VMware Cloud on AWS: Network Architecture

VMware Architecture

Table of contents

VMware Cloud on AWS: Network Architecture

Overview

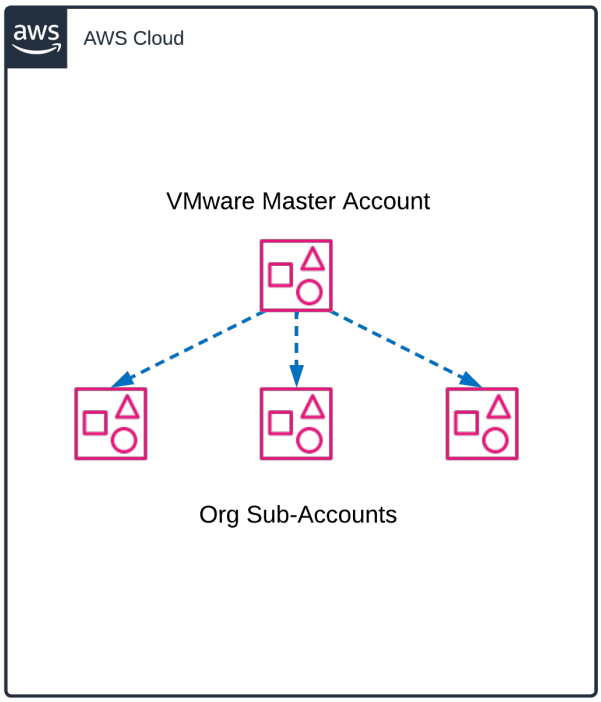
The networking architecture of a VMware Cloud on AWS Software-Defined Data Center (SDDC) is arguably one of the most complex components of VMware Cloud on AWS. While it isn't necessary to fully understand all aspects of the internal details of this architecture, it is important to have an understanding of fundamentals in order to properly design and manage the solution.

This document provides an overview of the network architecture of VMware Cloud on AWS, starting with the AWS base layer that underpins the infrastructure. We delve into the SDDC Underlay and NSX Overlay networks. The network aspects of a multi-AZ SDDC are also discussed, where a vSAN stretched cluster is utilized to ensure data redundancy and to protect against AZ failures. The section on VMware Transit Connect details how this service offers high bandwidth and resilient connectivity for SDDCs within an SDDC Group, and simplifies the connectivity to AWS VPCs and on-premises data centers via an AWS Direct Connect Gateway in multi-SDDC deployments. The document concludes with a summary of the network services available in VMware Cloud on AWS SDDC.

The AWS Base Layer

The VMware AWS Account

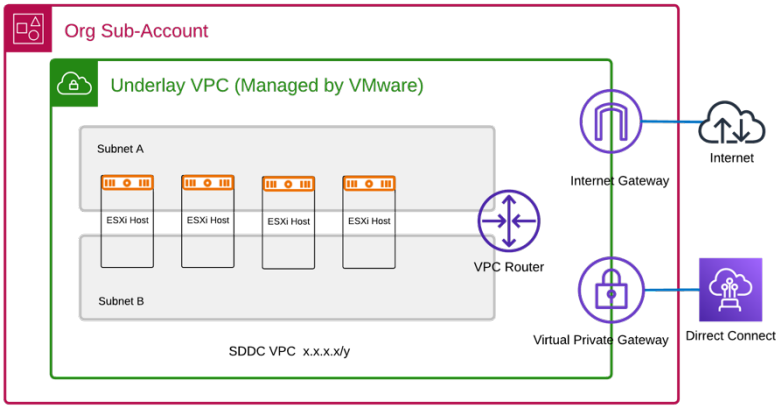
As part of the partnership with AWS, VMware maintains a master AWS account which is used as part of the VMware Cloud on AWS service. Whenever a new cloud services Organization (Org) is created, VMware creates a sub-account within this master account which acts as the parent for all AWS resources used by that Org.



Whenever an SDDC is provisioned, resources for that SDDC are created within the AWS sub-account for the Org. This model allows VMware to manage billing for SDDC consumption (hardware, bandwidth, elastic IPs, etc.).

The SDDC Underlay VPC

As part of the process for provisioning an SDDC it is required to provide an IP address range to use for SDDC management. This management Classless Inter-Domain Routing (CIDR) size must be one of three available sizes: /16, /20 or /23. Using this information, VMware Cloud on AWS will create a new Amazon Virtual Private Cloud (VPC) within the VMware-owned AWS account for that SDDC's Org. The management CIDR will be carved out into multiple subnets. Some of these subnets are allocated for the VPC and others will be used for the NSX overlay networks. VMware Cloud on AWS will then deploy hardware instances for the SDDC and connect them to the subnets of that VPC.



An internet gateway (IGW) and virtual private gateway (VGW) will also be created for this VPC. These gateways enable internet and Direct Connect connectivity to the VPC. It's important to note that customers cannot directly see or manage these AWS components. Any interaction with or configuration of these components must be done through VMware-provided options such as the NSX UI, API, and Cloud Console.

The SDDC Underlay

ESXi Networking

Once the underlay VPC has been created for the SDDC management cluster and hardware hosts have been provisioned, the SDDC is bootstrapped with ESXi and the remaining VMware software stack is installed. Once the SDDC is up and running, it is possible to get a glimpse of the underlay networking design by viewing the networking configuration of a single ESXi host. This setup is fairly complex, so it is not entirely necessary that it be well understood; however, having some insight into the inner workings of the underlay network will help when it comes to understanding the networking behaviors of the SDDC.

The hosts of an SDDC are provisioned with an Elastic Network Adapter (ENA) which provides their connectivity to the AWS underlay VPC. Within the host, this ENA is attached to a Virtual Distributed Switch (VDS), where it is made available to the various virtual switches which provide network connectivity to the SDDC. The hosts themselves have a number of VMkernel interfaces which they use for the following purposes:

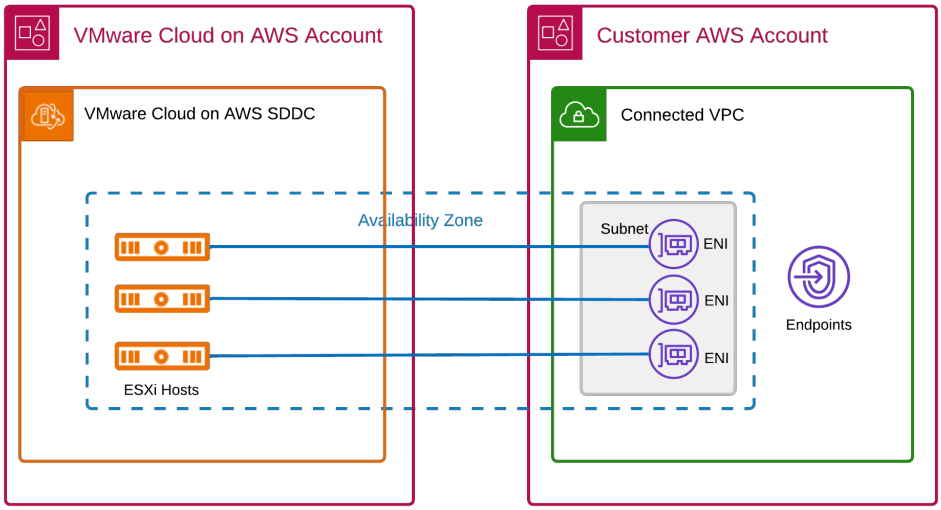
- management (vmk0)
- vSAN (vmk1)
- vMotion (vmk2)
- AWS API (vmk4)
- NSX Tunnel End Point

These VMkernel adapters are visible from the networking section of the configuration tab of the ESXi host. You can also get a sense of the setup by viewing the TCP/IP configuration for the host.

The last portion of the underlay network are the virtual switches. Again, these are visible from the networking section of the configuration tab of the ESXi host. Here, you will see a mix of switches; some that are part of the underlay network and others which represent NSX network segments. The switches which are part of the underlay network are there to provide connectivity to the various management appliances of the SDDC. Of particular note is the management switch, which exists in order to provide the appliances with access to the management network. You can get a sense of some of the other important virtual switches in a host by examining their names as well as the management appliances which are attached to them.

The VPC Cross-Link and Connected VPC

Every production SDDC must be cross-linked to a VPC within the customer-owned AWS account, referred to as the connected VPC. This cross-linking is accomplished using the Cross-Account Elastic network interface (ENI) feature of AWS and creates a connection between every host within the management cluster of the SDDC to a subnet within the cross-linked VPC. This cross-link provides the SDDC with a high-bandwidth and low-latency network forwarding path to services maintained within the customer-owned AWS account.



The Cross-Account ENIs are visible from the customer-owned AWS account (by viewing network interfaces within EC2). You will notice that there are several ENIs created when the SDDC is deployed but not all are active. In addition to ENIs created for active hosts of the SDDC, there will also be ENIs created for future expansion and for upgrades/maintenance. Even though it is possible, you should avoid modifying or deleting these ENIs since doing so may impact the cross-link to the SDDC.

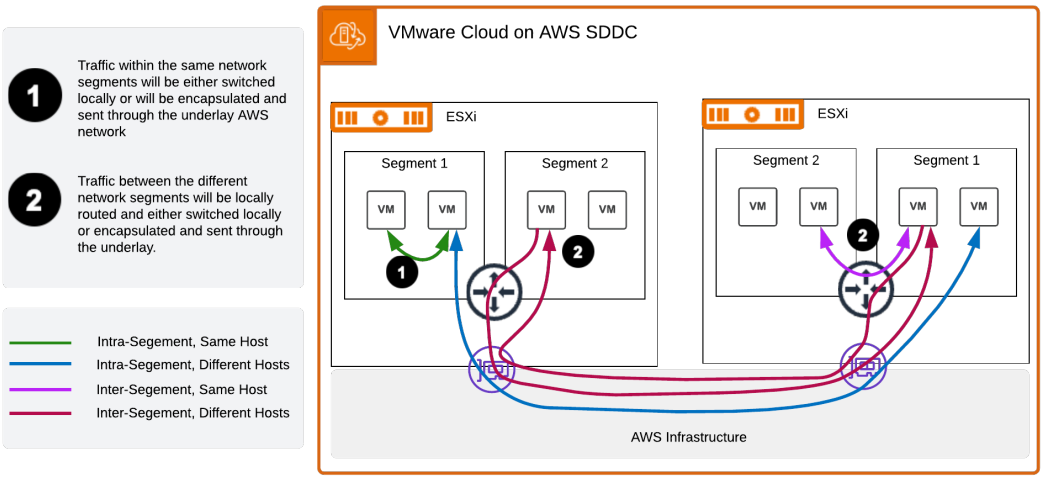
The SDDC Overlay

An Overview of NSX Networking

As part of the standard SDDC software stack, VMware Cloud on AWS utilizes VMware NSX to create an overlay network atop the base-layer provided by AWS. The end result is a logical network architecture which is completely abstracted from the underlying infrastructure.

