress PEs resulting in different E2E performance.

In the diagram below, E1 <-> E2 SDWAN (most likely IPsec encrypted tunnel) path can traverse C1 <-> C4, C1<->C6, C3<->C6, or C3<->C4 within the VPN. There are many flows (by different Apps) between E1 <-> E2. Some flows may need to traverse C1<->C4, others may need to traverse C3<->C6 or other segments within the VPN, which are determined by the SDWAN controller based on the characteristics & need of the Apps, such as cost, available bandwidth, latency, or special functions only available at specific locations, etc.

Even with the same ingress/egress, some flows may need different segments across the SR Domain. It is not practical, or even possible, for PEs (e.g. C1, C2, C3 in this example) to determine which Apps' flows should egress C4 or C6 where both C4&C6 can reach E2.

Segment Routing can be used to steer packets (or path) to traverse the explicit egress node (C4 or C6), or explicit segments through

the SR Domain based on the SLA requested by the SDWAN head-end nodes.

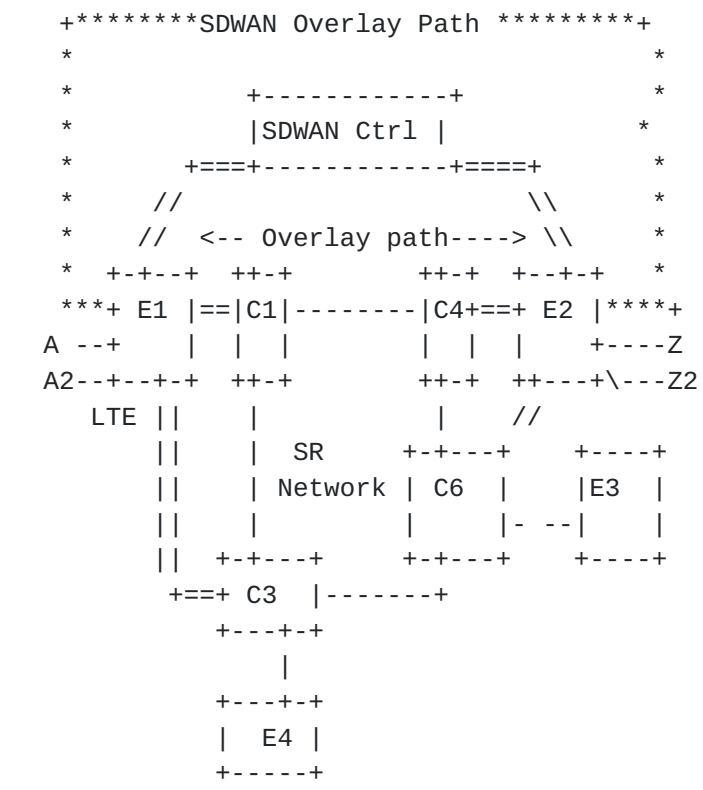


Figure 2: SDWAN end points not directly attached to PEs of SR Domain

4. Mechanism for SDWAN path over one SR Domain and existing access

This section describes the mechanism to enforce a SDWAN path' head-end selected route traversing through a list of specific nodes of multiple network segments without requiring the nodes in each network segment to have the intelligence (or maintaining states) of selecting next hop or next domain.

There may be two approaches here:
1) Controller installs the entire SID stack at E1.

2) Controller delivers to E1 a "Key" that the SR ingress PE can use to map to the SID stack for the packets arriving at the SR Ingress PE. Section 4.2 & 4.3 will describe how the "Key" is carried by the packets.

The Approach 1) requires less processing at the SR Ingress PE nodes, but only works if the remote CPEs are in the same administrative domain as the SR domain. SR domain usually is not willing to expose its internal binding SIDs to devices in different administration domains. This approach also requires more changes to SDWAN end nodes and need more header bytes added to the packets when traversing rd through 3 party internet. Some SDWAN nodes might not be capable of supporting encapsulating packets with the SID stack.

The Approach 2) above requires SR Ingress PE nodes to map the "Key" to the SID Stack and prepend the SID stack to the packets (Same processing for other traffic except the mapping is from the received "Key" carried in the payload).

