Network Working Group L. Dunbar
Internet Draft A. Malis
Intended status: Informational Huawei
Expires: January 2019 C. Jacquenet
Orange
August 6, 2018

Gap Analysis of Interconnecting Underlay with Cloud Overlay
draft-dm-net2cloud-gap-analysis-01

Abstract

This document analyzes the technological gaps when using SD-WAN to interconnect workloads & apps hosted in various locations, especially cloud data centers when the network service providers do not have or have limited physical infrastructure to reach the locations [Net2Cloud-problem].

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 various locations, especially in cloud data centers that the enterprise existing VPN service providers might not have the physical infrastructure to reach.

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: 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.

3. Gap Analysis of CPEs Registration Protocol

SD-WAN, conceived in ONUG (Open Network User Group) a few years ago as a means to aggregate multiple connections between any two points, has emerged as an on-demand technology to securely interconnect the OnPrem branches with the workloads instantiated in Cloud DCs that do not connect to BGP/MPLS VPN PEs or have very limited bandwidth.

Some SD-WAN networks use the NHRP protocol [RFC2332] to register SD-WAN endpoints 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 among SD-WAN endpoints.

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

- Location identifier, such as Site Identifier, System ID, and/or Port ID.
- CPE-attached GW information. When a CPE is instantiated within a Cloud DC, the Cloud DC operator's Gateway to which the CPE is attached.
- Private <-> Public address mapping, which is needed when the CPEs use private addresses.
- IPsec configuration parameters (sent by the controller to the CPEs)

4. Gap Analysis in aggregating VPN paths and Internet paths

Most likely, enterprises, especially large ones, already have their CPEs interconnected by providers' VPNs, such as EVPN, L2VPN, or L3VPN. The VPN can be PE based or CPE based which is shown in the following diagram. The commonly used CPE-based VPNs have CPE directly attached to PEs, therefore the communication is considered as secure. BGP can (also) be used to distribute routes among CPEs, even though sometimes routes among CPEs are statically configured.

```
                              +---+
                              |RR | EVPN MAC/IP BGP updates
                       +======+---+==========+
                      //                      \\
                     //  <-----EVPN-VxLAN----> \\
                  +-+--+  ++-+        ++-+  +--+-+
                  | CPE|--|PE|        |PE+--+ CPE|
               +--|  1 |  |1 |        |x |  | c  |---+
                  +-+--+  ++-+        ++-+  +----+
                           |           |
                           |  VPN    +-+---+    +----+
          +---------+      | Network | PE3 |    |CPE |
          | CPE     |      |         |     |- --| 3  |
          |   c     |      +------+  +-+---+    +----+
          +------+-+-------+ PE4 |-----+
                           +---+-+

         === or \\ indicates control plane communications
```

Figure 1: L2 or L3 VPNs over IP WAN

To use SD-WAN to aggregate Internet routes with the VPN routes, the CPEs need to have some ports connected to PEs and other ports connected to the Internet. NHRP & DSVPN/DMVPN can be used for the CPEs to be registered with their SD-WAN Controllers to establish secure tunnels among relevant CPEs.

That means the CPEs need to participate in two separate control planes: EVPN&BGP for CPE-based VPNs via links directly attached to PEs and NHRP & DSVPN/DMVPN. Two separate control planes not only add complexity to CPEs, but also increase operational cost.

```
               +---------Internet paths--------------+
               |                                     |
               |              +----+                 |
               |              |RR |                  |
               |       +======+---+===========+      |
               |     //                     \\       |
               |    //  <-----EVPN-VxLAN----> \\     |
               |  +-+---+  ++-+        ++-+  +--+-+  (|)
               |  | CPE|--|PE|        |PE+--+ CPE|  (|)
               +--|  1 |  |1 |        |x |  | c  |---+
                  +-+---+  ++-+        ++-+  +-----+
                           |           |
                           |  VPN    +-+---+    +-----+
          +---------+      | Network | PE3 |    |CPE |
          | CPE     |      |         |     |- --| 3  |
          |   c     |      +------+  +-+---+    +-----+
          +-------+-+--------+ PE4 |------+
                           +---+-+
```

Figure 2: CPEs interconnected by VPN paths and Internet Paths

4.1. Gap analysis of Using BGP to cover SD-WAN paths

Since BGP is widely deployed, it is desirable to consider using BGP to control the SD-WAN paths instead of NHRP, DSVPN/DMVPN. This section analyzes the gaps of using BGP to control SD-WAN.

RFC5512 and [Tunnel-Encap] describe methods for endpoints to advertise tunnel information and to trigger tunnel establishment. RFC5512 & [Tunnel-Encaps] have the Endpoint Address to indicate IPv4 or IPv6 address format, the Tunnel Encapsulation attribute to indicate different encapsulation formats, such as L2TPv3, GRE, VxLAN, IP-in-IP, etc. There are sub-TLVs to describe the detailed tunnel information for each of the encapsulations.

