Network Working Group L. Dunbar
Internet Draft Futurewei
Intended status: Informational A. Malis
Expires: March 22, 2024 Malis Consulting
C. Jacquenet
Orange
M. Toy
Verizon
K. Majumdar
Microsoft
September 22, 2023

Dynamic Networks to Hybrid Cloud Data Centers~~DCs~~: A Problem Statement and Mitigation Practices

draft-ietf-rtgwg-net2cloud-problem-statement-30

Abstract

This document describes ~~the~~ a set of network-related problems that enterprises face at the moment of writing this specification (2023) when interconnecting their branch offices with dynamic workloads in third-party data centers (DC) ~~(a.k.a. Cloud DCs)~~. These ~~Net2Cloud~~ problems ~~statements~~ are mainly for enterprises with ~~traditional~~ conventional VPN services ~~who~~ that want to leverage those networks (instead of altogether abandoning them). Other problems are out of the scope of this document.

This document also describes ~~the~~ various mitigation ~~practices~~ actions to ~~alleviate the issues caused by the identified~~soften issues induced by these problems.

Status of this Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress." The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html

This Internet-Draft will expire on March 22, 2024.

Copyright Notice

Commenté [BMI1]: To be changed to RTGWG

Commenté [BMI2]: I don't think that you claim to be exhaustive.

Commenté [BMI3]: This is not introduced at this stage?

Commenté [BMI4]: To make inclusive-checks happy 😊

Commenté [BMI5]: Not sure this is useful, especially with the proposed change s/the/a set of.

Copyright (c) 2023 IETF Trust and the persons identified as the
document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal
Provisions Relating to IETF Documents
(http://trustee.ietf.org/license-info) in effect on the date of
publication of this document. Please review these documents
carefully, as they describe your rights and restrictions with
respect to this document. Code Components extracted from this
document must include Simplified BSD License text as described in
Section 4.e of the Trust Legal Provisions and are provided without
warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction

With the advent of widely available Cloud data centers (DCs) providing services in ~~diverse~~ various geographic locations and advanced tools for monitoring and predicting application behaviors, it is desirable for enterprises to instantiate applications and workloads in Cloud DCs. Some enterprises prefer that ~~their~~ for specific applications be located close to the~~ir~~ ~~end~~ users accessing these services, as the proximity ~~can improve~~is a means to act on the ~~end~~-to-end latency. ~~and overall user experience.~~ In addition, applications and workloads in Cloud DCs can be shut down or moved along with end users in motion (thereby modifying the networking connection of subsequently relocated applications and workloads).

Commenté [BMI6]: This is the call of this enterprises. I don't see the causality effect here, especially that the motivation for the use of external resources (in general) is not the set of features listed here.

If you maintain this sentence, please change "desirable" to "tempting" or something else.

Commenté [BMI7]: I would simply delete this sentence or replace it with a statement that basically state that cloud resources are used to host enterprise app + motivation for such sue is local to each enterprise.

Commenté [BMI8]: No, proximity does not ensure this.

~~Key characteristics of~~ Cloud ~~Services~~ services are generally exposed on-demand, claim to be scalable, highly available, and usage-based billing. Most Cloud Operators provide Cloud network functions, such as, virtual Firewall services, virtual private network services, virtual ~~PBX (~~Private Branch eXchange (PBX)) services including voice and video conferencing systems, ~~etc~~. A Cloud DC is a shared infrastructure that hosts services to many customers.

This document describes the network-related problems enterprises face at the moment of writing this specification when interconnecting their branch offices with dynamic workloads in Cloud DCs and the mitigation practices.

2. Definition of T~~t~~erms

Cloud DCs: Third party Data Centers that usually host applications and workloads owned by different organizations or tenants.

Heterogeneous Cloud: applications and workloads split among Cloud DCs owned or managed by different operators.

Hybrid Clouds: A hybrid C~~c~~loud is a mixed computing environment where applications are run using a combination of computing, storage, and services in different environments-public clouds and private clouds, including on-premises data centers or "edge" locations. [https://cloud.google.com/learn/what-is-hybrid-cloud].

IXPs: Internet eX~~x~~change ~~points~~ Points (IXes or IXPs) ~~are common grounds of IP networking, allowing~~ are an interconnect facility used by ~~participant~~ Internet service providers (ISPs) to exchange data destined for their respective networks. [https://en.wikipedia.org/wiki/Internet_exchange_point].

SD-WAN An overlay connectivity service that ~~optimizes~~ T~~transport~~makes forwarding decisions of IP ~~Packets~~ packets over one or more ~~Underlay~~ underlay ~~Connectivity~~connectivity ~~Services~~ networks by recognizing applications (Application Flows) and ~~determining forwarding behavior by~~ applying relevant ~~Policies~~policies to them~~.~~ [MEF-70.1].

