Network Working Group
Internet Draft
Intended status: Informational
Expires: December 19, 2019

L. Dunbar
A. Malis
Futurewei
C. Jacquenet
Orange

June 19, 2019

Gap Analysis of Dynamic Networks to Hybrid Cloud DCs
draft-ietf-rtgwg-net2cloud-gap-analysis-02

Abstract

This document analyzes the technological gaps when using SD-WAN to dynamically interconnect workloads and applications hosted in various 3rd party cloud data centers.

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction

[Net2Cloud-Problem] describes the problems that enterprises face today in transitioning their IT infrastructure to support digital economy, such as connecting enterprises' branch offices to dynamic workloads in different Cloud DCs.

This document analyzes the technological gaps to interconnect dynamic workloads & apps hosted in cloud data centers that the enterprise's VPN service provider may not own/operate or may be unable to provide the enterprise with the required connectivity to access such locations. When VPN service providers have insufficient bandwidth to reach a location, SD-WAN techniques can be used to aggregate bandwidth of multiple networks, such as MPLS VPNs or the Public Internet to achieve better performance. This document primarily focuses on the technological gaps raised by using SD-WAN techniques to connect enterprise premises to cloud data centers operated by third parties.

For the sake of readability, a SD-WAN edge, a SD-WAN endpoint, C-PE, or CPE are used interchangeably throughout this document. However, each term has some minor emphasis, especially when used in other related documents:

- SD-WAN Edge: could include multiple devices (virtual or physical);
- SD-WAN endpoint: to refer to a WAN port of SD-WAN devices or a single SD-WAN device;
- C-PE: more for provider owned SD-WAN edge, e.g. for SECURE-EVPN's PE based VPN, when PE is the edge node of SD-WAN;
- CPE: more for enterprise owned SD-WAN edge.

2. Conventions used in this document

Cloud DC: Third party 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 Network designed and deployed from 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: Software Defined Wide Area Network, "SD-WAN" refers to the solutions of pooling WAN bandwidth from multiple underlay networks to get better WAN bandwidth management, visibility & control. When the underlay is a private network, traffic may be forwarded without any additional encryption; when the underlay networks are public, such as the Internet, some traffic needs to be encrypted when passing through (depending on user-provided policies).

3. Gap Analysis of C-PEs WAN Port Registration

SD-WAN technology has emerged as means to dynamically and securely interconnect the OnPrem branches with the workloads instantiated in Cloud DCs that do not have direct connectivity to BGP/MPLS VPN PEs or have very limited bandwidth.

Some SD-WAN networks use the NHRP protocol [RFC2332] to register WAN ports of SD-WAN edges with a "Controller" (or NHRP server), which then has the ability to map a private VPN address to a public IP address of the destination node. DSVPN [DSVPN] or DMVPN [DMVPN] are used to establish tunnels between WAN ports of SD-WAN edge nodes.

NHRP was originally intended for ATM address resolution, and as a result, it misses many attributes that are necessary for dynamic endpoint C-PE registration to the controller, such as:

- Interworking with the MPLS VPN control plane. A SD-WAN edge can have some ports facing the MPLS VPN network over which packets can be forwarded without any encryption and some ports facing the public Internet over which sensitive traffic needs to be encrypted before being sent.

- Scalability: NHRP/DSVPN/DMVPN works fine with small numbers of edge nodes. When a network has more than 100 nodes, these protocols do not scale well.
- NHRP does not have the IPsec attributes, which are needed for peers to build Security Associations over the public internet.
- NHRP messages do not have any field to encode the C-PE supported encapsulation types, such as IPsec-GRE or IPsec-VxLAN.
- NHRP messages do not have any field to encode C-PE Location identifiers, such as Site Identifier, System ID, and/or Port ID.
- NHRP messages do not have any field to describe the gateway(s) to which the C-PE is attached. When a C-PE is instantiated in a Cloud DC, it is desirable for C-PE's owner to be informed of how/where the C-PE is attached.
- NHRP messages do not have any field to describe C-PE's NAT properties if the C-PE is using private addresses, such as the NAT type, Private address, Public address, Private port, Public port, etc.

[BGP-SDWAN-PORT] describes how SD-WAN edge nodes use BGP to register their WAN ports properties to the SD-WAN controller, which then propagates the information to other SD-WAN edge nodes that are authenticated and authorized to communicate with them.

