AWS Well-Architected Framework

# Hybrid Networking Lens

# Hybrid Networking Lens: AWS Well-Architected Framework

# Table of Contents

# Hybrid Networking Lens

Publication date: **November 22, 2021** (*Document history*)

This whitepaper describes the Hybrid Networking Lens for the AWS Well-Architected Framework, which helps customers review and improve their cloud-based architectures and better understand the business impact of their design decisions. The document describes general design principles, as well as specific best practices and guidance for the six pillars of the Well-Architected Framework.

This paper is intended for those in technology roles, such as chief technology officers (CTOs), architects, developers, and operations team members. After reading this paper, you will understand AWS best practices and the strategies to use when designing hybrid networking architectures.

## Introduction

The AWS Well-Architected Framework helps you understand the pros and cons of decisions you make while building systems on AWS. Using the Framework helps you learn architectural best practices for designing and operating reliable, secure, efficient, and cost-effective systems in the cloud. The Framework provides a way for you to consistently measure your architectures against best practices and identify areas for improvement. We believe that having well-architected systems greatly increases the likelihood of business success.

In this Lens, we focus on how to design, deploy, and architect hybrid networking for workloads in the AWS Cloud. For brevity, we only cover details from the Well-Architected Framework that are specific to hybrid networking. We recommend that you consider best practices and questions from the AWS Well-Architected Framework whitepaper when designing your architecture.

## Custom lens availability

Custom lenses extend the best practice guidance provided by AWS Well-Architected Tool. AWS WA Tool allows you to create your own custom lenses, or to use lenses created by others that have been shared with you.

To determine if a custom lens is available for the lens described in this whitepaper, reach out to your Technical Account Manager (TAM), Solutions Architect (SA), or Support.

# Definition

Hybrid Networking refers to a network that spans AWS and an on-premises data center. Hybrid networking architectures help organizations integrate their on-premises data center and AWS operations to support a broad spectrum of use cases, by using a common set of cloud services, tools, and APIs across on-premises and cloud environments. To establish a hybrid networking environment, there are specific services that connect your on-premises and AWS resources through a common network. For example, you can use Amazon Virtual Private Cloud (Amazon VPC) to gain control over your virtual networking environment in AWS, AWS Site-to-Site VPN to provide secure encrypted remote office to AWS connectivity over the internet in minutes, and AWS Direct Connect to establish a dedicated private network connection between AWS and your on-premises environment.

The AWS Well-Architected Framework is based on six pillars: operational excellence, security, reliability, performance efficiency, cost optimization, and sustainability. When architecting technology solutions, you must make informed tradeoffs between pillars based upon your business context. AWS provides multiple core components that enable you to design robust architectures for your hybrid networking workload applications. This section presents an overview of the AWS components that are used throughout this document to architect hybrid networking workloads. There are three specific areas to consider when designing hybrid network connectivity for your workload:

- Data layer
- Monitoring and config management
- Security

# Data layer

The data layer of your hybrid networking environment is important because it provides a data path for network traffic between applications hosted on AWS and an on-premises data center. As part of your hybrid deployment, it's important to carefully consider the hybrid connectivity options between AWS and an on-premises data for the forwarding of your network traffic. The Hybrid Networking Lens recommends using AWS Virtual Private Network (AWS VPN), AWS Direct Connect, and Amazon VPC to support your application team's agility and speed to market by using AWS as a data center extension. AWS VPN and Direct Connect are used as network paths for providing connectivity for these hybrid networking workloads.

**AWS Virtual Private Network** (AWS VPN) establishes a secure and private tunnel from your network or device to the AWS Cloud. AWS Site-to-Site VPN allows you to securely connect your on-premises network or branch office site to your Amazon VPC.

**AWS Direct Connect** (DX) makes it easy to establish a dedicated network connection from your on-premises environment to AWS. Using Direct Connect, you can establish private connectivity between AWS and your data center, office, or colocation environment. In many cases, this can reduce your network costs, increase bandwidth throughput, and provide a more consistent network experience than internet-based connections.

A **virtual private gateway** (VGW) is part of a VPC that provides edge routing for AWS managed VPN connections and Direct Connect connections. You associate a Direct Connect gateway with the virtual private gateway for the VPC.

**AWS Transit Gateway** (TGW) connects VPCs and on-premises networks through a central hub. It is a fully managed AWS gateway that acts as a cloud router and enables rich routing scenarios. With AWS Transit Gateway, you can quickly add Amazon VPCs, AWS accounts, VPN capacity, or AWS Direct Connect gateways to meet unexpected demand, without having to wrestle with complex connections or massive routing tables.

## Monitoring and config management

Monitoring and config management is an important way to gain insights and improve the performance of your hybrid networking environment. AWS provides the following monitoring and config management services that enable you to monitor your AWS services and resolve the root causes of performance issues based on your business needs.

**Amazon CloudWatch** enables you to access system metrics on the AWS services that are being used, consolidate system and application level logs, and create business KPIs as custom metrics for your specific needs. CloudWatch provides dashboards and alerts that can trigger automated actions on the platform.

**AWS Health Dashboard** provides alerts and remediation guidance when AWS is experiencing events that may impact you and The **Service Health Dashboard** provides public information about the regional availability of a service. While the Service Health Dashboard displays the general status of AWS services, AWS Health Dashboard gives you a personalized view into the performance and availability of the AWS services underlying your AWS resources.

**VPC Reachability Analyzer** is a configuration analysis tool that enables you to perform connectivity testing between a source resource and a destination resource in your virtual private clouds (VPCs). When the destination is reachable, reachability analyzer produces hop-by-hop details of the virtual network path between the source and the destination. When the destination is not reachable, reachability analyzer identifies the blocking component. While VPC reachability analyzer doesn't test end to end hybrid connectivity, it can analyze connectivity between VPN gateways (VGW or TGW) and a target within a VPC.

**Transit Gateway Network Manager** (Network Manager) enables you to centrally manage your networks that are built around transit gateways. You can visualize and monitor your global network across AWS Regions and on-premises locations.

**Route Analyzer** is a part of Transit Gateway Network Manager that allows you to perform an analysis of the routes in your transit gateway route tables. The Route Analyzer analyzes the routing path between a specified source and destination, and returns information about the connectivity between components. You can use the Route Analyzer to perform the following actions:

- Verify that the transit gateway route table configuration will work as expected before you start sending traffic
- Validate your existing route configuration
- Diagnose route-related issues that are causing traffic disruption in your global network

**AWS Config** is a service that enables you to assess, audit, and evaluate the configurations of your AWS resources. Config continuously monitors and records your AWS resource configurations and allows you to automate the evaluation of recorded configurations against desired configurations. With Config, you can review changes in configurations and relationships between AWS resources, dive into detailed resource configuration histories, and determine your overall compliance against the configurations specified in your internal guidelines. This enables you to simplify compliance auditing, security analysis, change management, and operational troubleshooting.

# Security

The security of your hybrid networking environment is crucial and should span protection across your on-premises and cloud environments. It is recommended to implement security controls for traffic within your cloud environment, and particularly, traffic flowing from the Internet and from your on-premises network to AWS. The following services can help secure your hybrid networking environment on AWS.

A **security group** (SG) acts as a virtual firewall for your network interfaces in a VPC and allows you to control inbound and outbound traffic including hybrid traffic.

A **network access control list** (ACL) is an optional layer of security for your VPC that acts as a firewall for controlling traffic in and out of one or more subnets. You might set up network ACLs with rules similar to your security groups in order to add an additional layer of security to your VPC.

**AWS Network Firewall** is a managed service that makes it easy to deploy essential network protections such as URL and domain names, IP addresses, and content-based traffic filtering to secure traffic traversing to and from your VPCs.

**AWS CloudTrail** is a service that enables governance, compliance, operational auditing, and risk auditing of your AWS account. With CloudTrail, you can log, continuously monitor, and retain account activity related to actions across your AWS infrastructure. CloudTrail provides event history of your AWS account activity, including actions taken through the AWS Management Console, AWS SDKs, command line tools, and other AWS services. This event history simplifies security analysis, resource change tracking, and troubleshooting. In addition, you can use CloudTrail to detect unusual activity in your AWS accounts. These capabilities also help simplify operational analysis and troubleshooting.

**Amazon GuardDuty** is a threat detection service that continuously monitors for malicious activity and unauthorized behavior to protect your AWS accounts and workloads. With the cloud, the collection and aggregation of account and network activities is simplified, but it can be time consuming for security teams to continuously analyze event log data for potential threats. With GuardDuty, you now have an intelligent and cost-effective option for continuous threat detection in AWS. GuardDuty analyzes tens of billions of events across multiple AWS data sources, such as AWS CloudTrail event logs, Amazon VPC Flow Logs, and DNS logs.

# General design principles for hybrid networking

This section outlines two key scenarios that are common in hybrid networking and how they influence the design and architecture of your workloads on AWS. It also presents the assumptions that were made for each scenario, the common drivers for the design, and a reference architecture of how these scenarios are implemented.

This whitepaper covers the following two hybrid networking scenarios:

- Getting started with hybrid connectivity using Site-to-Site VPN
- Designing a reliable, dedicated global hybrid networking setup using fiber connectivity for critical workloads

# Getting started with hybrid connectivity using Site-to-Site VPN

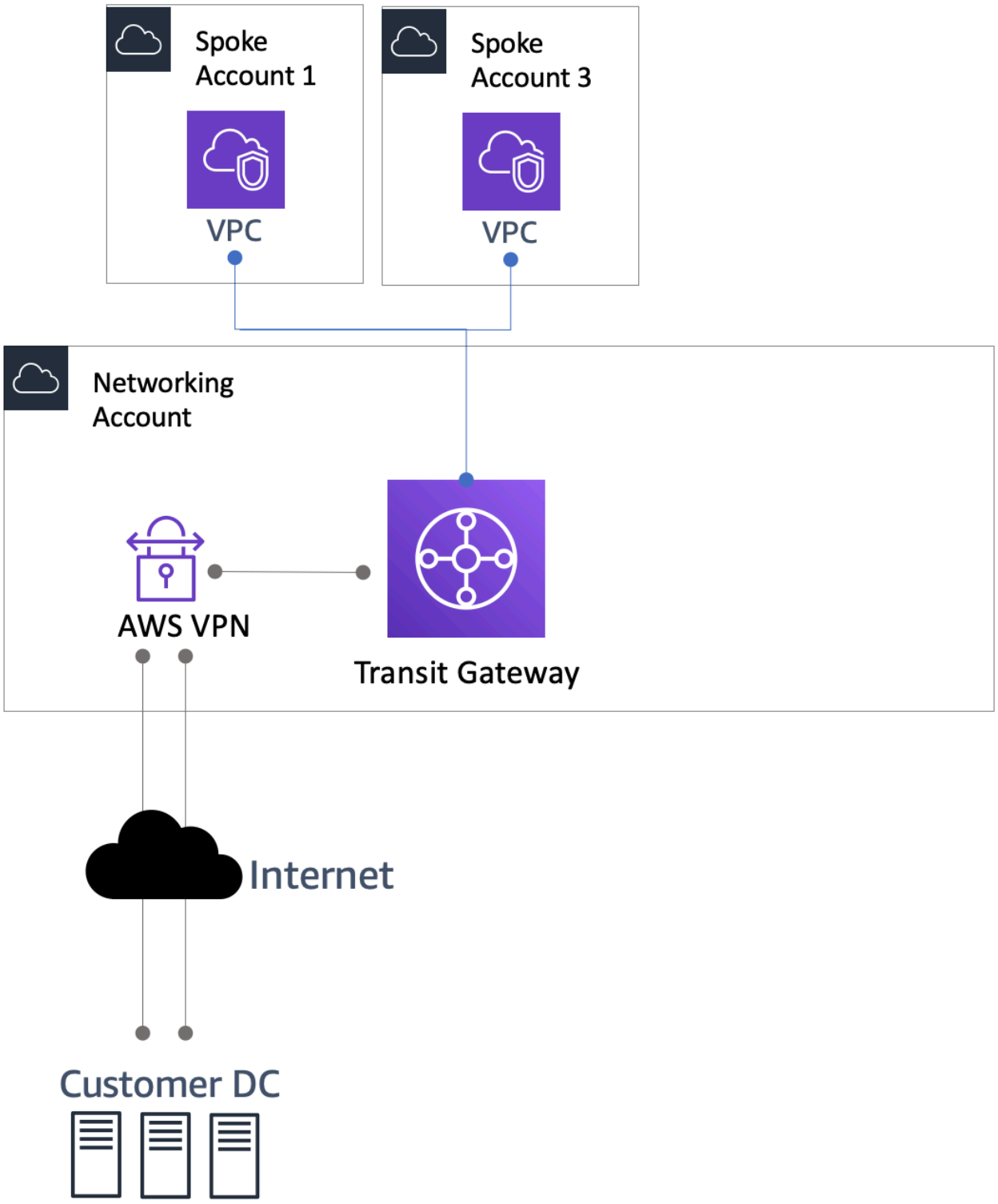
The easiest way to get started with hybrid connectivity is to establish site-to-site VPN over the internet. AWS Site-to-Site VPN extends your data center or branch office to the cloud using IP Security (IPsec) tunnels. You can configure routing using Border Gateway Protocol (BGP) over the IPsec tunnel or configure static routes. Traffic in the tunnel is encrypted with AES128 or AES256 and uses Diffie-Hellman groups for key exchange, providing Perfect Forward Secrecy.

Each AWS Site-to-Site VPN connection consists of two VPN tunnel endpoints for redundancy. For high-availability, it's important to terminate a VPN tunnel to both of the endpoints.  Each tunnel terminates in a different Availability zone within the AWS global network, but must also terminate on the same equipment on-premises.  It's also important that you have a similar highly-available configuration set up at the on-premises equipment and terminate the VPN on two different physical devices in your data center.

AWS Site-to-Site VPN supports terminating IPSEC tunnels to both virtual private gateway and AWS Transit Gateway at the AWS end. When terminating a VPN on a virtual private gateway, you can access the VPC that the gateway is attached to. For every other VPC that you want to connect to, you must create a separate VPN tunnel to a separate virtual private gateway attached to that VPC. With AWS Transit Gateway you get connectivity to thousands of VPCs over a pair of VPN tunnels. Additionally, Transit Gateway supports Equal Cost Multipath (ECMP routing strategy, allowing you to load balance traffic across multiple VPN tunnels for high-availability and bandwidth aggregation.

When leveraging Transit Gateway, you can optionally enable acceleration for your Site-to-Site VPN connection. An accelerated Site-to-Site VPN connection uses AWS Global Accelerator to route traffic from your on-premises network to an AWS edge location that is closest to your customer gateway device. AWS Global Accelerator optimizes the network path, using the congestion-free AWS global network to route traffic to the endpoint that provides the best application performance.

When using Site-to-Site VPN you are charged for each VPN connection-hour that your VPN connection is provisioned and available. Data transfer out on AWS Site-to-Site VPN incurs data transfer out charges. For more information, refer to the EC2 On-Demand pricing page. When using Transit Gateway, in addition to the AWS Site-to-Site VPN costs, you also pay for transit gateway VPN attachment. For more information, refer to AWS Transit Gateway pricing. To summarize, terminating VPN at TGW gives you a lot more flexibility into the number of VPCs you can connect to over single tunnel and provides added functionality like ECMP, accelerated VPN and hence is a recommended default starting point for your architectures. That being said, for some unique use cases involving large data transfers, leveraging the VGW termination endpoint can be cheaper and hence can be a viable alternative.

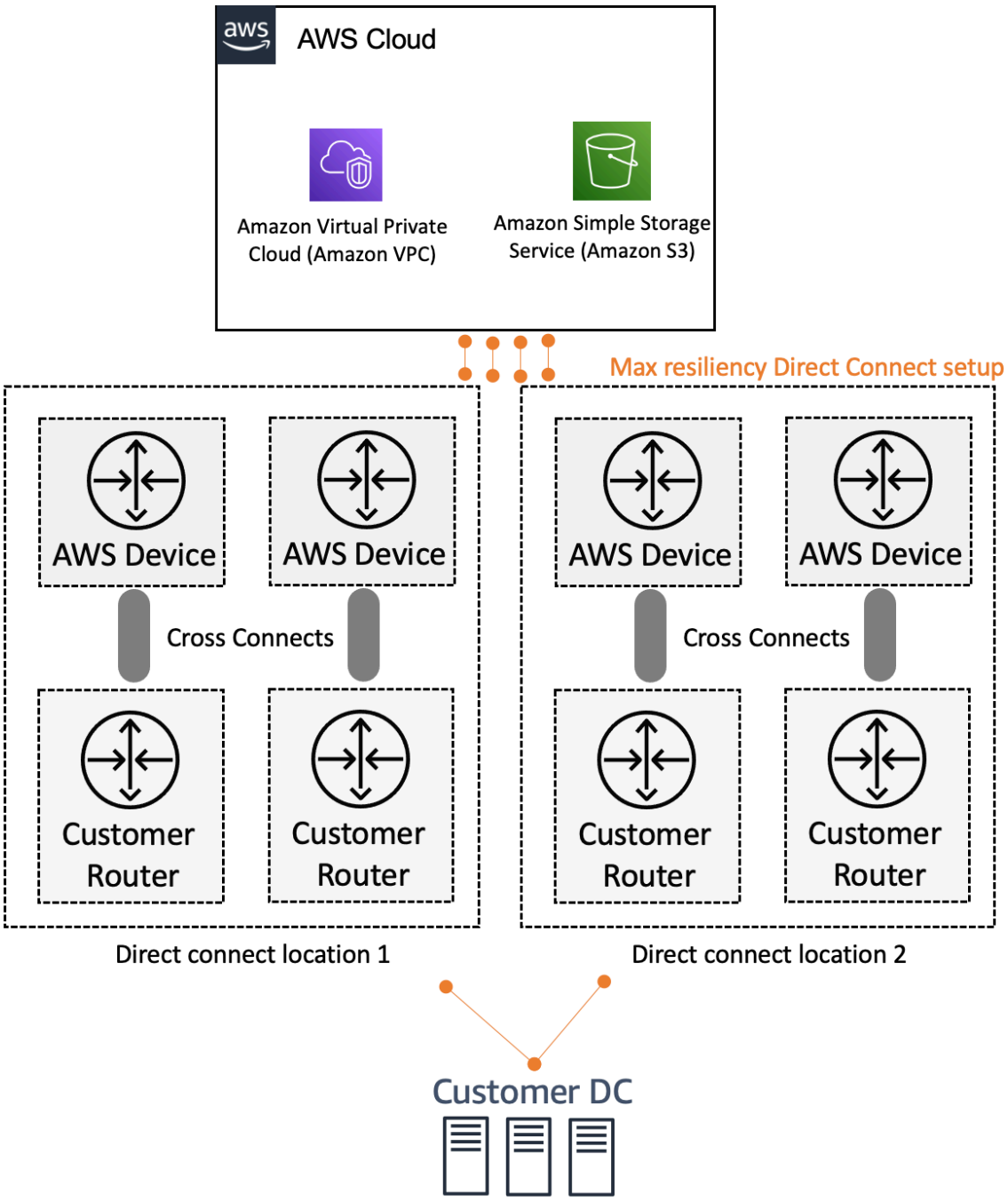



*Site-to-Site VPN reference architecture*

# Designing a reliable, dedicated hybrid networking setup using fiber connectivity for critical workloads

Network latency over the internet varies because it is constantly changing how data gets from point A to point B. While accelerated VPN helps with network latency, the last mile from an AWS edge location to your data center is still over the internet. To achieve consistent end-to-end network performance, you can leverage AWS Direct Connect to enable consistent, low latency, high-bandwidth dedicated fiber connectivity between your data centers and AWS. Direct Connect provides dedicated connections at bandwidths of 1 Gbps, 10 Gbps, and 100 Gbps. Hosted connections are provided by AWS Direct Connect Partners using pre-established network links between themselves and AWS and are available from 50 Mbps up to 10 Gbps.