Network overlays operate on the notion of encapsulation; they hide network traffic between VMs within the overlay from the underlying infrastructure. VMware NSX uses GENEVE as its overlay networking protocol within the SDDC.

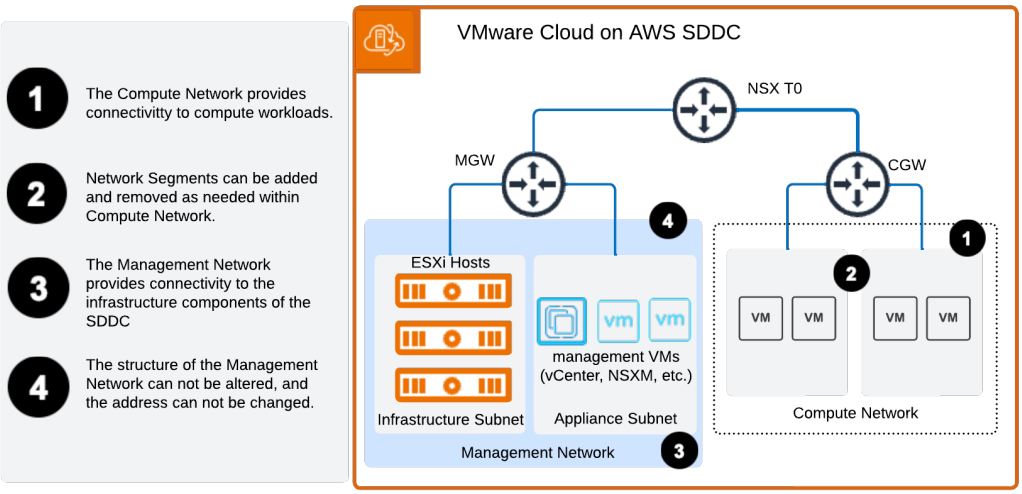
Higher-level constructs aside, software defined networking defines two types of objects: logical switches and logical routers. These objects are designed to mimic the behavior of their counterparts in traditional hardware-based networks. Logical switches operate at layer-2 and will forward traffic between nodes within the same network segment. Logical routers operate at layer-3 and will route traffic between network segments. With NSX, logical switches and logical routers are distributed. This means that each host of the SDDC maintains enough information to understand which VMs belong to which logical switch(es) and how to forward traffic through the underlay network. When two VMs communicate with one another, the exact path through the underlay becomes a function of where the VMs reside (i.e. on which host they reside).



In the figure above, we see several examples of different types of traffic flows. As illustrated, intra-segment traffic will either be switched locally for VMs on the same host or, for VMs on different hosts, encapsulated and sent through the underlay. Similarly, for inter-segment traffic the routing will take place locally before being switched or encapsulated and sent through the underlay.

The NSX Overlay Network

The SDDC utilizes NSX to create an overlay network with two tiers of routing. At the first tier of the network is an NSX tier-0 (T0) router which acts as the north-south border device for the entire SDDC. At the second tier are the NSX tier-1 (T1) routers which are known as the Management Gateway (MGW), Compute Gateway (CGW). Customers are allowed to add new T1 routers (aka custom T1s). These tier-1 routers act as the gateways for their respective networks. The figure below shows the network topology of a newly provisioned SDDC.

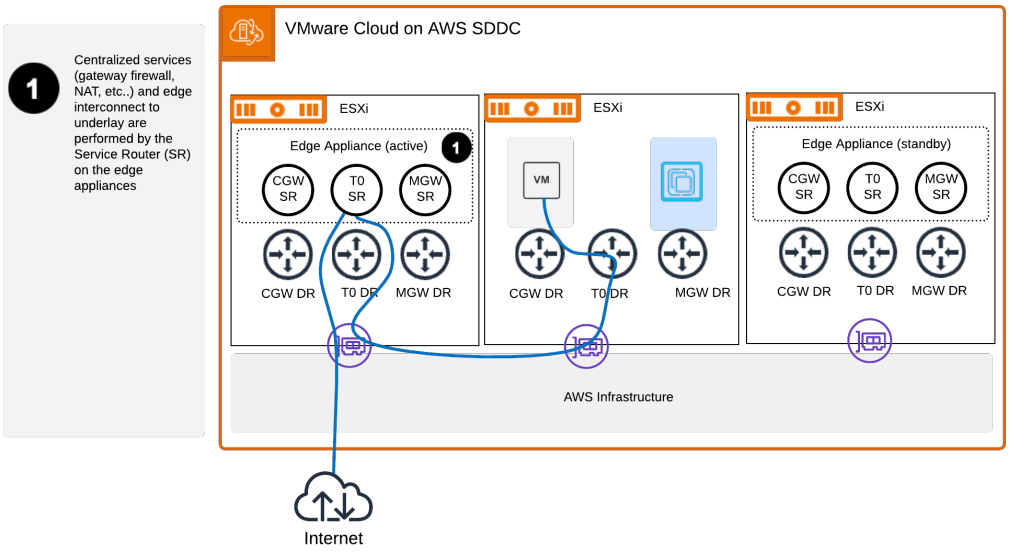


There are two distinct types of networks within the SDDC: the management network and the compute network. The management networks are considered to be part of the SDDC infrastructure and provides network connectivity to the various infrastructure components of the SDDC. Within these, the appliance subnet is used by the vCenter, and NSX appliances in the SDDC. Other appliance-based services added to the SDDC also connect to this subnet. The infrastructure subnet is utilized by the ESXi hosts in the SDDC. Due to the permissions model of the service, the IP address space is fixed for this network and its layout may not be altered.

Compute networks are used by the compute workloads of the SDDC. Customers have the ability to add and remove network segments within the compute network as needed.

NSX Logical Routers

All routers within the SDDC are distributed. This means that routing between segments is performed locally on each ESXi host by the appropriate distributed router (DR). Certain functions, however, are not distributed and must be handled centrally by a Service Router (SR) component on the NSX edge appliances. Specifically, gateway firewall and Network Address Translation (NAT) operations are handled in a centralized manner. Also, any traffic which passes between the overlay and underlay networks must be handled by the tier-0 edge SR on the edge appliances. The edge appliances are deployed in a redundant pair, with all SRs running on the active appliance and with SRs on the standby appliance sitting idle. In the event of a failure of one or more SRs on the active appliance, all SRs will fail over to the standby appliance. This is an optimization strategy designed to prevent unnecessary traffic flows between the edge appliances.



In the figure above, we see an example traffic flow where a workload VM connected to CGW is communicating to the internet. It

this example, the traffic is routed between the CGW and T0 DRs locally before being sent through the underlay to the T0 SR in the active edge appliance. From there, the traffic is sent out to the internet. This routing pattern would be visible in a traceroute from the VM.

NSX Network Segments

VMware Cloud on AWS supports three types of compute network segments: routed, extended and disconnected.

The difference of three network types is primarily in their connectivity capability:

- Routed – routed (default type) network segments have connectivity to other segments within an SDDC and to external networks.
- Extended – extended network segments are used with the NSX L2VPN to create a common broadcast domain between a VMware Cloud on AWS SDDC and on-premises networks.
- Disconnected – disconnected network segments have no uplinks and create an isolated segment. Disconnected networks are primarily created by HCX, that can be changed to routed networks.

For more information, please refer to Understanding Segments in VMC on AWS.

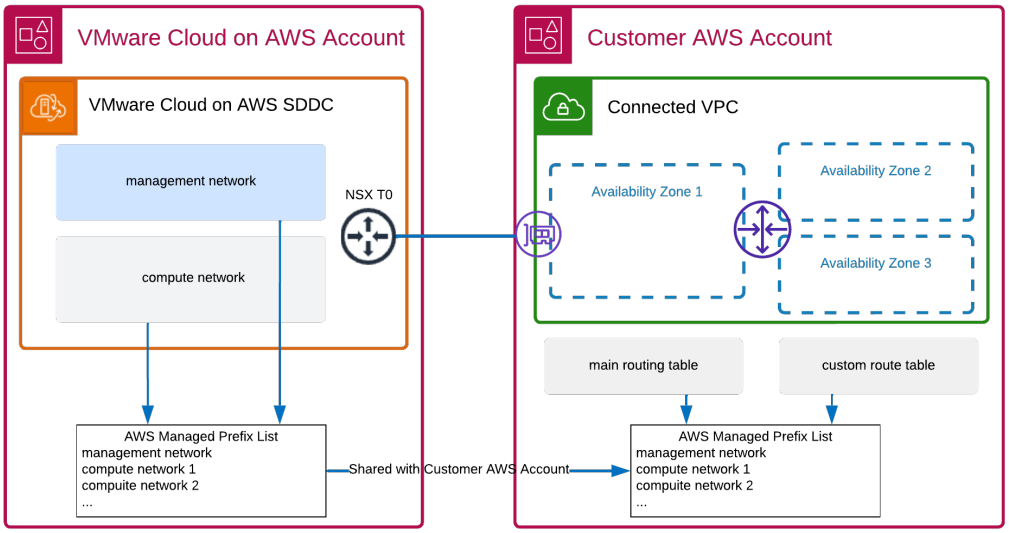
Connected VPC Connectivity

Since we must always pass through the Edge SR whenever traffic leaves the SDDC, all traffic to and from the connected VPC must pass through the active edge appliance of the SDDC. Since this edge appliance is a VM, it resides on a specific ESXi host. As such, it will always use the Cross-Account ENI for that host. It should be noted that the hosts of the SDDC will be deployed within the same AZ as the cross-link subnet (the subnet which contains the Cross- Account ENIs). This mechanism controls the AZ placement of the SDDC (customer chooses the subnet and thus controls the AZ placement), but also serves to eliminate cross-AZ bandwidth charges between the edge and any native AWS resources within that same AZ.