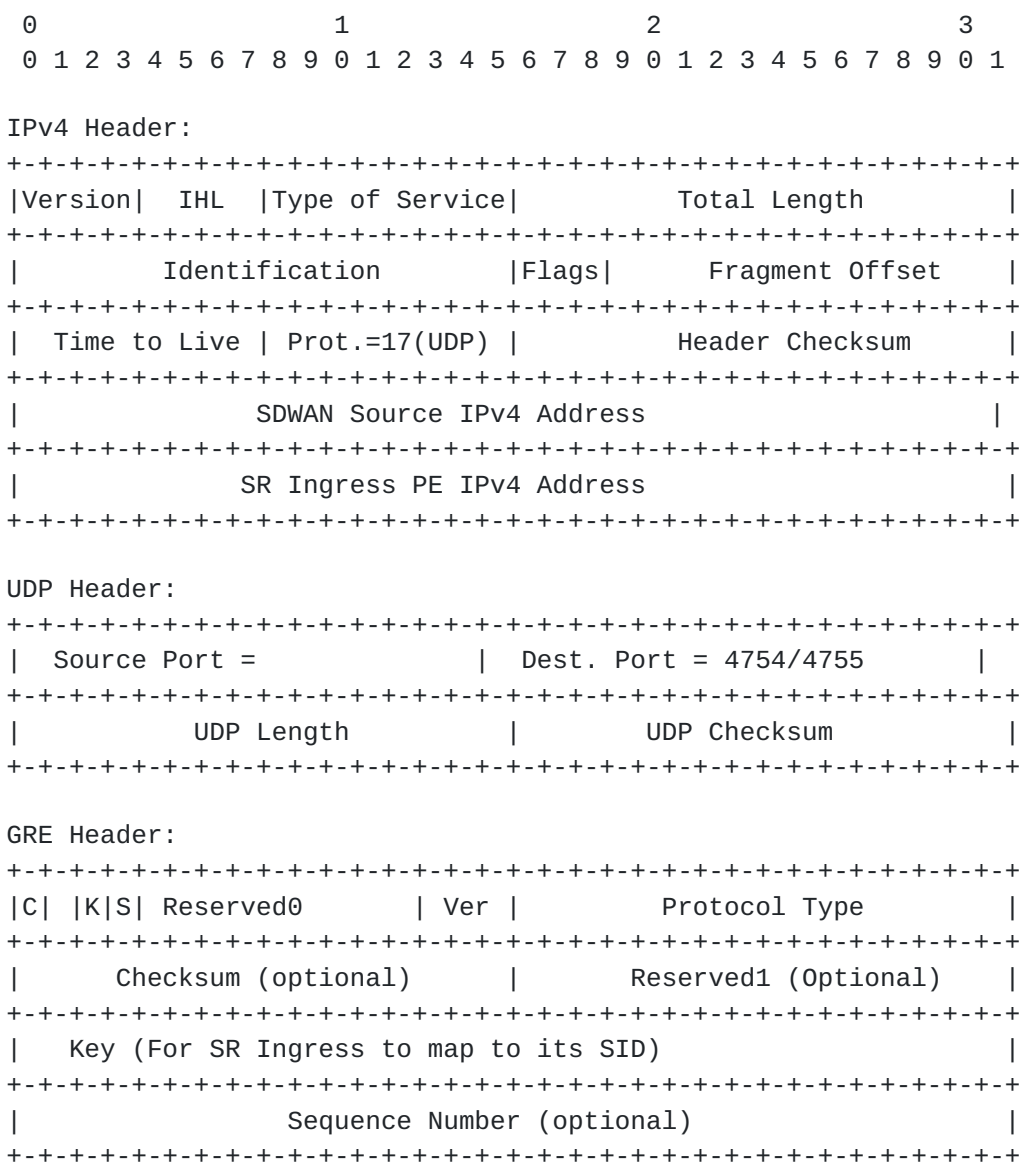
4.1. Controller Delivers SID Stack to SDWAN Head-end

This approach is straightforward.

```
      E1  -------------------------- > SDWAN controller
         request for a SDWAN path E1<->E2 with a specific SLA

      E1  <--------------------------  SDWAN controller
         Reply with the Ingress PE Node ID or address
         & the Binding SID.
```

Here is the packet header for SDWAN Source Node to prepend to the payload:



To traverse SRv6 domain, SRv6 Header is appended after the GRE header [SRv6-SRH]:

To traverse MPLS-SR domain, a stack of MPLS labels is appended after GRE Header [MPLS-SR].