AWS Direct Connect is available at over 100 locations around the world. Building highly resilient, fault-tolerant connections are key to a well-architected system when connecting to AWS Direct Connect locations. AWS recommends connecting from multiple data centers for physical location redundancy as well as establishing multiple connections at a direct connect location for device redundancy. When designing WAN connections, investigate using redundant hardware and telecommunications providers with redundant paths. A best practice is to use dynamic routing with Direct Connect, Active/Active connections for automatic load balancing, and failover across redundant network connections. Additionally, provision sufficient network capacity to ensure that the failure of one network connection does not overwhelm and degrade redundant connections. To achieve the best availability, you can leverage the resiliency architecture for AWS Direct Connect as shown in the following diagram.




*AWS Direct Connect maximum resiliency architecture*

If you need more than 100 Gbps of bandwidth, you can provision a link aggregation group (LAG) bundle with AWS Direct Connect. A LAG is a logical interface that uses the Link Aggregation Control protocol (LACP) to aggregate multiple connections at a single Direct Connect location, allowing you to treat them as a single, managed connection. You can have a maximum of two 100G connections, or four connections with a port speed less than 100G in a LAG. You can create a LAG from existing connections, or you can provision new connections. However, a LAG only

includes ports on the same AWS device. AWS doesn't support multi-chassis LAG, this means all of your Direct Connect connections terminate on the same hardware on the AWS side. A LAG is not recommended for a high-availability strategy.

Once the physical connectivity is established at the Direct Connect location, you can create virtual interfaces which are logical connections on top of physical Direct Connect connections that enable access to AWS resources. These virtual interfaces are tagged with 802.1Q VLANs and require the use of Border Gateway Protocol (BGP).

AWS Direct Connect provides the following virtual interfaces:

**Public virtual interfaces –** Provide global connectivity to public AWS resources, including AWS

public service endpoints public Amazon EC2 IP addresses, and public Elastic Load Balancing addresses.

**Private virtual interfaces –** Provide connectivity to the private IP range of your VPC.

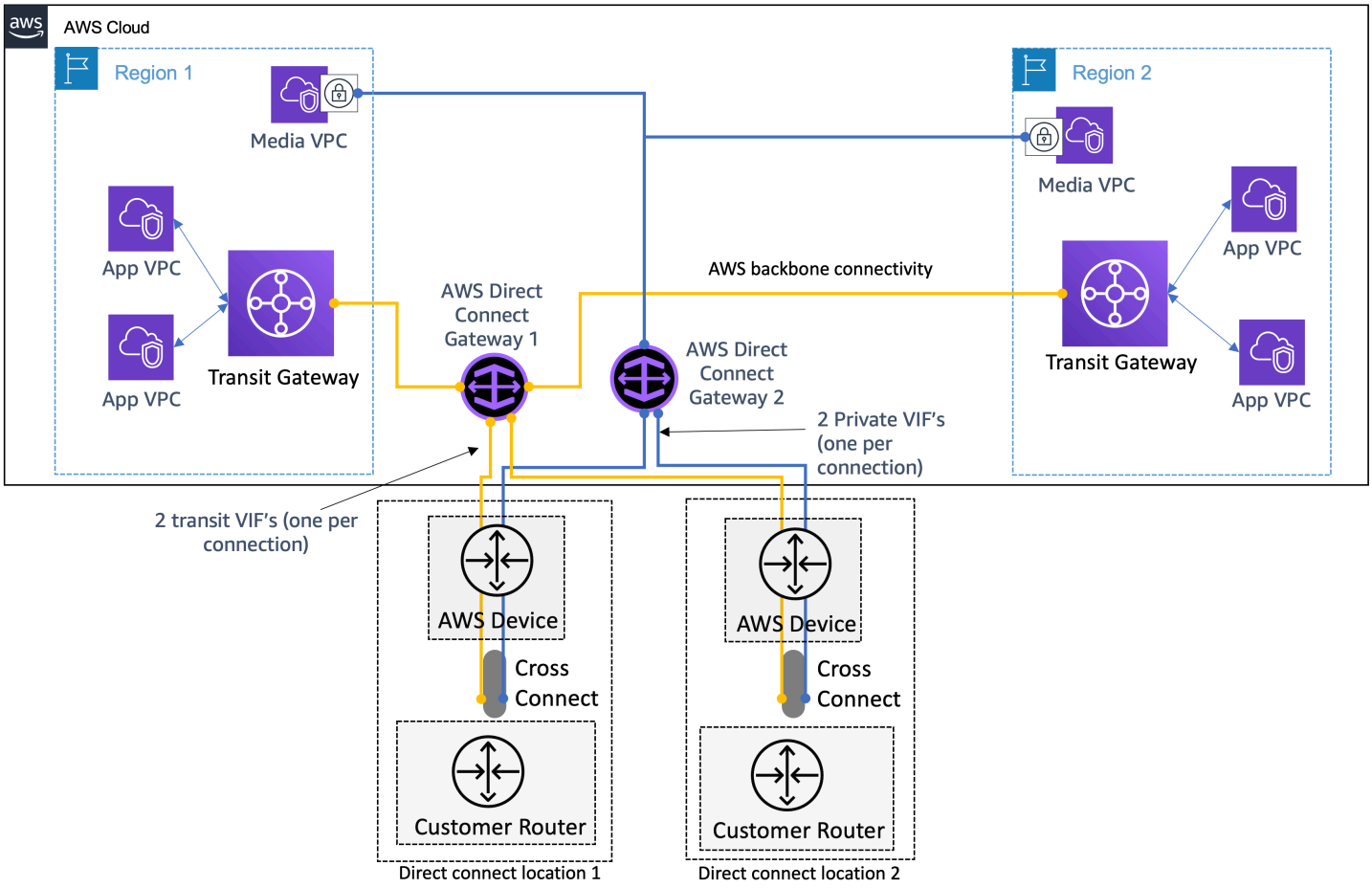
When you use private virtual interfaces, your VPC becomes a logical layer-3 extension of your network. For information about pricing, refer to the AWS Direct Connect pricing.

**Transit virtual interfaces –** Enables connectivity to Transit Gateway(s). While this virtual interface enables scaling you pay for the cost associated with AWS Direct Connect and Transit Gateway. For more information about pricing, refer to the AWS Transit Gateway pricing.

In the digital world of today, most customers are establishing global presence. There is a need to deploy resources within a large number of VPCs across multiple AWS Regions and connect to them from datacenters spread across geographies. By leveraging AWS Direct Connect gateway, a global resource that allows you to use your Direct Connect connections to connect to resources in VPCs in most AWS Regions, you can more easily connect to your resources. To enable global connectivity, you can create and associate a Direct Connect gateway with the virtual private gateway of the VPC you want connectivity into, and then create a private virtual interface to the Direct Connect gateway. You can associate up to 10 virtual private gateways (each attached to a VPC) in different AWS Regions, directly to a Direct Connect gateway. Alternatively, you can create a transit virtual interface and attach a total of three transit gateways (each attached to thousands of VPCs) across different AWS Regions to a Direct Connect gateway.

For standard use cases, we recommend starting with a transit virtual interface. However, if your data transfer volume is high, for example on-premises data backup to a VPC, or if you have 100 Gbps connections and want full 100 Gbps bandwidth to a VPC, we recommend using a private

virtual interface. Additionally, you can use a hybrid approach for multiple use cases, as showing in the following diagram.




*Global Direct Connect reference architecture – high resiliency*

# The pillars of the Well-Architected Framework

This section describes each of the pillars, and includes definitions, best practices, questions, considerations, and key AWS services that are relevant when building solutions for hybrid networking and is intended to be read as a companion to the AWS Well-Architected Framework whitepaper. When designing your hybrid networking architecture, we recommend that you read both papers and examine the full set of considerations.

**Pillars**

- Operational excellence pillar
- Security pillar
- Reliability pillar
- Performance efficiency pillar
- Cost optimization pillar
- Sustainability pillar

# Operational excellence pillar

The operational excellence pillar includes the ability to run and monitor systems to deliver business value, and to continually improve supporting processes and procedures. It provides an overview of design principles, best practices, and questions

## Best practices

There are three best practice areas for operational excellence in the cloud:

- Preparation
- Operation
- Evolution

To drive operational excellence in hybrid networking architectures, operations teams need to understand their business and customer needs so that they can effectively and efficiently support business outcomes. Operations creates and uses procedures to respond to operational events and validates their effectiveness to support business needs. Operations collects metrics that are used to measure the achievement of desired business outcomes. Everything continues to change—your

business context, business priorities, customer needs, and etc. It's important to design operations to support evolution over time in response to change and to incorporate lessons learned through their performance.

## Preparation

Preparation is essential to drive operational excellence in your hybrid networking environment. Many operational issues can be avoided by following best practices when designing the workload, and fixes are less expensive to implement in design phases rather than in production. It is important to assess the current state of your on-premises network and understand solutions for establishing your hybrid networking capabilities.

**HN_OPS1: How do you ensure efficient IP address allocation across your VPCs and on-premises networks?**

To prepare for operational excellence, you have to understand your hybrid workloads and their requirements. One key aspect is IP addressing. A well defined IP address allocation scheme has the following benefits:

- Enables efficient routing structure where you can summarize routes based on a network boundary. For example, if you were hosting workloads in VPCs in `us-east-1`, you can allocate CIDR ranges to these VPCs from a defined block like 10.1.1.0/22. You can then configure the Transit Gateway association a Direct Connect gateway to advertise this block over transit virtual interface instead of advertising individual prefixes associated with individual VPCs. Well summarized CIDR ranges can also help with security and firewall configuration when defining security groups and NACLs.
- Reduces the risk of over-lapping CIDR ranges between VPC and on-premises networks. You should avoid re-using the same CIDR range between your on-premises and Amazon VPC network, since overlapping CIDR ranges will make host to host communication very difficult.

We recommend that you keep track of the IP prefixes you currently have and allocate CIDR ranges for your deployment in a systematic manner. You can utilize one of many IPAM solutions available from the AWS marketplace. When provisioning VPCs, you need to ensure that you allocate right-sized CIDR ranges. If you overprovision your VPC CIDR ranges you will later face IP exhaustion

as you grow the number of VPCs. If you under provision the VPC CIDR range, you can associate secondary IPv4 CIDR blocks to your VPC, but with restrictions. It's important to understand your workload requirements such as their scalability patterns (and how scalability impacts IP usage), their reliability requirements (how many AZs the workload gets deployed in and how many IPs per AZ is utilized), and factor this information in when allocating IP ranges to the VPCs running these workloads. Its better to start with a conservative approach than overprovision the size of CIDR range allocation. In addition, ensure that VPC ranges don't overlap with on-premises IP ranges.

## Operation

It's important to define standards, procedures, and monitoring capabilities for your on-premises network environment that can provide you with real-time metrics important for your specific business needs. Aggregate these metrics, visualize them in a dashboard, and set automated alerts that can notify the operations team. In addition, develop a runbook that provides procedures for different alerts and alarms.

**HN_OPS2: How do you understand the health of your hybrid network?**

**HN_OPS3: How do you manage operational events?**

**AWS Direct Connect:** Enables you to monitor physical AWS Direct Connect connections using Amazon CloudWatch to collect and process raw data from AWS Direct Connect into readable, near real-time metrics. You can consolidate these metrics in CloudWatch, and build dashboards and alerts to notify your operations team based on the defined conditions.

AWS Direct Connect schedules planned maintenance and notifies you. To help you manage these events, you can leverage AWS Health Dashboard to display relevant information and provide proactive notifications so that you can plan for scheduled activities. We recommend using the AWS Health Dashboard to receive notifications for scheduled maintenance or events that will affect Direct Connect.

Your operations team should be prepared for unplanned outages with networking while connecting from on-premises to AWS. For example, to be prepared for an unplanned outage

like an AWS Direct Connect connection failure, you should establish a second Direct Connect connection. Traffic will fail over to the second link automatically if the BGP prefixes advertised are same over both connections. We recommend enabling Bidirectional Forwarding Detection (BFD) when configuring your connections to ensure fast detection and failover. Ensure that you test your high-availability design and configuration periodically using AWS Direct Connect Resiliency toolkit failover testing. Additionally, you can configure a back-up IPsec VPN connection, in which case all VPC traffic will fail over to the VPN connection automatically when direct connect connections fails. Traffic to and from public resources, such as Amazon Simple Storage Service (Amazon S3), can be routed over the internet if they were previously being routed over Direct Connect public virtual interface.

**AWS site-to-site VPN:** Enables you to monitor VPN tunnels using CloudWatch, leveraging near real-time metrics. You can monitor the state of your VPN tunnels and the data retrieved in/out of the tunnels. These metrics are recorded for 15 months, so you can access historical information and gain a better perspective on how your hybrid setup performed. VPN metric data is automatically sent to CloudWatch as it becomes available.

To ensure operational stability in case of failures, AWS VPN has built in high-availability. AWS Site-to-Site VPN connection has two tunnels, with each tunnel using a unique virtual private gateway public IP address. It is important to configure both tunnels for redundancy, if one tunnel becomes unavailable (for example, if it is down for maintenance), network traffic is automatically routed to the available tunnel for that specific Site-to-Site VPN connection. However, to protect against a loss of connectivity if your customer gateway becomes unavailable, you can set up a second Site-to-Site VPN connection to your VPC and virtual private gateway by using a second customer gateway. By using redundant Site-to-Site VPN connections and customer gateways, you can perform maintenance on one of your customer gateways while traffic continues to flow over the second customer gateway Site-to-Site VPN connection.

**AWS Transit Gateway:** Leveraging a transit gateway as a central hub enables access between your VPC resources and on-premises using AWS Direct connect or AWS VPN. AWS Transit Gateway provides statistics and logs that can be used by services such as Amazon CloudWatch and Amazon VPC Flow Logs. You can start by tracking health data and manage operations by building dashboards/alarms off transit gateway attachment level CloudWatch metrics. You can use Amazon CloudWatch to retrieve bandwidth usage between Amazon VPCs and a VPN connection, packet flow count, and packet drop count. Additionally, you can enable Amazon VPC Flow Logs on AWS Transit Gateway to capture information on the IP traffic routed through the AWS Transit Gateway.

**AWS Transit Gateway Network Manager:** Provides a single global view of your private network including hybrid connectivity. It enables you to see network activity in many locations from one single dashboard. It also includes the following data to help you monitor and troubleshoot the quality of your global network.

- **Events:** Describes changes in your global network. Transit Gateway Network Manager sends the following type of events to CloudWatch Events:
  - Topology changes: For example, an AWS Direct Connect gateway was attached to a transit gateway
  - Routing updates: For example, a VPN attachment's route table association changed
  - Status updates: For example, a VPN tunnel's BGP session went up (after being down)

For more information on tracking and getting notified of events relevant to your use case, refer to Transit Gateways User Guide.

- **Metrics:** Enables you to view CloudWatch metrics in your global network for your registered transited gateways, your associated Site-to-Site VPN connections, and your on-premises resources. You can view metrics per transit gateway and per transit gateway attachment, per global network. For more information, refer to Monitoring CloudWatch metrics.

- **Route Analyzer:** Enables you to perform an analysis of the routes in your transit gateway route tables in your global network. The Route Analyzer analyzes the routing path between a specified source and destination, and returns information about the connectivity between components. You can use the Route Analyzer to do the following:

- Verify that the transit gateway route table configuration will work as expected before you start sending traffic.
- Validate your existing route configuration.
- Diagnose route-related issues that are causing traffic disruption in your global network.

When building a hybrid network leveraging Transit Gateway, we recommend using Route Analyzer to verify and resolve network connectivity issues.

### Evolution

There are no operational practices unique to Hybrid lens for the **evolve** practice area, you can review the corresponding section in the AWS Well-Architected Framework whitepaper.

## Resources

Refer to the following resources to learn more about AWS best practices for operational excellence.

### Documents

- Amazon CloudWatch metrics for AWS Direct Connect
- Amazon CloudWatch metrics for AWS Site-to-site VPN

### AWS Support

- How can I get notifications for AWS Direct Connect Scheduled maintenance or events?
- How do I monitor AWS VPN tunnels using Amazon CloudWatch alarms?
- How do I check the current status of my VPN tunnel?
- How can I get notifications for AWS Direct Connect scheduled maintenance or events?
- How should I prepare for maintenance on my Direct Connect connection?

# Security pillar

The security pillar describes how to take advantage of cloud technologies to protect data, systems, and assets in a way that can improve your security posture.

## Best practices

There are five best practice areas for security in the cloud:

- Identity and access management
- Detective controls
- Infrastructure protection
- Data protection
- Incident response

Securing your hybrid network includes identifying security incidents, protecting your systems and services, and maintain the confidentiality and integrity of data through data protection. Unauthorized access to systems can cause financial loss and loss of compliance with regulatory obligations. Before creating a hybrid network, we recommend having a well-defined and practiced process for responding to security incidents.

The AWS Shared Responsibility Model enables organizations that adopt the cloud to achieve their security and compliance goals. Because AWS physically secures the infrastructure that supports our cloud services, AWS customers can focus on using services to accomplish their goals. The AWS also provides greater access to security data and an automated approach to responding to security events.

In this section, we provide principles that help you strengthen your system's security and hybrid networking workloads, it's expected that you will also be following the security best practices of the AWS Well-Architected Framework whitepaper.