VPC: Virtual Private Cloud is a virtual network dedicated to one client account. It is logically isolated from other virtual networks in a Cloud DC. ~~Each c~~Clients can launch ~~his/her~~their desired resources, such as compute, storage, or network functions into ~~his/her~~their VPC. At the moment of ~~of~~ writing this specification (2023), most Cloud operators' VPCs only support private addresses, some support IPv4 only,

Commenté [BMI9]: What does that mean?

Commenté [BMI10]: Please follow a consistent style when expanding. In the previous para you went with "Data Center (DC)" and here "PBX (xxx)".

Commenté [BMI11]: This is not an example.

Commenté [BMI12]: You may move this sentence to be positioned right before the one list the functions.

a mis en forme : Surlignage

Commenté [BMI13]: ?

Commenté [BMI14]: Move to the ref section, if really needed.

a mis en forme : Français (France)

a mis en forme : Anglais (États-Unis)

a mis en forme : Anglais (États-Unis)

Commenté [BMI15]: Move to the ref section, if really needed.

Commenté [BMI16]: Optimization may or may not be there as this depends on how efficient are forwarding decisions.

Commenté [BMI17]: How this is defined?

Commenté [BMI18]: To be defined.

Commenté [BMI19]: What does that mean for IPv6?

others support IPv4/IPv6 dual stack.

3. Issues and Mitigation Methods of Connecting to Cloud DCs

This section identifies some of the high-level problems that IETF can address, especially ~~by~~ with the Routing area. Other Cloud DC problems (e.g., managing cloud spending) are out of the scope of this document~~, e.g., managing cloud spending is not discussed here~~.

3.1. Increased BGP Peering Errors and Mitigation Methods

Where ~~traditional~~ conventional ISPs view peering as a means to improve network operations, Public Cloud DCs offer direct ~~peering~~ interconnections to get more customers to use their ~~data centers~~DCs and services. As such, there is pressure to peer more widely and to peer with customers, including those who lack the expertise and experience in running complex BGP ~~peering~~ interconnection relationships. All those can contribute to increased BGP peering errors such as capability mismatch, unwanted route leaks, missing Keepalives, and errors causing BGP ceases. Capability mismatch can cause BGP sessions not to be adequately established. Those issues are more acute to Cloud DCs than they have ~~traditionally~~ been, even though they may apply to ~~traditional~~ conventional ISPs, just to a lesser degree.

Here are the recommended mitigation practices:

- If a Cloud GW (BGP speaker) receives from its BGP peer a capability that it does not itself support or recognize, it must ignore that capability, and the BGP session must not be terminated per Section 5 of [RFC5492]. When receiving a BGP UPDATE with a malformed attribute, the revised BGP error handling procedure [RFC7606] should be followed instead of session resetting.
- When a Cloud DC doesn't support multi-hop eBGP peering with external devices (as many don't), enterprise GWs must establish tunnels (e.g., IPsec) to the Cloud GWs to form the IP adjacency.
- When a Cloud DC eBGP session supports a limited number of routes from external entities, the on-premises DCs need to set up default routes and filter as many routes as practical replacing them with that default in the eBGP advertisement to minimize the number of routes to be exchanged with the Cloud DC eBGP peers.
- When a Cloud GW receives the inbound routes exceeding the maximum routes threshold for a peer, the currently common practice is generating out-of-band alerts (e.g., Syslog) via the management system or terminating the BGP session (with cease notification messages [RFC4486] being sent). More discussion is needed in the IETF IDR WG for potential in-band or autonomous notification directly to the peers when the inbound routes exceed the maximum routes threshold.

3.2. Site Ffailures and Methods to Minimize Impacts

Commenté [BMI20]: It is not how clear how this mention is useful in the rest of the document.

Commenté [BMI21]: Do you mean "interconnection" or the specific "peering" free form of it?

a mis en forme : Surlignage

Commenté [BMI22]: FWIW, please refer to https://datatracker.ietf.org/doc/html/draft-ietf-opsawg-teas-attachment-circuit-03#name-connecting-a-virtualized-en about an example to automate such interconnection

Commenté [BMI23]: You may indicate how the reco are defined. Are those used in practice, etc.

Commenté [BMI24]: To be defined, first

Commenté [BMI25]: Do you mean section 3?

Commenté [BMI26]: FWIW, some cloud DCs uses MD5 for BGP sessions.

Commenté [BMI27]: To be defined.

Commenté [BMI28]: How this is different from the current practices? Should you recommend a secure form of tunnels?