Note - The cross-link architecture is slightly different for stretched cluster SDDCs (covered in later sections).

Routing between the SDDC and the VPC is facilitated through static routes, which are created on-demand as networks are added to the SDDC. By default, these static routes are incorporated into the main routing table of the customer-owned connected VPC, utilizing the active Cross-Account ENIs as the next hop for the routes. It is crucial to note that the next-hop ENI designated for the static routes will consistently be that of the ESXi host hosting the active edge appliance within the SDDC. Consequently, should the edge appliance migrate to a different host (as can occur during failover events or when the SDDC undergoes upgrades), the next-hop for the static routes will be updated to reflect this change.

When the AWS Managed Prefix List Mode is activated, it enables customers to leverage custom route tables within their connected VPCs for AWS to SDDC network connectivity. VMware Cloud on AWS generates an AWS managed prefix list populated with the default Compute Gateway prefixes along with any additional prefix list aggregations created by customers. This list is then shared with the customer's AWS Account that owns the Connected VPC. Once this AWS resource share is accepted, the shared prefix lists can be utilized as destinations when defining routes to VMware Cloud on AWS SDDC networks updated to point to the correct ENI whenever the active Edge instance's host changes.



For more information about the AWS Managed Prefix List Mode, please refer to Understanding Managed Prefix List Mode for Connected VPC in VMC on AWS.

Edge Uplinks

There are currently three uplinks from the tier-0 routers of the SDDC. These are described below.

Internet Uplink

The internet uplink is connected to AWS IGW within the underlay VPC. Since the default from other sources such as Direct Connect, IPSec VPN or VMware Transit Connect (VTC) has a higher priority, the IGW path would be only used when no other default routes are learned.

Please note that customers will still use this Internet uplink for VPN public endpoint termination or Hybrid Cloud Extension (HCX) using the public IP uplink profile, even if the default route is not pointing to the IGW.

Traffic over this uplink is billable and the charges will be passed through as part of the billing for the SDDC, except when the destination of traffic is AWS services within the same AWS region.

The MTU size for Internet Uplink is 1500 bytes, which is not configurable.

Services Uplink

The services uplink connects the SDDC edge appliances to the connected VPC in the customer-owned AWS account. A static route on the SDDC edge appliances is configured for the primary CIDR of the VPC which points to the VPC router as a next hop during the SDDC provisioning. It enables compute workloads to leverage this uplink to consume AWS services and resources in the connected VPC. For services not running in the connected VPC, it is still possible to using the uplink to get connected through VPC endpoints.

By default, compute workloads in an SDDC will consume the Amazon S3 service using SDDC Internet connectivity. VMware Cloud on AWS also allows customers to utilize the service uplinks for S3 access. Once this option is enabled, static routes for the regional S3 prefixes will be automatically added to the NSX-T T0 routers.

Traffic over this uplink is non-billable only for AWS resources which are within the same availability Zone as the SDDC. Traffic to resources in other Availability Zones is billable and charges will be accrued on the customer-owned AWS account.

The default MTU size for Services Uplink is 8900 bytes, which is not configurable.

Intranet Uplink

The Intranet uplink is used when Direct Connect private virtual interface (VIF) or VMware Transit Connect is attached into the SDDC. The SDDC edge will use this uplink for whatever network prefixes are received via Direct Connect or VMware Transit Connect over this uplink. Since Direct Connect is a resource which is managed by the customer-owned AWS account, bandwidth charges over the Direct Connect will be accrued on the customer-owned AWS account. Regarding the cost of VMware Transit Connect, please refer to Designlet: VMware Transit Connect for VMware Cloud on AWS.

The default MTU size of the Intranet uplinks is 1500 bytes. The MTU for the Intranet Uplink can be increased to up to 8900 bytes for Direct Connect) or 8500 for VMware Transit Connect to support jumbo frames. It is always recommended to enable jumbo frames support on all network devices before the end systems can take advantage of it.

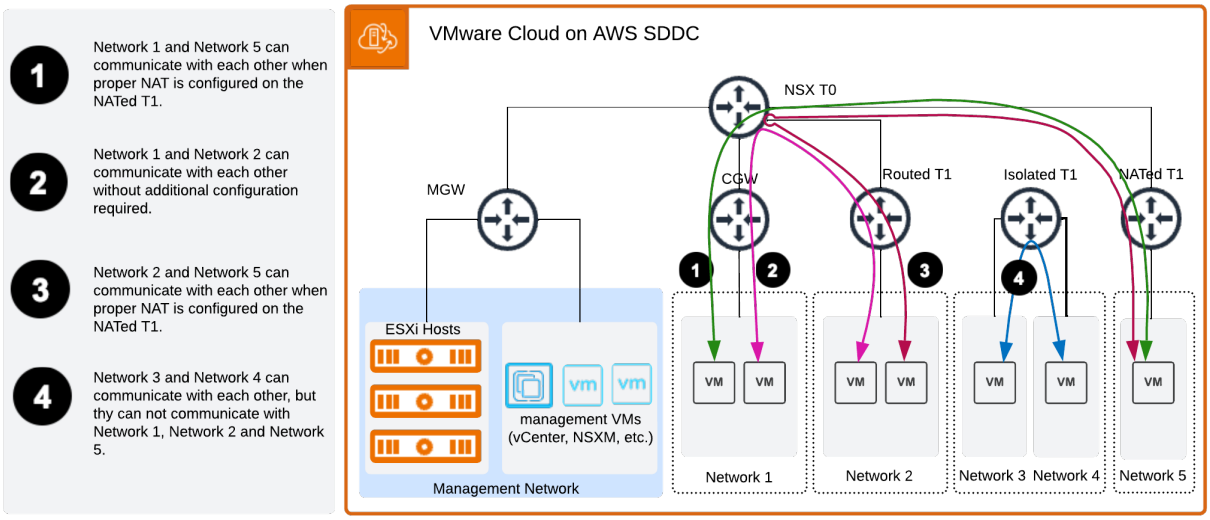
VMware Cloud on AWS also allows customers to disable the default Unicast Reverse Path Forwarding setting (URPF) strict mode at Services Uplink or Intranet Uplink to support non-local source-IP traffic.

Custom T1 Gateway

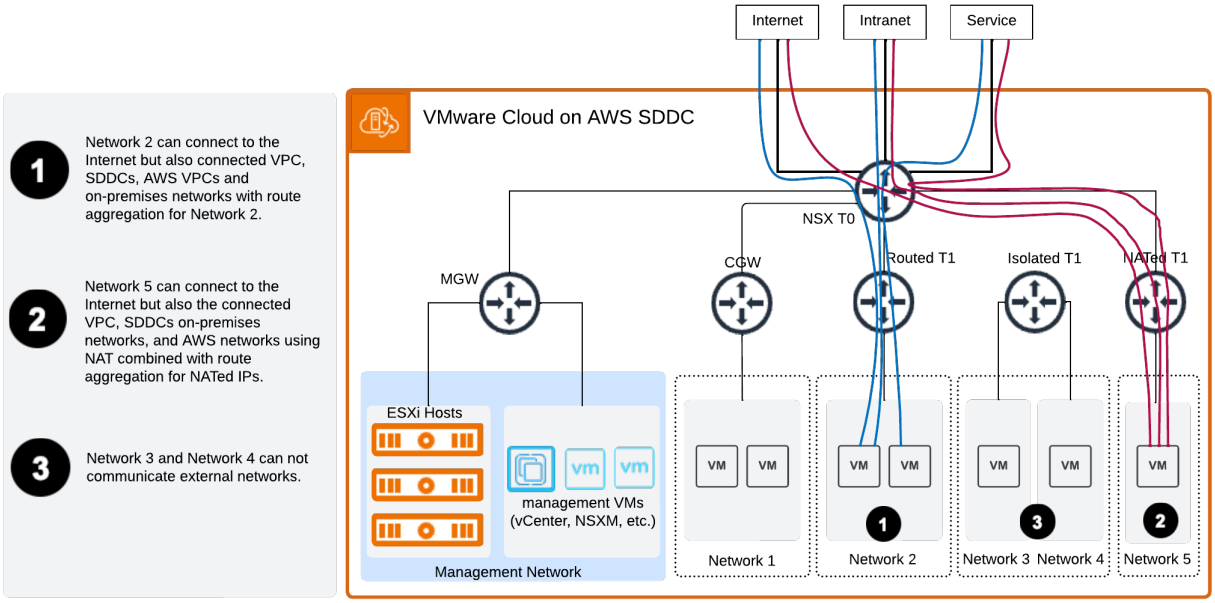
When an SDDC is provisioned, only the MGW and CGW are initially constructed. Additional custom T1 routers can be added to allow customers to extend the default network topology and support new use cases, such as multi-tenancy and disaster recovery testing. Three types of T1 gateways are supported: Routed, Isolated and NATed, each offering distinct networking functionalities:

- Routed T1: Traffic from network segments connected to a routed T1 is routed through it.
- Isolated T1: Traffic from network segments connected to an isolated T1 can't traverse the gateway. However, local segments connected to the same isolated T1 can communicate with each other.
- NATed T1: Traffic from network segments connected to a NATed T1 can't traverse until NAT rules are created for the segments.

Both the routed T1 and NATed T1 are connected to the NSX T0, while the isolated T1 is not. Routes for compute networks on the routed T1 and NATed IPs on the NATed T1 will be added to the NSX T0 route table. Conversely, routes for compute networks on the NATed T1 will not be added, and no routes from the isolated T1 will be included in the NSX T0 router. This behaviour determines communication within a VMware Cloud on AWS SDDC, as shown in the diagram below.



In terms of external connectivity, routes from custom T1 routers will not be advertised over the Intranet and Service uplink interfaces. Route aggregation is necessary to enable external connectivity for networks on routed T1s and for NATed IPs configured on NATed T1s. Route aggregation for both scenarios will be covered in a later section.



