Network Working Group
Internet Draft
Intended status: Standards Track
Expires: January 2019

L. Dunbar
H. Wang
W. Hao
Huawei

November 7, 2018

BGP Extension for SDWAN Overlay Networks
draft-dunbar-idr-bgp-sdwan-overlay-ext-03

Abstract

The document defines a new BGP SAFI with a new NLRI in order to advertise a SD-WAN node's properties with other SD-WAN nodes through third party untrusted networks. The goal is for SD-WAN network to scale; enabling SD-WAN overlay among large number of SD-WAN nodes with minimal provisioning, and allow services to be carried by different SD-WAN transport networks based on user specified criteria.

A "SD-WAN" network consists of many segments of IPsec tunnels between SD-WAN nodes. An "end-point" is referring to a port on a SD-WAN node throughout this document.

This document specifies new sub-TLVs for the SD-WAN End Point's Attributes.

Copyright Notice

Table of Contents

1. Introduction

The document defines a new BGP SAFI with a new NLRI in order to advertise a SD-WAN node's properties with other SD-WAN nodes through third party untrusted networks. The goal is for SD-WAN network to scale; enabling SD-WAN overlay among large number of SD-WAN nodes with few provisioning needed, and allow services to be carried by different SD-WAN transport networks based on user specified criteria.

A "SD-WAN" network consists of many segments of IPsec tunnels between SD-WAN nodes. An "end-point" is referring to a port on a SD-WAN node throughout this document.

[Net2Cloud-Problem] describes the problems that enterprises face today in transitioning their IT infrastructure to support digital economy, such as the need to connect enterprises' branch offices to dynamic workloads in different Cloud DCs, or aggregating multiple paths provided by different service providers to achieve better experience.

Even though SD-WAN has been used as a flexible way to reach workloads in dynamic third party data centers or aggregate multiple underlay paths, scaling becomes a big issue when there are hundreds or thousands of nodes to be interconnected by the SD-WAN overlay paths.

BGP is widely used by underlay networks. This document expand the BGP to make SD-WAN overlay network scale better.

2. Conventions used in this document

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

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

CPE-Based VPN: Virtual Private Secure network formed among CPEs. This is to differentiate from most commonly used PE-based VPNs a la RFC 4364.

OnPrem: On Premises data centers and branch offices

SD-WAN End-point: a port (logical or physical) of a SD-WAN node.

SD-WAN: Software Defined Wide Area Network, which can mean many different things. In this document, "SD-WAN" refers to the solutions specified by ONUG (Open Network User Group), which build point-to-point IPsec overlay paths between two end-points (or branch offices) that need to intercommunicate.

SD-WAN Transport Network: One transport network between two SD-WAN nodes. There could be multiple transport networks between two SD-WAN nodes.

SD-WAN IPsec Tunnel: IPsec tunnel over a Transport Network.