Commenté [BMI29]: Do you assume that there is always one?

Commenté [BMI30]: We need first a means to share/set that max.

Commenté [BMI31]: FYI, the notification can be controlled using means such as: https://datatracker.ietf.org/doc/html/draft-ietf-opsawg-ntw-attachment-circuit-03#name-bgp

Failures within a Cloud site, which can be a building, a floor, a ~~POD~~pod, or a server rack, include capacity degradation or complete out-
of-service. ~~Here~~ Examples of~~are some~~ events that can trigger a site failure: a)
fiber cut for links connecting to the site or among pods within the site; b) cooling failures; c) insufficient backup power; d) cyber threat attacks; e) too many changes outside of the maintenance window; ~~etc~~. Fiber-cut is not uncommon in a Cloud site or between sites.

As described in RFC7938, a Cloud DC might not have a~~n IGP~~ dynamic routing protocol to route
around link/node failures within its domain. When a site failure happens, the Cloud DC GW visible to clients is running fine; therefore, the site failure is not detectable by the clients using Bidirectional Forwarding Detection (BFD).

When a site failure occurs, many instances can be impacted. When the impacted instances' IP prefixes in a Cloud DC are not adequately aggregated
~~nicely~~, which is very common, one single site failure can trigger a huge number of BGP UPDATE messages. There are proposals, such as [METADATA-PATH], to enhance BGP advertisements to address this problem.

[RFC7432] specifies a mass withdrawal mechanism for EVPN to signal a large number of routes being changed to remote PE nodes as quickly as possible.

3.3. Limitations of DNS-based Cloud DC Location Selection

Many applications have multiple instances running ~~instantiated~~ in different
Cloud ~~DCs~~sites. A commonly deployed solution has DNS server(s) responding
to a~~n~~ ~~FQDN (~~Fully Qualified Domain Name (FQDN) inquiry with an IP address
of the closest or lowest cost DC that can reach the instance. Here are some problems associated with DNS-based solutions:

- Dependent on client behavior
  - Misbehaving client can cache results indefinitely.
  - Clients may not ~~receive~~ access a service even though there are
    Server instances available in other Cloud DCs because the failing
    IP address is still cached in the DNS resolver and has not expired yet.
- No inherent leverage of proximity information present in the network (routing) layer, resulting in loss of performance.
- Inflexible traffic control:

  The Local DNS resolver becomes the unit of traffic management. This requires DNS to receive periodical update of the network condition, which is difficult.

One method to mitigate the problems listed above is to use ~~the~~

Commenté [BMI32]: To align with the term used in NVO-related RFCs.

Commenté [BMI33]: Cite as a reference

Commenté [BMI34]: Instances of what?

Commenté [BMI35]: Do we have any public reference to cite?

Commenté [BMI36]: TTL is expected to handle this.

Commenté [BMI37]: Which one? Stub or on-path forwarders?

~~ANYCAST~~anycast [RFC4786] for the services so that network proximity and conditions can be inherently considered in optimal path selection.

[SERVICE-METRICS] identifies the metrics that can be utilized for the ingress routers to make path steering selections, not only based on the routing cost but also the running environment of the edge services.

3.4. Network Issues for 5G Edge Clouds and Mitigation Methods

The 5G Edge Cloud~~s~~ DCs [3GPP-5G-Edge] may host edge computing applications for ultra-low latency services on virtual or physical servers. Those edge computing applications have low latency connections to the UEs (User Equipment) and might have other connections to backend servers or databases in other locations.

The low latency traffic to/from the UEs is transported through the 5G Core (gNB (Next Generation Node B))<-> UPFs (User Plane Function)) and the 5G Local Data Networks (LDN) to the edge Cloud DCs. The LDN's ingress routers connected to the UPFs might be co-located with 5G Core functions in the edge Clouds. The 5G Core functions include Radio Control Functions, Session Management Functions (SMF), Access Mobility Functions (AMF), User Plane Functions (UPF), and others.

Commenté [BMI38]: Do you really need these details?

Here are some network problems with connecting the services in the 5G Edge Clouds:

1) The difference in routing distances to server instances in different edge Clouds is relatively small. Therefore, the instance in the Edge Cloud with the shortest routing distance from ~~the~~ a 5G UPF might not be the best in providing the overall low latency service.
2) Capacity status at the Edge Cloud might play a more significant role in end-to-end performance.
3) Source (UEs) can ingress from different LDN Ingress routers due to mobility.

Commenté [BMI39]: Not sure how this is specific to 5G.

[METADATA-PATH] describes a mechanism to get around those problem. [METADATA-PATH] extends the BGP UPDATE messages for a Cloud GW to propagate the edge service-related metrics from Cloud GW to the ingress routers so that the ingress routers can incorporate the destination site's capabilities with the routing distance in computing the optimal paths.