Unlike the default CGW, static routes are permitted on these custom T1s. For example, a default route can be added to an isolated T1 to direct traffic to a customer-managed network appliance, such as a VMware NSX ALB load balancer. This setup makes the creation of an inline load balancing topology possible in VMware Cloud on AWS. For more information, please refer to Designlet: VMware Cloud on AWS Static Routing on Multiple CGWs.

These custom T1s support gateway firewall and VPN functionalities.

By default, the gateway firewall on these T1s is configured to allow all traffic. Since custom T1s utilize the NSX T0 for external connectivity and CGW firewall rules are applied on NSX T0's uplink interfaces, the CGW firewall must be configured to permit external communication traffic from custom T1 workloads.

Regarding VPN support, custom routed and NATed T1s support policy-based, route-based IPSec VPN (RBVPN), and L2VPN, serving as an L2VPN server. It is important to note that these T1s do not support dynamic routing protocols, such as BGP. Consequently, auto-failover between two RBVPNs is not possible.

It is worth pointing out that all T0 and T1 SR routers, including the MGW, CGW, and customer T1s, run on the same edge appliances, thereby sharing the same underlying computing resources.

Route Aggregation and Egress Filtering

In VMware Cloud on AWS, route aggregation and egress filtering can be used to control the set of routes advertised to SDDC network uplinks including AWS Direct Connect, VMware Transit Connect and the Connected VPC. The two features can help reduce the size of the route table of these external entities.

Route aggregation aggregates prefixes behind CGW and custom T1s. The aggregated routes will be advertised to the selected connectivity endpoints (Intranet or Services). When 'Intranet' is chosen as the endpoint, route aggregation applies the advertised routes over Direct Connect and VMware Transit Connect. Selecting 'Services' as the endpoint means applying the advertised routes to the connected VPC. Route aggregation is essential to provide connectivity for Routed and NATed T1 routes over AWS Direct Connect or VTC. One significant distinction between this route aggregation feature and those implemented by other vendors is that the aggregated routes are always advertised to their respective endpoint regardless of whether the SDDC contains any networks within the aggregation. It is important to note that the management CIDR of the SDDC cannot be included in an aggregation; it will always be advertised separately from any aggregations. Attempting to create a route aggregation that uses a prefix list with a CIDR overlapping with the management CIDR will result in failure.

On the other hand, route egress filtering stops the advertising of routes over the uplinks for the default CGW. In the default configuration, all compute networks within the SDDC of the default compute gateway are advertised to the connected VPC and external connections, such as AWS Direct Connect and VMware Transit Connect. Once route egress filtering is enabled, it removes routes for all compute networks of the default CGW from the selected uplink interface. The supported uplink interfaces for this feature are Intranet Uplink and Services Uplink.

If route egress filtering is enabled and route aggregation is configured, and the filtered routes from the default compute gateway are part of the aggregated subnet, then the aggregated subnets will take precedence.

In summary, once route egress filtering is enabled, only routes for prefixes defined in the route aggregation and management networks will be advertised.

Aggregated/filtered routes can be verified against the corresponding connectivity endpoints under the Advertised routes tab in NSX manager console.

For more information about route aggregation and egress filtering, please refer to Understanding Route Aggregation in VMC on AWS and Understanding Route Filtering in VMC on AWS.

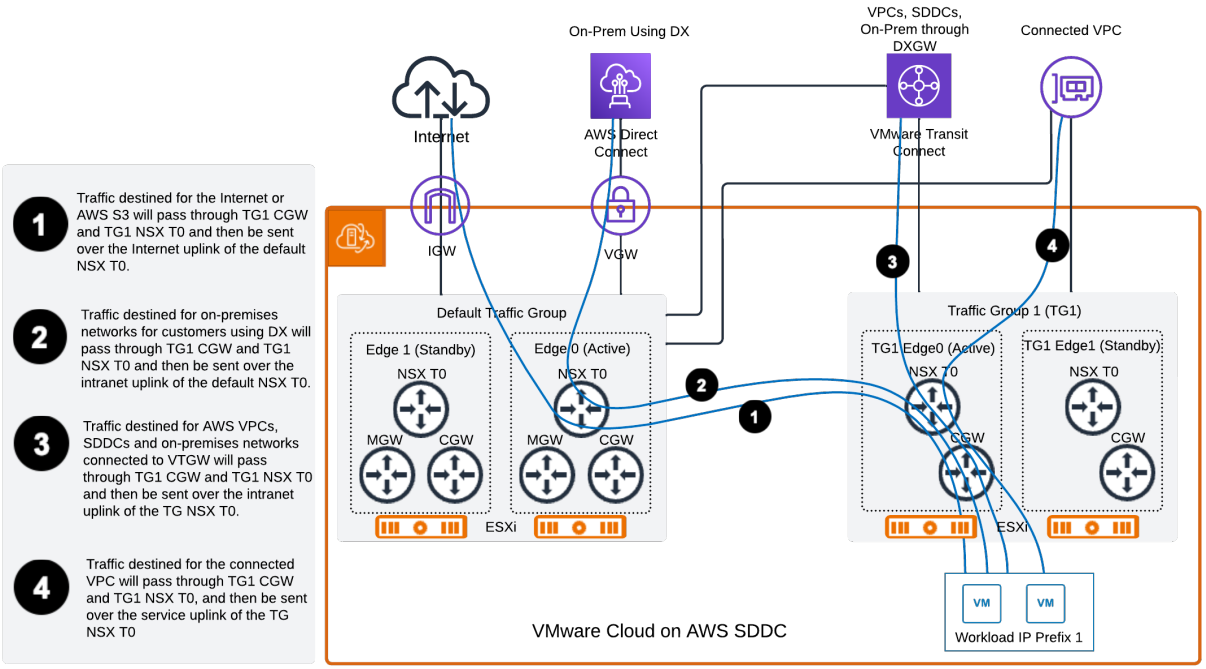
Multi-Edge SDDC

In the default configuration, an SDDC is deployed with a pair of NSX Edge appliances operating in an active/standby mode. The default edge pair is also known as the default traffic group. The Edge appliances provide the platform for the default T0 and T1 routers to run. All north-south traffic for the SDDC workloads is routed through the default T0 router. Additional edges or traffic groups can be added for certain use cases. For example, large SDDCs may require high north-south throughput or need traffic isolation, such as separating workloads backup traffic or HCX migration from application traffic.

Before adding additional NSX edges, the SDDC must connect to the VMware Managed Transit Gateway. The new edges will be deployed in the form of a traffic group, which includes two edge appliances operating in active-standby mode. VMware Cloud on AWS leverages Source-Based Routing (SBR) to direct traffic from workloads to the desired traffic group edge and destination-based routing for traffic destined to workloads. To map workloads to a traffic group, customers must define an IP prefix or multiple prefixes to include these workloads' IPs then associate these prefixes with the traffic group.

Each non-default edge pair or traffic group requires two dedicated ESXi hosts in the management cluster. A VMware-managed anti-affinity rule is in place to ensure that no two edge appliances operate on the same host.

It's worth noting that not all types of traffic can leverage the non-default edges for external connections. This category includes Internet traffic, VPN traffic, NATed traffic, and SDDC management traffic, all of which must be forwarded from the non-default edge to the default edge. The Figure below illustrates the different data flows for workloads mapped to a non-default traffic group. For more information about multi-edge SDDC, please refer to Edge scale out with SDDC Multi-Edge in VMware Cloud on AWS.



IPv6

VMware Cloud on AWS supports dual-stack (IPv4 and IPv6) networking for an SDDC. In a dual-stack SDDC network, IPv6 is supported for workload communications on network segments connected to a custom T1 gateway, and IPv6 VMs can now communicate with management appliances such as vCenter (on IPv4) using SRE-configured NAT64 rules. IPv6 is also supported for SDDC communication over AWS Direct Connect and VMware Transit Connect.

For more information of using IPv6 in VMware Cloud on AWS, please refer to Designlet: Understanding IPv6 in VMware Cloud on AWS.

Stretched Clusters

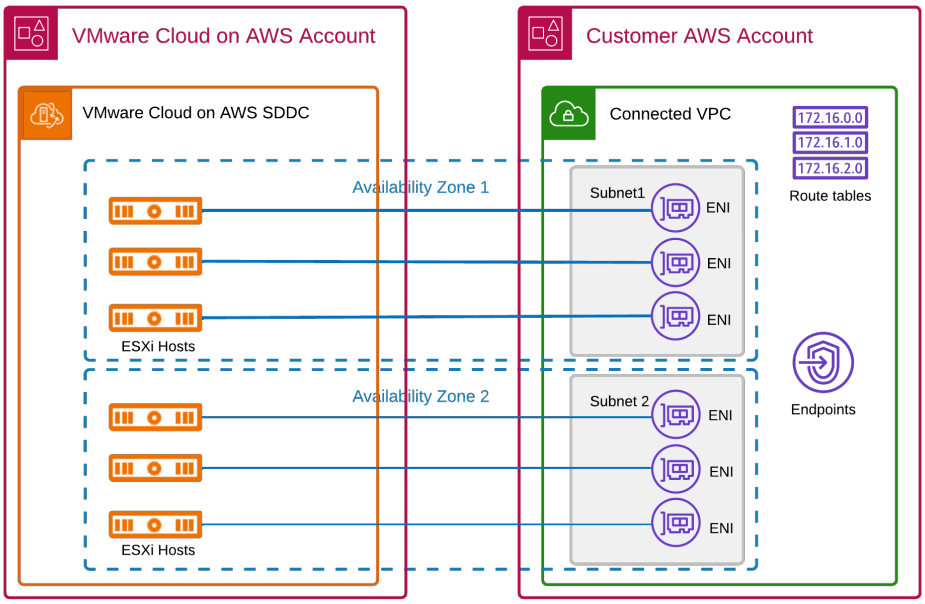
As a resilience solution, stretched clusters for VMware Cloud on AWS are designed to provide data redundancy and protect against an AZ failure. It also enables workloads to vMotion (live migrate) between AZs through the use of NSX overlay networking.

Stretched cluster SDDCs are implemented using a vSAN feature of the same name. Per the requirements of vSAN, the SDDC provides two data sites and one witness host per cluster. The data sites are composed of two groups of hosts, evenly split between two AZs. The witness host is implemented behind the scenes, using a custom EC2 instance, and is deployed to a 3rd AZ that is separate from the data sites. It's important to note that the witness EC2 instance is managed by VMware Cloud on AWS at no extra cost.