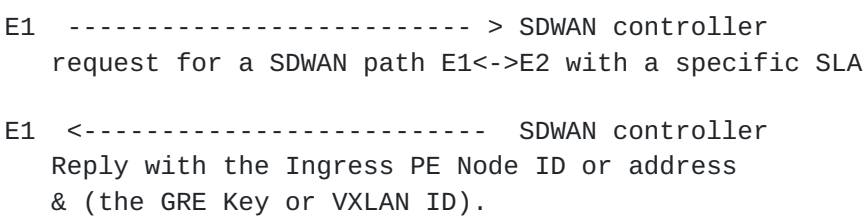
4.2. Using GRE Key or VXLAN ID to Differentiate Flows

This section describes a method of SDWAN head-end node using GRE Key or VXLAN Network ID (VNI) to indicate the desired property for specific flows between SDWAN end-points (E1<->E2 in the figure above): such as different desired routes through the SR Domain, different egress PEs based on cost, performance or other factors. It might be difficult or impossible for DiffServ bits carried by the packets to describe those flow properties because there can be more than what DiffServ bits can represent.

The SR Domain ingress can map the GRE key to different SID through the SR Domain.

We assume that the SDWAN Controller can determine which ingress PE can lead to the optimal path between E1<->E2. It is beyond the scope of this document on how SDWAN controller computes the paths and how & what SDWAN controller communicates with the SR Domain controller.

Here is the sequence of the flow:



Note: the GRE key (or VXLAN ID) from the SDWAN controller is for the ingress PEs to correlate desired Path with the list of SIDs to prepend the packet across the SR domain.

When SDWAN Controller get the E1<->E2 path request, it will communicate with the VPN Controller to get the optimal Ingress PE Node ID (or IP address) and the GRE key (or VXLAN ID) to encapsulate

the original packets between E1 <-> E2 (assuming IPsec Tunnel mode is used).

Upon receiving the GRE encapsulated packets, the provider ingress Edge C1/C3 uses the GRE key (or VXLAN ID) to map to the pre-defined (by the network controller) Binding SIDs, prepend the Binding SIDs to the packets, and forward its desired paths across the provider VPN.

Depending on how the SDWAN path destination can be reached by the egress PE, the egress PE has different processing procedure:

- If the destination of the SDWAN path is directly attached to the egress VPN PE node, the egress VPN PE decapsulates SR header and forward the packets to SDWAN path destination node, such as the E2 in the figure above.
- If the destination of the SDWAN path is IP reachable via IPv4 network from the egress VPN PE node, the egress VPN PE node decapsulates SR header and forward the packets to SDWAN path destination node via its internet facing port to the SDWAN path destination (i.e. the E2 node in the figure above).
- If the SDWAN path is traversing multiple domains owned by different network operators, the egress PE processing is described in the next session.

4.3. Using UDP Source Port Number to Differentiate Flows

[RFC8086] describes how to use GRE-in-UDP source port number as entropy for better ECMP performance. When the remotely attached CPEs is within very close proximity to the PEs, e.g. only one or two hopes away like in LTE access, there is less issue if ECMP put all flows with same traffic classifier into one path. Then, those UDP numbers can also be used as a key to SR PE nodes to map to the appropriate SID to the packets.

Same as RFC8086, UDP source port values used as a key for SR PEs to map to appropriate SIDs SHOULD be chosen from the ephemeral port range (49152-65535) [RFC8085].

The GRE-in-UDP encapsulation format contains a UDP header [RFC768] and a GRE header [RFC2890]. The format is shown as follow (presented in bit order):