The IETF CATS working group is examining general aspects of this space, and may come up with protocol recommendations for this information exchange.

3.5. DNS Practices for Hybrid Workloads

DNS name resolution is essential for on-premises and cloud-based resources. For customers with hybrid workloads, which include on-premises and cloud-based resources, extra steps are necessary to configure DNS to work seamlessly across both environments.

Cloud operators have their own DNS to resolve resources within their Cloud DCs and to well-known public domains. Cloud's DNS can be configured to forward queries to customer managed authoritative DNS servers hosted on-premises and to respond to DNS queries forwarded by on-premises DNS servers.

For enterprises utilizing Cloud services by different Cloud operators, it is necessary to establish policies and rules on how/where to forward DNS queries. When applications in one Cloud need to communicate with applications hosted in another Cloud, DNS queries from one Cloud DC could be forwarded to the enterprises' on-premises DNS, which in turn be forwarded to the DNS service in another Cloud. Configuration can be complex depending on the application communication patterns.

However, collisions can still occur even with carefully managed policies and configurations. If an organization uses an internal name like ".internal" and wants its services to be available via or within some other Cloud provider that also uses ".internal", collisions might occur. Therefore, using the global domain name is better even when an organization does not make all its namespace globally resolvable. An organization's globally unique DNS can include subdomains that cannot be resolved outside certain restricted paths, zones that resolve differently based on the origin of the query, and zones that resolve the same globally for all queries from any source [Split-Horizon-DNS].

Globally unique names do not equate to globally resolvable names or even global names that resolve the same way from every perspective.

Globally unique names can prevent any possibility of collisions at present or in the future, and they make DNSSEC trust manageable. Consider using a registered and ~~fully qualified domain name (~~FQDN~~)~~ from global DNS as the root for enterprise and other internal namespaces.

Commenté [BMI40]: Already expanded

3.6. NAT Practice for Accessing Cloud Services

Cloud resources, such as VMs (Virtual Machine) or application instances, are usually assigned private IP addresses. By configuration, some private subnets can have the NAT function to reach out to external networks, and some private subnets are internal to Cloud only.

Commenté [BMI41]: IPv4? Or ULA like IPv6 @?

Commenté [BMI42]: Many DCs are IPv6-only and are using techniques such as https://www.rfc-editor.org/rfc/rfc8512.html#appendix-A.5

Different Cloud operators support different levels of NAT functions. For example, AWS NAT Gateway does not currently support connections towards, or from VPC Endpoints, VPN, AWS Direct Connect, or VPC Peering [AWS-NAT]. AWS Direct Connect/VPN/VPC Peering does not currently support any NAT functionality.

Google's Cloud NAT [Google-NAT] allows Google Cloud VM instances without external IP addresses and private Google Kubernetes Engine (GKE) clusters to connect to the Internet. Cloud NAT implements outbound NAT in conjunction with a default route to allow instances to reach the Internet. It does not implement inbound NAT. Hosts outside the VPC network can only respond to established connections initiated by instances inside the Google Cloud; they cannot initiate new connections to Cloud instances via NAT.

For enterprises with applications running in different Cloud DCs, proper configuration of NAT must be performed in Cloud DCs and their on-premises DC.

Commenté [BMI43]: More concretely?

3.7. Cloud Discovery Practices

One of the concerns of enterprises using Cloud services is the lack of awareness of the locations of their services hosted in the Cloud, as Cloud operators can move the service instances from one place to another. While the geographic locations are usually exposed to the enterprises, such as Availability Zones or Regions, the topological location is usually hidden. When applications in Cloud communicate with on-premises applications, it may not be clear where the Cloud applications are located or to which VPCs they belong.

Being able to detect Cloud services' location can help on-premises gateways (routers) to connect the services in a more optimal site when the enterprise's end users or policies change.

For enterprises that instantiate virtual routers in Cloud DCs, metadata can be attached (e.g., GENEVE header or IPv6 optional header) to indicate additional properties, including useful information about the sites where they are instantiated.

Commenté [BMI44]: May add a pointer

4. Dynamic Connecting Enterprise Sites with Cloud DCs

For many enterprises with established private VPNs (e.g., private circuits, MPLS-based L2VPN/L3VPN) interconnecting branch offices & on-premises data centers, connecting to Cloud services will be a mix of different types of networks. When an enterprise's existing VPN service providers do not have direct connections to the desired cloud DCs that the enterprise prefers to use, the enterprise faces additional infrastructure and operational costs to utilize the Cloud services.