The vCenter, NSX Managers and NSX edge appliances are typically deployed in the same AZ. In the event of an AZ failure, if vCenter, NSX Managers, and NSX edge appliances are in the impacted AZ, vSphere HA will restart these appliances and customers' workloads in the surviving AZ. These management appliances and NSX edge appliances will continue to operate there, even after the affected AZ is restored, until a failure or other scenario necessitates their recovery in another AZ.

Stretched Cluster Connected VPC

Stretched clusters require an Amazon VPC with two subnets, one subnet per availability zone. The subnets determine ESXi hosts placement between the two AZs. In the event that the active edge appliance migrates between hosts or AZs, the route tables of the connected VPC will be updated to route traffic through the appropriate cross-link ENI.



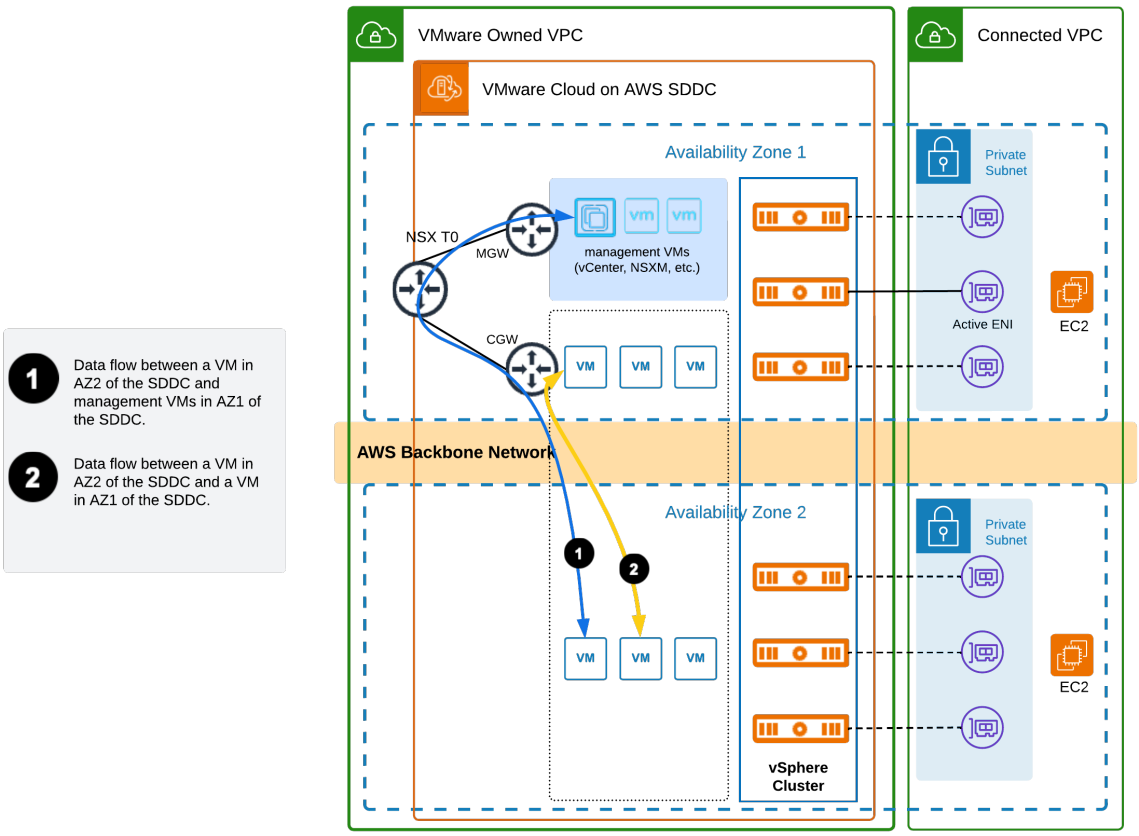
Cross-AZ Data Flows

As the NSX edge appliances operate in active/standby mode, the active edge is confined to a single AZ. Consequently, this configuration results in cross-AZ data flows within stretched cluster SDDCs.

Cross-AZ East-West Flows

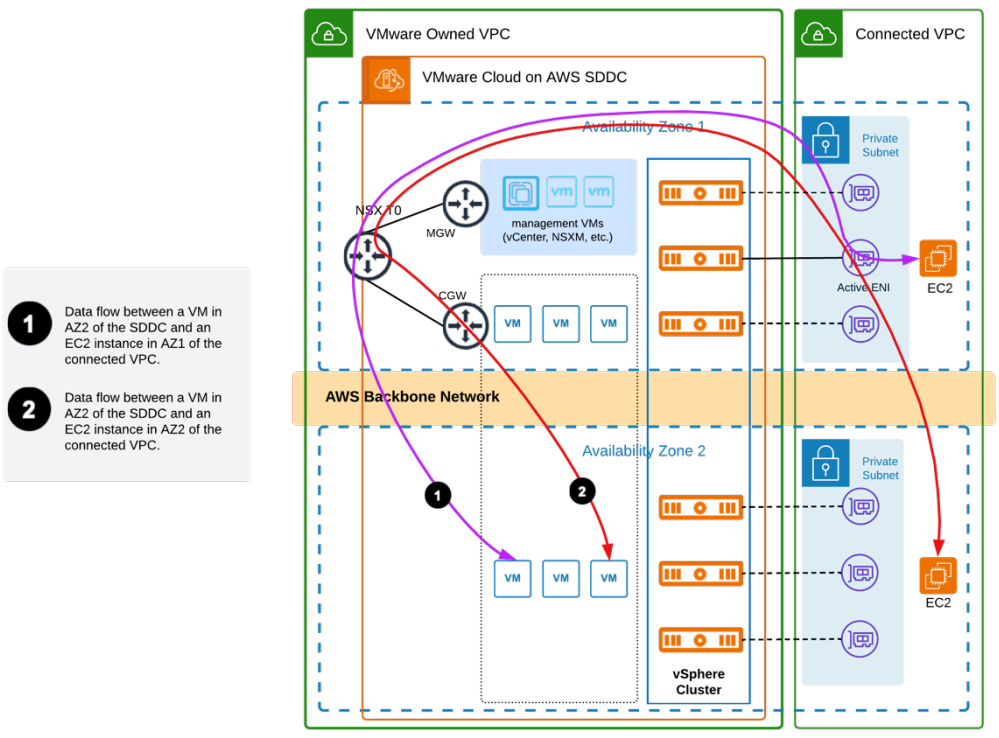
East-west communications between the workloads, as well as between SDDC management appliances such as vCenter and NSX Manager. within an SDDC whenever the source and destination reside in different AZs.

The figure below shows the cross-AZ data flows within a stretched cluster SDDC.

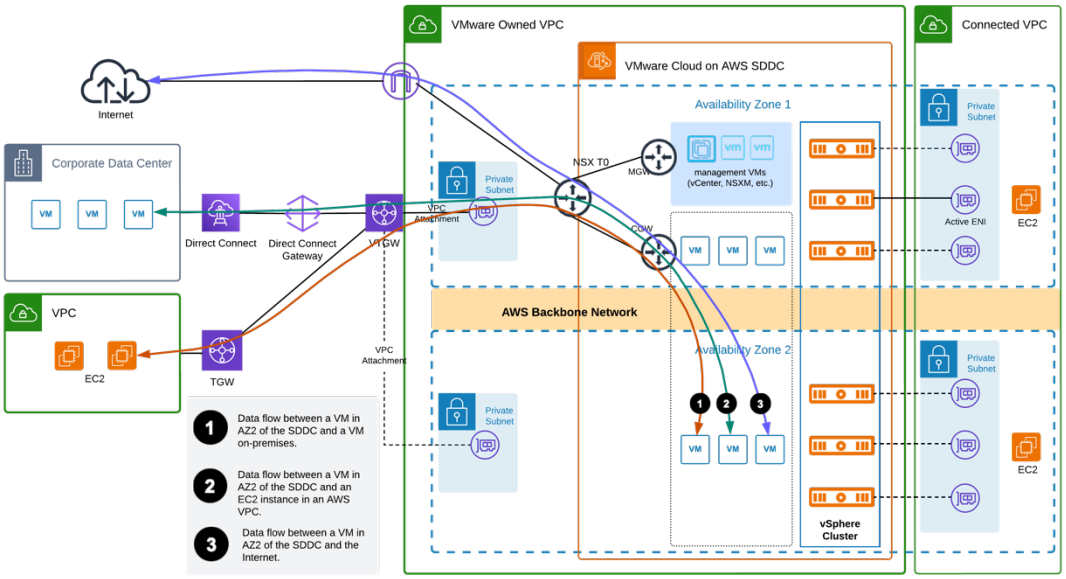


Cross-AZ North-South Flows

North-south traffic for workloads which reside in a different AZ than the NSX active edge appliance. This includes all traffic which must pass through the edge, including internet, Direct Connect, and traffic to the connected VPC or other customer owned VPCs. The subsequent two figures illustrate the above data flows. The first figure shows data flows from a VM in AZ2 to EC2 instances in AZ1 and AZ2.



The second figure details data flows from a VM in AZ2 to the Internet, on-premises, and Amazon VPCs.



vSAN Replication Traffic

Traffic for vSAN synchronous writes between the hosts of a stretched cluster SDDC will result in cross-AZ traffic flows. The amount of traffic heavily depends on the storage I/O profile of your workloads. The new vSAN Express Storage Architect (ESA) has reduced the replicated traffic with the redesign of data compression mechanism.

Please note that any stretched cluster deployed on VMware Cloud on AWS includes ten petabytes per month of Cross-AZ data transfer free of charge.

Cross-AZ Network Latency

The AWS cross-AZ network latency is typically less than 10ms. It's crucial to note that this latency affects both network performance and the write performance of vSAN when designing applications on VMware Cloud on AWS.

VMware Transit Connect

SDDC Group

An SDDC deployment group (SDDC Group) is a logical entity that contains multiple VMware Cloud on AWS SDDC members. It is designed to simplify the management of your organization's VMware Cloud on AWS resources at scale. SDDC groups are organization-level objects and cannot contain SDDCs from more than one organization. Additionally, an SDDC group can include members from up to three AWS regions.

