RTG Working Group
Internet Draft
Intended status: Informational
Expires: October 12, 2024

L. Dunbar
Futurewei
A. Malis
Malis Consulting
C. Jacquenet
Orange
M. Toy
Verizon
K. Majumdar
Microsoft

April 15, 2024

Dynamic Networks to Hybrid Cloud DCs: Problems and Mitigation Practices

draft-ietf-rtgwg-net2cloud-problem-statement-39

Abstract

This document describes a set of network-related problems enterprises face at the time of writing this document (2023) when interconnecting their branch offices with dynamic workloads in third-party data centers (DCs) (a.k.a. Cloud DCs). These problems are mainly from enterprises with conventional VPN services that want to leverage those networks (instead of altogether abandoning them). This document also describes various mitigation practices and actions to soften the issues induced by these problems.

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

This Internet-Draft will expire on October 15, 2024.

Copyright Notice

Table of Contents

1. Introduction

With the advent of widely available Cloud data centers (DCs) providing services in various geographic locations and advanced tools for monitoring and predicting application behaviors, it is tempting for enterprises to instantiate applications and workloads in Cloud DCs. Some enterprises prefer specific applications to be located close to the end users accessing these services, as the proximity can improve end-to-end latency. 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.

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

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

2. Definition of Terms

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 cloud is a mixed computing environment where applications are run using a combination of computing, storage, and services in different public clouds and private clouds, including on-premises data centers or "edge" locations. <https://cloud.google.com/learn/what-is-hybrid-cloud>.

IXPs: Internet exchange points (IXes or IXPs) are the common grounds of IP networking, allowing participating 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 transport of IP Packets over one or more Underlay Connectivity Services by recognizing applications (Application Flows) and determining forwarding behavior by applying Policies to them. [MEF-70.1]

VPC: A 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 client can launch his/her desired resources, such as compute, storage, or network functions into his/her VPC. At the time of writing this document, most Cloud operators' VPCs only support private addresses, some support IPv4 only, others support IPv4/IPv6 dual stack.

3. Issues and Mitigation Methods of Connecting to Cloud DCs

This section identifies some high-level problems that the IETF could address, especially within the Routing area. Other Cloud DC problems (e.g., managing cloud spending) are out of the scope of this document.

3.1. Increased BGP Peering Errors and Mitigation Methods

Where conventional ISPs view BGP peering as a means to improve network operations, Public Cloud DCs offer direct BGP peering to attract more customers to use their DCs and services. As such, there is pressure to peer more widely with more customers, including those who may lack the expertise and experience in running complex BGP peering relationships. This can contribute to increased BGP peering errors such as capability mismatch, unwanted route leaks, missing Keepalives, and errors causing BGP session resets. Capability mismatch can cause BGP sessions not to be adequately established. These issues are more acute for Cloud DCs than they have been, even though they may apply to conventional ISPs, just to a lesser degree. Here are the recommended mitigation practices:

- If a Cloud Gateway (GW), a BGP speaker, receives from its BGP peer a capability that it does not itself support or recognize, it need to ignore that capability, and the BGP session need not be terminated per [RFC5492]. When receiving a BGP UPDATE with a malformed attribute, the revised BGP error handling procedure in [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 need to establish tunnels (e.g., IPsec) to the Cloud GWs to form an 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 a 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 inbound routes exceeding the maximum routes threshold for a peer, the currently common practice is generating out-of-band alerts (e.g., Syslog entries) via the management system or terminating the BGP session (with cease notification messages [RFC4486] being sent). Although out of the scope of this document, more discussion is needed in the IETF Inter-Domain Routing (IDR) Working Group for potential in-band or autonomous notification directly to the peers when the inbound routes exceed the maximum routes threshold.
- Leveraging YANG models to programmatically synchronize configurations between BGP peers (e.g., [SVC-AC]) and to adjust the local configuration accordingly (e.g., [NTW-AC] or [DATAMODEL-BGP]). This proactive approach reduces the likelihood of BGP configuration issues and ensures that both BGP peers operate with synchronized and compatible settings.

3.2. Site Failures and Methods to Minimize Impacts

Failures within a Cloud site, which can be a building, a floor, a pod, or a server rack, include capacity degradation or complete out-of-service failure. Here 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 during a power failure; d) cyber threat attacks; e) too many changes outside of the maintenance window; etc. A fiber-cut is not uncommon in a Cloud site or between sites.

As described in [RFC7938], a Cloud DC might not have an IGP 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)[RFC5880].

When a site failure occurs, many services can be impacted. When the impacted services' IP prefixes in a Cloud DC are not aggregated nicely, which is 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 in different Cloud DCs. A commonly deployed solution has DNS server(s) responding to a Fully Qualified Domain Name (FQDN) inquiry with an IP address of the instance in the closest or lowest cost DC. Here are some problems associated with DNS-based solutions:

- Dependent on client behavior
    - A misbehaving client can cache results indefinitely.
    - Clients may fail to access a service even though there are servers available in other Cloud DCs because the failing IP address is still cached in the DNS resolver and has not expired yet.