## Identity and access management

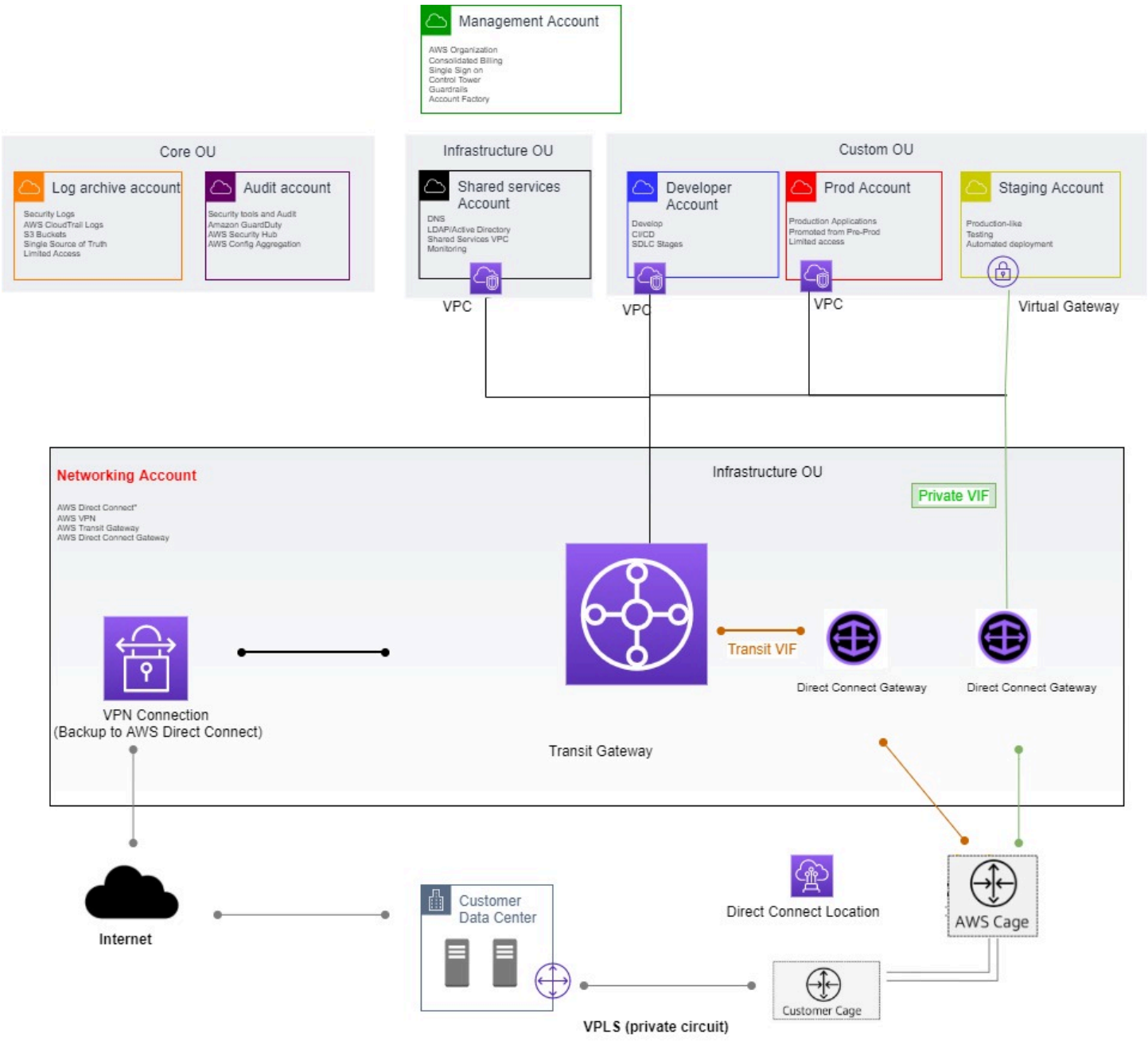
Deploying hybrid networking requires the creation of constructs that must be controlled to prevent unauthorized access to your VPCs and services. To secure and control access to your hybrid networking environment, it's important to consider the roles and responsibilities of your teams managing and operating your workloads using the principle of least privilege. Additionally, it's important to isolate your networking services and implement separation of duties between the network specialists and application owners. Regardless of your operating model, this will allow the different teams to have required access to network services based on their roles. For example, it's best practice to create a separate Central Networking cloud account managed by the networking specialist team to centrally manage network policies, gateways and internet-based VPN and dedicated Direct Connect connections. Separation of accounts will prevent your application teams from negatively impacting your hybrid network.

**HN_SEC1: How do you control access to your resources and workloads across your Hybrid networking environment?**

AWS recommends implementing a landing zone, which is a preconfigured, secure, scalable, and multi-account AWS environment based on best practice blueprints. AWS Control Tower can automate the provisioning of your landing zone and accounts and help you manage the level of

separation. AWS Control Tower provides an initial set of guardrails to help enhance the security of your overall AWS environment.

The following reference architecture diagram shows an example of a central Networking account which hosts all of the hybrid networking resources and enables demarcation of network administrative boundaries. There are additional AWS accounts owned by the various application teams for example, production account, staging account, and dev account, some are associated to the central networking account for hybrid connectivity. Hybrid networking constructs such as Direct Connect connections, Direct Connect gateway, Networking services, Shared VPCs, and Transit Gateway resources should be deployed in the Central Networking account. In order to share AWS Direct Connect connectivity with the rest of your Landing Zone, you can share the Transit Gateway through RAM with VPCs that reside in other accounts.




*Sample architecture for AWS hybrid network connectivity using Transit Gateway*

In a hybrid networking environment, the networking and security teams may share some of the responsibilities for securing network boundaries. To implement the principle of least privilege for example, the networking and/or security teams should have control over creating and modifying resources to enable hybrid connectivity while the developer teams should not have permissions to create or make changes to any network or security settings defined by the networking or security teams. The networking team should own the management of circuits as well as the provisioning of private dedicated Direct Connect virtual interfaces and/or internet-based Site-to-Site VPN

connections, even though the development teams have dependencies on the various shared networking resources.

Sensitive APIs for setting up hybrid connectivity for example, the creation and deletion of AWS resources such as Direct Connect virtual interfaces, Transit Gateway, Direct Connect Gateway, and Direct Connect Gateway associations should be restricted to the networking account and specialists. If most Hybrid Networking connections are established with these APIs from specific networks or geographic areas, you should restrict access to hybrid networking connections based on location where possible. For example, AWS Direct Connect APIs which support resource-based access policies based on the CIDR range or source IP address. This can isolate network access to a given Direct Connect resource from only the specific VPC within the central networking account. For more information, refer to AWS Direct Connect resource-based policy examples. Additionally, the networking teams should own all of the API calls for creation and deletion of connections while the application teams should be allowed to use describe calls for APIs on the hybrid network resources. Direct Connect supports specific actions, resources, and condition keys. To learn about all of the elements that you use in a JSON policy, refer to IAM JSON Policy Elements Reference in the *IAM User Guide*.

We also recommend that the networking team tag your AWS hybrid network resources upon creation, for example, Site-to-Site VPN connections, virtual private gateways, and customer gateways from the central networking account. You can enforce the use of tagging and control which tag keys and values are set on networking resources. For example, you can tag a Direct Connect connection based on the different business units or AWS account and limit control over who can create a private or transit VIF on that connection. As shown in the previous reference architecture diagram, if Direct Connect is connected to a Transit Gateway, you should restrict control to the networking team to allow changes to the Transit Gateway route table association and propagation, failing to do so could potentially give other teams access to hybrid connectivity.

For accounts with multiple VPCs that need to be shared with two or more accounts, VPC sharing can be used to improve security and compliance of your different accounts. With VPC sharing, the AWS account that creates and owns a VPC can choose to share particular subnets with other AWS accounts, and that account can then create, view, and modify resources it owns within those particular subnets. For example, using VPC sharing, you won't have to write complicated IAM policies to prevent developers from altering the VPC that connects to your on-premises network or interact with resources with which they are not associated. For more information, refer to working with shared VPCs.

**HN_SEC2: How do you segment access between AWS and your on-premises network?**

It's also important to segment access between your on-premises network and AWS networks. Your networking team should verify that your customer gateway router and firewall configurations are aligned with how you expect to separate traffic between your on-premises and AWS environments. For example, you might need to constrain your production VPC to accessing only allowed on-premises production services and data. Additionally, you might need to configure your on-premises routers to help ensure that only certain on-premises networks or specific clients can access your networks. For more information, refer to the Infrastructure Protection section of the Hybrid Networking Lens Security pillar for ways to constrain access in your hybrid networking environment.

It is common that your application owners in your development, staging, and production environments in AWS will need to have connectivity to some of your on-premises resources such as Active Directory services, DNS, or network security proxying services. It is also possible that your on-premises users and workloads have dependencies on AWS workloads and data. For example, there may be existing test and production data services in your on-premises environment that your workloads in AWS will need to access. It is important that access to the AWS platform be controlled via role-based access control and permissions assigned to roles with AWS Identity and Access Management (IAM).

## Detection controls

Detective controls can be used to identify a potential security threat or incident. You can get detailed insights into your hybrid network performance and use that information to detect misconfigurations or potential malicious activity, and further optimize your deployment.

**HN_SEC3: How are you capturing and analyzing metrics in your hybrid networking environment?**

A recommended best practice is to monitor and implement an immediate response process that detects and reacts to any suspicious or malicious activity. Monitoring workloads is important especially when investigating a security incident. At a minimum, the metadata of logs should be

captured for hybrid network connections with private connections like AWS Direct Connect for 1GB and higher. You can leverage Amazon GuardDuty, a threat detection service with built in VPC flow logs that continuously monitors your workloads for malicious activity.

We also recommended having a central logging and analytics setup for your hybrid environment. With AWS Hybrid Networking, you can implement detective controls using Amazon CloudWatch Logs, Amazon CloudWatch metrics, and Transit Gateway Route Analyzer. For an example, refer to AWS Central Logging and Analytics architecture example for Hybrid Networking.

For dedicated Direct Connect deployments with multiple virtual interfaces, you can leverage CloudWatch metric math to configure specific ingress and egress metrics and send a CloudWatch alarm for all virtual interfaces if the threshold for a metric is breached.

If you deploy AWS Transit Gateway for dedicated Direct Connect or internet based Site-to-Site VPN environments, you can gain focused insights into the amount of data flowing in and out over a transit gateway connection using CloudWatch metrics. Verify routes in your hybrid networking environment for traffic reachability to prevent faulty route configurations that could lead to the exposure of sensitive environments on AWS to untrusted networks on-premises and on AWS. The Transit Gateway Network Manager dashboard can be used to visualize and monitor your AWS resources and on-premises networks, and can help you identify whether issues in your hybrid network are caused by AWS resources, on-premises resources, or the connections between resources such as network topology changes, routing updates, and connection status. Using the Route Analyzer feature of Transit Gateway, you can validate that the AWS Transit Gateway route table configurations work as expected before sending live network traffic over Site-to-Site VPN connections or dedicated AWS Direct Connect.

Hybrid network connectivity leverages IP prefixes which network administrators should monitor to ensure that the number of prefixes injected from customer gateways such as routers or firewalls stay within the allowed limits. Staying within the allowed limits ensures that you avoid resource starvation of your AWS hybrid networking services and also protects them from abuse.

For best practices in detection controls for hybrid networking, refer to the AWS Well-Architected Security pillar whitepaper.

## Infrastructure protection

Infrastructure protection for hybrid networking involves securing all networking resources from your on-premises deployment to the cloud. Enforcing boundary protection, monitoring points of ingress and egress, and comprehensive logging, monitoring, and alerting are all essential to an effective information security plan.

**HN_SEC4: How do you protect network resources in your hybrid network environment?**

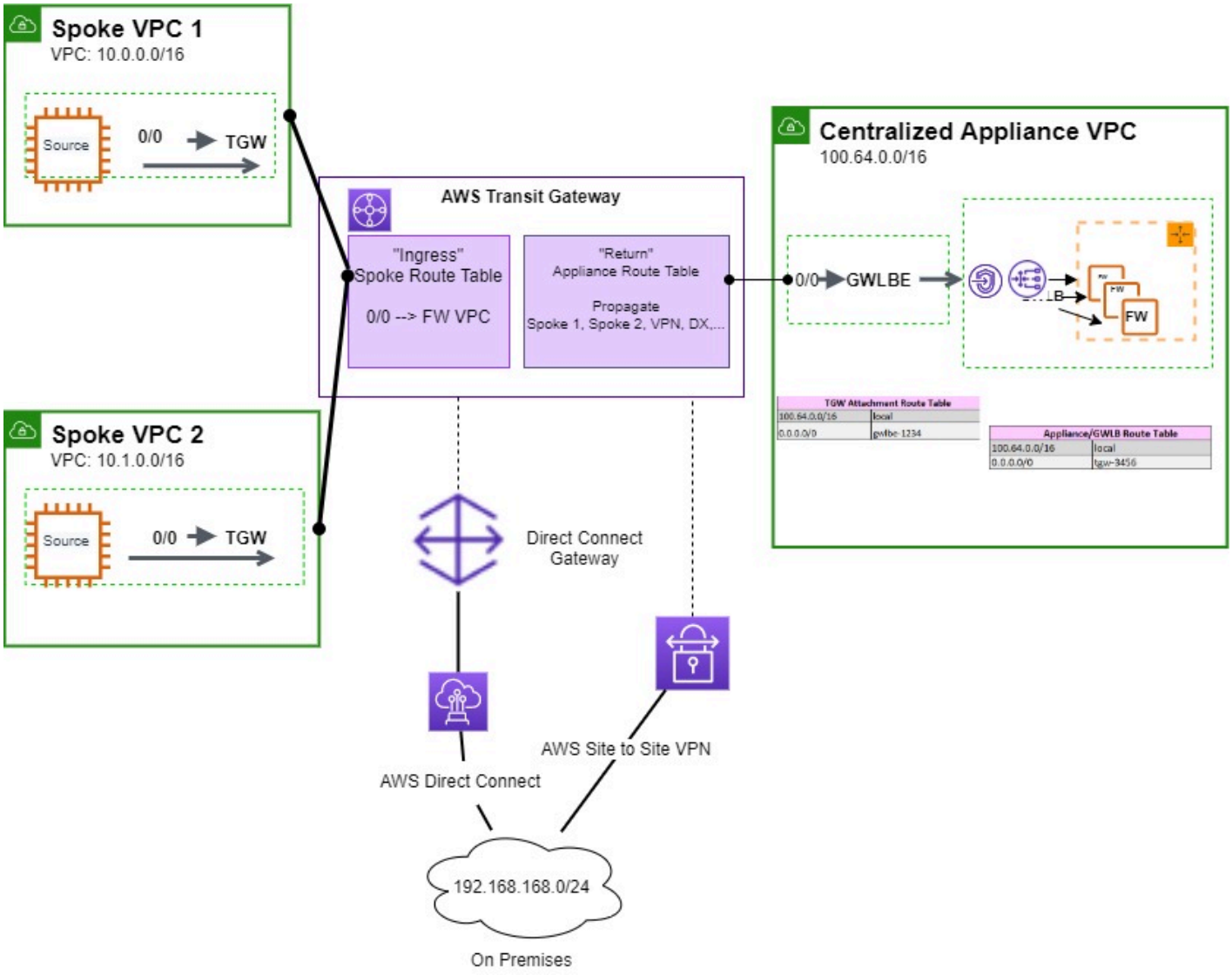
It's important to ensure that networking boundaries are considered and enforced for hybrid networking environments. Secure your hybrid environment using multiple layers of defense and segmentation. There are five key methods to consider when protecting your AWS hybrid networking boundaries:

- **Security Groups**: Utilize security groups as a stateful (layer 4) firewall to allow access to instances in your VPC from your on-premises network and as a first line of defense. When defining security group rules for your AWS Direct connect virtual interface or Site-to-Site VPN, ensure that you are only allowing inbound and outbound traffic for your on-premises network prefix. For example, refer to Security Group Rules. Consider using other security groups as sources for a security group rules instead of configuring multiple CIDRs.
- **Network Access Control Lists** (NACLs): You can use NACLs as an optional stateless (layer 4) firewall that allows defining the port, protocol, and source of traffic that should be explicitly denied at the subnet level. Security groups and NACLs mutually complement each other, consider using NACLs with rules similar to your security groups in order to add an additional layer of security to your VPC. When using NACLs, ensure that the outbound rules that allow traffic from all ports limit access to the required ports or port ranges across your hybrid network. For more information, refer to Network ACL Rules.
- **AWS Transit Gateway Route Tables**: Transit Gateway route tables can be used to enable defined connectivity between AWS VPCs and your on-premises network. Based on the configuration of the transit gateway routing tables, you have control over which VPCs have connectivity with each other and with your on-premises network. Transit Gateway route tables can also use a routing mechanism known as *null routing*, which drops traffic that matches a particular route and can be used to achieve security isolation and optimal traffic flow. It prevents the source attachment from reaching a specific route by dropping traffic that matches the route. For example, by configuring a null route in your Transit Gateway route table for a particular destination, you can block hybrid traffic for that destination to flow from the VPC spokes via the VPN attachment and/or Direct Connect GW attachment to your on-premises Data Center.