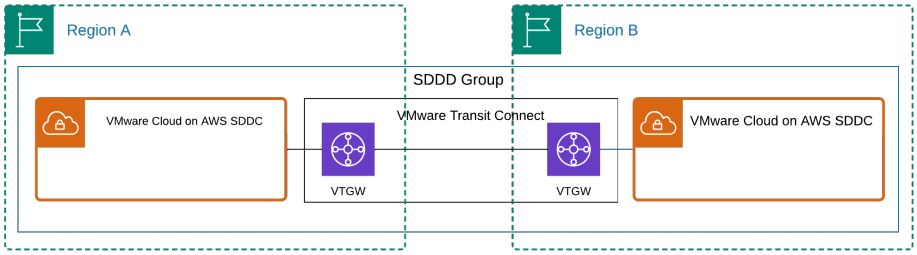
VMware Transit Connect

VMware Transit Connect provides high bandwidth, resilient connectivity to SDDCs in an SDDC Group. As a VMware managed service for SDDC Groups connectivity, VMware Transit Connect automates the provisioning and controls to interconnect SDDCs. It simplifies connectivity to Amazon VPCs as well as on-premises networks over an AWS Direct Connect Gateway. With a centralized, yet highly available connectivity design, VMware Transit Connect scales easily as you add SDDCs, VPCs and on-premises networks to the Group.

VMware Transit Connect Connectivity

Under the hood, VMware Transit Connect uses the AWS Transit Gateway (TGW) to construct the required connectivity. These VMware-managed transit gateways are also known as VMware Transit Gateway (VTGW).

As a regional resource, one VTGW will be deployed in each region where there are member SDDCs in the SDDC group. VMware Transit Connect will interconnect these VTGW through AWS TGW inter-region peering and add necessary routes to enable interconnectivity for all member SDDCs across all regions.



A VMware Cloud on AWS SDDC will be connected to the VTGW through a VPC attachment. The attachment is completed when the SDDC is added as a member of the SDDC group. In a multi-region SDDC group, each member SDDC will connect to its respective VTGW in its own AWS region.

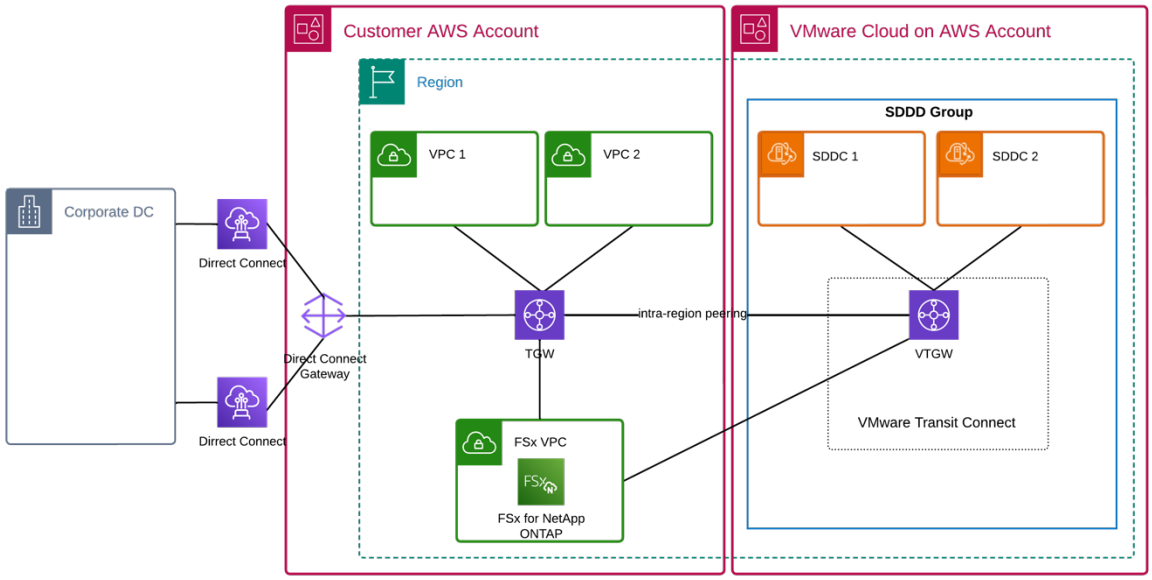
Additionally, customers VPCs, AWS Transit Gateway, and AWS Direct Connect Gateway (DXGW) can all be connected as external connections to VTGW.

- Customers VPC connections are achieved through resource sharing from the VMware AWS account to the customer-owned AWS account. Once the VTGW resource sharing request is accepted in the customers AWS Resource Access Manager, customer VPCs can be attached to VTGW through a VPC attachment from the AWS console. The attachment request must be accepted in VMware Cloud on AWS before the attachment is completed.
- In terms of interconnectivity for customers AWS Transit Connect Gateways, these TGWs would be interconnected to VTGW through TGW peering. In a multi-region SDDC group, the choice of region for AWS TGW connection to is configurable.
- Similarly, VTGW can be associated with customers DXGW through a TGW association. In the association configuration, 'allow prefixes' will be defined to specify which routes will be advertised to on-premises networks over the DXGW, with an AWS limit of up to 20 prefixes.

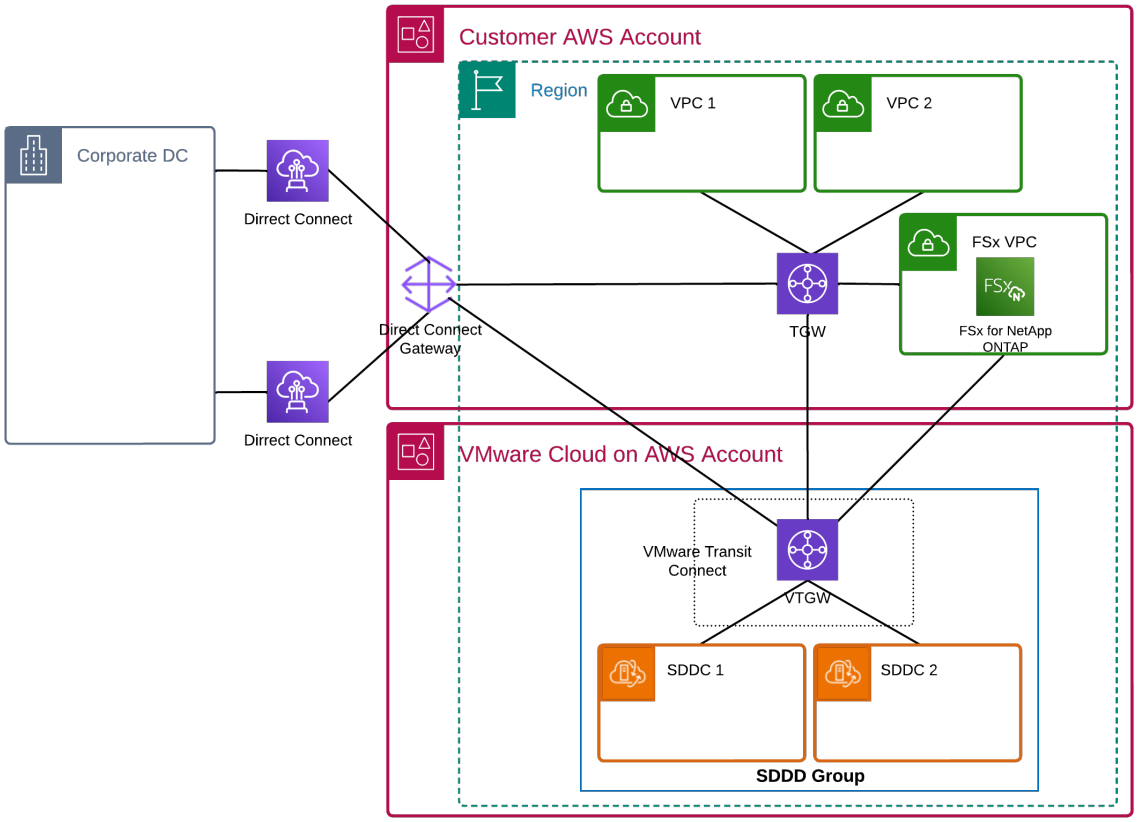
The two figures below illustrate how SDDCs, Amazon VPCs, and the AWS Direct Connect Gateway are interconnected using

VMware Transit Connect in a single-region SDDC group setup. In this configuration, the Amazon FSx for NetApp ONTAP service, deployed in an Amazon VPC named “FSx VPC”, provides NFS data stores to SDDC 1 and SDDC 2.

In the first connectivity design option, the TGW is connected to a DXGW, and VTGW uses the TGW as its next-hop for connectivity to on-premises networks.



When the VTGW requires its own DXGW attachment, the second design option, as depicted in the subsequent figure, can be used to achieve the desired connectivity. In this design, both TGW and VTGW are attached to the same DXGW. It is important to note that a configuration where the TGW and VTGW are attached to different Direct Connect Gateways is also a supported design.



VMware Transit Connect Route Tables

VMware Transit Connect defines two route tables on each VTGW: the Members route table and the External route table. VMware Cloud on AWS SDDCs will use the Members route table for traffic forwarding. Conversely, Amazon VPCs and the associated AWS Direct Connect Gateway will use the External route table. Routing between VTGW and AWS TGW also uses the external route table.

By default, routes for an SDDC's management CIDR, all compute networks on CGW (excluding custom T1s), and route aggregation prefixes will be added to both Members and External route tables. Employing route egress filtering with route aggregation can help reduce the size of VTGW route tables while maintaining the necessary connectivity for a large-scale environment, where the default limit of 1000 route entries may not meet the requirements.