Commenté [BMI45]: Add ref to L2/L3 VPN Framework RFCs

This section describes some mechanisms for enterprises with private VPNs to connect to Cloud services dynamically.

4.1. Sites to Cloud DC

Commenté [BMI46]: I'm not sure calling out products is OK in an RFC (AWC, Microsoft, etc.). May be consider extracting common considerations and move product-specific matter from the I-D or to an appendix.

Most Cloud operators offer some type of network gateway through which an enterprise can reach their workloads hosted in the Cloud DCs. For example, AWS ~~(Amazon Web Services)~~ offers the following options to reach workloads in AWS Cloud DCs [AWS-Cloud-WAN]:

- AWS Internet gateway allows communication between instances in AWS VPC and the Internet.
- AWS Virtual gateway (vGW) where IPsec tunnels [RFC6071] are established between an enterprise's own gateway and AWS vGW, so that the communications between those gateways can be secured from the underlay (which might be the public Internet).
- AWS Direct Connect, which allows enterprises to purchase direct connect from network service providers to get a private leased line interconnecting the enterprises gateway(s) and the AWS Direct Connect routers. In addition, an AWS Transit Gateway can be used to interconnect multiple VPCs in different Availability Zones. AWS Transit Gateway acts as a hub that controls how

traffic is forwarded among all the connected networks which act like spokes.

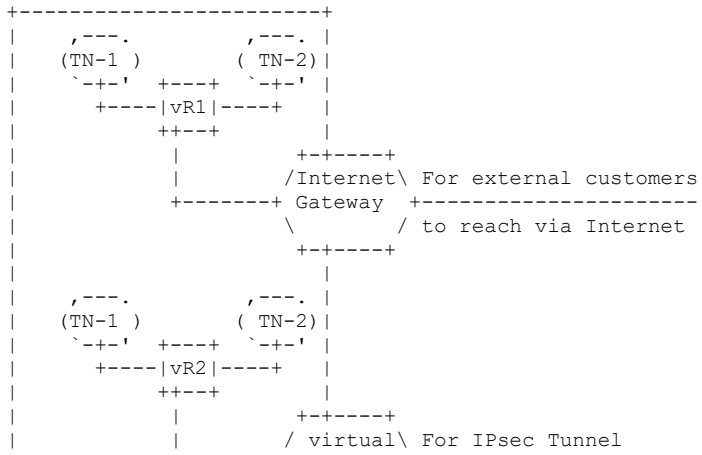
Microsoft Azure's Virtual WAN [Azure-SD-WAN] allows extension of a private network to any of the Microsoft Cloud services, including Azure and Office365. ExpressRoute is configured using Layer 3 routing. Customers can opt for redundancy by provisioning dual links from their location to two Microsoft Enterprise edge routers (MSEEs) located within a third-party ExpressRoute peering location. The BGP routing protocol is then setup over WAN links to provide redundancy to the cloud. This redundancy is maintained from the peering data center into Microsoft's cloud network.

Google's Cloud Dedicated Interconnect offers similar network connectivity options as AWS and Microsoft. One distinct difference, however, is that Google's service allows customers access to the entire global Cloud network by default. It does this by connecting the on-premises network with the Google Cloud using BGP and Google Cloud Routers to provide optimal paths to the different regions of the global cloud infrastructure.

Figure 1 below shows an example of a portion of workloads belonging to one tenant (e.g., TN-1) are accessible via a virtual router connected by AWS Internet Gateway; some of the same tenant (TN-1) are accessible via AWS vGW, and others are accessible via AWS Direct Connect. The workloads belonging to one tenant can communicate within a Cloud DC via virtual routers (e.g., vR1 and~~,~~ vR2).

Different types of access require different level of security functions. Sometimes it is not visible to end customers which type of network access is used for a specific application instance. To get better visibility, separate virtual routers (e.g., vR1 ~~&~~ and vR2) can be deployed to differentiate traffic to/from different Cloud GWs. It is important for some enterprises to be able to observe the specific behaviors when connected by different connections.

Customer Gateway can be customer owned router or ports physically connected to AWS Direct Connect GW.



Commenté [BMI47]: Please refer to the Attachement Circuit Spec to see how bearers/ACs can be used to provide this feature.

Commenté [BMI48]: We can map this to the concept of "bearer" and "attachment circuit" in OPSAWG.