3. Key Characteristics of SD-WAN Overlay Network

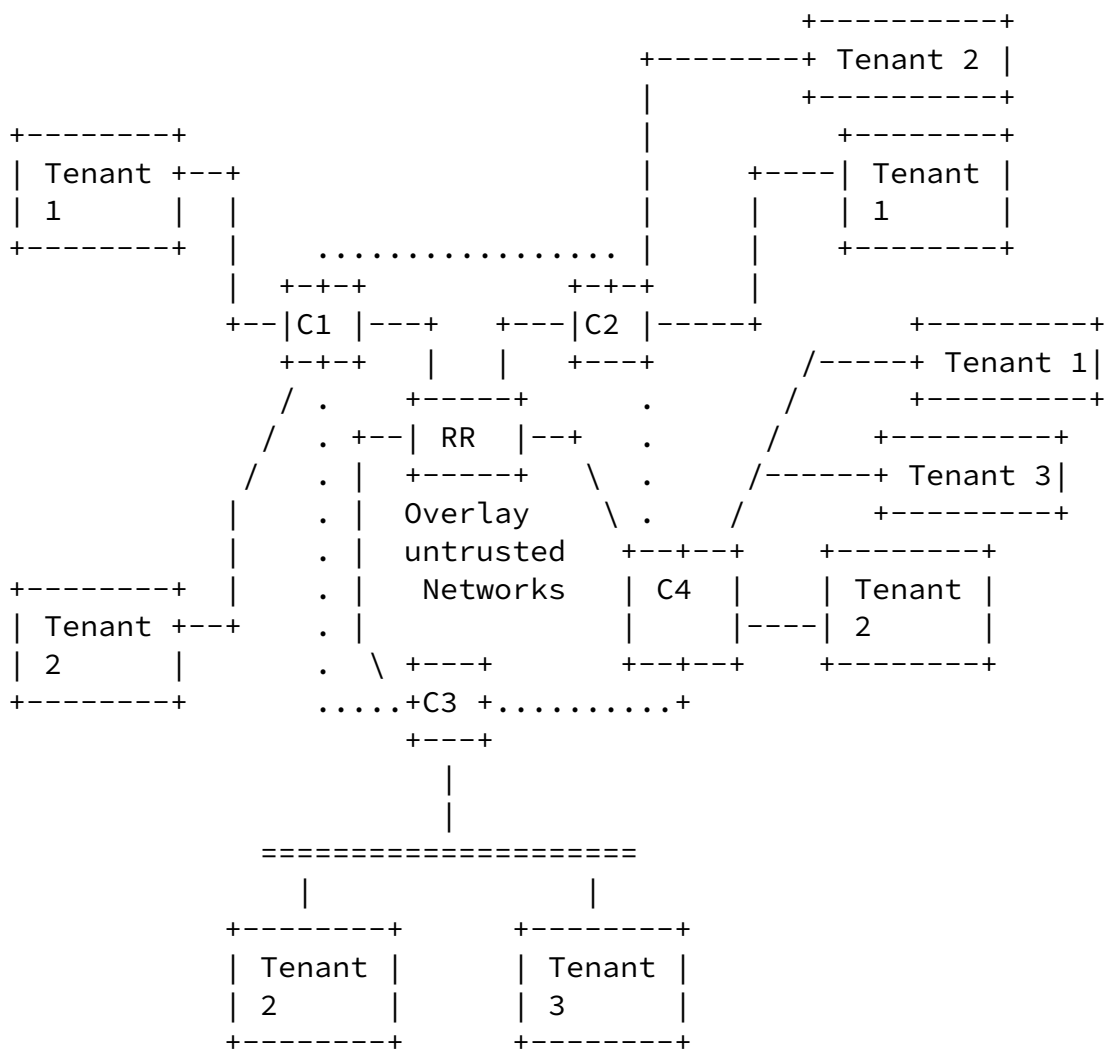
A SD-WAN Overlay consists of many SD-WAN nodes that are interconnected by multiple untrusted underlay transport networks. For a small sized SD-WAN network, traditional hub & spoke model using NHRP or DSVPN/DMVPN with a hub node (or controller) managing SD-WAN node end-points (e.g. local & public addresses and tunnel identifiers mapping) can work reasonably well. However, for a large SD-WAN network, say more than 100 nodes with different types of topologies, the traditional approach becomes very messy, complex and error prone.

Here are some characteristics of SD-WAN Overlay network:

1. SD-WAN IPsec - Transport establishment needs to be separate from Routes/services attached to each site.
2. Route distribution has to be independent from multiple transport networks between sites
   - Site based routes (instead of port based routes)
3. Transport selection between sites are local section. Same service can use different Transport networks between sites.
   - Different services, routes, or VLANs can be carried by one SD-WAN Transport; same service/routes/VLAN can be carried by different SD-WAN Transport at different time depending on the policies specified by users.
4. Managing IPsec Keys and re-Keys are complicated, it does not scale well if a SD-WAN end point has to manage many fine-grained tunnels with its peers, such as per route, per VLAN based SD-WAN IPsec tunnel.