- No inherent use 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 periodic updates of the network condition, which is difficult.

One method to mitigate the problems listed above is to use anycast [RFC4786] for the services so that network proximity and conditions can be automatically considered in optimal path selection.

[METADATA-PATH] identifies some of 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.

[RFC8490] and [RFC8765] on stateful DNS can be used to achieve better performance in refreshing the cache and handling session idle timeouts.

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

5G Edge Cloud 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.

Here are some network problems with connecting to 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 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.

[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 (Computing-Aware Traffic Steering) 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 provided 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 can be forwarded to the DNS service in another Cloud. Configuration can be complex depending on the application communication patterns.

However, name collisions can still occur even with carefully managed policies and configurations. If an organization uses internal names like those under a .internal top level domain name, and wants its services to be available via or within some other Cloud provider that also uses .internal, collisions might occur. Therefore, using a 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, and they make DNSSEC trust manageable. Consider using a registered and FQDN from global DNS as the root for enterprise and other internal namespaces.

3.6. NAT Practices for Accessing Cloud Services

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

Different Cloud operators support different levels of NAT functionality. 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 need to be performed in Cloud DCs and their on-premises DC.

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 DCs 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 to 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 [RFC8926] header or IPv6 optional header) to indicate additional properties, including useful information about the sites where they are instantiated.

4. Dynamic Connecting Enterprise Sites with Cloud DCs

For many enterprises with established private VPNs (e.g., private circuits, MPLS-based L2VPN[RFC6136]/L3VPN[RFC4364]) interconnecting branch offices and 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.

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

4.1. Sites to Cloud DC

Most Cloud operators offer multiple types of network gateways (GWs) through which an enterprise can reach their workloads hosted in the Cloud DCs:

- Internet GW for services hosted in the Cloud DCs to be accessed by external requests via Internet routable addresses. E.g., AWS Internet GW [AWS-Cloud-WAN].
- IPsec tunnels terminating GW for establishing IPsec SAs [RFC6071] with an enterprise's own gateway, so that the communications between those gateways can be secured from the

underlay (which might be the public Internet). E.g., AWS Virtual gateway (vGW).

- Direct connect GW for enterprises to connect with Cloud services via private leased lines provided by Network Service Providers. E.g., AWS Direct Connect. 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) that are accessible via a virtual router connected by AWS Internet Gateway; some of the same tenant (TN-1) services 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, 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 & 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.

A CPE (Customer Premises Equipment) can be a customer owned router or ports physically connected to an AWS Direct Connect GW.

```
     +------------------------+
     |    ,---.         ,---. |
     |   (TN-1 )       ( TN-2)|
     |    `-+-'  +---+  `-+-' |
     |      +----|vR1|----+   |
     |           ++--+        |
     |            |         +-+----+
     |            |        /Internet\ For external customers
     |            +-------+ Gateway  +----------------------
     |                     \        / to reach via Internet
     |                      +-+----+
     |                        |
     |    ,---.         ,---. |
     |   (TN-1 )       ( TN-2)|
     |    `-+-'  +---+  `-+-' |
     |      +----|vR2|----+   |
     |           ++--+        |
     |            |         +-+----+
     |            |        / virtual\ For IPsec Tunnel
     |            +-------+ Gateway  +----------------------
     |            |        \        /  termination
     |            |         +-+----+
     |            |           |
     |            |    + - - - - - - - - - - - - - - - --+
     |            |    |    +-+----+         +----+
     |            |        /        \ Direct /      \   |
     |            +----|--+ Gateway  +------+ Fabric|--VPN-- CPE
     |                     \        / Connect\ edge /   |
     |                 |    +-+----+         +----+
     |                        |         IXP              |
     |                 + - - - - - - - - - - - - - - - --+
     +------------------------+
```

TN: Tenant Network. One TN can be attached to both vR1 and vR2.
Figure 1: Examples of Multiple Cloud DC connections.

4.2. Inter-Cloud Connection

The connectivity options to Cloud DCs described in Section 4.1 are for reaching Cloud providers' DCs, but not between cloud DCs. For example, 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 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 intermediary 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 costs 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, have been widely deployed as an effective way to support businesses and organizations that require network performance and reliability although such services may be considered premium, available only at additional cost. Connecting an enterprise's on-premesis 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].

An enterprise can connect to multiple Cloud DC locations and establish different BGP peering 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-premises 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-premises 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 Geolocation of the Cloud DC, the delay measurement, or other relevant data.

5. Methods to Scale IPsec Tunnels to Cloud DCs

As described in Section 4.3, IPsec tunnels can be used to dynamically establish connection between private VPN PEs with Cloud GWs. 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 problems. This document assumes that endpoints manage the appropriate MTU sizes, therefore, not requiring VPN PEs to perform fragmentation when encapsulating user payloads in the IPsec packets.