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

To express the support of multiple Encap types, multiple Extended BGP communities with SAFI value = 7 can be used.

Here are some of the gaps using RFC5512 and [Tunnel-Encap] to

control SD-WAN:

- Doesn't have fields to carry detailed information of the remote CPE: such as Site-ID, System-ID, Port-ID

- Does not have the proper field to express IPsec configuration information sent by the "Controller" [CJ1]to the CPEs.
- Does not have proper way for two peer CPEs to negotiate IPSec keys, based on the configuration sent by the Controller.
- UDP NAT private address <-> public address mapping
- CPEs tend to communicate with a few other CPEs, not all the CPEs need to form mesh connections. Using BGP, CPEs can easily get dumped with too much information of other CPEs that they never need to communicate with. NHRP only sends the relevant information for the endpoints interested in establishing tunnels. Therefore, there is a need for some form of "Registration" methods.

[VPN-over-Internet] describes a way to securely interconnect CPEs via IPsec using BGP. This method is useful, however, it still misses some aspects to aggregate CPE-based VPN routes with internet routes that interconnect the CPEs. In addition:

- The draft assumes that CPE "registers" with the RR. However, it doe say how. Should "NHRP" (modified version) be considered? In SD-WAN, Zero Touch Provisioning is expected. It is not acceptable to require manual configuration on RR which CPEs are controlled.
- For RR communication with CPE, this draft only mentioned IPSec. Mis TLS/DTLS.
- The draft assumes that CPEs and RR are connected with an IPsec tunn With zero touch provisioning, we need an automatic way to synchronize the IPsec SA between CPE 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. How to synchronize refreshment among multiple nodes?
- IPsec usually only sends configuration parameters to two endpoints let the two endpoints negotiate the KEY. Now we assume that RR is responsible for creating the KEY for all endpoints. When one endpoint is confiscated, all other connections are impacted.

4.2. Gaps in preventing attacks to CPEs from their Internet ports

When CPEs have ports facing Internet, it brings in the security risks of potential DDoS attacks to the CPEs from the ports facing internet.

To mitigate security risks, it is absolutely necessary to enable Anti-DDoS features on those CPEs to prevent major DDoS attacks.

5. Gap analysis of 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 the direct links to the Cloud DCs that are optimal 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 over public Internet can have unpredictable performance, especially over long distances and across domains. Therefore, it is highly desirable to place 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/segments over public Internet.

MEF Cloud Service Architecture [MEF-Cloud] also describes a use case of network operators needing to use SD-WAN over LTE or public Internet for last mile accesses where they are not present.

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

Using SD-WAN to connect the enterprise existing sites with the workloads in Cloud DC, the enterprise existing sites' CPEs have to be upgraded to support SD-WAN. If the workloads in Cloud DC 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 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 in Cloud DC 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 (due to proximity, capacity, pricing, or other criteria chosen by the enterprises) does not happen to have a direct connection to the PEs

of the MPLS VPN that interconnects the enterprise's existing sites.

5.1. Gap Analysis of 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.

Gap:

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 or the ability to support multiple cloud DCs scattered geographically), 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

Most cloud DCs only assign private IPv4 addresses to the workloads instantiated. Therefore, traffic to/from the workload usually need to traverse NAT.

5.3. Complication 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/gaps in extending BGP from MPLS VPN PEs to remote CPEs via any access paths (e.g., Internet):

EBGP Multi-hop scheme 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 addresses of the CPEs. When remote CPEs, especially remote virtualized CPEs are dynamically

instantiated or removed, the configuration on the PE Multi-Hop EBGP has to be changed accordingly.

Gap:

Egress peering engineering (EPE) is not enough. Running BGP on virtualized CPEs in Cloud DC requires GRE tunnels being established first, which in turn requires address and key

management for the remote CPEs. RFC 7024 (Virtual Hub & Spoke) and Hierarchical VPN is not enough.

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

EBGP Multi-hop scheme does not have an embedded security mechanism. The PE and remote CPEs need secure communication channels when connecting via the public Internet.

Remote CPEs, if instantiated in Cloud DC, might have to traverse NAT to reach PE. It is not clear how BGP can be used between devices outside the NAT and the entities 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 multiple floating PEs available for a remote CPE, multicast traffic from the remote CPE towards the MPLS VPN can be forwarded back to the remote CPE due to the PE receiving the multicast data frame forwarding the multicast/broadcast frame to other PEs that in turn send to all attached CPEs. This process may cause a traffic loop.

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

Gap: the MPLS VPN does not have features like TRILL's Appointed Forwarders.

5.5. Traffic Path Management

When there are multiple floating PEs that have established IPsec tunnels to the remote CPE, the remote CPE can forward the outbound traffic to the Designated Forwarder PE, which in turn forwards the traffic to egress PEs to the 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

TBD

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, as is the general intention 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.

[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

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
Huawei
Email: Linda.Dunbar@huawei.com

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

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