4. Aggregating VPN paths and Internet paths

Most likely, enterprises (especially the largest ones) already have their CPEs interconnected by providers' VPNs, based upon VPN techniques such as EVPN, L2VPN, or L3VPN. The VPN can be PE-based or CPE-based. The commonly used PE-based VPNs have CPE directly attached to PEs, therefore the communication between CPEs and PEs is considered as secure. MP-BGP is used to learn & distribute routes among CPEs, even though sometimes routes among CPEs are statically configured on the CPEs.

To aggregate paths over the Internet and paths over the VPN, the C-PEs need to have some WAN ports connected to the PEs of the VPNs and other WAN ports connected to the Internet. It is necessary for the CPEs to use a protocol so that they can register the WAN port properties with their SD-WAN Controller(s): this information conditions the establishment and the maintenance of IPsec SA associations among relevant C-PEs.

When using NHRP for registration purposes, C-PEs need to run two separate control planes: EVPN&BGP for CPE-based VPNs, and NHRP & DSVPN/DMVPN for ports connected to the Internet. Two separate control planes not only add complexity to C-PEs, but also increase operational cost.

```
                                       +---+
                        +--------------|RR |----------+
                       /  Untrusted    +-+-+           \
                      /                                 \
                     /                                   \
     +----+  +---------+  packets encrypted over     +------+  +----+
     | TN3|--|         A1-----+ Untrusted    +------ B1     |--| TN1|
     +----+  | C-PE    A2-\                          | C-PE |  +----+
     +----+  |  A      A3--+--+              +---+---B2  B  |  +----+
     | TN2|--|         |   |PE+--------------+PE |---B3     |--| TN3|
     +----+  +---------+   +--+   trusted    +---+   +------+  +----+
                              |      WAN     |
     +----+  +---------+   +--+   packets    +---+   +------+  +----+
     | TN1|--|         C1--|PE| go natively  |PE |-- D1     |--| TN1|
     +----+  | C-PE    C2--+--+ without encry+---+   | C-PE |  +----+
             |  C      |      +--------------+       |  D   |
             |         |                             |      |
     +----+  |         C3--|  without encrypt over   |      |  +----+
     | TN2|--|         C4--+---- Untrusted  --+------D2     |--| TN2|
     +----+  +---------+                             +------+  +----+
```

Figure 1: CPEs interconnected by VPN paths and Internet Paths

4.1. Key Control Plane Components of SD-WAN

As described in [BGP-SDWAN-Usage], the SD-WAN Overlay Control Plane has three distinct properties:

- SD-WAN node's WAN Port Property registration to the SD-WAN Controller.
  - o To inform the SD-WAN controller and authorized peers of the WAN port properties of the C-PE [SDWAN-Port]. When the WAN ports are assigned private addresses, this step can register the type of NAT that translates private addresses into public ones.

- Controller facilitated IPsec SA management and NAT information distribution

  - o It is for SD-WAN controller to facilitate or manage the IPsec configuration and peer authentication for all IPsec tunnels terminated at the SD-WAN nodes.

- Establishing and Managing the topology and reachability for services attached to the client ports of SD-WAN nodes.
  - o This is for the overlay layer's route distribution, so that a C-PE can populate its overlay routing table with entries that identify the next hop for reaching a specific route/service attached to remote nodes. [SECURE-EVPN] describes EVPN and other options.

4.2. Using BGP Tunnel-Encap

RFC5512 and [Tunnel-Encap] describe methods to construct BGP UPDATE messages that advertise endpoints' tunnel encapsulation capability and the respective attached client routes, so that the peers that receive of the BGP UPDATE can establish appropriate tunnels with the endpoints for the aforementioined routes. RFC5512 uses the Endpoint Address subTLV, whereas [Tunnel-Encap] uses Remote Endpoint Address subTLV to indicates address of the tunnel endpoint which can be an IPv4 or an IPv6 address. There are Tunnel Encapsulation attribute subTLVs to indicate the supported encapsulation types, such as L2TPv3, GRE, VxLAN, IP-in-IP, etc.

[Tunnel-Encap] removed SAFI =7 (which was specified by RFC5512) for distributing encapsulation tunnel information. [Tunnel-Encap] requires that tunnels need to be associated with routes.

There is also the Color sub-TLV to describe customer-specified information about the tunnels (which can be creatively used for SD-WAN).