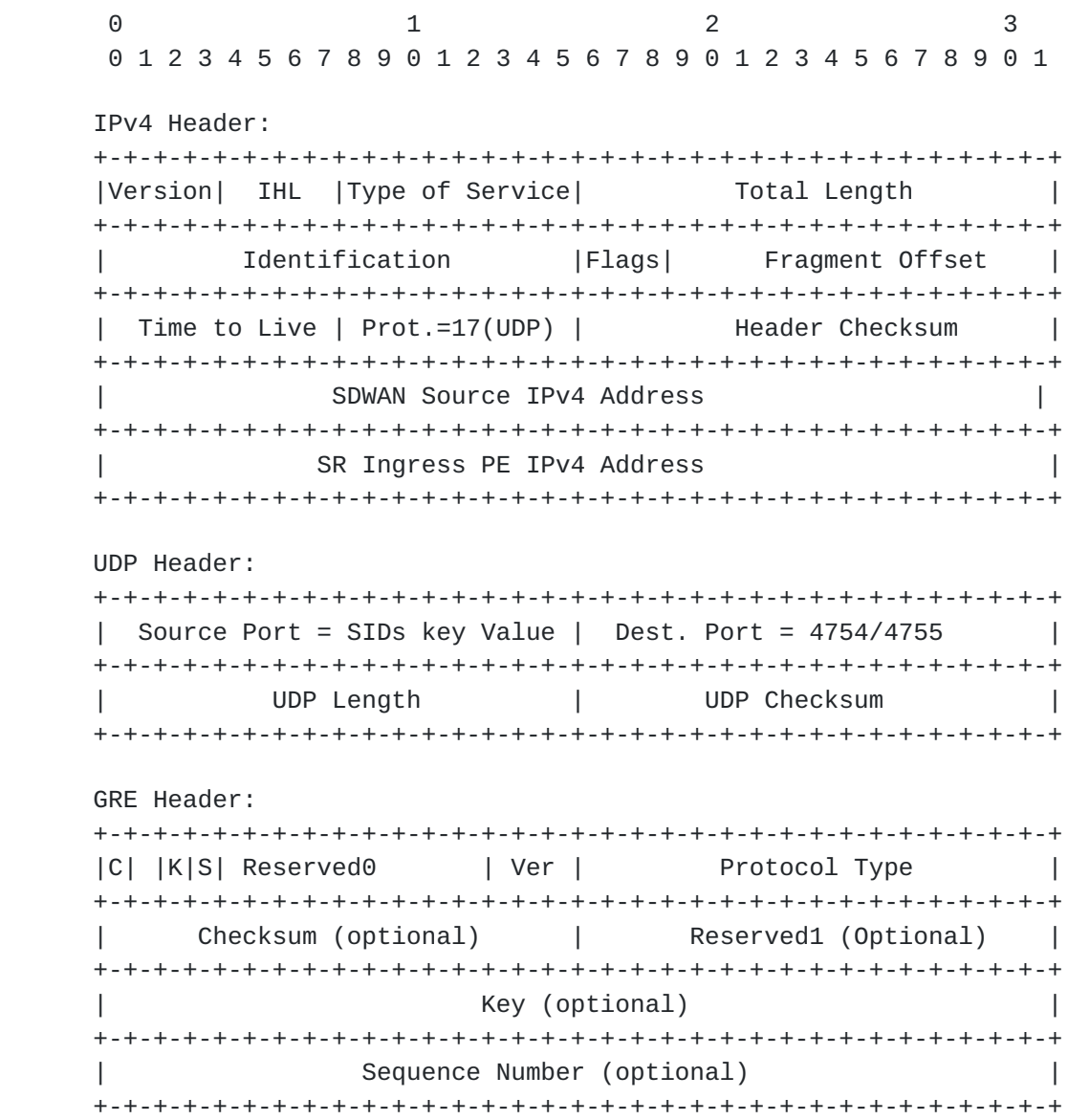
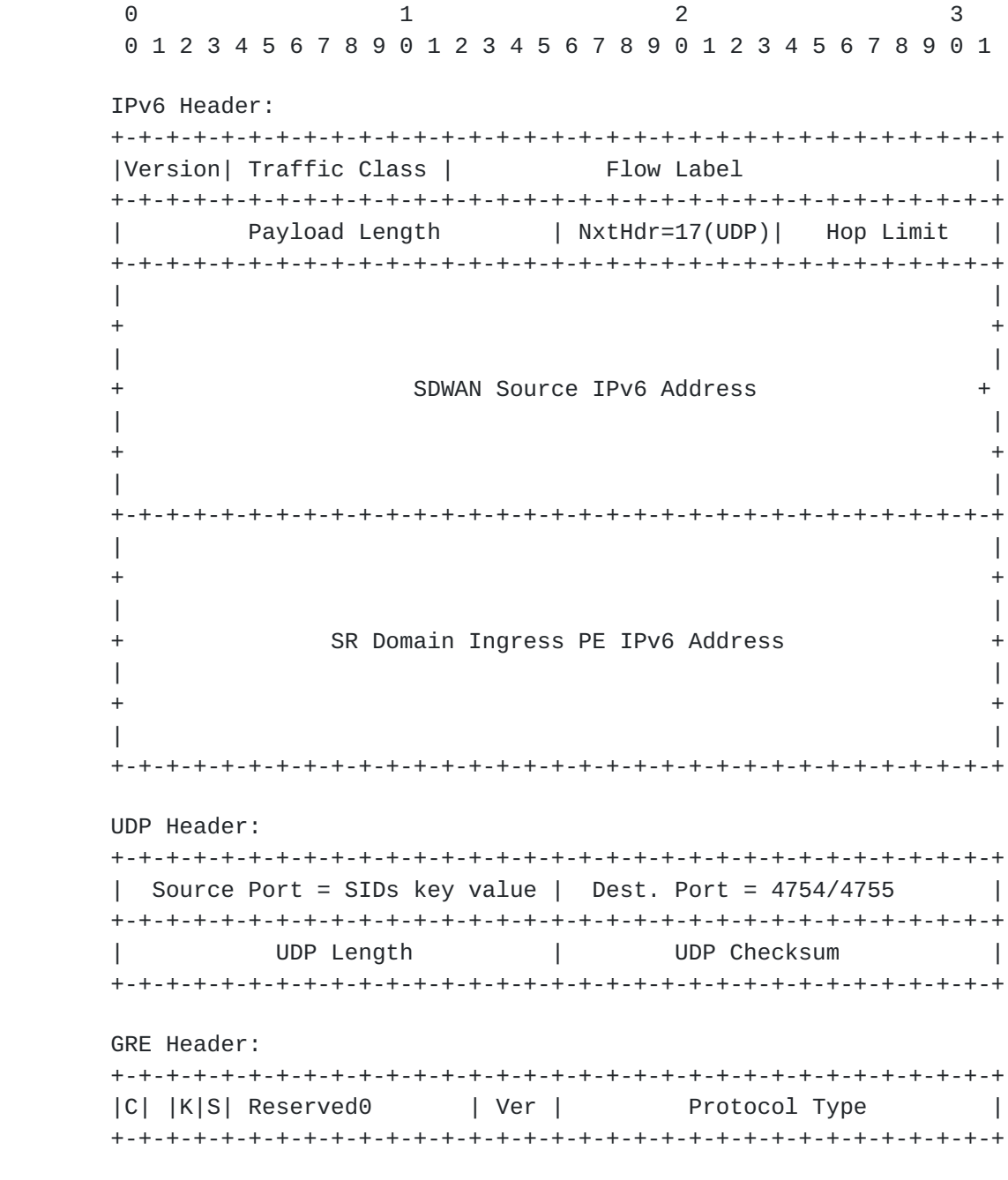