  Avoid using a single transit gateway route table per VPC as a security feature. To keep your hybrid environment easy to manage and to stay within the transit gateway route tables limits, keep the number of route tables to a minimum by associating VPCs with the same routing behavior to the same transit gateway route table. For example, you can group your development

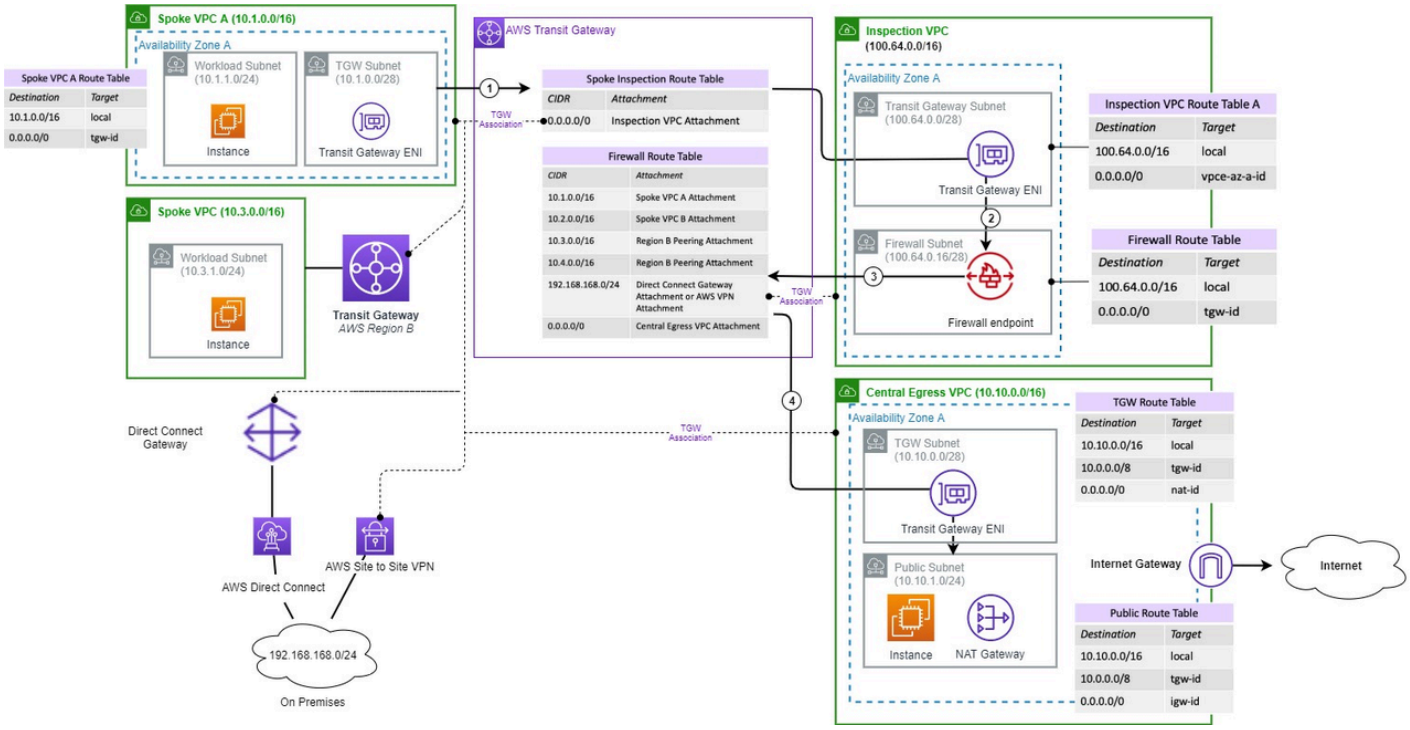
VPCs and associate them to a development transit gateway route table while the production VPCs can be associated to a production transit gateway route table. The development and production route tables can have different routes to your hybrid environments based on your security posture. For more information, refer to working with Transit Gateway Route Tables.

- **Gateway Load Balancer** (GWLB): Enables easy configuration for adding third- party software appliance/IPS/IDS/firewalls running on Amazon Elastic Compute Cloud (Amazon EC2) instances in your AWS network. You can route all internet traffic using only private subnets in a VPC, using an IDS/IPS leveraging transit gateway and establishing centrally managed egress/ingress security capabilities in AWS. If you need a firewall for inline inspection of traffic, out of multiple VPCs over Direct Connect or VPN, you can use Transit Gateway with a centralized appliance VPC model for the firewall. Transit Gateway helps control separation of duties between accounts that perform the inline functionality. For more details on deployment options inline inspection with the AWS Gateway Load Balancer, refer to this blog. The following diagram depicts a Gateway Load Balancer deployment for North/South inline inspection of traffic.




*Firewall inline inspection for VPCs with centralized North/South connectivity*

- **AWS Network Firewall**: AWS Network Firewall secures AWS Direct Connect and VPN traffic from client devices and on-premises environments for deployments supported by AWS Transit Gateway. With AWS Network Firewall, you can filter traffic at the perimeter of your VPC(s) for VPN or Direct Connect. A key requirement for this support is to connect AWS Direct Connect using Transit virtual interfaces to AWS Transit Gateway or establishing the AWS Site-to-Site VPN directly to AWS Transit Gateway. Using AWS Transit Gateway routing tables functionality, AWS Direct Connect/VPNs are attached to the spoke route table. For more information on deployment options for the AWS Network Firewall, refer to this blog. The following diagram depicts a Network Firewall deployment for a hybrid networking environment.




*Traffic between VPC and on-premises protected by centrally deployed AWS Network Firewall*

When VPC security groups, Network ACLs, and route tables are used in AWS, use managed prefix lists to provide consistency for a list of external prefixes to be shared via VPC sharing and IAM access control.

- **Amazon Route 53 Resolver DNS Firewall**: Helps to protect against DNS-level threats such as data exfiltration attempts where malicious actors can use DNS queries to smuggle sensitive data out of your hybrid network. With Amazon Route 53 Resolver DNS Firewall, you can create domain name blocklists for domains that you don't want your VPC resources to communicate with via DNS and also create domain name allow lists that permit outbound DNS queries only to domains your organization trust and specify.

  >For best practices in the Infrastructure Protection area for security in hybrid networking, refer to the AWS Well-Architected Security pillar whitepaper.

## Data Protection

Encrypting sensitive data traffic to connect to AWS over the internet or over a private network connection for their hybrid networking workloads, is important in to ensure that an unauthorized person or entity is unable to gain access to your data.

**HN_SEC5: How will you provide support for encryption of customer data?**

For hybrid network connectivity over the internet, AWS Site-to-Site VPN can be used to create encrypted tunnels using IPSec VPN.

For hybrid network connectivity over a private network connection using AWS Direct, configure MACsec (802.1ae) encryption (Layer 2) for dedicated 10Gbps and 100Gbps connections to encrypt data across your hybrid network. Encrypting traffic using MACsec enables you to securely pass high bandwidth workloads with native, point-to-point encryption, ensuring that data communications between AWS and your data center, office or colocation facility remain protected. To enable MACsec, both dedicated connection and on-premises resources must support MACsec.

For hybrid network connectivity using AWS Direct Connect for hosted connections and speeds lower than 10Gbps, use application-level encryption or VPN to secure your sensitive data. Encrypt at the application layer (Layer 7) using TLS. For network layer (L3) encryption, establish an AWS Site-to-Site VPN connection to create an IPSec VPN over a Direct Connect public virtual interface. You can also create an IPSec VPN over your Direct Connect private virtual interface using VPN or firewall software from the AWS Marketplace on EC2 instances. Leverage certificates for authentication where available for Data Protection.

For additional considerations and solutions for data protection in hybrid networking, refer to the Security section of the Hybrid Connectivity AWS Whitepaper.

For best practices in data protection area for security in Hybrid Networking, refer to the AWS Well-Architected Framework whitepaper .

## Incident Response

Security incidents such as ransomware attacks on a hybrid networking environment can be harmful to any organization with potential loss of critical data and reputational damage to the organization's brand.

**HN_SEC6: How do you isolate a hybrid networking environment from a security incident that originates from your on-premises network?**

Responding to any cyber incident requires that you're able to detect the threat's existence and establish a baseline for what normal looks like in an AWS hybrid environment. To help identify threats, Amazon GuardDuty continuously monitors your AWS environment for malicious behavior, with threat detected from multiple sources such as VPC Flow Logs, API activity, and DNS logs to help protect your AWS accounts and workloads.

Some ransomware incidents are designed to use a company account to perform the attack. Always follow the principle of least privilege to provision access for all of your users. As a quick containment measure for your users across the hybrid network, deny access to IAM principals (users and roles) with privileged access across accounts in your AWS Organization using Service Control Policies (SCPs) until a thorough investigation is done or the malware attack is over.

If ransomware is spreading, quickly isolate your AWS hybrid network environment by preventing any inbound traffic from your on-premises environment to your AWS accounts. To block all incoming and outgoing traffic into your subnets and significantly diminish attackers from moving laterally within your AWS network, implement Network Access Control Lists. Additionally, you can also use an Intrusion Prevention System (IPS) and Intrusion Detection system (IDS) such as the AWS Network Firewall to block communication with known malware hosts and secure AWS Direct Connect and AWS VPN traffic running through the AWS Transit Gateway from client devices and your on-prem environment. For additional ways to prevent traffic across your hybrid network environment, refer to the Infrastructure Protection section in this whitepaper.

Automate incident response rather than leveraging manual processes to monitor your security posture and manually react to events. Automating responses improve manual processes, reduce containment time, and prevent alert fatigue by the incident response teams, leaving your human processes to handle the sensitive and unique incidents. Leverage AWS Security Hub to automate and detect security incidents. Security Hub continuously monitors your environment using automated checks and you can take action on the security findings with event based automation tools such as AWS Lambda, AWS Step Functions, and AWS Config Rules. The automated response approach should be tested in your non-production environment before deployment in your production accounts.

For additional best practices in the Incident Response area for security in hybrid networking, refer to the AWS Well-Architected Security pillar whitepaper.

## Resources

Refer to the following resources to learn more about our best practices for security.

### Documentation and Blogs

- Logging and Monitoring AWS Site-to-site VPN
- IAM for AWS Site-to-site VPN
- Data protection in AWS Site-to-site VPN
- Monitoring AWS Direct Connect
- IAM for AWS Direct Connect
- Data protection in AWS Direct Connect
- Deployment models for AWS Network Firewall
- Hybrid Connectivity Whitepaper

### AWS Support

- How do I create a certificate-based VPN using AWS Site-to-Site VPN?
- How do I establish an AWS VPN over an AWS Direct Connect connection?

# Reliability pillar

The reliability pillar encompasses the ability of a system to recover from infrastructure or service disruptions, dynamically acquire computing resources to meet demand, and mitigate disruptions such as misconfigurations or transient network issues.

## Best Practices

There are three best practice areas for reliability in the cloud:

- Foundations
- Change management
- Failure management

To achieve reliability, a system must have a well-planned foundation and monitoring in place, with mechanisms for handling changes in demand or requirements. The system should be designed to detect failure and automatically heal itself.

# Foundations

**HN_REL1: How do you manage AWS service quotas for hybrid networking services?**

AWS sets service quotas (also called service limits) to protect you from accidentally over-provisioning resources. A service quota is an upper limit on the number of each resource your team can request. You will need to have governance and processes in place to monitor and change these quotas to meet your business needs. As you adopt the cloud, plan integration with existing on-premises resources (a hybrid approach). A hybrid model allows the gradual transition to an all-in cloud approach over time. Therefore, it's important to have a design for how your AWS and on-premises resources will interact as a network topology.

If you are using AWS Direct Connect, there are quotas on the amount of data that can be transferred on each connection. Currently, you can have a dedicated connection of 1Gbps, 10Gbps, or 100 Gbps bandwidth and if you need more bandwidth you can order Link Aggregation Groups ( LAG ) of two 100 Gbps, totaling a total of 200 Gbps aggregate bandwidth. If you are using an AWS Site-to-site VPN connection to access resources in a VPC, then you are cumulatively bound by the virtual gateway throughput of 1.25 Gbps.

Service quotas need to be increased from the default values to handle the requirements of a large deployment per your business needs. For supported services, you can proactively manage your quotas by configuring Amazon CloudWatch alarms that monitor usage and alert you to approaching quotas by accessing the Service Quotas in the AWS Management Console. Contact AWS Support to request an increase for services not currently supported by Service Quotas.

You can also proactively raise quotas if you anticipate exceeding them in your workloads. When raising these quotas, ensure that there is a sufficient gap between your service quota and your maximum usage to accommodate scale.

## Change Management

Being aware of how change affects a system enables you to plan proactively, and monitoring enables you to quickly identify trends that could lead to capacity issues or SLA breaches.

**HN_REL2: How do you prepare for AWS Direct Connect scheduled maintenance or events?**

**HN_REL3: How do you regulate bandwidth usage for Direct Connect connections and executing changes?**

**HN_REL4: How do you monitor your Direct Connect connections and Site-to-Site VPN?**

Logs and metrics are a powerful tool to gain insight into the health of your workloads. Configure your workload to monitor CloudWatch logs and metrics to send notifications when thresholds are crossed or significant events occur. For example, with AWS Direct Connect, `ConnectionBpsIngress`, `ConnectionBpsEgress`, `ConnectionPpsEgress`, and `ConnectionPpsIngress` metrics help track the connection capacity utilization. For a comprehensive list of metrics, refer to the AWS Direct Connect User Guide. For information about VPN metrics, refer to the AWS VPN User Guide.

When an AWS Direct Connect connection is down for maintenance, that connection can be down from a few minutes to a few hours based on the level of maintenance required. To prepare for this downtime, consider one or more of the following options:

- Request a redundant Direct Connect connection.
- Configure a virtual private network (VPN) connection as a backup.

Monitor the bandwidth usage on the Direct Connect connection and increase bandwidth for active traffic on a Direct Connect connection by ordering more Direct Connect connections and aggregating them to form a Link Aggregation Group (LAG). To increase the bandwidth with minimal downtime using LAG and migrating virtual interface from a single Direct Connect connection to LAG, refer to Knowledge Center.

## Failure management

**HN_REL5: How does your system withstand component failures?**

### HN_REL6: How are you testing for resiliency?

### HN_REL7: How are you planning for disaster recovery?

In any system of reasonable complexity, it's expected that failures will occur. Know how to become aware of these failures and respond to them automatically, to ensure your network can withstand the failures and not affect the existing workload over it.

Highly resilient network connections are key to a well-architected system. AWS recommends connecting from multiple data centers for physical location redundancy. For more information, refer to AWS Direct Connect resiliency recommendations. When designing remote connections, consider using redundant hardware and telecommunications providers. Your telecommunication provider should have a diverse fiber and path connectivity for your last mile connection or offer any SLA. Ensure that the physical infrastructure configuration you set up is in accordance with the requirements to meet AWS Direct Connect SLA and AWS VPN SLA. Additionally, use dynamically routed, Active/Active connections for automatic load balancing and failover across redundant network connections. Provision sufficient network capacity to ensure that the failure of one network connection does not overwhelm and degrade redundant connections.

Each Site-to-Site VPN connection has two tunnels, with each tunnel using a unique virtual private gateway public IP address. It's important to configure both tunnels for redundancy by preferably using dynamic routing, Active/Active setup. When one tunnel becomes unavailable (for example, is down for maintenance or unplanned outage), network traffic is automatically routed to the available tunnel for that specific Site-to-Site VPN connection.

Test the Direct Connect failover scenarios to help you find any latent bugs that could surface in production. Exercise these tests regularly to ensure that your configurations are appropriate for failovers and verify the impact on workload during these tests. These tests help in validating your recovery procedures. You can use the Resiliency Toolkit - Failover Testing feature to test the resiliency of the Direct Connect connections. The failover testing feature allows you to test resiliency by disabling one or more Border Gateway Protocol (BGP) sessions using the AWS Management Console, Command Line Interface, or AWS Direct Connect API. It allows you to shut down BGP sessions for a configurable time period. You can also cancel failover tests at any time during the testing period to return to the pre-test configuration. Alternatively, use automation

to simulate different failures or to recreate scenarios that led to failures before. This exposes failure pathways that you can test and rectify before a real failure scenario, reducing the risk of components failing that have not been tested before.

## Resources

Refer to the following resources to learn more about our best practices related to reliability.

### Documents

- Maximizing resiliency with AWS Direct Connect

### AWS Support

- How can I get notifications for AWS Direct Connect scheduled maintenance or events?
- How should I prepare for maintenance on my Direct Connect connection?

# Performance efficiency pillar

The performance efficiency pillar focuses on the efficient use of computing resources to meet requirements and maintain efficiency as demand changes and technologies evolve.

## Best Practices

There are four best practice areas for Performance Efficiency in the cloud:

- Selection
- Review
- Monitoring
- Tradeoffs

Use a data-driven approach to select a high-performance architecture. Gather data on all aspects of the architecture, from the high-level design to the selection and configuration of resource types. Review your choices on a cyclical basis to ensure that you are taking advantage of the continually evolving AWS platform. Monitor your workload to ensure that you are aware of any deviance from expected performance. Understand where you can make architecture tradeoffs to improve performance, such using VPN over internet vs dedicated circuits via AWS Direct Connect for your

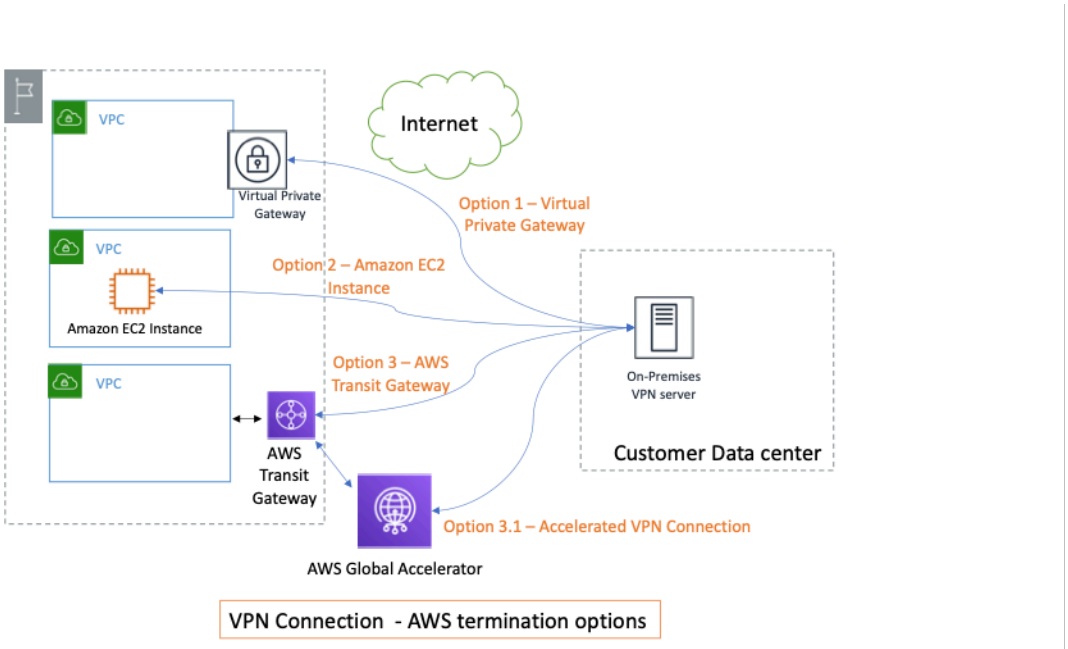
hybrid connectivity or terminating your hybrid connectivity on Virtual Private gateway instead of Transit Gateway.

## Selection

There are multiple technology and design choices to consider when setting up hybrid networking connectivity on AWS. Each option has its own performance characteristics and considerations. Understand your performance requirements to make the right selection.

