Network Working Group   L. Dunbar
Internet Draft   A. Malis
Intended status: Informational   Huawei
Expires: September 25, 2019   C.
Jacquenet

                                                           Orange
                                                  March 25, 2019

            Gap Analysis of Interconnecting Underlay with Cloud Overlay
                 draft-ietf-rtgwg-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 and in Cloud
DCs that the enterprise's VPN service provider may not own/operate
or may be unable to provide the required connectivity to access
these locations. When enterprise' 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,
the Public Internet, to achieve better performance and visibility.
This document primarily focuses on the technological gaps of SD-WAN.

For ease of description, a SD-WAN edge, a SD-WAN end-point, C-PE, or
CPE are used interchangeably throughout this document.

[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, "SDWAN" refers to
             the solutions of pooling WAN bandwidth from multiple
             underlay networks to get better WAN bandwidth
             management, visibility & control. When the underlay is
             private network, 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).

[3](#). Gap Analysis of C-PEs WAN Ports Registration

The SD-WAN WG stemmed out from ONUG (Open Network User Group) in
2014 and was the placeholder to define SD-WAN as a means to
aggregate multiple underlay networks between any two points. SD-WAN
technology 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 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 controller, such as:

- Interworking with MPLS VPN control plane. A SD-WAN edge can have
  some ports facing MPLS VPN network over which packets can be sent
  natively without 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, the protocol
  does not work well.
- NHRP does not have the IPsec attributes, which are needed for
  peers to build Security Associations over 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, to establish connection to the C-PE, it is necessary to know
   the Cloud DC operator's Gateway to which the CPE 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 to use BGP for SD-WAN edge nodes to
   register their WAN ports properties to the SD-WAN controller, which
   then disseminates the information to other SD-WAN edge nodes that
   are authenticated before the SD-WAN controller and the other SD-WAN
   edge nodes can communicate with them.

4. Gap Analysis in 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 & PEs is
   considered as secure. MP-BGP is used to learn & distribute routes
   among CPEs, even though sometimes routes among CPEs are statically
   configured.

   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.