Figure 3: UDP + GRE Headers in IPv4

Here is the GRE Header for IPv6 network, i.e. the SDWAN Source SDWAN Destination, and SR PEs are all in IPv6 domain:



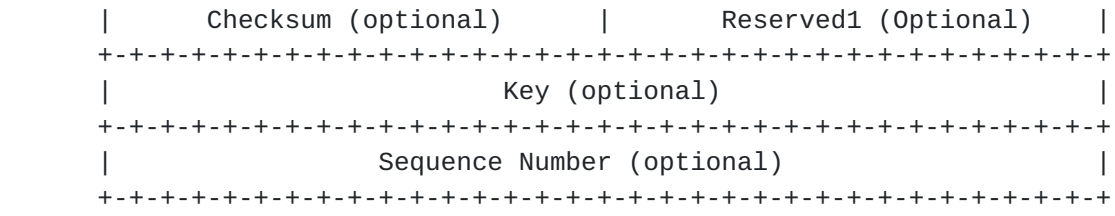
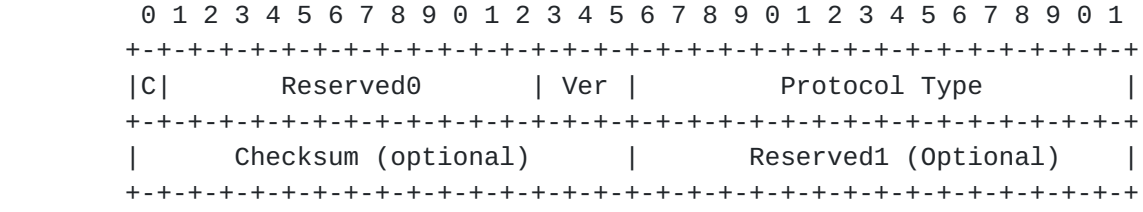