In addition, hosts/applications attached to SD-WAN nodes can belong to different tenants, requiring SD-WAN nodes to establish different tunnels to different SD-WAN nodes. As shown in the figure below, C1 node alone has to establish following SD-WAN tunnels:

- two SD-WAN tunnels to C2: one for Tenant 1, another one for Tenant 2,
- One SD-WAN tunnel to C3 for Tenant 1, and
- two SD-WAN tunnels to C4: one for Tenant 1, another one for Tenant 2,



This document proposes a method of using BGP for a SD-WAN node to advertise its SD-WAN capabilities and SD-WAN end-point properties to other SD-WAN nodes.

[Tunnel-Encaps] removed SAFI =7 (which was specified by RFC5512) for distributing encapsulation tunnel information. [Tunnel-Encap] require Tunnels being associated with routes.

The mechanisms described by [Tunnel-Encap] cannot be effectively used for SD-WAN overlay network because:

- SD-WAN Tunnel needs to be established before data arrival because it takes several rounds of negotiation between two end-points to agree upon the encryption algorithms, exchange public keys, and be authorized to communicate with each other. Unlike an EVPN PE, which can always establish VxLAN tunnel to a another PE, an IPsec tunnel

between two points might fail to be established due to no agreed upon encryption mechanism.
- Different services, routes, or VLANs can be carried by one SD-WAN tunnel; same service/routes/VLAN can be carried by different SD-WAN tunnels at different time depending on the policies specified by users. When one SD-WAN tunnel encounter more congestion or delay, a subset of the services/routes/VLAN carried by the SD-WAN tunnel have to be switched to a different SD-WAN tunnel.
- When a VLAN or a route is deleted/added from/to an SD-WAN node, the SD-WAN Tunnel between this node and another node should not be impacted.
- Managing IPsec Keys and re-Keys are complicated, it does not scale well if a SD-WAN end point has to manage many fine-grained tunnels with its peers, such as per route, per VLAN based SD-WAN IPsec tunnel.
- Sometimes, a SD-WAN tunnel has to traverse a specific location due to policies or running environment.

There is a suggestion on using a "Fake Route" for a SD-WAN node to use [Tunnel-Encap] to advertise its SD-WAN tunnel end-points properties. However, using "Fake Route" can create deployment complexity for large SD-WAN networks with many tunnels. For example, for a SD-WAN network with hundreds of nodes, with each node having many ports & many end-points to establish SD-WAN tunnels to their corresponding peers, the node would need many "fake addresses". For large SD-WAN networks (such as has more than 10000 nodes), each node might need 10's thousands of "fake addresses", which is very difficult to manage and needs lots of configuration to get the nodes provisioned.

The key value proposition of SD-WAN is its dynamic nature. Most SD-WAN deployment requires the following key properties:

- Zero Touch Provisioning: meaning a SD-WAN node needs to be plug and play. The huge amount of "fake addresses" configurations required by the [Tunnel-Encap] mechanism make it not possible to be used for SD-WAN tunnels.
- The IP address of ports to a SD-WAN node can be dynamic (e.g. assigned by DHCP); therefore, there is no fixed IP address that can be used to uniquely to represent a SD-WAN tunnel end-point. "System-ID + PortID" can usually uniquely identify a SD-WAN end-point. That means the nexthop of a SD-WAN tunnel can be "System-ID + Port ID". Sometimes, a SD-WAN tunnel end-point can be associated with "private IP" + "public IP" (if NAT is used.)

Another very important reason for needing a specific SAFI for SD-WAN Overlay is for many intermediate nodes that do not terminate SD-WAN tunnels to ignore the NLRI SD-WAN Overlay SAFI update messages, to avoid the extra processing incurred.

[Net2Cloud-gap] has more in-depth analysis of the gaps of available protocols in support SD-WAN overlay networks.

4. Overview of the BGP Extension for SD-WAN

To avoid confusion of different interpretation of SD-WAN, the BGP SD-WAN Overlay NLRI extension described in this document is for a SD-WAN deployment with the following characteristics:

- There is a Central Controller, which can be reached by an SD-WAN node

upon power up, and a TLS or SSL secure channel can be established between the SD-WAN node and the Central Controller.