Here are some of the gaps using [Tunnel-Encap] to control SD-WAN Tunnels:

- [Tunnel-Encap] doesn't have the functionality that would help the C-PE to register its WAN Port properties.
- A SD-WAN tunnel, e.g. IPsec-based, requires a negotiation between the tunnel's end points for supported encryption algorithms and tunnel types before it can be properly established, whereas [Tunnel-Encap] only allow the announcement of one endpoint's supported encapsulation capabilities for specific attached routes

and no negotiation between tunnel end points is needed. The establishment of a SD-WAN tunnel can fail, e.g., in case the two endpoints support different encryption algorithms. That is why a SD-WAN tunnel needs to be established and maintained independently from advertising client routes attached to the edge node.

- [Tunnel-Encap] requires all tunnels updates are associated with routes. There can be many client routes associated with the SD-WAN IPsec tunnel between two C-PEs' WAN ports; the corresponding destination prefixes (as announced by the aforementioned routes) may also be reached through the VPN underlay without any encryption. A more realistic approach to separate SD-WAN tunnel management from client routes association with the SD-WAN tunnels.
- When SD-WAN tunnel and clients routes are separate, the SD-WAN Tunnel establishment may not have routes associated.
  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 raise some design 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 endpoints to establish SD-WAN tunnels with their corresponding peers, the node would need as many "fake addresses". For large SD-WAN networks (such as those comprised of more than 10000 nodes), each node might need 10's thousands of "fake addresses", which is very difficult to manage and requires lots of configuration tasks to get the nodes properly set up.
- [Tunnel-Encap] does not have any field to carry detailed information about the remote C-PE, such as Site-ID, System-ID, Port-ID
- [Tunnel-Encap] Does not have any field to carry IPsec attributes for the SD-WAN edge nodes to establish IPsec Security Associations with others. It does not have any proper way for two peer CPEs to negotiate IPsec keys either, based on the configuration sent by the Controller.
- [Tunnel-Encap] does not have any field to indicate the UDP NAT private address <-> public address mapping
- C-PEs tend to communicate with a subset of the other C-PEs, not all the C-PEs need to be connected through a mesh topology. Without any BGP extension, many nodes can get dumped with too much

information coming from other nodes that they never need to communicate with.

4.3. SECURE-L3VPN/EVPN

[SECURE-L3VPN] describes how to extend the BGP/MPLS VPN [RFC4364] capabilities to allow some PEs to connect to other PEs via public networks. [SECURE-L3VPN] introduces the concept of Red Interface & Black Interface used by PEs, where the RED interfaces are used to forward traffic into the VPN, and the Black Interfaces are used between WAN ports through which only IPsec-protected packets are forwarded to the Internet or to other backbone network thereby eliminating the need for MPLS transport in the backbone.

[SECURE-L3VPN] assumes PEs using MPLS over IPsec when sending traffic through the Black Interfaces.

[SECURE-EVPN] describes a solution where point-to-multipoint BGP signaling is used in the control plane for SDWAN Scenario #1. It relies upon a BGP cluster design to facilitate the key and policy exchange among PE devices to create private pair-wise IPsec Security Associations without IKEv2 point-to-point signaling or any other direct peer-to-peer session establishment messages.

Both [SECURE-L3VPN] and [SECURE-EVPN] are useful, however, they both miss the aspects of aggregating VPN and Internet underlays. In summary:

- These documents do not address the scenario of C-PE having some ports facing VPN PEs and other ports facing the Internet.

- The [SECURE-L3VPN] assumes that a CPE "registers" with the RR. However, it does not say how. It assumes that the remote CPEs are pre-configured with the IPsec SA manually. In SD-WAN, Zero Touch Provisioning is expected. Manual configuration is not an option, as it contradicts the objectives of SD-WAN to automate configuration tasks.
- For RR communication with C-PEs, this draft only mentions IPsec. Missing TLS/DTLS.
- The draft assumes that C-PEs and RR are connected with an IPsec tunnel. With zero touch provisioning, we need an automatic way to synchronize the IPsec SAs between C-PEs and RR. The draft assumes:

A CPE must also be provisioned with whatever additional information is needed in order to set up an IPsec SA with each of the red RRs

- IPsec requires periodic refreshment of the keys. The draft does not provide any information about how to synchronize the refreshment among multiple nodes.
- IPsec usually sends configuration parameters to two endpoints only and lets these endpoints negotiate the key. Let us assume that the RR is responsible for creating the key for all endpoints: When one endpoint is compromised, all other connections will be impacted.