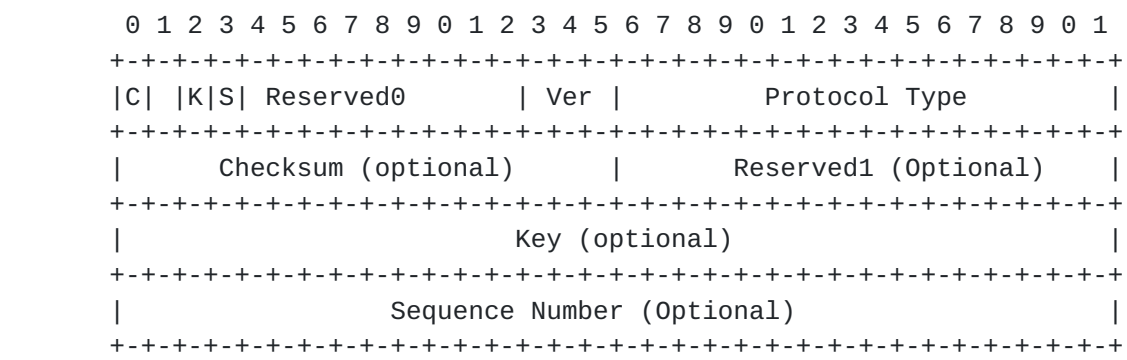
Figure 4: GRE+UDP for IPv6

4.4. GRE Header Extension

A new protocol type can be added to the GRE header [RFC2890] to make it easier for the SR PE to do the proper actions:



The proposed GRE header will have the following format:



New protocol type (value to be assigned by IANA):
UDP-Key: Using UDP source port value as a Key for SR Ingress PE to map to the appropriate SIDs.

GRE-KEY: Using GRE Key value as a key for SR ingress PE to map to the appropriate SIDs

5. SDWAN path over multiple SP managed domains

The following figure shows a SDWAN Path E1<->E2 over two SP domains which are interconnected by public internet.

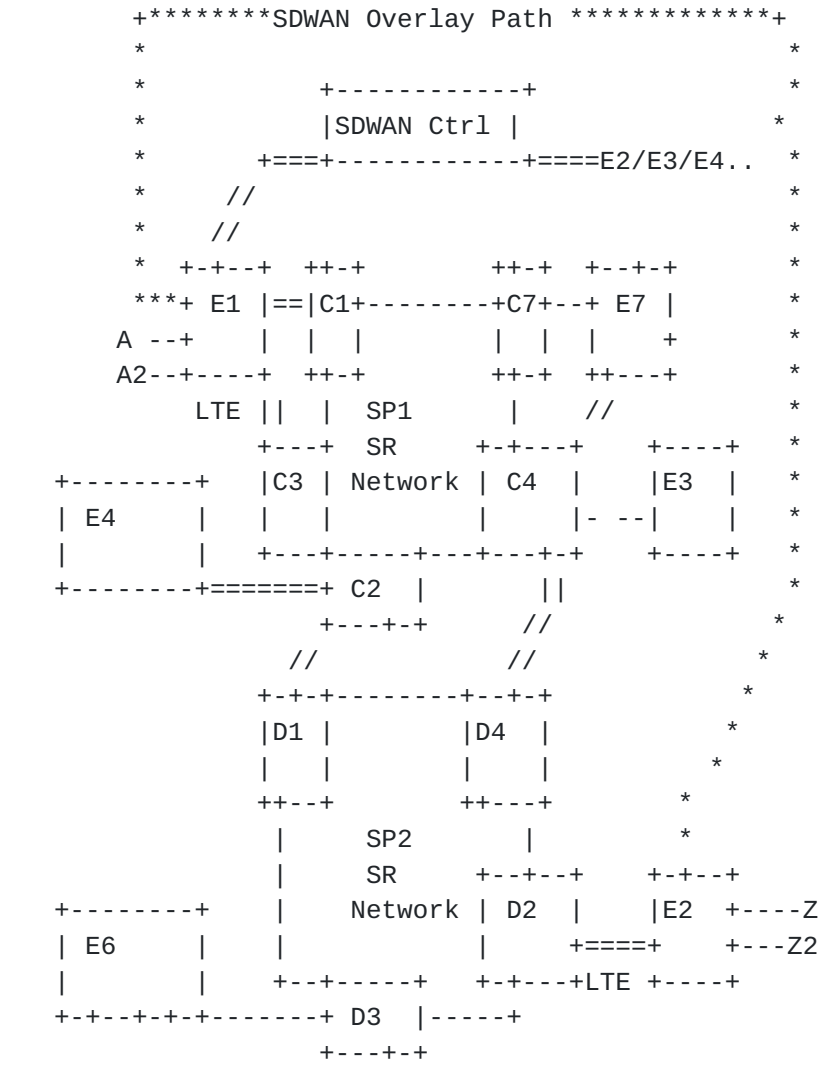


Figure 5: SDWAN path over two different SP domains

Let's assume that the SP-1 domain's egress node for the SDWAN path E1<->E2 is C2, which can reach D1 or D4 of SP-2 via public IP network (say IPv4 network).