- The Central Controller can designate a Local Controller in the proximity of the SD-WAN node; the Local Controller and the SD-WAN nodes might be connected by third party untrusted network. In the context of using BGP to control the SD-WAN overlay network, Route Reflector (RR, [RFC4456]) can act as a Local Controller. The SD-WAN node can establish a secure connection (TLS, SSL, etc) to the Local Controller (RR).

The BGP SD-WAN Overlay NLRI extension described in this document is for SD-WAN nodes to advertise their SD-WAN capabilities & tunnel end-points attributes to peers belonging to the same tenant, such as

a. to advertise the identifiers of ports that support establishing SD-WAN overlay tunnels to other peers,
b. to advertise ports private addresses (or dynamically assigned IP addresses),
c. to advertise its supported IPsec capability, such as the supported encryption algorithms, etc.

Since there are secure channels (TLS, SSL, etc.) established between the Local Controller (i.e. RR) and SD-WAN nodes, the NLRI can be advertised to their peers belonging to the same tenants via the secure channel to/from the RR.

The BGP extension for the advertisement of SD-WAN tunnels includes following components:

- A new Subsequent Address Family Identifier (SAFI=74) whose NLRI identifies a (SD-WAN) overlay tunnel, the properties of the tunnel end-points, and the associated policies.
- A new Route Type that defines the encoding of the rest of the SD-WAN Overlay NLRI, and a set of sub-TLVs to specify the tunnel & its end-point attributes, policies associated with the tunnel, etc. Here are the sub-TLVs needed for SD-WAN tunnel:
  - o Tunnel IPsec configuration attributes, such as public keys, the encryption algorithms, etc.
  - o Tunnel Encap Extension, which is for specify specific attributes associated for SD-WAN tunnel end-points.
- Port Distinguisher: one (SD-WAN) node can have multiple ports, and each port can support multiple SD-WAN tunnels to different peers. The

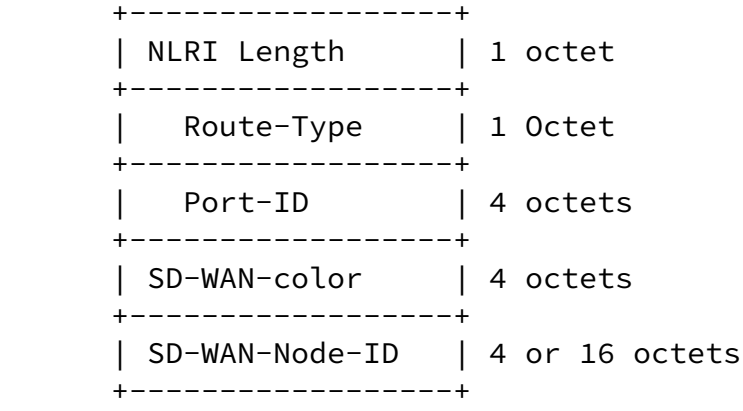
Port Distinguisher is used to describe port (or link identifier).

- SD-WAN Color: used to identify a common property shared by a set of SD-WAN nodes, such as the property of a specific geographic location. The property is used to steer an overlay route to traverse specific geographic locations for various reasons, such as to comply regulatory rules, to utilize specific value added services, or others.

5. SD-WAN Over Tunnel NLRI Format

The new SAFI=74, the SD-WAN Overlay SAFI, has been assigned by IANA, from the "Subsequent Address Family Identifiers (SAFI) Parameters" registry.

The SD-WAN Overlay SAFI (=74) uses a new NLRI defined as follows:



where:

- NLRI Length: 1 octet of length expressed in bits as defined i [RFC4760].
- Route-Type: to define the encoding of the rest of the SD-WAN NLRI.
- Port ID: one (SD-WAN) node can have multiple ports, and each support multiple SD-WAN tunnels to different peers. The Port ID is used to identify the port, a.k.a. link identifier.
- SD-WAN-color: used to identify a common property shared by a SD-WAN nodes, such as the property of a specific geographic location.
- SD-WAN Node ID: the SD-WAN NLRI advertisement is sent out by

WAN node to indicate all the available ports supporting SD-WAN tunnels. The SD-WAN Node ID can be the node's system ID, such as the loopback address of the SD-WAN node.