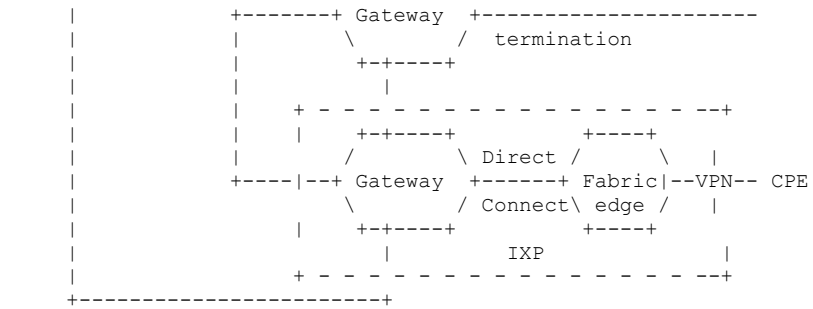


Figure 1: Examples of Multiple Cloud DC connections.

4.2. Inter-Cloud Connection

The connectivity options to Cloud DCs described in ~~the previous section~~Section 4.1 are for reaching Cloud providers' DCs, but not between cloud DCs. When applications in AWS Cloud need to communicate with applications in Azure, today's practice requires a third-party gateway (physical or virtual) to interconnect the AWS's Layer 2 DirectConnect path with Azure's Layer 3 ExpressRoute.

Enterprises can also instantiate their virtual routers in different Cloud DCs and administer IPsec tunnels among them. In summary, here are some approaches, available to interconnect workloads among different Cloud DCs:

a) Utilize Cloud DC provided inter/intra-cloud connectivity services (e.g., AWS Transit Gateway) to connect workloads instantiated in multiple VPCs. Such services are provided with the Cloud gateway to connect to external networks (e.g., AWS DirectConnect Gateway).
b) Hairpin all traffic through the customer gateway, meaning all workloads are directly connected to the customer gateway, so that communications among workloads within one Cloud DC must traverse through the customer gateway.
c) Establish direct tunnels among different VPCs (AWS' Virtual Private Clouds) and VNET (Azure's Virtual Networks) via client's own virtual routers instantiated within Cloud DCs. NHRP (Next Hop Resolution Protocol) [RFC2735] based multi-point techniques can be used to establish direct Multi-point-to-Point or multi-point-to multi-point tunnels among those client's own virtual routers.
d) Utilize a Cloud Aggregator or Cloud Services Broker (CSB) who acts as an intermediator among cloud service providers and network service providers to offer a combined total package for enterprises. The Cloud Aggregator can provide the network connections among one enterprise's services instantiated in multiple Clouds.

Approach a) usually does not work if Cloud DCs are owned and managed by different Cloud providers.

Approach b) creates additional transmission delay plus incurring

cost when exiting Cloud DCs.

For Approach c), [SDWAN-EDGE-DISCOVERY] describes a mechanism for virtual routers to advertise their properties for establishing proper IPsec tunnels among them. There could be other approaches developed to address the problem.

Approach d) is a method of third-party multi-cloud management business model.

4.3. Extending Private VPNs to Hybrid Cloud DCs

Traditional private VPNs, including private circuits or MPLS-based L2/L3 VPNs, when purchased with premium paid services, have been widely deployed as an effective way to support businesses and organizations that require network performance and reliability.

Connecting an enterprise's on-prem CPEs to a Cloud DC via a private VPN requires the private VPN provider to have a direct path to the Cloud GW. When the user base changes, the enterprise might want to migrate its workloads/applications to a new cloud DC location closer to the new user base. The existing private VPN provider might not have circuits at the new location. Deploying PEs routers at new locations takes a long time (weeks, if not months).

When the private VPN network can't reach the desired Cloud DCs, IPsec tunnels can dynamically connect the private VPN's PEs with the desired Cloud DCs GWs. As the private VPNs provide higher quality of services, choosing a PE closest to the Cloud GW for the IPsec tunnel is desirable to minimize the IPsec tunnel distance over the public Internet.

In order to support Explicit Congestion Notification (ECN) [RFC3168] usage by private VPN traffic, the PEs that establish the IPsec tunnels with the Cloud GW need to comply with the ECN behavior specified by ~~RFC6040~~ [RFC6040].

An enterprise can connect to multiple Cloud DC locations and establish different BGP peers with Cloud GW routers at different locations. As multiple Cloud DCs are interconnected by the Cloud provider's own internal network, its topology and routing policies are not transparent or even visible to the enterprise customer's on-prem routers. One Cloud GW BGP session might advertise all of the prefixes of the enterprise's VPC, regardless of which Cloud DC a given prefix resides, which can cause improper optimal path selection for on-prem routers. To get around this problem, virtual routers in Cloud DCs can be used to attach metadata (e.g., in the GENEVE header or IPv6 optional header) to indicate the Geo-location of the Cloud DC, the delay measurement, or other relevant data.