4.4. Preventing attacks from Internet-facing ports

When C-PEs have Internet-facing ports, additional security risks are raised.

To mitigate security risks, in addition to requiring Anti-DDoS features on C-PEs, it is necessary for CPEs to support means to determine whether traffic sent by remote peers is legitimate to prevent spoofing attacks.

5. CPEs not directly connected to VPN PEs

Because of the ephemeral property of the selected Cloud DCs, an enterprise or its network service provider may not have direct connections to the Cloud DCs that are used for hosting the enterprise's specific workloads/Apps. Under those circumstances, SD-WAN is a very flexible choice to interconnect the enterprise on-premises data centers & branch offices to its desired Cloud DCs.

However, SD-WAN paths established over the public Internet can have unpredictable performance, especially over long distances and across operators' domains. Therefore, it is highly desirable to steer as much as possible the portion of SD-WAN paths over service provider VPN (e.g., enterprise's existing VPN) that have guaranteed SLA to minimize the distance or the number of segments over the public Internet.

MEF Cloud Service Architecture [MEF-Cloud] also describes a use case of network operators that uses SD-WAN over LTE or the public Internet for last mile access where the VPN service providers cannot necessarily provide the required physical infrastructure.

Under those scenarios, one or two of the SD-WAN endpoints may not be directly attached to the PEs of a VPN Domain.

When using SD-WAN to connect the enterprise's existing sites with the workloads hosted in Cloud DCs, the corresponding CPEs have to be upgraded to support SD-WAN. If the workloads hosted in Cloud DCs need to be connected to many sites, the upgrade process can be very expensive.

[Net2Cloud-Problem] describes a hybrid network approach that integrates SD-WAN with traditional MPLS-based VPNs, to extend the existing MPLS-based VPNs to the Cloud DC Workloads over the access paths that are not under the VPN provider's control. To make it work properly, a small number of the PEs of the MPLS VPN can be designated to connect to the remote workloads via SD-WAN secure IPsec tunnels. Those designated PEs are shown as fPE (floating PE or smart PE) in Figure 3. Once the secure IPsec tunnels are established, the workloads hosted in Cloud DCs can be reached by the enterprise's VPN without upgrading all of the enterprise's existing CPEs. The only CPE that needs to support SD-WAN would be a virtualized CPE instantiated within the cloud DC.

```
   +--------+                                             +--------+
   | Host-a +--+                                     +----| Host-b |
   |        |  |                                    (')   |        |
   +--------+  |           +-----------+           (   )  +--------+
               |  +-+--+  ++-+        ++-+  +--+-+  (_)
               |  | CPE|--|PE|        |PE+--+ CPE|   |
               +--|    |  |  |        |  |  |    |---+
                  +-+--+  ++-+        ++-+  +----+
                   /       |           |
                  /        |  MPLS   +-+---+    +--+-++--------+
          +------+-+       | Network |fPE-1|    |CPE || Host   |
          | Host   |       |         |     |- --|    ||   d    |
          |   c    |       +-----+   +-+---+    +--+-++--------+
          +--------+       |fPE-2|-----+
                           +---+-+    (|)
                              (|)     (|) SD-WAN
                              (|)     (|) over any access
                              +=\======+=========+
                             //   \    | Cloud DC \\
                            //      \ ++-----+       \\
                                      +Remote|
                                      |  CPE |
                                      +-+----+
                            ----+-------+-------+-----
                                |               |
                            +---+----+      +---+----+
                            | Remote |      | Remote |
                            | App-1  |      | App-2  |
                            +--------+      +--------+
```

Figure 3: VPN Extension to Cloud DC

In Figure 3, the optimal Cloud DC to host the workloads (as a function of the proximity, capacity, pricing, or other criteria chosen by the enterprises) does not have a direct connection to the PEs of the MPLS VPN that interconnects the enterprise's existing sites.

5.1. Floating PEs to connect to Remote CPEs

To extend MPLS VPNs to remote CPEs, it is necessary to establish secure tunnels (such as IPsec tunnels) between the Floating PEs and the remote CPEs.

Even though a set of PEs can be manually selected to act as the floating PEs for a specific cloud data center, there are no standard protocols for those PEs to interact with the remote CPEs (most likely virtualized) instantiated in the third party cloud data centers (such as exchanging performance or route information).

When there is more than one fPE available for use (as there should be for resiliency purposes or the ability to support multiple cloud DCs geographically scattered), it is not straightforward to designate an egress fPE to remote CPEs based on applications. There is too much applications' traffic traversing PEs, and it is not feasible for PEs to recognize applications from the payload of packets.