These are our general recommendations:

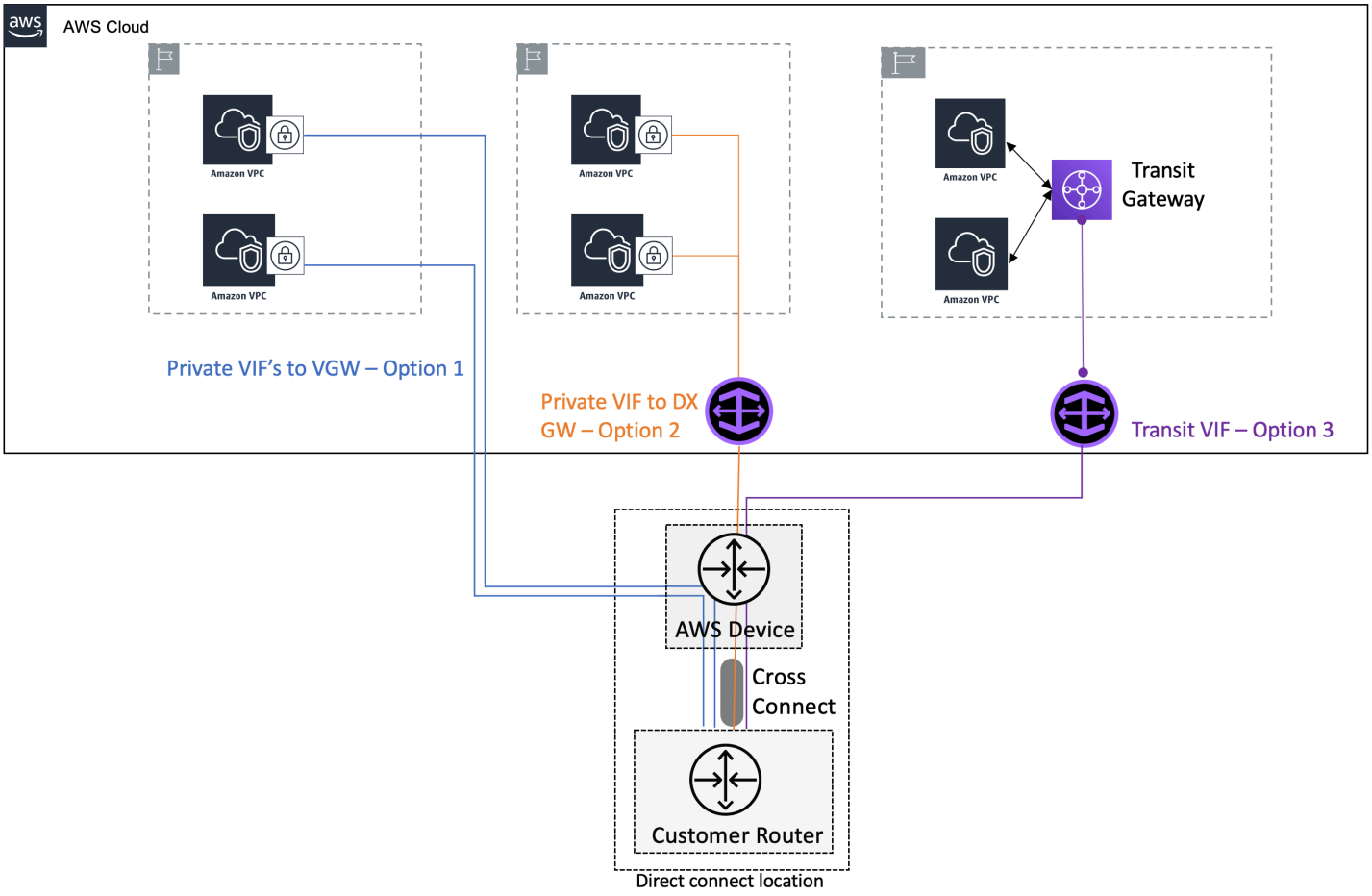
1. Choose VPN (encrypted tunnels over the internet), AWS Direct Connect (dedicated fiber connectivity), or both.
2. Select the right termination endpoints in AWS.
   - VPN options include AWS Transit Gateway, a customer-managed EC2 instance, and a virtual private gateway. You can optionally enable acceleration for your Site-to-Site VPN connection to AWS Transit Gateway.




*VPN connectivity options*

- AWS Direct Connect options include virtual private gateway, Direct Connect gateway and virtual private gateway, or Direct Connect gateway and AWS Transit Gateway. Select Direct Connect locations where you perform a standard cross-connect between customer/service

provider router and AWS device. For a list of Direct Connect locations, refer to AWS Direct Connect Locations.




*Direct Connect connectivity options*

Based on your bandwidth requirements, a single VPN or Direct Connect connection might not be sufficient and you will have to architect the hybrid networking setup to enable traffic load balancing across multiple connections.

**HN_PERF1: How do you decide between AWS Direct Connect and AWS VPN as your connectivity option?**

**HN_PERF2: Determine and define your performance requirements using bandwidth, latency and jitter values.**

Before you design the best performing architecture, define what performance means for you and the parameters involved. Typically, performance metrics are based around bandwidth (rate of data transfer), latency (round trip time for a network packet to travel form source to destination), and jitter (variation in latency). Start by estimating the bandwidth and latency requirements of your hybrid networking applications. For existing apps that are moving to AWS, a good way to get these estimates is to rely on data from your internal monitoring systems. For new apps or existing apps for which you don't have monitoring data, talk to the app/product owners to understand what traffic load is expected on the system, what network dependencies does the application have and what are the acceptable latency numbers to ensure good customer experience. Match these estimates with the options available from AWS to determine which technology you should choose, and the appropriate configuration.

**Deciding between AWS Direct Connect and AWS VPN as your connectivity option**

Based on your requirements (bandwidth, latency, jitter), you can either choose to establish VPN connectivity using AWS VPN or AWS Direct Connect (or both). The following information will help you guide the path to take.

*Table 1 - Guided path for deciding between AWS VPN and AWS Direct Connect*

| Characteristic | AWS VPN | AWS Direct Connect |
| --- | --- | --- |
| Bandwidth | Low-Medium: Depends on customer internet connection and VPN device constraints. At the AWS end, you can scale the VPN bandwidth by creating multiple VPN connections | High: Each Direct Connect connection can be up to 100 Gbps with option to scale bandwidth by adding additional connections. |

| Characteristic | AWS VPN | AWS Direct Connect |
| --- | --- | --- |
| Latency | Medium-high: Traffic traverses the internet and can flow through multiple hops.<br><br>Note: Leveraging accelerated VPN could result in lower latency. | Low-Medium: Traffic traverses private circuit between AWS, third-party cloud provider, customer and has the minimal number of hops. In certain circumstances (outside customers control) like when there is a failure, higher latency can be possible. |
| Jitter | Medium: It's hard to predict the number of hops the traffic traverses over the internet and the congestion (or lack thereof) at these hops.<br><br>Note: Leveraging accelerated VPN could result in lower jitter. | Low-medium: Traffic traverses private circuit and does not have the same uncertainties of internet. In certain rare circumstances (outside customers control) like when there is failure, higher jitter can be possible. |

As depicted in the previous table, the best performing and recommended option is AWS Direct Connect for production workloads. To get started quickly while getting good performance for your development/sandbox environments, choose AWS VPN. It is very important to always work backwards for your use case and requirements to make the right technology choice. You can also choose a hybrid design where they leverage both AWS VPN and AWS Direct Connect.

**HN_PERF3: How do you select the best performing hybrid VPN architecture?**

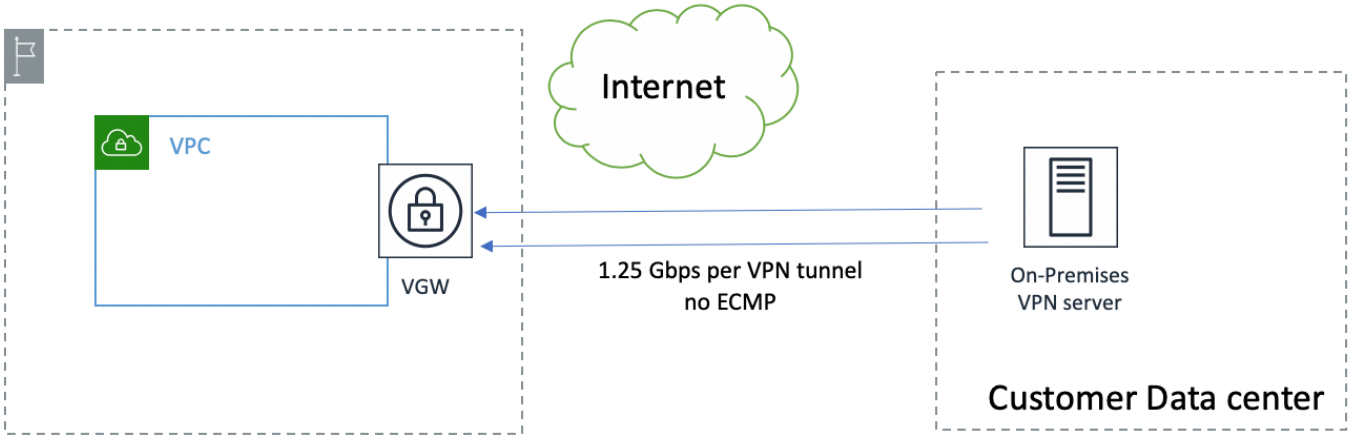
**Selecting the right VPN termination endpoint at the AWS end:**

There are four termination options at the AWS end. The VPN performance and scalability will vary based on which option you choose. For each option it's important to understand the bandwidth and scalability characteristics.

1. **Termination at the virtual private gateway**:

   *Bandwidth* - Up to 1.25 Gbps per VPN connection.

   *Scalability* - Load balancing traffic across multiple VPN connections for a given prefix is not enabled on the virtual private gateway. Since the gateway can only attach to a single VPC, you can get up to 1.25 Gbps per VPC for a given on-premises prefix you advertise.

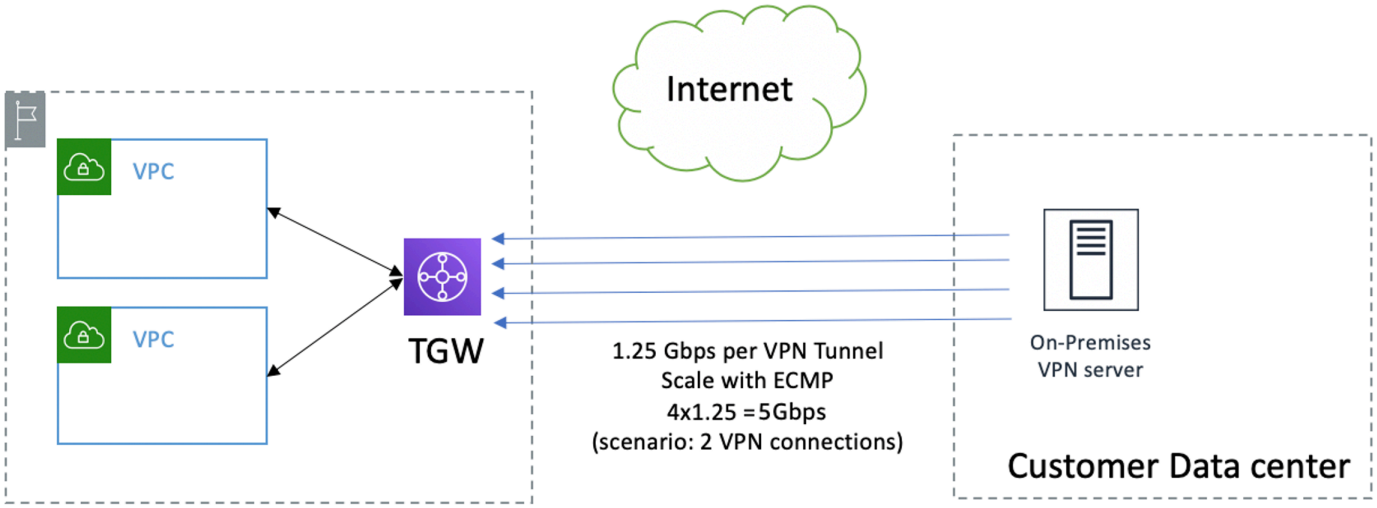



   *Termination at virtual gateway*

2. **Termination on AWS Transit Gateway**:

   *Bandwidth* - Up to 1.25 Gbps per VPN tunnel. Each VPN attachment allows you to create two VPN tunnels, with BGP ECMP load balancing enabled you can get a total of 2.5 Gbps per VPN attachment.

   *Scalability* - Create multiple VPN attachments and scale the aggregate bandwidth by load balancing traffic across these VPN connections. To make sure that ECMP is active, the BGP path cost of all the links should be the same that is, AS PATH length should be the same and you should have no or same MED values tagged to a given prefix across all connections. The bandwidth you get is aggregate for all the VPCs that AWS Transit Gateway attaches to.




*Termination at Transit Gateway*

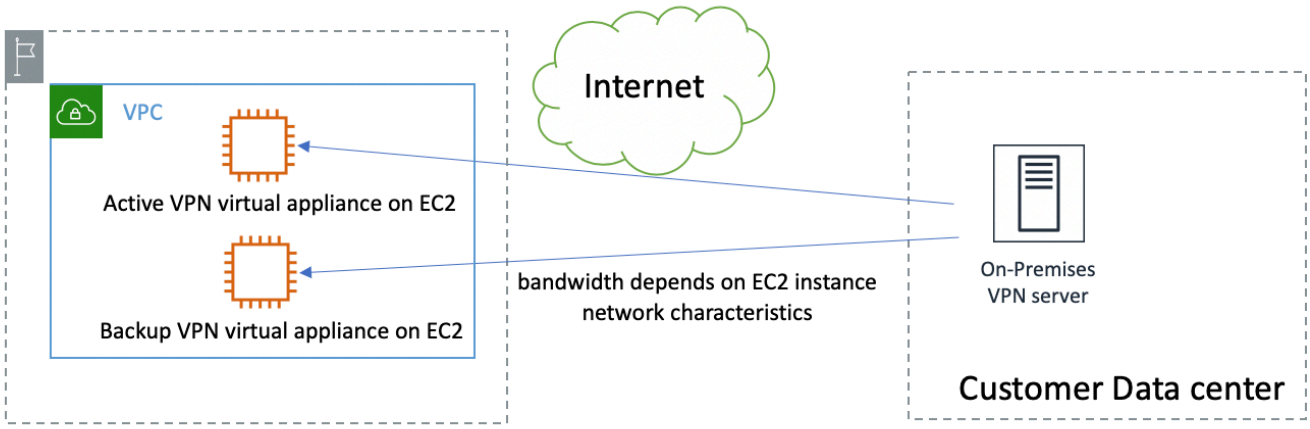
Boost performance with accelerated Site-to-Site VPN – Optionally, you can enable acceleration for your Site-to-Site VPN connection. An accelerated Site-to-Site VPN connection (accelerated VPN connection) uses AWS Global Accelerator to route traffic from your on-premises network to an AWS edge location that is closest to your customer gateway device. AWS Global Accelerator optimizes the network path, using the congestion-free AWS global network to route traffic to the endpoint that provides the best application performance. You can use an accelerated VPN connection to avoid network disruptions that might occur when traffic is routed over the public internet. VPN tunnel bandwidth remains same as when terminating directly on AWS Transit Gateway, but you get better latency, jitter, and overall performance with an accelerated VPN connection.

3. **Termination on a customer-managed EC2 instance running virtual VPN appliance**:

   *Bandwidth* - The bandwidth is dependent on the type/size of the EC2 instance and the capabilities of the VPN software running on the EC2 instance. The maximum network throughput to the internet from an EC2 instance varies based on EC2 instance type and size. Instance flow limits (10 Gbps within a placement group and 5 Gbps otherwise) should be considered as well.

   *Scalability* – You can scale the number of EC2 instances and create multiple VPN tunnels to these virtual appliances. Within a VPC route table you can only have single ENI as next hop for a destination prefix, in order to distribute load across EC2 instances you need to designate

different EC2 instances as next hop targets for different prefixes. You are responsible for the overall management of the EC2 instances and ensuring availability.



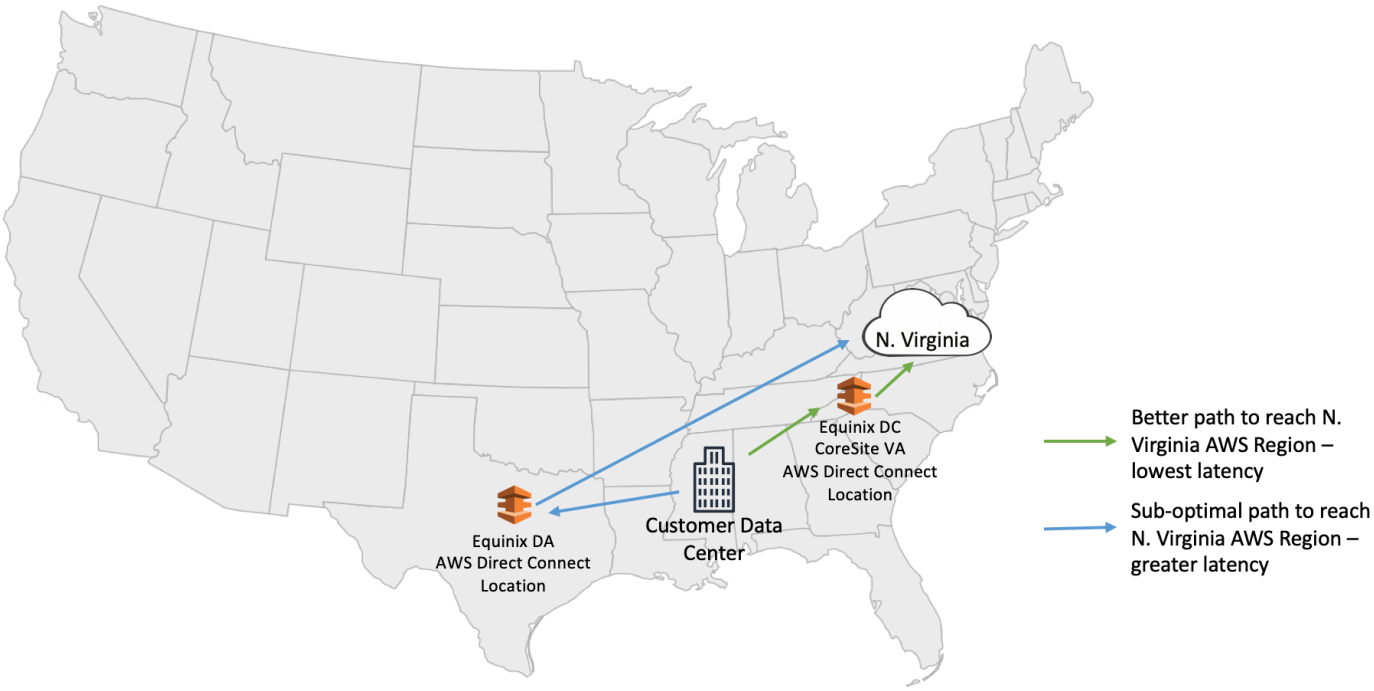

*Termination on an EC2 instance*

**HN_PERF4: How do you select the best performing hybrid architecture leveraging AWS Direct Connect?**

**Choosing the right AWS Direct Connect location:**

Getting an AWS direct connect connection is to decide on a direct connect location where you will establish a cross connect with AWS. You can connect to AWS at any of the direct connect locations and get access to all AWS regions globally (except China). A key factor in deciding on which direct location to choose is latency. The latency you get when choosing Direct Connect as your hybrid connectivity option is dependent on two factors – the distance between your data center and the Direct Connect location where you connect into and the distance between the Direct Connect location and the AWS Region you are connecting into.