   If using NHRP for registration purposes, C-PEs need to participate
   in 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.
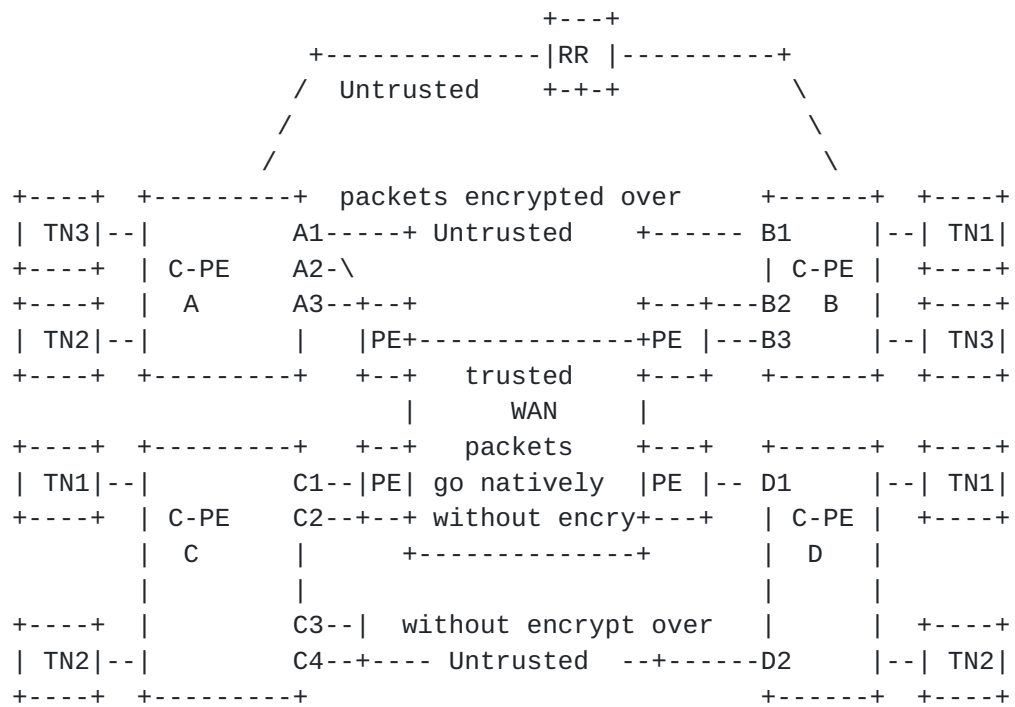
```
                            +---+
               +--------------|RR |----------+
              /  Untrusted    +-+-+           \
             /                                 \
            /                                   \
   +----+  +---------+  packets encrypted over    +------+  +----+
   | 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. Gap analysis of Using BGP for SD-WAN

   This section analyzes the gaps of using BGP to control SD-WAN.

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

     - SD-WAN node's WAN Ports Property registration to the SD-WAN
        Controller.
          o It is to inform the SD-WAN controller and potential peers
            of the WAN ports property of the C-PE [SDWAN-Port]. When
            the WAN ports are using private addresses, this step can
            register the type of NAT that translate 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 SDWAN 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.

     RFC5512 and [Tunnel-Encap] describe methods for endpoints to
     advertise tunnel information and trigger tunnel establishment.
     RFC5512 & [Tunnel-Encap] use the Endpoint Address that indicates an
     IPv4 or an IPv6 address, and 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 encapsulation types.

     [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:

     - Lacking C-PE WAN Port Property Registration functionality
     - Lacking IPsec Tunnel type
     - [Tunnel-Encap] has Remote Address SubTLV, but does not have any
       field to indicate the Tunnel originating interface, as defined in
       RFC5512.
     - The mechanisms described by [Tunnel-Encap] cannot be effectively
       used for SD-WAN overlay network because a SD-WAN Tunnel can be
       established between Internet-facing WAN ports of two C-Pes. This
       tunnel needs to be established before data arrival because the
       tunnel establishment can fail, e.g., in case the two end-points
       support different encryption algorithms.
     - Client traffic can either be forwarded through the MPLS network
       natively without any encryption for better performance, or through
       the Internet-facing ports with IPsec encryption.
     - There can be many client routes associated with the SD-WAN IPsec
       tunnel between two C-PE's Internet-facing WAN ports, but the

corresponding destination prefixes (as announced by the
aforementioned routes) can also be reached over the VPN underlay
natively without encryption. A more realistic approach to separate
IPsec SA management from client routes association with IPsec.
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 end-points 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 needs lots of configuration to get the nodes provisioned.
- Does not have any field to carry detailed information about the
  remote C-PE, such as Site-ID, System-ID, Port-ID