5. Methods to Scale IPsec T~~t~~unnels to Cloud DCs

As described in Section 4.3, IPsec tunnels can be used to dynamically establish connection between private VPN PEs with Cloud GW~~s~~. Enterprises can also instantiate virtual routers within Cloud DCs to connect to their on-premises devices via IPsec tunnels.

As described in [Int-tunnels], IPsec tunnels can introduce MTU

Commenté [BMI49]: What about provisioning matters?

Commenté [BMI50]: Still some provisioning will be required for BGP session.

problems. This document assumes that endpoints manage the appropriate MTU sizes, therefore, not requiring VPN PEs to perform the fragmentation when encapsulating user payloads in the IPsec packets.

5.1. Improvement of IPsec Tunnels Management

IPsec tunnels are a very convenient solution for an enterprise with a small number of locations to reach a Cloud DC. However, for a medium-to-large enterprise with multiple sites and data centers to connect to multiple cloud DCs, there are N*C*2 bi-directional IPsec SAs (tunnels) between Cloud DC gateways and all those sites, with N being the number of enterprise sites and C being the number of Cloud sites. Each of those IPsec Tunnels requires pair-wise periodic key refreshment. For a company with hundreds or thousands of locations, managing hundreds (or even thousands) of IPsec tunnels can be very processing intensive. That is why many Cloud operators only allow a limited number of (IPsec) tunnels ~~&~~ and bandwidth to each customer.

To scale the IPsec key management, a solution like group encryption can be considered. But the drawback of the group encryption is higher security risk of the key distribution and maintenance of a key server.

[SECURE-EVPN] leverages the BGP point-to-multipoint signaling to create private pair-wise IPsec Security Associations among peers without IKEv2 point-to-point signaling or any other direct peer-to-peer session establishment messages.

5.2. Improving CPEs ~~interconnection~~ Interconnection Over the Public Internet

When enterprise CPEs are far away from each other, e.g., across country/continent boundaries, the performance of IPsec tunnels over the public Internet can be problematic and unpredictable. Even though there are many monitoring tools available to measure delay and various performance characteristics of the network, the measurement for paths over the Internet is passive and past measurements may not represent future performance.

[MULTI-SEG-SDWAN] describes some methods to utilize the Cloud backbone to interconnect enterprise CPEs in dispersed geographic locations without requiring the Cloud GW to decrypt and re-encrypt the traffic from the CPEs.

6. Requirements for Networks Connecting Cloud Data Centers

To address the issues identified in this document, network solutions for connecting enterprises with their dynamic workloads or applications in Cloud DCs should satisfy the following requirements:

- Should support scalable policy management for the traffic to and from the newly instantiated application instances at any Cloud DC location.
- Should allow enterprises to take advantage of the current state-of-the-art private VPN technologies, including the ~~traditional~~ conventional circuit-based, MPLS-based VPNs, or IPsec-based VPNs (or any combination thereof) that run over the public Internet.

Commenté [BMI51]: I would characterize what is meant here.

- Should support scalable IPsec key management among all nodes involved in DC interconnect schemes.
- Should support easy and fast, on-demand network connections to dynamic workloads and applications in Cloud DCs and easily reach these workloads when they migrate within or across data centers.
- Should support traffic engineering to distribute loads across regions/AZs based on performance/availability of workloads in addition to the network path conditions to the Cloud DCs.
- Should support network traffic traceability, logging, and diagnostics.
- Should support transit/spoke gateways interconnection scalability and consistent policy enforcement as workloads are increased/migrated. This requirement is mainly for the Cloud Aggregators or Cloud Service Brokers who provide managed services to enterprises over multiple Cloud service providers.

Commenté [BMI52]: Automated?

Commenté [BMI53]: Steering?

7. Security Considerations

The security issues in terms of networking to ~~clouds~~ Cloud DCs include:

- Service instances in Cloud DCs are connected to users (enterprises) via Public IP ports which are exposed to the following security risks:

  a) Potential DDoS attack to the ports facing the untrusted network (e.g., the public internet), which may propagate to the cloud edge resources. To mitigate such security risk, it is necessary for the ports facing internet to enable Anti-DDoS features.

  b) Potential risk of augmenting the attack surface with inter-Cloud DC connection by means of identity spoofing, man-in-the-middle, eavesdropping or DDoS attacks. One example of mitigating such attacks is using DTLS to authenticate and encrypt MPLS-in-UDP encapsulation ~~([~~RFC 7510~~).]~~.

- Potential attacks from service instances within the cloud. For example, data breaches, compromised credentials, and broken authentication, hacked interfaces and APIs, account hijacking.