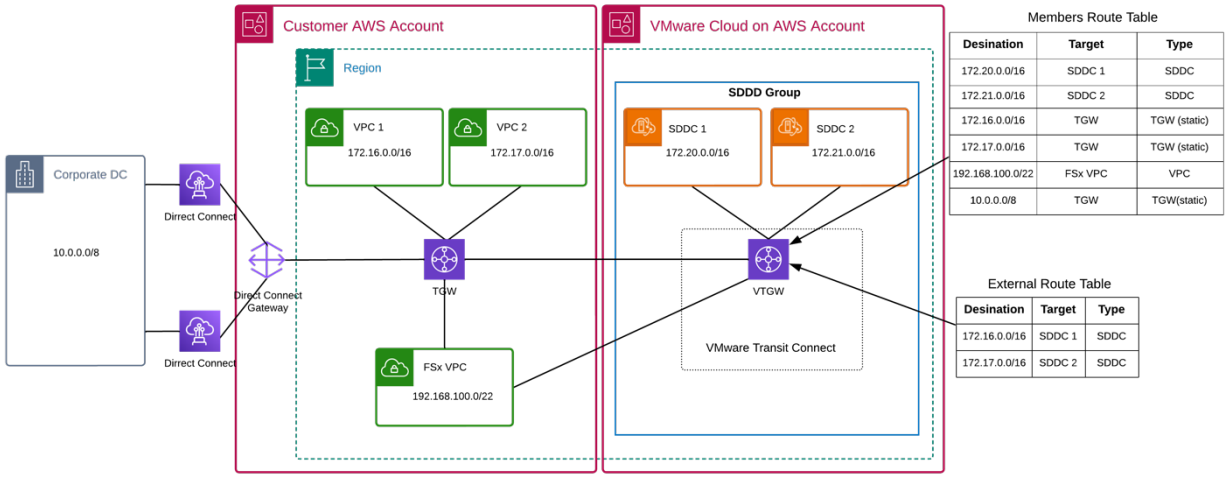
Once an Amazon VPC is attached to a VTGW, both the primary CIDR and secondary CIDRs of the VPC will be automatically added to VTGW's Members route table by default. If the attached VPC is designed as a transit VPC or a security VPC for secure Internet access, static routes or a default route can be added. These added static or default routes will stop the VPC CIDRs from being automatically included in the Members route table. It is important to note that static routes must also be configured in the VPC route tables to direct traffic from the VPCs to the SDDCs.

AWS DXGW will add its learned routes from on-premises network into the VTGW's Member route table automatically.

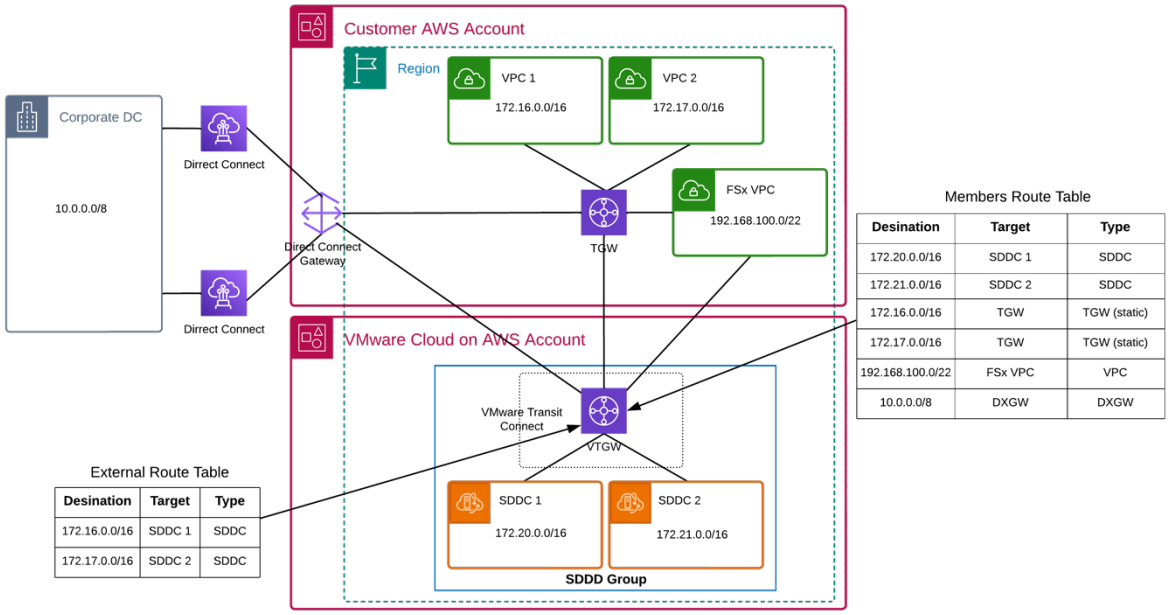
When adding static routes with the next-hop pointing to an AWS Transit Gateway on a VTGW, the static routes will be only added to the VTGW Members route table.

In summary, the Members route table includes routes to all member SDDCs, Amazon VPCs, on-premises networks, and external networks such as the Internet. The External route table only includes routes to SDDCs.

The figure below illustrates the route tables of VMware Transit Connect in a single-region SDDC group setup for the first network design.

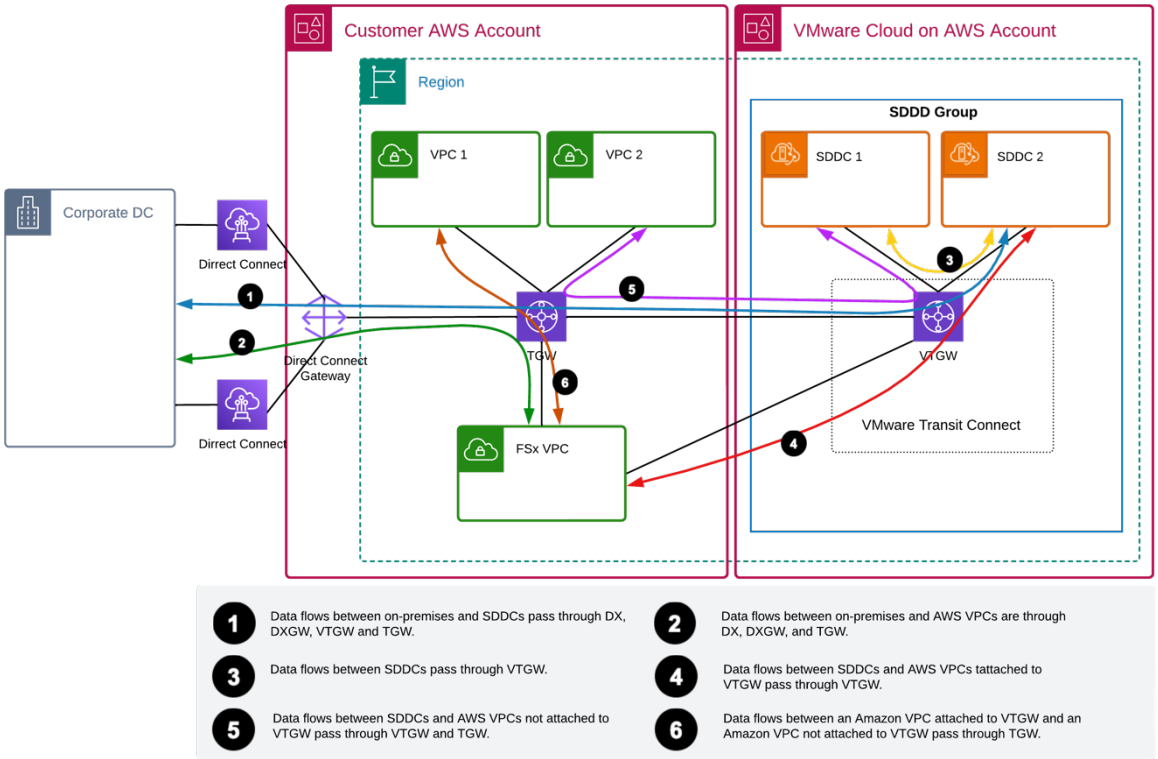


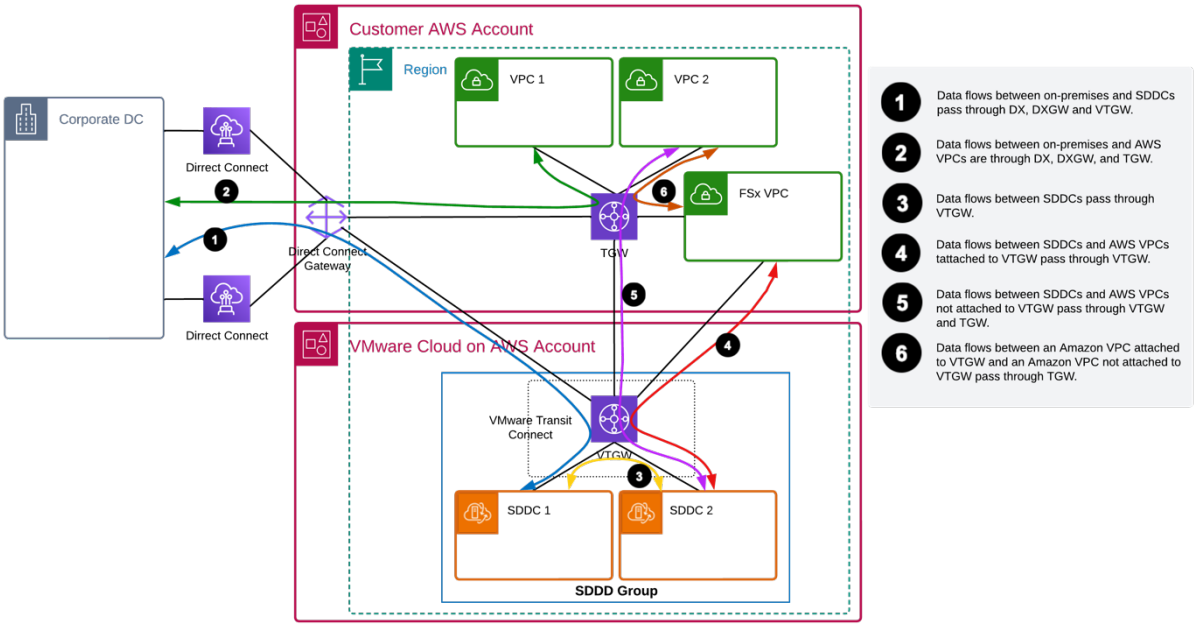
The route tables of the second network design are displayed in the following figure.



VMware Transit Connect Data Flows

The figures below illustrate the data flows between on-premises networks, SDDCs, and VPCs for network design option 1 and 2. In these designs, a VPC dedicated to Amazon FSx is used to provide NFS services to the SDDCs.