Let's also assume that the optimal route for some flows over SDWAN path E1<->E2 are C1->C2->D1 and other flows are over C1->C2->D4 (out of the scope of this document on how the path is calculated).

If SP-1 is SR enabled, the mechanism described in Section 4 is applicable to the SDWAN path source node E1 and the SP-1's ingress PE (e.g. C1 or C3 in the figure).
However, the processing at egress node might be different depending on how the SP-1's egress edges are connected to the SP-2's ingress edge nodes.

5.1. When Both SP domains support SR

There may be three approaches here:

1) Controller installs the entire SID stack at E1, and the SID list contains SID entries belong to both domains.

2) Controller delivers to E1 the SID stack that only for the first domain, but delivers to C6 (egress node of first domain) the binding SID of the second domain.

3) Controller delivers a "Key" to E1, which can be encoded as GRE KEY or represented by the Source UDP port of the GRE encapsulation, for Ingress PE of the first SR Domain to map to its own SID stack as described in Section 4. The first SR Domain will reserve the "Key" through its domain and pass the "Key" to the second SR domain. The second SR Domain Ingress node will use the same method to map the "Key" to its SID stack.

5.2. When SP-2 does not support SR

Under this circumstance (which can be caused by SP-2 not supporting SR or not willing to share Binding SIDs to SP-1), if the packets arriving at SP-1 egress node C6 do not have any metadata indicating the types of encrypted payload, C6 does not really have much choice other than simply forwarding the packets to E2 via public IP

network. This way, the packets may or may not traverse through the SP-2 domain. If the distance between C6 and E2 is far, the quality of service can be unpredictable.

5.3. When SP-1 and SP-2 don't want to share network information

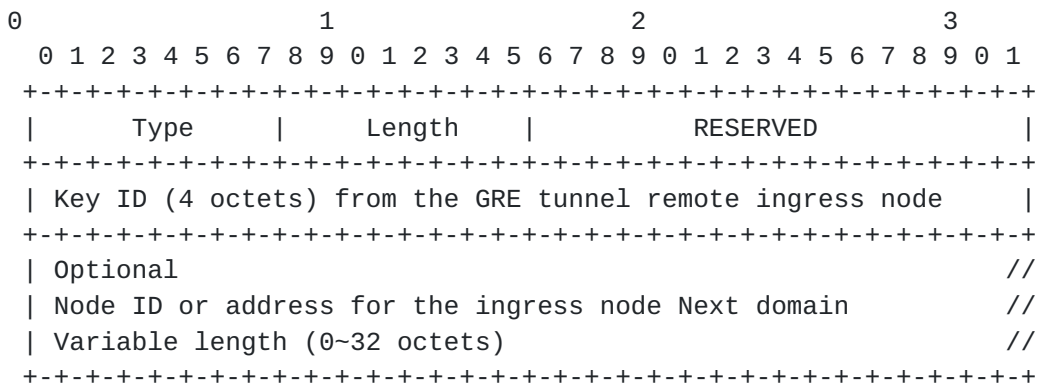
If SP-1's ingress node C1 can include the GRE KEY it receives from E1 in the data packets' SR header, the SP-1's egress node can map the Key to the SP-2's Ingress node and encapsulate the data packet in a new GRE header destined towards the SP-2's Ingress node. Then the SP-2's Ingress node can follow the procedure described in the Section 4 to forward the data packets across its domain.

If the first SR Domain does not support adding metadata to carry the "key" through its domain, the controller can deliver the "key" to SP-1's egress node the same time as it delivers the key to E1, knowing the SDWAN path will need to traverse two domains with the second one does support SR but the two SPs don't want to exchange network information.

5.4. TLV to pass Metadata through SRv6 Domain

If SP-1 is SRv6 based, the ingress node C1 can append a TLV to the end of the SR Header [SRv6-SRH] to carry the KEY it receives from E1.

The SP-1 egress node C6 can get the mapping between the KEYs and the Node-IDs (or Addresses) of the next domain's ingress edge node (i.e. D1 or D4 in the figure 3 above) from its network controller ahead of time.