When deciding on a Direct Connect location, minimize the combined latency of these two factors. Latency is directly proportional to geographical distance and hence you should choose locations that minimize the overall distance. As you choose multiple Direct Connect locations for High-availability it's key to choose direct connect locations that are geographically apart; striking a balance between and latency and high-availability is important.




*Choosing the correct AWS Direct Connect location*

**Choosing the right termination endpoint on AWS end:**

Once you have established cross connect to an AWS device at an AWS Direct Connect location, there are two options on how you can access VPC resources within an AWS Region.

3. Private virtual interface to AWS Direct Connect Gateway which is associated to a virtual private gateway.
4. Transit virtual interface to AWS Direct Connect Gateway which is associated to a virtual private gateway.

Both options offer high bandwidth, scalability, and low latency. However, when using transit virtual interface to connect to AWS Transit Gateway, you are limited by the bandwidth of the AWS Transit Gateway attachment (up to 50 Gbps). Of the two options, we recommended using transit virtual interface as it enables a hub and spoke topology which scales better and is easier to manage, unless you have a requirement for bandwidth speeds in the range of hundreds of Gbps.

**Choosing the right AWS infrastructure location for deploying your workloads:**

While it has been assumed throughout this document that your infrastructure will be deployed in an AWS region, for workloads that require very low-latency or local data processing, you can bring AWS infrastructure closer to you by leveraging AWS Local Zones.

AWS Local Zones are a new type of AWS infrastructure designed to run workloads that require single-digit millisecond latency, such as video rendering and graphics intensive, and virtual desktop applications. AWS Local Zones have their own connection to the internet and support AWS Direct Connect, so resources created in the Local Zone can serve local end-users with very low-latency communications.

When connecting to a Local Zone via AWS Direct Connect, you can create a private virtual interface (leveraging Direct Connect gateway) which allows connectivity directly to the VPC associated with the Local Zone. For more information, refer to the Amazon VPC User Guide. This creates a direct traffic path between your data centers and the local zone allowing you to achieve latency as low as 1-2ms.

**Note:** If you are using a transit virtual interface, traffic is first sent to the transit gateway in the AWS region before being forwarded to the local zone. For low-latency traffic we recommend creating a private virtual interface when connecting to a Local Zone.

**Scaling your direct connect connection bandwidth:**

AWS currently offers Direct Connect connections with speeds up to 100 Gbps. You can aggregate up to two 100 Gbps (or up to four 10Gbps) connections in a LAG, to get up to 200 Gbps of bandwidth. A link aggregation group (LAG) is a logical interface that uses the Link Aggregation Control Protocol (LACP) to aggregate multiple connections at a single AWS Direct Connect endpoint, allowing you to treat them as a single, managed connection. All connections in a LAG operate in Active/Active mode.

Increase bandwidth by load balancing traffic across multiple Direct Connect connections at a single Direct Connect location using BGP Equal-cost multipath (ECMP). Advertise the same prefixes with the same BGP attribute values (ex: AS PATH) on virtual interfaces you create over multiple connections to enable this behavior. When you have Direct Connect connections across multiple Direct Connect locations, by default, AWS uses the distance from the local Region to the AWS Direct Connect location to determine the virtual interface/connection to send the traffic (assuming you are advertising same prefixes) over the different VIF/connections' across Direct Connect locations. You can modify this behavior by tagging the prefixes you advertise over virtual interfaces with BGP communities. To load balance traffic across multiple AWS Direct Connect connections, apply the same community tag to the prefixes you advertise across the connections.

**Note:** When multiple paths exist for the same advertised prefix, BGP community preference (high, medium, low) value takes priority over AS PATH length when making a routing decision. If you tag your prefixes across multiple connections with the same BGP community, then the AS PATH length is looked at to determine which connection to send the traffic to. Same AS PATH lengths across multiple connections in this case will result in ECMP load balancing.

## Review

See the AWS Well-Architected Framework whitepaper for best practices in the review area for performance efficiency that apply to hybrid networking .

## Monitoring

Monitoring and tracking the performance of your hybrid networking connectivity is important. Often, you deploy your hybrid networking connectivity for an initial set of applications but as time progresses more and more apps start using this existing connectivity. This can lead to a low-performing or over-subscribed link. Rely on Amazon CloudWatch metrics and your on-premises (and/or service provider) device and router metrics for tracking the performance of your VPN and Direct Connect connection. If your applications are experiencing less than ideal performance, you must identify the root cause and fix it. For AWS service related issues, you can open an AWS Support ticket, while in other cases, re-architecting for scalability might be needed.

**HN_PERF5: How do you monitor and scale your hybrid connectivity post launch to ensure they are performing as expected?**

**Tracking and estimating growing usage of VPN or Direct Connect connectivity:**

Every time you deploy a new hybrid networking application, estimate the bandwidth requirements ahead of time to ensure that you are not oversubscribing the existing hybrid network connectivity link. To track the usage of a VPN connection, rely on Amazon CloudWatch metrics for VPN (we recommend `TunnelDataIn` and `TunnelDataOut` metrics), and your on-premises VPN device metrics. For Direct Connect, rely on Amazon CloudWatch metrics (we recommend the `ConnectionBpsIngress` and `ConnectionBpsEgress` metrics). Additionally, look at customer device metrics and, if applicable, metrics provided by your network service provider or circuit provider.

**Increasing VPN or Direct Connect bandwidth:**

When using VPN, if excessive latency and jitter are seen due to internet congestion, the best option might be moving to accelerated VPN or AWS Direct Connect. If VPN bandwidth is the limiting factor, depending on what limit is, you might want to get a more capable VPN termination device on premises or move to a different VPN termination endpoint in AWS. If your VPN is shutting down on a virtual private gateway, you should move the VPN endpoint to AWS Transit Gateway, which enables you to achieve higher VPN throughput at the AWS end using BGP ECMP load balancing across different VPN tunnels. You can do this without having to make any changes on your customer gateway. Using the ModifyVpnConnection API or the AWS Management Console, you can update the target gateway of a VPN connection. This preserves the endpoint's IP addresses on AWS, and the tunnel options such as inside-tunnel Classless Inter-Domain Routing (CIDR) and pre-shared keys. After you have moved to AWS Transit Gateway, add additional VPN connections and enable ECMP load balancing across all the VPN tunnels.

When using Direct Connect, you can add more capacity and load balance traffic across the old and new connections. You can add capacity one of two ways:

- Add more connections in a Link Aggregation Group (LAG) – you can request additional Direct Connect connections at the same Direct Connect location and place them in a LAG. All connections in the LAG shut down on the same AWS Direct Connect device. You can bundle up to four Direct Connect connections in a LAG. Since we implement same-chassis LAG at the AWS end, your new connections also land on the same AWS device as your previous connections. This might not always be possible, especially if there are no empty ports on the existing AWS device.

**Note:** All connections in the LAG must have the same bandwidth.

- Add more individual Direct Connect connections (no LAG) and load balance traffic across all the connections using BGP ECMP. When the added connections are in the same Direct Connect location, ensure that the ECMP is active, the BGP path cost of all the links should be the same, and the AS PATH length should be the same. When the added connections are spread across multiple direct connect locations, tag your prefixes with the same community tag in addition to ensuring same BGP path cost across the connections.

**Estimating how AWS Direct Connect connection failures impact your application performance:**

Failures often leave your hybrid networking connectivity in a degraded state, which affects performance. To ensure that failures don’t affect performance, architect and implement your connectivity in a way that upon failure, your connectivity can still meet the load of your workloads.

You can either set up your Direct Connect connections to be Active/Active or Active/Passive. If using Active/Passive, your passive connection should have the same performance metrics as your primary connectivity. Essentially, you are provisioning twice the capacity of what your requirement is and using half of it at any given time. This is same for Active/Active, with the only difference being that you are using both the links at half capacity.

As your connectivity requirements change and you need to scale your primary connection bandwidth, scale the passive or secondary links as well. It's also important to frequently test for failure scenarios to make sure that your application experience is not degraded when switching to the backup connectivity links.

## Tradeoffs

**HN_PERF6: How do you use tradeoffs to improve network performance?**

When deciding on which technology to choose (VPN vs dedicated circuits) or which termination endpoint to choose (EC2 instance vs TGW for VPN termination), consider the tradeoff between performance, cost, and ease/time to setup. Understanding the tradeoffs will help you choose the right tool for the right job.

**Tradeoffs between AWS VPN and AWS Direct Connect on cost and time to setup:**

The following tables show different factors to consider when making a choice between AWS VPN and Direct Connect.

*Table 2 - Tradeoffs between AWS VPN and AWS Direct Connect on cost and time to setup*

| Tradeoff | AWS VPN | AWS Direct Connect |
| --- | --- | --- |
| Cost | Low: If you have an active already paid for internet connection at the customer end, you pay for AWS VPN costs. | Low-Medium: Data transfer out cost over Direct Connect is lower than that over VPN. If you don't have existing circuits to a Direct Connect location, you may have to pay for circuit costs to your |

| Tradeoff | AWS VPN | AWS Direct Connect |
| --- | --- | --- |
| | | service provider in addition to AWS Direct connect hourly charge. |
| Time to Setup | Low: If you have an internet connection at the customer end, VPN can be established in minutes. | Low-Medium: If you already have a circuit to a Direct Connect location, AWS Direct connect can be setup in few days. If you don't have a circuit, circuit setup times can vary in the range of weeks. |
| Performance | Medium: medium bandwidth (based on internet speeds), medium-high latency (unpredictable number of hops), medium-high jitter (unknown hops, internet congestion).<br><br>Note: Leveraging accelerated VPN could lead to reduced (low-medium) latency and jitter. | High: high bandwidth (upto 100 Gbps), low latency and minimal jitter (private circuit with predictable number of hops). |

**Example: Where you can tradeoff performance for time to setup** - If you want to get your developers connectivity to development or sandbox VPC's quickly to start experimenting, then VPN is the recommended approach.

**Example: Where you can tradeoff time to setup for performance** - For production workloads, where performance is critical and you have a longer time to plan things, AWS Direct Connect is the recommended approach.

*Table 3 - Tradeoffs between AWS VPN and VPN termination on EC2 instances*

| Tradeoff | Virtual appliance on EC2 instance | AWS VPN (on AWS Transit Gateway) |
|---|---|---|
| Cost | Medium - High: You pay for EC2 instance pricing (2 or more instances for HA) in addition to any third-party licensing fees. | Low: You pay for AWS VPN costs (hourly charge + data transfer). |
| Setup complexity and management overhead | High: Management and maintenance of EC2 instances is customer responsibility. | Low: AWS VPN is a fully managed solution. |
| Scalability and performance | Low-medium: You can vertically scale an EC2 instance to get higher bandwidth. Horizontal scalability is limited due to how egress traffic can be load balanced to multiple ENI's in a VPC route table. | High: You can scale VPN bandwidth by provisioning more VPN connections and load balancing traffic using ECMP. |
| 3rd party features | High: You can choose a third-party vendor software to get additional functionality like DMVPN (Dynamic Multipoint VPN, a Cisco VPN protocol). | Low-medium: AWS VPN doesn't support third-par ty proprietary features like DMVPN. |

**Example: Where you can tradeoff performance for third-party features** - If you standardize on using DMVP for all your sites and want to quickly setup connectivity to a dev environment, the recommended approach is to use a third-party virtual appliance like Cisco CSR to terminate VPN.

**Example: Where you can tradeoff third-party features for performance** - If you standardize on using DMVP for all your sites and but want a fully managed, scalable solution for your production workloads, the recommended approach is to take time to move to IPSEC VPN connectivity leveraging AWS VPN.

## Resources

Refer to the following resources to learn more about our best practices related to performance efficiency.

### Documents

- Hybrid Connectivity Whitepaper
- AWS Direct Connect Locations
- AWS Direct Connect Link Aggregation Groups
- Accelerated Site-to-site VPN
- AWS Global Accelerator
- Routing policies and BGP communities

### Videos

- AWS Networking Series | Episode 2 | Connectivity to AWS and Hybrid AWS Network Architectures
- AWS re:Invent 2019: The right AWS network architecture for the right reason (NET320)
- AWS re:Invent 2019: Connectivity to AWS and hybrid AWS network architectures (NET317)
- AWS re:Invent 2020: Go global with AWS multi-Region network services

# Cost optimization pillar

The cost optimization pillar includes the continual process of refinement and improvement of a system over its entire lifecycle. From the initial design of your first proof of concept to the ongoing operation of production workloads, adopting the practices in this whitepaper will enable you to build and operate cost-aware systems that achieve business outcomes and minimize costs, thus allowing your business to maximize its return on investment.

## Best practices

There are four best practice areas for cost optimization in the cloud:

- Practice Cloud Financial Management
- Expenditure and usage awareness

- Cost-effective resources
- Manage demand and supply resources
- Optimize over time

As with the other pillars, there are trade-offs to consider. For example, do you want to optimize for speed to market or for cost? In some cases, it's best to optimize for speed—going to market quickly, shipping new features, or simply meeting a deadline—rather than investing in upfront cost optimization. Design decisions are sometimes guided by haste as opposed to empirical data, as the temptation always exists to overcompensate *just in case* rather than spend time benchmarking for the most cost-optimal deployment. This often leads to drastically over-provisioned and under-optimized deployments.

The following sections provide techniques and strategic guidance for the initial and ongoing cost optimization of your deployment.

## Practice Cloud Financial Management

For best practices in the Practice Cloud Financial Management area for cost optimization in hybrid networking, refer to the AWS Well-Architected Framework .

## Expenditure and usage awareness

**HN_COST1: How are you monitoring usage of your hybrid networking solution?**

**HN_COST2: How are you identifying data transfer costs for shared resources?**

Integrate existing hybrid network monitoring solutions with Cloud Monitoring solutions to enable complete end to end visibility. Data transfer costs are included as part of your AWS bill but it can be a challenge to understand exactly what type of data transfer charges are represented. AWS provides tools such as the AWS Cost & Usage Report and AWS VPC Flow Logs that can be used to track the costs associated with hybrid networking components as well as Amazon Athena and QuickSight for cost analysis and visualization.

Implementing solutions, or leveraging services that enable the identification of data transfer costs in particular, complements the use of cost-effective resources and optimal architectures. For example, customers can use Amazon Athena queries to analyze usage information in Amazon S3 and QuickSight to visualize the Amazon Athena analysis of the AWS Cost and Usage reports to identify data transfer costs. To setup QuickSight and Athena for analysis of the AWS Data Transfer cost details in the AWS CUR, refer to AWS Well-Architected Lab on Data Transfer Cost Analysis.

The Cost optimization pillar of the Well Architected Framework provides additional best practices for Expenditure and usage awareness.

## Cost-effective resources

**HN_COST3: How do you determine your data connectivity requirements for the most cost effective hybrid networking option?**

Choose hybrid networking connectivity for workloads with varying requirements for throughput and consistency as indicated in the reliability pillar. You may decide it would be more cost effective to use an Internet-based VPN connection for non-mission critical workloads that have no strict resiliency or latency requirements. You may also decide to use a private dedicated connectivity for your production traffic to achieve a more consistent and higher bandwidth connectivity between your Data Center and AWS.

Start off with Internet-based hybrid networking connections while in the testing phase of any workload migration or deployment and then migrate to more permanent connections only after baseline bandwidth requirements have been identified. For example, leveraging an internet based solution like AWS Site-to-Site VPN and then migrating to a dedicated connection like AWS Direct Connect. This enables you to start-off with a cost effective solution that can also be easily decommissioned prior to deploying a more permanent hybrid networking connection.

If you have a small hybrid network setup with a few VPCs and cost-saving intent and want to quickly establish on-premises network integration with your emerging AWS environment, you can use AWS Virtual Private Gateway to terminate your internet based AWS Site-to-Site VPN or private AWS Direct Connect connections. If you have multiple VPCs and there's a need to enable your development, test, production, and other VPCs to have network connectivity to your on-

premises environment, use an AWS Transit Gateway in addition to AWS Site-to-Site VPN or AWS Direct Connect.

While AWS Transit Gateway scales your VPN and Direct Connect connections and simplifies management, you are charged for the number of connections that you make to the AWS Transit Gateway per hour and the amount of traffic that is processed by the AWS Transit Gateway. The lower operational overhead cost savings benefits that Transit Gateway provides can outweigh the additional cost of AWS Transit Gateway data processing charges. For use cases where you require a transfer of very large amounts of data into AWS, consider a design approach where AWS Transit Gateway is in the traffic path to most VPCs but not all. This approach avoids the AWS Transit Gateway data processing fees. A cost-effective comparison for different hybrid networking scenarios is shown in the following table.

*Table 5 - Cost-Effectiveness comparison for AWS Hybrid Networking scenarios*

| **Category** | **Customer-managed VPN or SD-WAN** | **AWS S2S VPN** | **AWS Accelerated S2S VPN** | **AWS Direct Connect Hosted Connection** | **AWS Direct Connect Dedicated Connection** |
|---|---|---|---|---|---|
| Requires customer internet connection | Yes | Yes | Yes | No | No |
| Provisioned resources cost | EC2 instance and software licensing | AWS S2S VPN | AWS S2S VPN and AWS Global Accelerator | Hosted Connection port cost | Dedicated port cost |
| Data transfer cost | Internet rate | Internet rate or DIRECT CONNECT rate | Internet with data transfer premium | DX rate | DX rate |
| Transit Gateway | Optional | Optional | Required | Optional | Optional |

| Category | Customer-managed VPN or SD-WAN | AWS S2S VPN | AWS Accelerated S2S VPN | AWS Direct Connect Hosted Connection | AWS Direct Connect Dedicated Connection |
|---|---|---|---|---|---|
| Data processing cost | N/A | Only with AWS Transit Gateway | Yes | Only with AWS Transit Gateway | Only with AWS Transit Gateway |
| Can be used over AWS Direct Connect? | Yes | Yes | No | N/A | N/A |

## Manage demand and supply resources

**HN_COST4: How do you match the supply of resources for your hybrid networking with demand?**

**HN_COST5. How do you prioritize traffic across your hybrid networking connections?**

It is important to plan and accurately forecast your hybrid connectivity demands because of the length of time it can take to establish and scale the connectivity if requirements change. For example, you cannot dynamically self-provision to increase the speed from 1 Gbps to 10 Gbps or 100 Gbps for a private dedicated connection like AWS Direct Connect. Since there can be lead times of several weeks or more to create and configure increased speeds for AWS Direct Connect connectivity, we recommend starting the process soon even if you don't intend to depend on AWS Direct Connect in support of your initial few productions workload on AWS.