6. SD-WAN Tunnel Encapsulation Attribute sub-TLV:

The SD-WAN overlay tunnel end-points property is encoded in the Tunnel Encapsulation Attribute originally defined in [Tunnel-Encap]using a new Tunnel-Type TLV (SD-WAN Tunnel Type, with the code point to be assigned by IANA) from the "BGP Tunnel Encapsulation Attribute Tunnel Types".

The SD-WAN Tunnel End-Point Property Encoding structure is as follows:

Overlay SAFI (=74) NLRI: < Route-Type, Length, Port-ID, SD-WAN-color, SD-WAN-Node-ID>
Attributes:
   Tunnel Encaps Attribute
       Tunnel Type: SD-WAN-Tunnel
            EncapExt SubTLV
            IPsec-SA Attribute SubTLV

Where
- Encap-Ext SubTLV is for describing additional information about the SD-WAN tunnel end-points, such as NAT property.
- IPsec-SA SubTLV is for the node to establish IPsec SA with other peers.

The Tunnel Encaps Attribute are defined as follows:

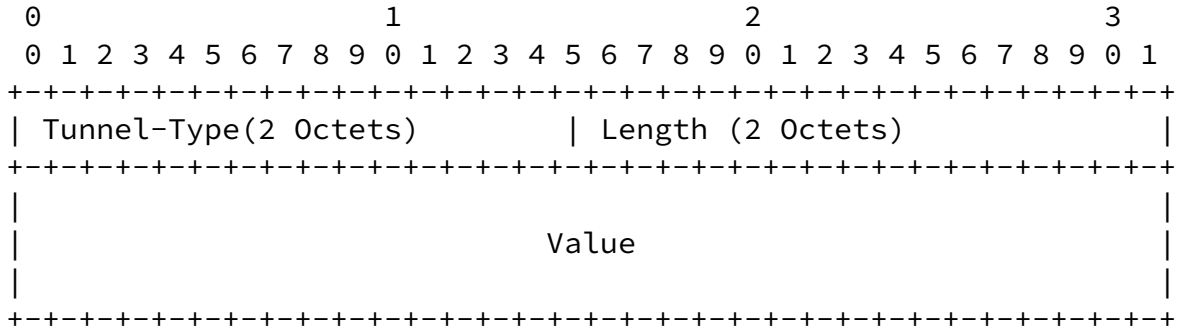


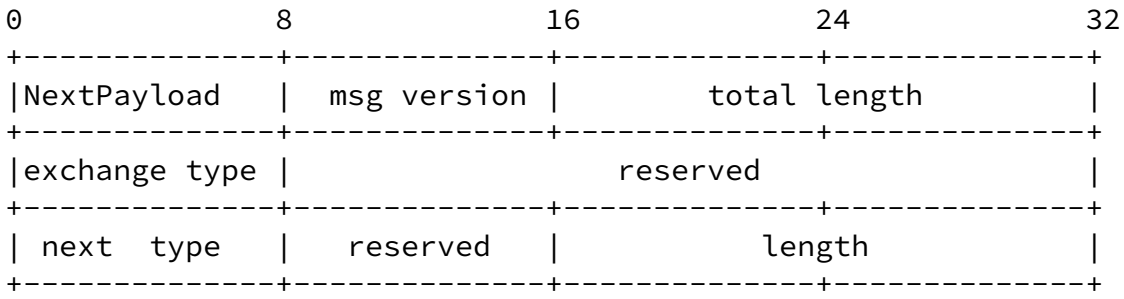
Figure 1: SD-WAN Tunnel Encapsulation TLV Value Field