- When IPsec tunnels established from enterprise on-premises CPEs are terminated at the Cloud DC gateway where the workloads or applications are hosted, traffic to/from an enterprise's workload can be exposed to others behind the data center gateway (e.g., exposed to other organizations that have workloads in the same data center).

  To ensure that traffic to/from workloads is not exposed to unwanted entities, IPsec tunnels may go all the way to the workload (servers, or VMs) within the DC.

Commenté [BMI54]: Cite as a ref

The Cloud DC operator's security practices can affect the overall security posture and need to be evaluated by customers. Many Cloud operators offer monitoring services for data stored in Clouds, such as AWS CloudTrail, Azure Monitor, and many third-party monitoring tools to improve the visibility of data stored in Clouds.

Solution drafts resulting from this work will address security concerns inherent to the solution(s), including both protocol aspects and the importance (for example) of securing workloads in cloud DCs and the use of secure interconnection mechanisms.

A full security evaluation will be needed before [SECURE-EVPN] can be recommended.

8. IANA Considerations

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

9. References

9.1. Normative References

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

[RFC3168] K. Ramakrishnan, et al, "The Addition of Explicit Congestion Notification (ECN) to IP", RFC3168, Sept. 2001.

[RFC4486] E. Chen and V. Gillet, "Subcodes for BGP Cease Notification Message", RFC4486, April 2006.

[RFC4786] J. Abley and K. Lindqvist, "Operation of Anycast Services", RFC4786, Dec. 2006.

[RFC5492] J. Scudder and R. Chandra, "Capabilities Advertisement with BGP-4", RFC5492, Feb. 2009.

[RFC6040] B. Briscoe, "Tunnelling of Explicit Congestion Notification", RFC6040, Nov 2010.

[RFC7606] E. Chen, et al "Revised Error Handling for BGP UPDATE Messages". Aug 2015.

[RFC7432] A. Sajassi, et al "BGP MPLS-Based Ethernet VPN", RFC7432, Feb. 2015.

Commenté [BMI55]: Not sure that all these are normative ones. Please double check.

9.2. Informative References

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

[3GPP-5G-Edge] 3GPP TS 23.548 v18.1.1, "5G System Enhancements for Edge Computing", April 2023.

[SDWAN-EDGE-DISCOVERY] L. Dunbar, S. Hares, R. Raszuk, K. Majumdar, G. Mishra, V. Kasiviswanathan, "BGP UPDATE for SD-WAN Edge Discovery", draft-ietf-idr-sdwan-edge-discovery-10, June 2023.

[AWS-NAT] NAT gateways - Amazon Virtual Private Cloud.

[AWS-Cloud-WAN] Introducing AWS Cloud WAN (Preview) | Networking &

Content Delivery (amazon.com).

[Azure-SD-WAN] Architecture: Virtual WAN and SD-WAN connectivity - Azure Virtual WAN | Microsoft Learn.

[Google-NAT] Cloud NAT overview | Google Cloud.

[Int-tunnels] J. Touch and W Townsley, "IP Tunnels in the Internet Architecture", draft-ietf-intarea-tunnels-13.txt, March, 2023.

[MEF-70.1] MEF 70.1 SD-WAN Service Attributes and Service Framework. Nov. 2021.

[METADATA-PATH] L. Dunbar, et al, "BGP Extension for 5G Edge Service Metadata" draft-ietf-idr-5g-edge-service-metadata-09, Sept. 2023.

[MULTI-SEG-SDWAN] K. Majumdar, et al, "Multi-segment SD-WAN via Cloud DCs", draft-dmk-rtgwg-multisegment-sdwan-02, Sept 2023.

[SECURE-EVPN] A. Sajassi, et al, "Secure EVPN", draft-ietf-bess-secure-evpn-00, June 2023.

[SERVICE-METRICS] L. Dunbar, et al, "5G Edge Services Use Cases and Metrics", draft-dunbar-cats-edge-service-metrics-01, July 2023.

[Split-Horizon-DNS] K. Tirumaleswar, et al, "Establishing Local DNS Authority in Validated Split-Horizon Environments", draft-ietf-add-split-horizon-authority-04, Mar. 2023.

10. Acknowledgments

Many thanks to Joel Halpern, Aseem Choudhary, Adrian Farrel, Alia Atlas, Chris Bowers, Paul Vixie, Paul Ebersman, Timothy Morizot, Ignas Bagdonas, Donald Eastlake, Michael Huang, Liu Yuan Jiao, Katherine Zhao, and Jim Guichard for the discussion and contributions.

Authors' Addresses

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

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

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

Mehmet Toy

Commenté [BMI56]: Please add an url

Commenté [BMI57]: Please add a link

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

Kausik Majumdar
Microsoft Azure
kmajumdar@microsoft.com