Obtain a comprehensive profile of the application demand to be placed on the network to prevent under-provisioning or over-provisioning hybrid network connectivity resources available or

supplied. As previously mentioned, you should start with a hybrid networking connectivity solution that can be easily modified or terminated if the baseline traffic bandwidth requirements are not known at the time.

The baseline network demand would also dictate whether an expansive transit multi-site network like MPLS would be required in the customer site depending on the number of locations, which might require consistent access to AWS. Another strategy could be to create a Link Aggregation Groups (LAG) with the minimum initial Direct Connect connection and then add connections as the network demand grows. For AWS Site-to-Site VPN, you can gain more bandwidth by provisioning new VPN tunnels using Equal Cost Multipath (ECMP) with a Transit Gateway deployment.

Customers should leverage prioritization and queuing techniques on their on-premises network devices in situations where the traffic entering the hybrid networking connections exceed the available bandwidth. For example, ensuring that traffic that is very susceptible to delay and jitter such as voice is configured to be transmitted ahead of bulky traffic such as replication traffic. You should also have minimum bandwidth guaranteed throughout the portions of the network where this can be configured. If you are going to over subscribe, customers should ensure traffic prioritization can be deployed through the network.

## Optimize over time

**HN_COST6: How are you designing your architecture for data transfer?**

**HN_COST7: How are you optimizing your hybrid networking architecture for data transfer?**

Data transfer fees can be a hidden cost in a hybrid connectivity deployment if not properly tracked. It is important to understand what drives your data transfer costs in order to optimize the cost of running your AWS hybrid networking architecture. Although you may regard data transfer costs as high, data transfer costs are relatively inexpensive compared to typical bandwidth charges from Internet Service Providers (ISPs) and data center operating costs that are common to on-premises workloads. If you have multi-tenant Software-as-a-Service (SaaS) workloads and a hybrid

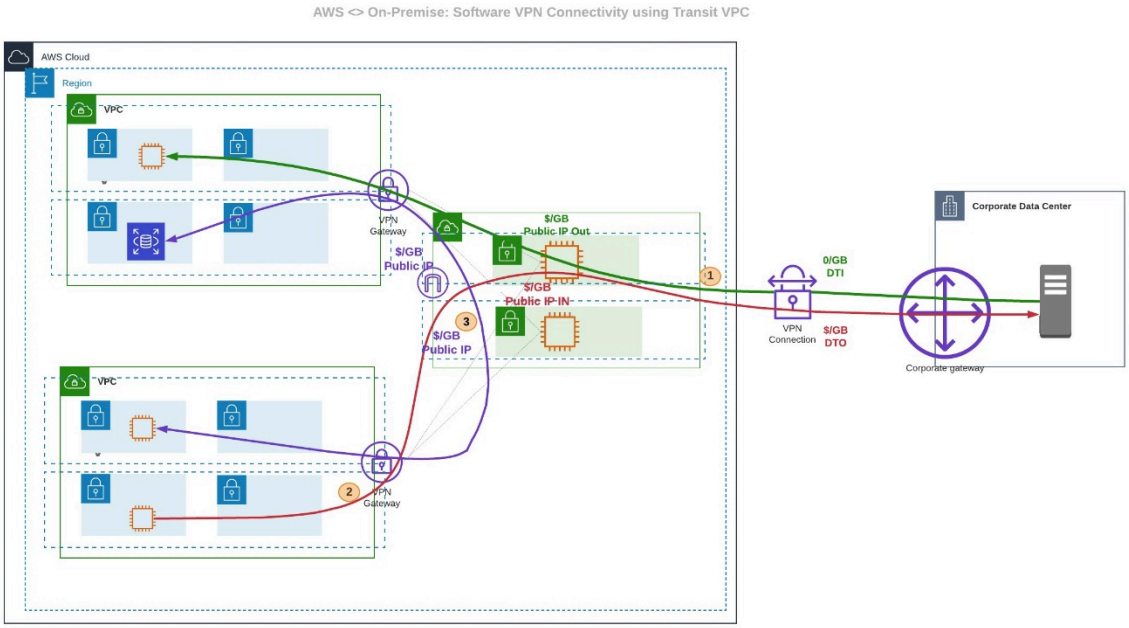
networking environment, it's also important to understand who pays for hourly and data transfer charges to enhance your pricing model.

You have different design choices on AWS when architecting your hybrid networking environment to optimize data transfer costs and it's important to understand the different data transfer pricing considerations. There are 4 main network connectivity options to consider on the AWS end and the data transfer costs will vary based on the option you choose.

1. **Customer managed VPN and Transit VPC**: For this internet-based connectivity model, you deploy commercial router virtual appliances in a transit VPC. This deployment option with Transit VPC was a common pattern used by customers prior to the advent of AWS Transit Gateway. The virtual appliances come with VPN licensing deployed on virtual machines (EC2) acting as an Intrusion Prevention System (IPS)/Intrusion Detection System (IDS), with all traffic flowing through these machines within a VPC. AWS Site-to-site VPN connections are established between the router virtual appliances and a virtual private gateway in each of your VPCs. Some customers terminate their customer managed VPN on an AWS Transit VPC, which is used to connect your VPC(s) and VPN connections via a Virtual Private Gateway (VGW) to your on-premises environment. Data Transfer cost considerations for a Customer Managed VPN and Transit VPC deployment are as follows and shown in the following diagram:

- (1) Data Transfer IN (DTI) from your on-prem environment to the VPN appliance within the Transit VPC is free. Data transferred out from the VPN appliance (EC2 public IP) through the Virtual Private Gateway (VGW) to the non-transit VPC is charged per GB. For example, data transfer from a server in your Data Center to the Transit VPC is free while there's a data transfer cost for data from the EC2 public IP via VGW to EC2 instances in your non-transit AWS VPC.
- (2) Data transferred from the VGW to the VPN appliance with public IP is charged per GB. Data Transfer OUT (DTO) over the internet to your on-premises environment is also charged per GB.
- (3) For inter-VPC connectivity through the Transit VPC, data transfer is charged per GB for traffic that flows in to and out of the Transit VPC VPN appliance with Public IP.
- This design provides multi-VPC connectivity as well as hybrid network connectivity to your on-premises environment. This option is a lower cost when compared to a virtual private gateway option for an AWS Site-to-Site VPN design mentioned in 2a, where every VPC has a VPN termination and data transfer costs. With this design, you mostly accrue data transfer costs between your Transit VPC and on-prem Data center.
- Customers with a Transit VPC and Software VPN deployment should consider replacing this model with an AWS Transit Gateway and Site-to-Site VPN deployment to reduce complexity and operational costs.

- For pricing information on the charges per GB of traffic sent, refer to the AWS VPN Pricing page and EC2 Data Transfer pricing page.



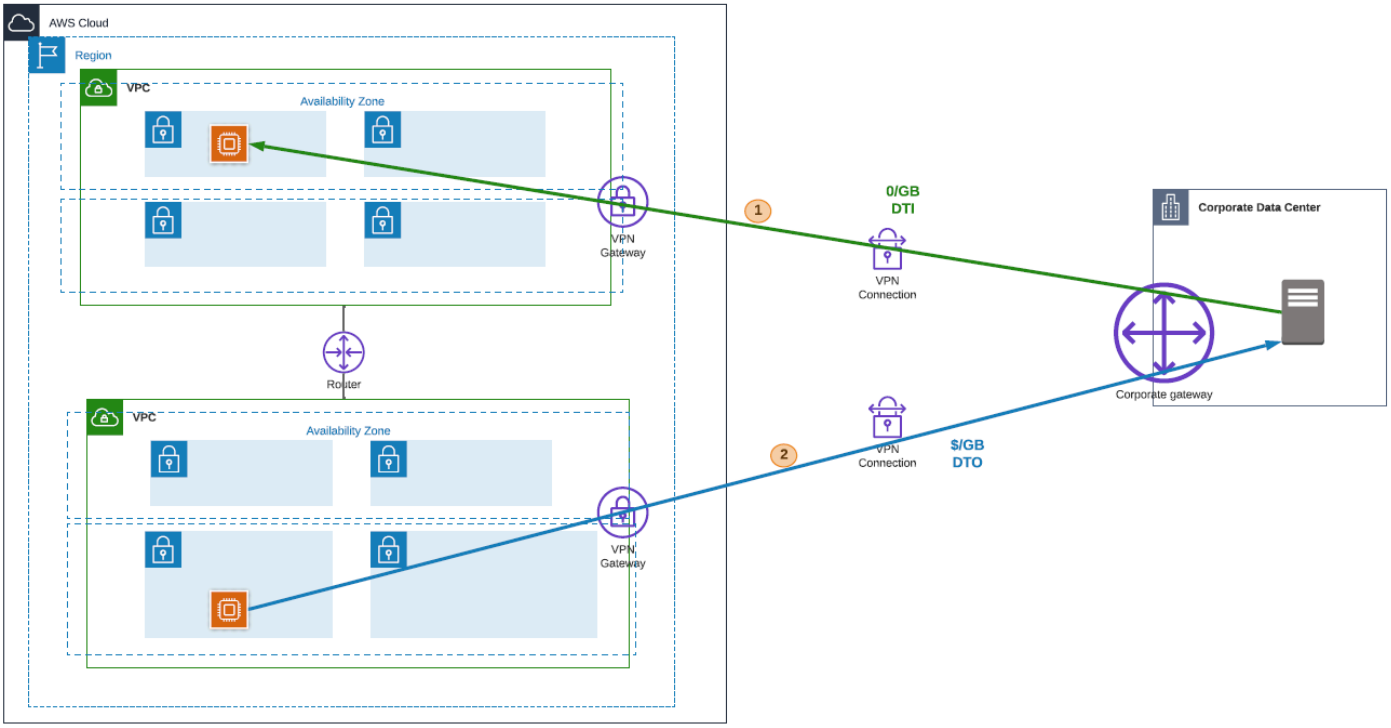

*Data Transfer: AWS to on-Premises: Software VPN Connectivity using Transit VPC*

2. **AWS Site-to-Site VPN**: You can deploy an internet-based hybrid option with AWS Site-to-Site-VPN using AWS Virtual Private Gateway or AWS Transit Gateway as termination points on AWS. For internet-based solutions like AWS Site-to-Site VPN, data transferred to an AWS VPC is free whereas Data transferred out from an AWS VPC is charged. VPN is charged on an hourly basis whether you use it or not so it's recommended to terminate inactive VPNs to avoid generating costs while not in use.

a. **Terminating via AWS Virtual Gateway** (VGW): When using an AWS virtual gateway, you need to connect your on-premises environment to each VPC due to the intransitive nature of virtual gateways. Data Transfer cost considerations for an AWS Site-to-Site VPN and AWS Virtual Gateway deployment are as follows and shown in the following diagram:

- (1) Data Transfer IN (DTI) from your on-prem environment to AWS VPC is free. For example, a server on-prem transfers data to an EC2 instance in your VPC.
- (2) Data Transfer OUT (DTO) from the AWS VPC to your on-prem environment is charged per GB. For example, an EC2 instance in your VPC putting out data to your on-prem server.

- For pricing information on the charges per GB of traffic sent, refer to the AWS VPN Pricing page

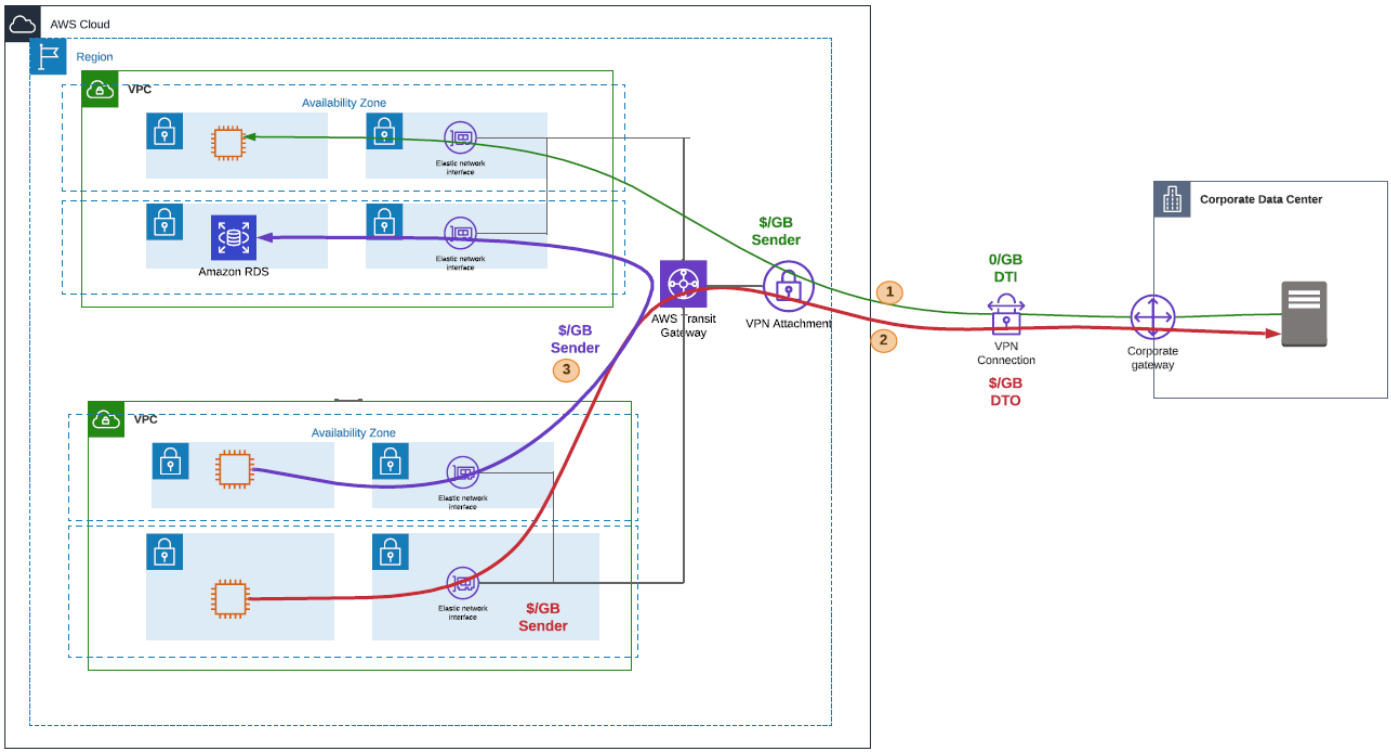



*Data Transfer: AWS to on-Premises: Site-to-Site VPN Connectivity using AWS Virtual Gateway*

b. **Terminating via AWS Transit Gateway** (TGW): The AWS Transit Gateway option provides a more scalable approach to using the Transit VPC (1) or Site-to-Site VPN using Virtual Gateway (2). While the data transfer costs for running AWS Transit Gateway may seem higher than the AWS Virtual Gateway option, Transit Gateway deployments help simplify VPC to VPC network connectivity, reducing operational overhead and total cost of ownership in the long term for AWS Site-to-Site VPN and AWS Direct Connect. Data Transfer cost considerations for an AWS Site-to-Site VPN and AWS Transit Gateway deployment are as follows and shown in the following diagram :

- (1) Data Transfer IN (DTI) from your on-prem environment to AWS VPC is free and you are charged per GB for AWS Transit Gateway data processing costs.
- (2) Data Processing from the sender, which is the VPC attachment for AWS Transit Gateway data processing costs is charged per GB in addition to Data Transfer OUT (DTO) charge per GB for data going out from your VPC over the internet to your on-prem environment.
- (3) For inter-VPC connectivity, you are charged per GB of traffic for AWS Transit Gateway data processing costs.

- For pricing information on the charges per GB of traffic sent, refer to the Transit Gateway pricing page.

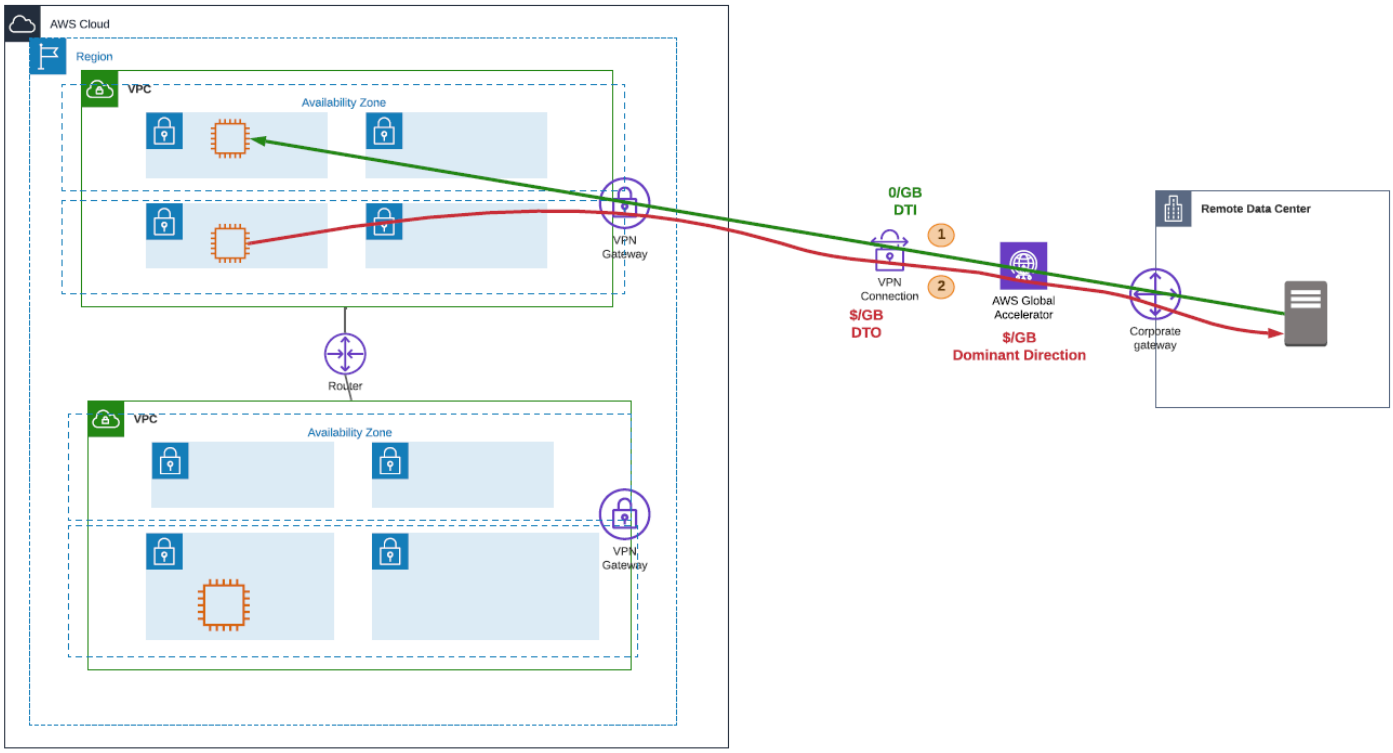



*Data Transfer: AWS to on-Premises: Site-to-Site VPN Connectivity using AWS Transit Gateway*

3. **Accelerated Site-to-Site VPN:** This option enables acceleration of your AWS Site-to-Site VPN connection using AWS Global Accelerator to avoid network disruptions that may occur as a result of using the public internet. With this option, AWS routes traffic from your on-prem network to an AWS edge location that is closest to your customer gateway device and provides consistency and reduced latency for data transfer. Data Transfer cost considerations for an Accelerated Site-to-Site VPN deployment are as follows and shown in the following diagram:

- (1) Data Transfer IN (DTI) from your remote on-prem environment to AWS VPC is free.