- Does not have any field to express IPsec attributes for the SD-WAN
  edge nodes to establish IPsec Security Associations with others.
- Does not have any proper way for two peer CPEs to negotiate IPsec
  keys, based on the configuration sent by the Controller.
- 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 form mesh connections. Without any BGP
  extension, many nodes can get dumped with too much information
  coming from other nodes that they never need to communicate with.

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

[SECURE-L3VPN] assumes PEs terminate MPLS packets, and use 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 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,
  given the dimensioning figures but also the purpose of SD-WAN to
  automate configuration tasks.
- For RR communication with CPEs, this draft only mentions IPsec.
  Missing TLS/DTLS.
- The draft assumes that CPEs and RR are connected with an IPsec
  tunnel. With zero touch provisioning, we need an automatic way to
  synchronize the IPsec SAs 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. 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.

4.2. Gaps in 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]. 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 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 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 that use SD-WAN over LTE or the public
   Internet for last mile access where the VPN 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 in Cloud DC, the corresponding CPEs have to be
   upgraded to support SD-WAN. If the workloads 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 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  |
                      +--------+      +--------+
```
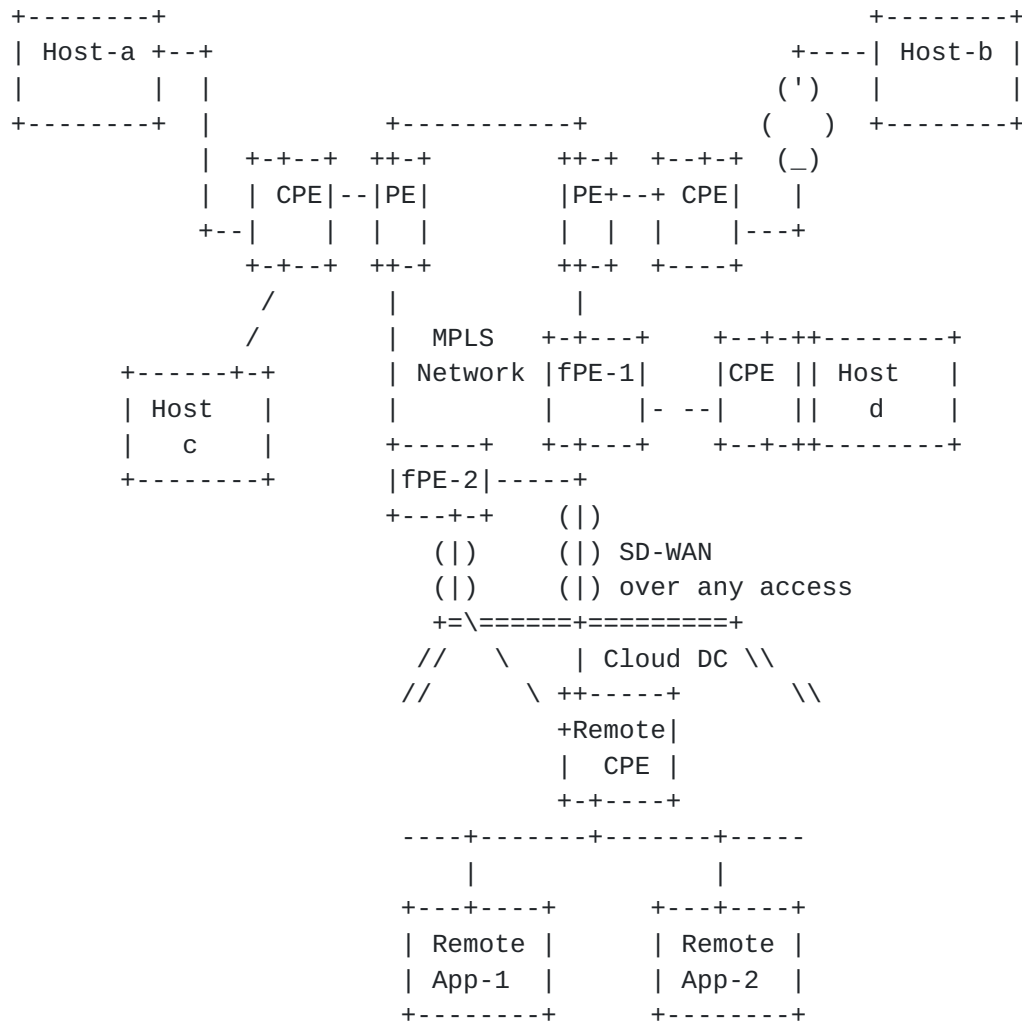
                  Figure 3: VPN Extension to Cloud DC

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 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 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

Most cloud DCs only assign private addresses to the instantiated workloads. Therefore, 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 to pass to peers.

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).

The path between the remote CPEs and VPN PE can 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.

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 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 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 the multiple floating PEs that are available for 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 data frame 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].

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 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 is expected in SD-WAN
deployment. 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-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
    Huawei
    Email: Linda.Dunbar@huawei.com

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

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