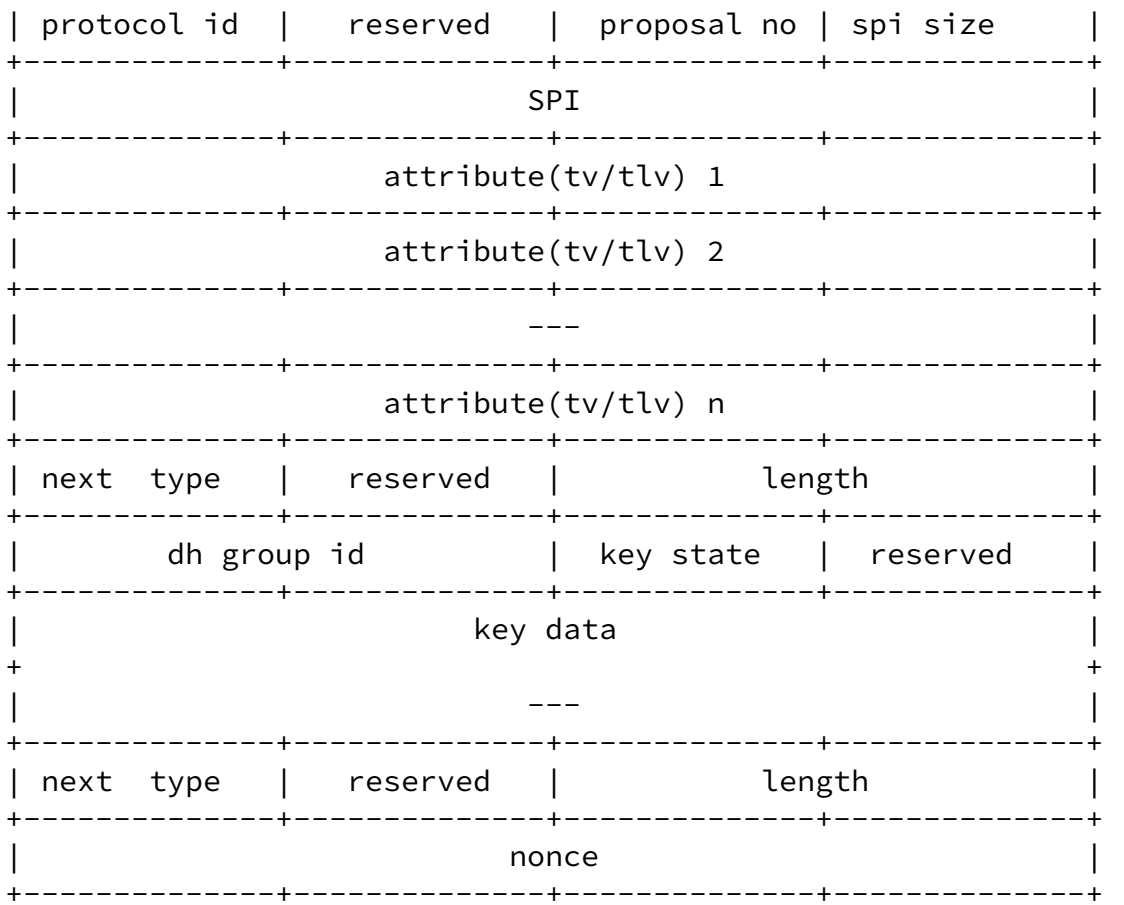
Where:

Tunnel Type is SD-WAN (to be assigned by IANA).

6.1. IPsec SA sub-TLV

The IPsecSA sub-TLV is for the SD-WAN node to establish IPsec security association with their peers:





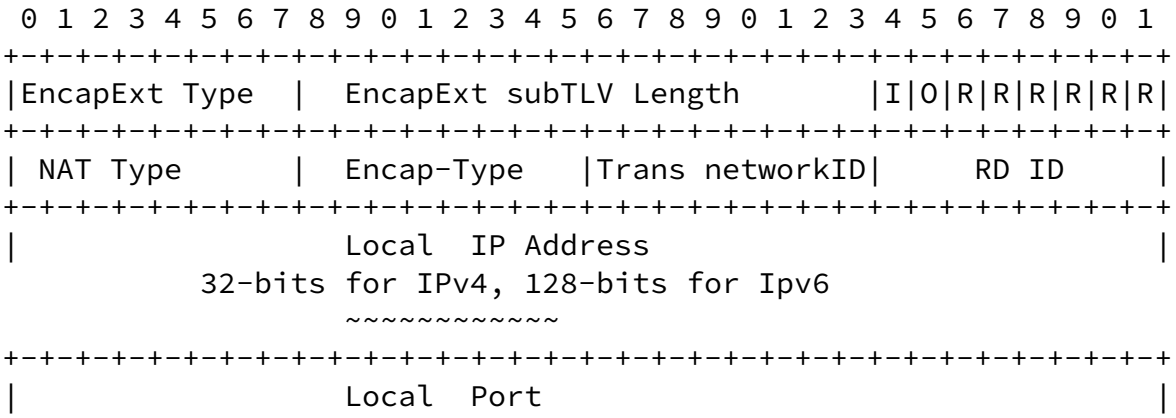
Where:

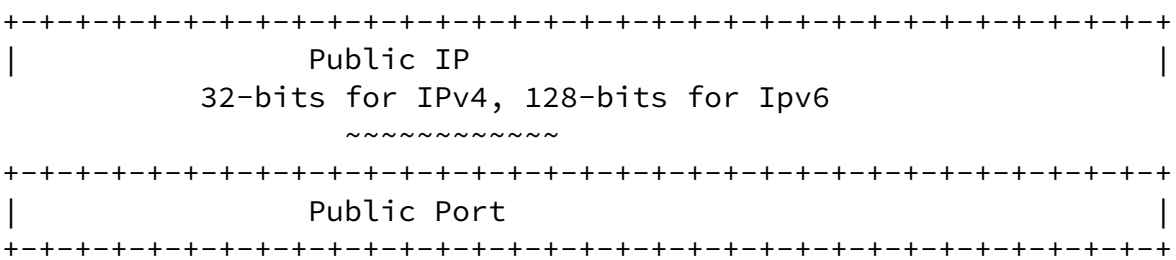
- o IPsec properties such as AH, ESP, or AH+ESP are encoded in the "Attributes", such as
  - o Tunnel Mode or Transport mode
  - o AH authentication algorithms supported, which can be md5 | sha1 | sha2-256 | sha2-384 | sha2-512 | sm3. Each SD-WAN node can have multiple authentication algorithms; send to its peers to negotiate the strongest one.
  - o ESP authentication algorithms supported, which can be md5 | sha1 | sha2-256 | sha2-384 | sha2-512 | sm3. Each SD-WAN node can have multiple authentication algorithms; send to its peers to negotiate the strongest one. Default algorithm is AES-256.

- o Duration: SA life span.

6.2. EncapsExt sub-TLV

EncapsExt sub-TLV is for describing additional information about the SD-WAN tunnel end-points, such as NAT property. A SD-WAN node can inquire STUN (Session Traversal of UDP Through Network Address Translation RFC 3489) Server to get the NAT property, the public IP address and the Public Port number to pass to peers.





Where:

- o EncapExt Type: indicate it is the EncapExt SubTLV.
- o EncapExt subTLV Length: the length of the subTLVE.
- o Flags:
  - I bit (CPE port address or Inner address scheme)
    If set to 0, indicate the inner (private) address is IPv4.
    If set to 1, it indicates the inner address is IPv6.

  - O bit (Outer address scheme):
    If set to 0, indicate the public (outer) address is IPv4.