5.2. NAT Traversal

Cloud DCs that only assign private IPv4 addresses to the instantiated workloads assume that traffic to/from the workload usually needs to traverse NATs.
A SD-WAN edge node can solicit a 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 so that such information can be communicated to the relevant peers.

5.3. Complexity of using BGP between PEs and remote CPEs via Internet

Even though an EBGP (external BGP) Multi-hop design can be used to connect peers that are not directly connected to each other, there are still some complications in extending BGP from MPLS VPN PEs to remote CPEs via any access path (e.g., Internet).

The path between the remote CPEs and VPN PEs that maintain VPN routes may very well traverse untrusted nodes.

EBGP Multi-hop design requires static configuration on both peers. To use EBGP between a PE and remote CPEs, the PE has to be manually configured with the "next-hop" set to the IP address of the CPEs. When remote CPEs, especially remote virtualized CPEs are dynamically instantiated or removed, the configuration of Multi-Hop EBGP on the PE has to be changed accordingly.

Egress peering engineering (EPE) is not sufficient. Running BGP on virtualized CPEs in Cloud DCs requires GRE tunnels to be established first, which requires the remote CPEs to support address and key management capabilities. RFC 7024 (Virtual Hub & Spoke) and Hierarchical VPN do not support the required properties.

Also, there is a need for a mechanism to automatically trigger configuration changes on PEs when remote CPEs' are instantiated or moved (leading to an IP address change) or deleted.

EBGP Multi-hop design does not include a security mechanism by default. The PE and remote CPEs need secure communication channels when connecting via the public Internet.

Remote CPEs, if instantiated in Cloud DCs, might have to traverse NATs to reach PEs. It is not clear how BGP can be used between devices located beyond the NAT and the devices located behind the NAT. It is not clear how to configure the Next Hop on the PEs to reach private IPv4 addresses.

5.4. Designated Forwarder to the remote edges

Among the multiple floating PEs that are reachable from a remote CPE, multicast traffic sent by the remote CPE towards the MPLS VPN can be forwarded back to the remote CPE due to the PE receiving the multicast packets forwarding the multicast/broadcast frame to other PEs that in turn send to all attached CPEs. This process may cause traffic loops.

Therefore, it is necessary to designate one floating PE as the CPE's Designated Forwarder, similar to TRILL's Appointed Forwarders [RFC6325].

MPLS VPNs do not have features like TRILL's Appointed Forwarders.

5.5. Traffic Path Management

When there are multiple floating PEs that have established IPsec tunnels with the remote CPE, the remote CPE can forward outbound traffic to the Designated Forwarder PE, which in turn forwards traffic to egress PEs and then to the final destinations. However, it is not straightforward for the egress PE to send back the return traffic to the Designated Forwarder PE.

Example of Return Path management using Figure 3 above.

- fPE-1 is DF for communication between App-1 <-> Host-a due to latency, pricing or other criteria.
- fPE-2 is DF for communication between App-1 <-> Host-b.

6. Manageability Considerations

Zero touch provisioning of SD-WAN edge nodes should be a major feature of SD-WAN deployments. It is necessary for a newly powered up SD-WAN edge node to establish a secure connection (by means of TLS, DTLS, etc.) with its controller.

7. 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, since this is the purpose of SD-WAN. See the individual sections above for further discussion of these security gaps.

8. IANA Considerations

This document requires no IANA actions. RFC Editor: Please remove this section before publication.

9. References

9.1. Normative References

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

9.2. 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.

[BGP-SDWAN-PORT]L. Dunbar, et al, "Subsequent Address Family Indicator for SDWAN Ports", draft-dunbar-idr-sdwan-port-safi-00, Work-in-progress, March 2019.

[BGP-SDWAN-Usage] L. Dunbar, et al, "Framework of Using BGP for SDWAN Overlay Networks", draft-dunbar-idr-sdwan-framework-00, work-in-progress, Feb 2019.

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

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

[SECURE-L3VPN] 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

10. Acknowledgments

Acknowledgements to xxx for his review and contributions.

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

Authors' Addresses

Linda Dunbar
Futurewei
Email: ldunbar@futurewei.com

Andrew G. Malis
Futurewei
Email: agmalis@gmail.com

Christian Jacquenet
Orange
Rennes, 35000
France
Email: Christian.jacquenet@orange.com