It's important to note that VMware Transit Connect only supports data flows that originate from or are destined for an SDDC. Connectivity between VPCs, as well as between on-premises networks and VPCs, requires the support of a TGW.

VMware Transit Connect Shared Prefix Lists

VMware Transit Connect allows network administrators to generate a prefix list for each AWS region and share this list with customer AWS accounts via resource sharing. The prefix list includes all advertised networks from member SDDCs within an SDDC group. Once the resource share accepted, it can be used to configure AWS security groups or routes within Amazon VPCs. This capability significantly reduces the complexity of AWS configurations. Furthermore, when customers make any network changes in VMware Cloud on AWS, this feature ensures that the relevant AWS configurations are automatically updated, thus eliminating the need for manual intervention.

For more information about VMware Transit Connect design, please refer to Designlet: VMware Transit Connect for VMware Cloud on AWS.

Route Priority

When the NSX learns the same route from different sources, the NSX T0 router will choose the route based on its source in the following order:

- Connected VPC
- Route-based VPN (“Use VPN as backup to Direct Connect” is OFF)
- Policy-based VPN *
- VMware Transit Connect
- Direct Connect Private VIF
- Route-based VPN (“Use VPN as Backup to Direct Connect” is ON) **
- AWS Internet Gateway ***

*IPsec VPN will never be used for traffic from ESXi hosts if SDDC is a member of an SDDC Group or has a DX Private VIF attached.

** Use VPN as Backup to Direct Connect option is not applied to routes learned through VMware Transit Connect.

*** Default Route only.

Network Services

VPC Peering for External Storage Connection

VMware Cloud on AWS SDDC supports the use of Amazon FSx for NetApp ONTAP as customer managed external storage. This integration enables customers to attach the NFS datastore to a vSphere cluster in an SDDC, providing a flexible and high-performance virtualized storage infrastructure that scales independently from compute resources.

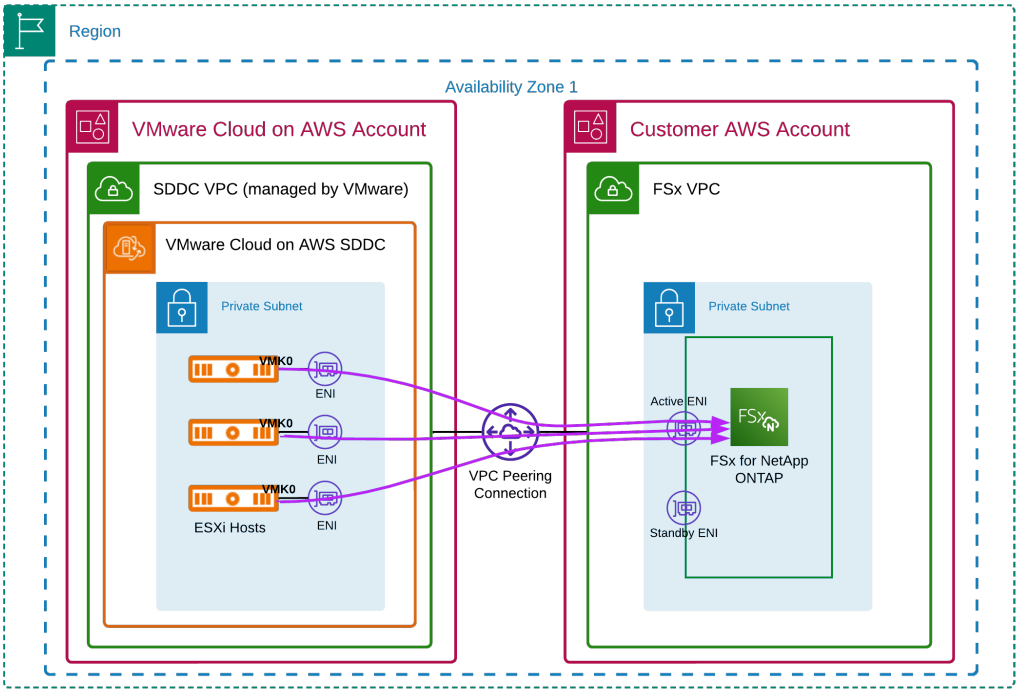
For optimal performance, a high-throughput and low-latency connection between the ESXi hosts and the NetApp ONTAP file system, running within a customer-owned VPC, is crucial. It's important to note that the SDDC and the peered Amazon VPC must not have any overlapped CIDRs, and the peered Amazon VPC must not contain the default 172.31.0.0/16 CIDR, to ensure a proper integration.

Two supported connectivity options are available:

- VPC Peering: only support for single-AZ deployments of FSx for NetApp ONTAP.
- VMware Managed Transit Gateway through SDDC groups, as detailed in the VMware Transit Connect section, is suitable for multi-AZ deployments of FSx for NetApp ONTAP.

A VPC peering connection is a networking connection between two VPCs that enables you to route traffic between them privately. In the integration between the SDDC and FSx for NetApp ONTAP, connectivity traffic is confined to the SDDC management subnet, permitting only NFS traffic. This results in a line-rate connection between the SDDC and the FSx for NetApp ONTAP VPC, enabling direct Elastic Network Interface ENI to ENI communication with sub-millisecond latency.

When setting up this cross-account VPC peering, the SDDC VPC act as the requestor, and the FSx NetApp ONTAP VPCs serves as the accepter. Once customers accept the connection, VMware Cloud on AWS automatically configures the routing to direct traffic from the SDDC to the peered VPC over the VPC peering connection. Additionally, security groups and network access control list are set up to permit NFS traffic through this VPC peering connection. In the FSx NetApp ONTAP VPC, customers must update the security group associated with the ONTAP system to allow NFS inbound traffic from the SDDC management network and add new routes for the SDDC management network pointing to the VPC peering connection.



VMware Cloud on AWS has enabled Jumbo Frames (MTU 8500) support on the VMKernel interface (vmk0) to improve NFS performance since version 1.24. Additionally, multiple TCP connections can be established by use of nConnect from each ESXi host to the NFS storage, increasing the aggregated throughput at the ESXi host level.

For more information on setting up VPC peering, please refer to VMware Cloud on AWS integration with Amazon FSx for NetApp

ONTAP Deployment Guide.

VPN

The SDDC's default NSX T0 router supports both policy-based VPN and route-based IPSec VPN. The VPN services run in an active-standby mode on the NSX edge appliance pair to minimize the down time in the event of an active edge failure. For route-based VPN, the supported dynamic routing protocol is the Border Gateway Protocol (BGP).

Similarly, both routed and NATed custom T1 gateways support policy-based VPN and route-based IPSec VPN. However, only static routes are supported for route-based VPN on custom T1 gateways, as stated in the custom T1 section.

Route based IPSec VPN is the preferred option as it offers better flexibility and scalability.

Regarding the NSX L2VPN, both the NSX T0 router and the custom T1 gateway support it, with VMware Cloud on AWS SDDC serving as an L2 VPN server.

Both IPSec VPN and NSX L2 VPN can be terminated at either a public or private IP. This means that the VPN tunnel can be established through the Internet or a private link, such as AWS Direct Connect.

VMware Cloud on AWS supports Equal-Cost Multi-Path routing (ECMP) with route-based IPSec VPN on T0 router for increased bandwidth and resiliency.

By default, routes learned via route-based VPN are preferred over AWS Direct Connect. This default behaviour can be changed by enabling “Use VPN as backup to Direct Connect” option in the Direct Connect Tab in the NSX manager.

VMware Cloud on AWS allows the use of either a preshared key or a certificate for the IPSec gateways to authenticate each other.

For more information about using VPN in a VMware Cloud on AWS SDDC, please refer to Designlet: VMware Cloud on AWS SDDC Connectivity With IPSec VPN.

Network Address Translation (NAT)

VMware Cloud on SDDC provides NAT services through the compute gateway and custom T1 gateways. By default, a pre-configured Source NAT (SNAT) rule exists for all workloads within the SDDC, enabling workloads to connect to the Internet once the specific data flows are explicitly permitted on the compute gateway firewall. This default SNAT can be overridden by a static NAT.

Destination NAT (DNAT) rules can be added to enable inbound connections from the Internet.

Since compute gateway NAT rules are applied to an SDDC’s Internet Uplink Interface, these rules won’t apply to the Internet outbound traffic if the default route does not utilize the Internet Uplink interface.

DHCP and DHCP Relay

NSX in VMware Cloud on AWS provides both DHCP and DHCP relay services. A default DHCP server is created during SDDC provisioning. Additional DHCP servers can be created as needed by creating a new DHCP server profile.

When a customer-managed DHCP service is preferred, DHCP relay can be configured to forward DHCP requests to specified DHCP servers.

DNS

VMware Cloud on AWS SDDC offers DNS services, where CGW and MGW will act as caching DNS forwarders, relaying the queries from virtual machines to the actual DNS servers specified. The DNS forwarders’ IPs are displayed under the DNS Service tab in NSX Manager.

MGW DNS Servers will be used by the management appliances like vCenter to resolve the on-prem fully qualified domain names (FQDNs). Features such as vCenter Hybrid link mode may not work until customer managed DNS servers are configured here.

Multiple DNS zones can be created for CGW DNS forwarder to forward the DNS queries for different domains to different DNS servers.

Broadcom owns and manages the DNS domain vmwarevmc.com, a publicly resolvable DNS zone used for SDDC management appliances. Unique DNS records are created for each SDDC, including those for vCenter and other management appliances. By default, the FQDNs of vCenter and HCX manager resolve to a public IP address. Customers can switch the resolved IP to the private IP of vCenter and HCX Manager to facilitate their access to the systems via a private link such as AWS Direct Connect. It's important to note that switching to private IPs for DNS resolution does not restrict access to the service; access control is managed through the MGW firewall.

Public IP

VMware Cloud on AWS customers can request public IPs for their workloads inbound or outbound Internet connectivity. These public IPs are AWS Elastic IPs. Broadcom passes the standard AWS cost of a public IP on to customers.

Author and Contributors

David Zhang

Ron Fuller

Michael Kolos

Oleg Ulyanov

Daniel Rethmeier

Gaurav Jindal