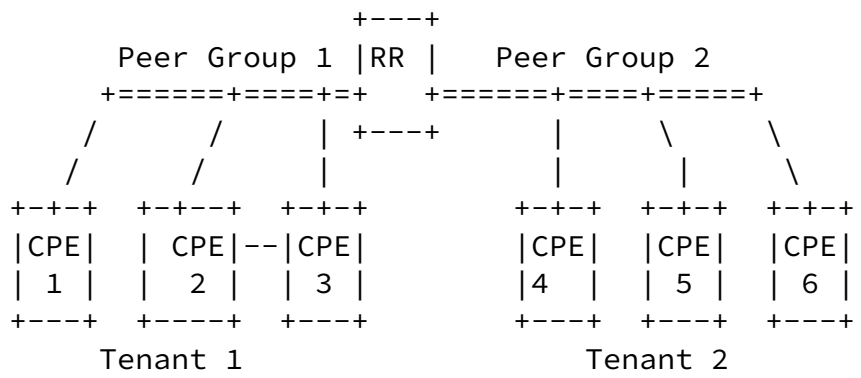
    If set to 1, it indicates the public (outer) address is IPv6.

  - R bits: reserved for future use. Must be set to 0 now.

- o NAT Type.without NAT; 1:1 static NAT; Full Cone; Restricted Cone; Port Restricted Cone; Symmetric; or Unknown (i.e. no response from the STUN server).
- o Encap Type.SD-WAN tunnel encapsulation types, such as IPsec+GRE, IPsec+VxLAN, IPsec without GRE, GRE (when tunnel is over secure underlay network)
- o Transport Network ID.Central Controller assign a global unique ID to each transport network.
- o RD ID.Routing Domain ID.Need to be global unique.
- o Local IP.The local (or private) IP address of the tunnel end-point.
- o Local Port.used by Remote SD-WAN node for establishing IPsec to this specific port.
- o Public IP.The IP address after the NAT.

o Public Port.The Port after the NAT.

7. SD-WAN Tunnel Advertisement Method:



For SD-WAN overlay network, the SD-WAN nodes (a.k.a. CPEs) belonging to the same Tenant can be far apart and can be connected by third party untrusted networks. Therefore, it is not appropriate for a SD-WAN node (CPE) to advertise its SD-WAN tunnel properties to its immediate neighbors. Each CPE

propagates its SD-WAN tunnel attributes via the secure channel established with RR.

The processing steps on CPE1 are as follow:

- Report the SD-WAN tunnel information, such as IPsec property, NAT, etc. to RR via the Overlay SAFI NLRI.
- RR propagate the information to CPE2 & CPE 3.
- CPE2 and CPE3 can establish IPsec SA with the CPE1 after receiving the Overlay SAFI NLRI from RR.

Tenant separation is achieved by different SD-WAN nodes being added to different Peer Group.

8. Manageability Considerations

TBD

9. Security Considerations

The intention of this draft is to identify the gaps in current and proposed SD-WAN approaches that can address requirements identified in [Net2Cloud-problem].

Several of these approaches have gaps in meeting enterprise security requirements when tunneling their traffic over the Internet, as is the general intention of SD-WAN. See the individual sections above for further discussion of these security gaps.

10. IANA Considerations

This document requires the following IANA actions.

- o SD-WAN Overlay SAFI = 74 assigned by IANA
- o SD-WAN Route Type
- o SD-WAN Tunnel Type
- o IPsec-SA Type
- o EncapExt Type

11. References

11.1. Normative References

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

11.2. Informative References

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

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

[Tunnel-Encap]E. Rosen, et al, "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-09, Feb 2018.

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

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

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

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

[Net2Cloud-Problem] L. Dunbar and A. Malis, "Seamless Interconnect Underlay to Cloud Overlay Problem Statement", draft-dm-net2cloud-problem-statement-02, June 2018

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

[Tunnel-Encap] E. Rosen, et al "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-10, Aug 2018.

12. Acknowledgments

Acknowledgements to Jim Guichard, John Scudder, Darren Dukes, Andy Malis and Donald Eastlake for their review and contributions.

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

Authors' Addresses

Linda Dunbar
Huawei
Email: Linda.Dunbar@huawei.com

Haibo Wang
Huawei
Email: rainsword.wang@huawei.com

WeiGuo Hao
Huawei Technologies Co.,Ltd.
Email: haoweiguo@huawei.com