TYPE: (to be assigned by IANA) is to indicate the TLV is for carrying the flow identifier of the packet encoded by the SDWAN source node.

Upon receiving the packet, the egress node (C6) can

- find the Node-ID (or the address) for the next domain's ingress node,
- construct a GRE header with the Key received from the TLV above and the destination address from the mapping given by the controller,
- encapsulate the GRE header to the data packet (which has decapsulated SR header),
- and forward the packet to the public internet.

6. Security Considerations

Remotely attached CPEs might brought the following security risks:

1) Potential DDoS attack to the PEs with ports facing internet.
  I.e. the PE resource being attacked by unwanted traffic.
2) Potential risk of provider VPN network bandwidth being stolen
  by the entities who spoofed the addresses of SDWAN end nodes.

To mitigate security risk of 1) above, it is absolutely necessary for PEs which accept remotely attached CPEs or simply have ports facing internet to enable Anti-DDoS feature to prevent major DDoS attack to those PEs.

To mitigate the security risk of 2) above, RFC7510 defines the use of DTLS to authenticate and encrypt the RFC7510 encapsulation.

However, for the scenario of SDWAN source node being remotely attached to PEs, using the method recommended by RFC7510 means the source node has to perform DTLS on top of the IPSec encryption between SDWAN end points E1<->E2. This can be too processing heavy for the SDWAN end nodes. In addition, if there are many SDWAN flows

to traverse through the ingress PE (e.g. C1, C2, C4 in the figure 1 above), heavy processing is required on the ingress PEs.

Since the payload between E2<->E2 is already encrypted, the confidentiality of the payload is already ensured. The network operators need to balance between how much they can tolerant some percentage of bandwidth being stolen and how much extra cost they are willing to pay for completely prevent any unpaid traffic traversing through its VPN networks. For operators who opt for lower cost ingress PEs and CPEs, but can tolerant some percentage of bandwidth being used by unpaid subscribers, a simple approach can be considered:

- Embed a key in the packets, which can be changed periodically, like the digital signature used by a certificate authority or certification authority (CA).
- The key can be encoded in the GRE Key field between SDWAN end node and Ingress PE. Since GRE has 24 bits, some fixed bits can be used to represent the signature of paid subscribers.

7. IANA Considerations

This document requires new protocol type:

Protocol type to be added to GRE header: SR_Route

8. References

8.1. Normative References

[RFC2890] G. Dommety "Key and Sequence Number Extensions to GRE". Sep. 2000.

8.2. Informative References

[RFC2735] B. Fox, et al "NHRP Support for Virtual Private networks". Dec. 1999.

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

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

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

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

[RFC4664] L. Andersson and E. Rosen, "Framework for Layer 2 Virtual Private Networks (L2VPNs)", Sept 2006.

[SR-SDWAN] D. Dukes, et al, "SR for SDWAN: VPN with Underlay SLA", draft-dukes-sr-for-sdwan-00, in progress, Oct 2017

[SRv6-SRH] S. Previdi, et al, "IPv6 Segment Routing Header (SRH)", draft-ietf-6man-segment-routing-header-13, in progress, April 2018.

[MPLS-SR] A. Bashandy, et al, "Segment Routing with MPLS data plane", draft-ietf-spring-segment-routing-mpls-13, in progress, April 2018.

[RFC7510] X. Xu, et al, "Encapsulating MPLS in UDP", April 2015.

[RFC8086] L. Yong, et al, "GRE-in-UDP Encapsulation", March 2017.

[BGP-SDWAN-Usage] L. Dunbar, et al, "BGP Usage for SDWAN Overlay Networks, draft-dunbar-bess-bgp-sdwan-usage-01, in progress, July 2019.

[SDWAN-Net2Cloud] L. Dunbar, et al, "Dynamic Networks to Hybrid Cloud DCs Problem Statement", draft-ietf-rtgwg-net2cloud-problem-statement-04, in progress, July 2019.

[MEF-Cloud] "Cloud Services Architecture Technical Specification", Work in progress, April 2018

9. Acknowledgments

Many thanks to Dean Cheng and Jim Guichard for the discussion and contributions.

Authors' Addresses

Linda Dunbar
Futurewei
Email: Linda.Dunbar@futurewei.com

Mehmet Toy
Verizon
One Verizon Way
Basking Ridge, NJ 07920
Email: mehmet.toy@verizon.com