- (2) Data Transfer OUT (DTO) from your AWS VPC to your remote on-prem environment is charged per GB in addition to the Global Accelerator charges, which is dependent on traffic flowing in the dominant direction in or out and the destination edge location. Refer to the AWS Global Accelerator pricing page for pricing details.

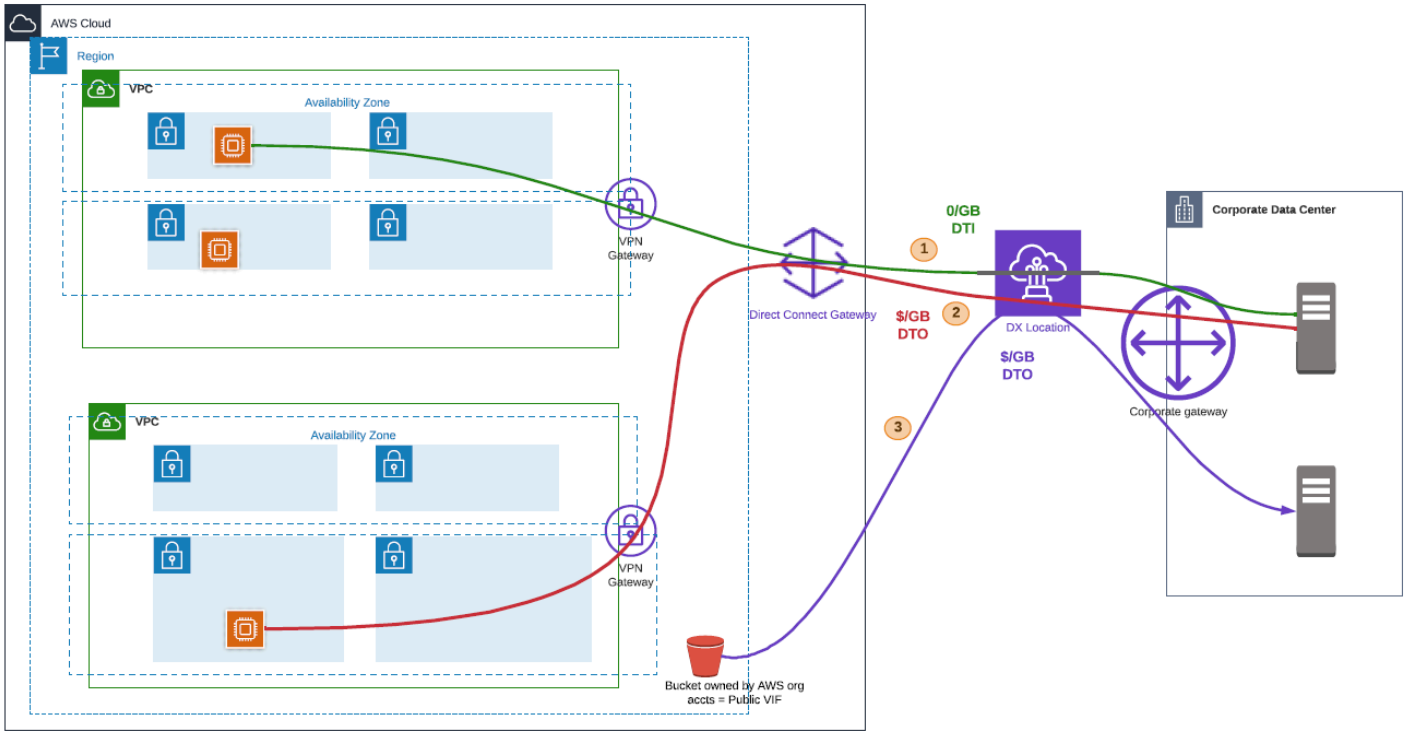



*Data Transfer: AWS to on-Premises: Accelerated Site-to-Site VPN Connectivity*

4. **AWS Direct Connect**: For situations where an internet-based connection like AWS Site-to-Site VPN is not sufficient, you can use a dedicated connection like AWS Direct Connect. For data transfer cost optimization between your on-premises and AWS environment, it's recommended to use a dedicated connection like AWS Direct Connect as it's usually multiple times less expensive than an internet based solution like AWS Site-to-Site VPN. There are two main ways to terminate AWS Direct Connect and for both options, the standard Internet data transfer charges do not apply. Data transferred out of AWS via a dedicated connection such as AWS Direct Connect is charged per GB. There is no data transfer charge for data coming into AWS Direct Connect from your data center or co-location facility. The cost of outbound data varies by region and Direct Connect location.

a. **AWS Direct Connect Gateway & Virtual Private Gateway (VGW)**: With this deployment model, every VPC will have its own AWS Virtual Private Gateway connected to an AWS Direct Connect Gateway. AWS Direct Connect Gateway is recommended for deployments that require connectivity between multiple VPCs in the same or different AWS Regions (except China) to their Direct Connect connection. AWS Direct Connect Gateway works with virtual private gateways or with Transit Gateway for multiple VPCs in the same region. Data Transfer cost considerations for an AWS Direct Connect Gateway and AWS Virtual Gateway deployment are as follows and shown in the following diagram:

- (1) Data Transfer IN (DTI) from the on-prem environment to AWS VPC is free
- (2) Direct Connect Data Transfer OUT (DTO) from your AWS VPC to on-prem environment is charged per GB according to the Direct Connect pricing per region.
- (3) If you have a resource such as an S3 bucket owned by one of your AWS organization accounts or Direct Connect public VIF (virtual interface), you are charged per GB for Data Transfer (DTO) charges based on Direct Connect data transfer out pricing and region for example if your on-prem server is pulling data out of the S3 bucket.
- (3) If your AWS resource such as a public S3 bucket is not owned by your AWS organization accounts or Direct Connect public VIF, the owner of that resource or S3 bucket is charged per GB for Data Transfer out (DTO) based on the internet data transfer charges.
- For pricing information on data transfer rates, refer to the Direct Connect pricing page.



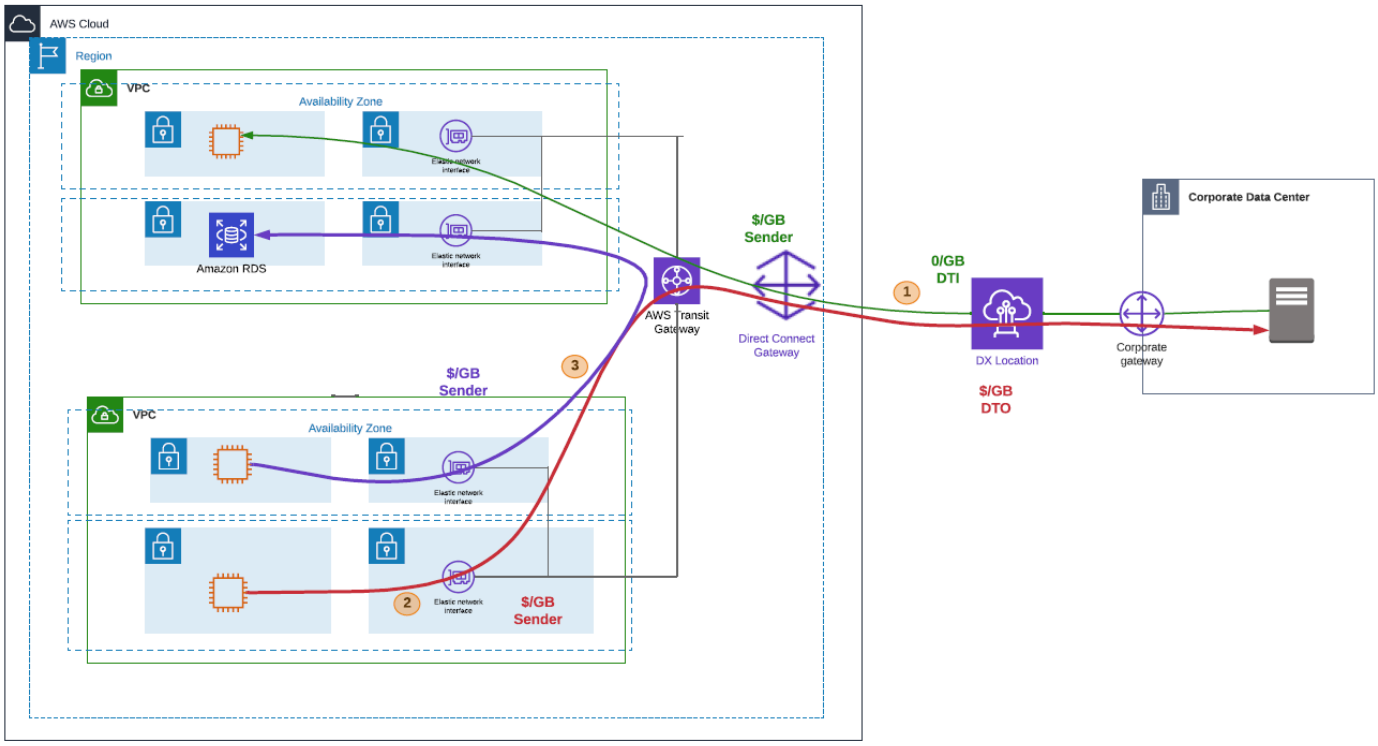

*Data Transfer: AWS to On-Premises: Direct Connect using a Virtual Gateway*

b. **Direct Connect Gateway & Transit Gateway Same-region**: With this scenario, you can replace the Virtual private gateways seen in Option 3(a) with AWS Transit Gateway for a more scalable design. Data Transfer cost considerations for an AWS Direct Connect Gateway and AWS Transit Gateway deployment in the same region are as follows and shown in the following diagram:

- (1) Data Transfer IN (DTI) from your on-prem environment to Direct Connect Location is free while Transit Gateway data processing charges for the Direct Connect Gateway attachment is charged per GB based on who the sender is (in this case Direct Connect Gateway). Note that there's a Transit Gateway attachment charge associated with using a Transit Gateway compared to the VGW option in 4(a).

- (2) For data sent from the EC2 instance in a VPC attached to the Transit Gateway, data Processing is charged per GB to the VPC owner who sends traffic to Transit Gateway. In addition, there are Direct Connect Data Transfer OUT (DTO) charges per GB from your AWS Region to your Direct Connect location.

- (3) For inter-VPC communication, you are charged per GB of traffic for Transit Gateway data processing costs based on the sender, which is the VPC owner who sends traffic to Transit Gateway. For pricing information on data transfer rates, refer to the Direct Connect pricing page

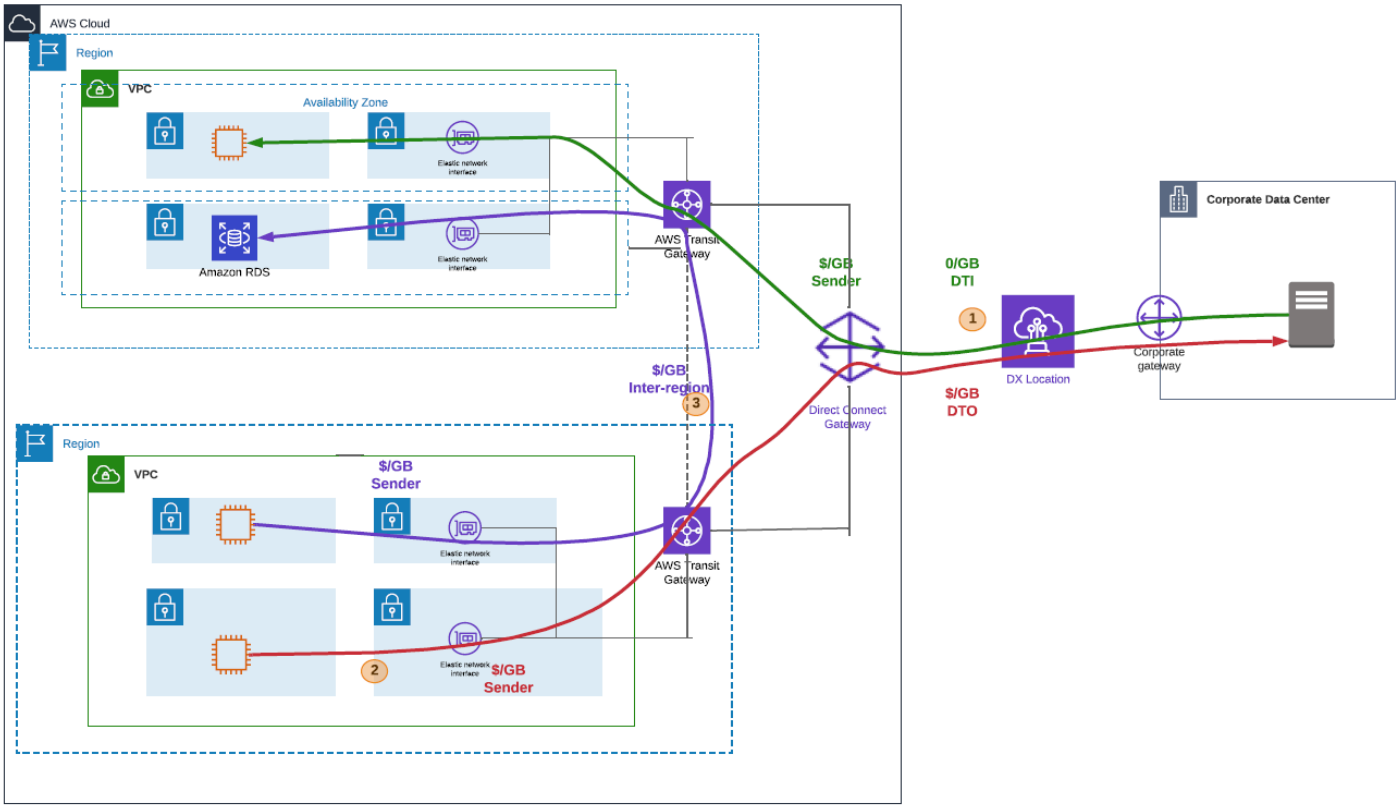



*Data Transfer: AWS to On-Premises: Direct Connect using Transit Gateway in the same AWS Region*

c. **Direct Connect Gateway & Transit Gateway cross-region:** With this connectivity model, Direct Connect is used to connect multiple regions to your on-prem location(s). Here, everything is

connected through a Direct Connect Gateway and between the regions, there is a Transit Gateway. Inter-VPC communication can happen through your AWS Transit Gateway peering connection but if there's a requirement to have connectivity from your AWS Direct Connect location to various VPCs in different AWS regions, that communication happens via the Direct Connect Gateway and Transit Gateway path. Data Transfer cost considerations for an AWS Direct Connect Gateway and AWS Transit Gateway deployment in different regions are as follows and shown in the following diagram:

- (1) Data Transfer IN (DTI) from your on-prem environment to Direct Connect Location is free Transit Gateway data processing charges for the Direct Connect Gateway attachment is charged per GB based on who the sender is (in this case Direct Connect Gateway). Note that there's a Transit Gateway charge associated with using a Transit Gateway compared to the VGW option in 4(a).
- (2) For data sent from the EC2 instance in a VPC attached to the Transit Gateway, data Processing is charged per GB to the VPC owner who sends traffic to Transit Gateway. In addition, there are Direct Connect Data Transfer OUT (DTO) charges per GB from your AWS Region to your Direct Connect location.
- (3) For inter-VPC communication across the region, you are charged per GB of traffic for data processing on traffic based on the sender, which is the VPC owner who sends traffic to Transit Gateway. In addition, you are charged for inter-region data transfer costs based on cross-region data transfer pricing.




*Data Transfer: AWS to On-Premises: Direct Connect using AWS Transit Gateway Cross-Region*

## Resources

Refer to the following resources to learn more about AWS best practices for cost optimization.

### Documents

- 10 things you can do today to reduce AWS costs
- Data Transfer Pricing
- AWS VPN FAQs
- Cost and Usage Analysis Well-Architected Lab
- Data Transfer Cost Analysis Well-Architected Lab
- Hybrid Connectivity Whitepaper

### Videos

- NET305 Behind the Scenes: Exploring the AWS Global Network

- Nine Ways to Reduce Your AWS Bill
- Pricing Model Analysis

## Sustainability pillar

The sustainability pillar includes the ability to continually improve sustainability impacts by reducing energy consumption and increasing efficiency across all components of a workload by maximizing the benefits from the provisioned resources and minimizing the total resources required.

There are no sustainability practices unique to this lens. For information on Sustainability, refer to the Sustainability Pillar whitepaper.

# Conclusion

The AWS Well-Architected Framework provides architectural best practices for designing and operating reliable, secure, efficient, and cost-effective systems in the cloud for Hybrid Networking. The framework provides a set of questions and best practices that allow you to review an existing or proposed hybrid networking architecture. Using the framework in your architecture helps you build stable and efficient systems, which enables you to focus on your functional requirements.

# Contributors

Contributors to this document include:

- Sidhartha Chauhan, Principal Solutions Architect, Amazon Web Services
- Ikenna Izugbokwe, Principal Solutions Architect, Amazon Web Services
- Jennifer Ihejimba, Solutions Architect, Amazon Web Services

# Document history

To be notified about updates to this whitepaper, subscribe to the RSS feed.

| Change | Description | Date |
| --- | --- | --- |
| Initial publication | Hybrid Networking Lens first published. | November 22, 2021 |

**Note**

To subscribe to RSS updates, you must have an RSS plug-in for the browser you are using.

# Notices

Customers are responsible for making their own independent assessment of the information in this document. This document: (a) is for informational purposes only, (b) represents current AWS product offerings and practices, which are subject to change without notice, and (c) does not create any commitments or assurances from AWS and its affiliates, suppliers or licensors. AWS products or services are provided “as is” without warranties, representations, or conditions of any kind, whether express or implied. The responsibilities and liabilities of AWS to its customers are controlled by AWS agreements, and this document is not part of, nor does it modify, any agreement between AWS and its customers.

# AWS Glossary

For the latest AWS terminology, see the AWS glossary in the *AWS Glossary Reference*.