5.1. Scale 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 fully 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.

A solution like group key management [RFC4535] has been used to scale the IPsec key management. The group key management protocol documented in [RFC4535] outlines the relevant security risks for any group key management system in Section 3 (Security Considerations). While this particular protocol isn't being suggested, the drawbacks and risks of group key management are still relevant.

[SDWAN-EDGE-DISCOVERY] leverages the peers communication polices on the SD-WAN controller and BGP Update messages to exchange IPsec Security Associations related parameters among peers without IKEv2 point-to-point signaling or any other direct peer-to-peer session establishment messages.

5.2. CPEs 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] outlines some approaches for leveraging the Cloud backbone to connect enterprise CPEs across diverse geographical areas, eliminating the need for the Cloud GW to decrypt and re-encrypt traffic from the CPEs. A thorough examination of the security implications associated with this proposed method is necessary. Alternative encapsulations, like SRH (Segment Routing Header) or others, can be considered for interconnecting enterprise 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. The scalable policy management, even though out of the scope of this document, can include centralized policy repositories and API-driven automation.
- Should allow enterprises to take advantage of the current state-of-the-art private VPN technologies, including the conventional circuit-based, MPLS-based VPNs, or IPsec-based VPNs (or any combination thereof) that run over the public Internet.
- 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 steering 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.

7. Security Considerations

The security issues in terms of networking to 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 (Distributed Denial of Service) 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 [RFC7510].

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

- Group key management [RFC4535] comes with security risks such as: keys being used too long, single points of compromise (one compromise affects the whole group), key distribution vulnerabilities, key generation vulnerabilities, to name a few.

  [RFC4535] outlines the security risks in Section 3 (Security Considerations). While this specific protocol isn't being suggested the risks and vulnerabilities apply to any group key management system.

- Striking a balance between scaling IPsec tunnel management outlined in this document and maintaining robust security is a delicate consideration. Simplifying the IPsec tunnel management to reduce management complexity for large SD-WAN networks might come with the inherent risk of decreased security. Careful consideration of the specific deployments, coupled with regular security assessments, is crucial to ensure the integrity and confidentiality of the transmitted data.

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 [MULTI-SEG-SDWAN] and [SDWAN-EDGE-DISCOVERY] can be recommended as a solution to some problems described in this document.

8. IANA Considerations

This document requires no IANA actions.

9. References

9.1. Normative References

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

[RFC4364] E. Rosen and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC4364, Feb. 2006.

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

[RFC4535] H. Harney, et a, "GSAKMP: Group Secure Association Key Management Protocol", RFC4535, June 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.

[RFC5880] D. Katz and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC5880, June 2010.

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

[RFC6136] A. Sajassi and D. Mohan, "Layer 2 Virtual Private Network (L2VPN) Operations, Administration, and Maintenance (OAM) Requirements and Framework", RFC6136, March 2011.

[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.

[RFC7510] X. Xu, et al, "Encapsulating MPLS in UDP", RFC7510, April, 2015.

[RFC7938] P. Lapukhov, "Use of BGP for Routing in Large-Scale Data Centers", RFC7938, Aug. 2016.

[RFC8490] R. Bellis, et al, "DNS Stateful Operations", RFC8490, March 2019.

[RFC8765] T. Pusateri and S. Cheshire, "DNS Push Notifications", RFC8765, June 2020.

[RFC8926] J. Gross and T. Sridhar, "Geneve: Generic Network Virtualization Encapsulation", RFC8926, Nov. 2020.

9.2. Informative References

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

[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-12, Oct 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.

[NTW-AC] M. Boucadair, et al, "A Network YANG Data Model for Attachment Circuits", draft-ietf-opsawg-ntw-attachment-circuit-08, March 2024.

[DATAMODEL-BGP] M. Jethanandani, K. Patel, S. Hares, "YANG Model for Border Gateway Protocol (BGP-4)", draft-ietf-idr-bgp-model-17, July 2023.

[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-16, March, 2024.

[MULTI-SEG-SDWAN] K. Majumdar, et al, "Multi-segment SD-WAN via Cloud DCs", draft-dmk-rtgwg-multisegment-sdwan-07, Feb 2024.

[SVC-AC] M. Boucadair, et al. "YANG Data Models for 'Attachment Circuits'-as-a-Service (ACaaS)", draft-ietf-opsawg-teas-attachment-circuit-10, April 2024.

[Split-Horizon-DNS] K. Tirumaleswar, et al, "Establishing Local DNS Authority in Validated Split-Horizon Environments", draft-ietf-add-split-horizon-authority-07, Mar. 2023.

10. Acknowledgments

Many thanks to Joel Halpern, Aseem Choudhary, Adrian Farrel, Alia Atlas, Chris Bowers, Mohamed Boucadair, 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
Verizon
One Verizon Way
Basking Ridge, NJ 07920
Email: mehmet.toy@verizon.com

Kausik Majumdar
Microsoft Azure
kmajumdar@microsoft.com