VMware Cloud on AWS: SDDC Design

VMware Architecture

Table of contents

VMware Cloud on AWS: SDDC Design 3
Key Deliverables 3
SDDC Sizing 4
Compute 5
- Selecting Host Instance Types 5
- Designing a Cluster Layout 5
- Determining an EDRS Strategy 5

Storage 6
Network 7
- IP Administration 7
- DNS 7
- DHCP 7
- SDDC Interconnectivity 7
- Considerations for IPSec VPN 7
- Considerations for Direct Connect (AWS) 8
- A Note on ECMP 8
- Network Capacity Planning 8
- Considerations for SDDC Edge 10
- Common Traffic Flows 11

Cloud Service Integration (AWS) 12
Network Security 13
Authors and Contributors 14

VMware Cloud on AWS: SDDC Design

Key Deliverables

There are a number of factors to consider when designing an SDDC. In addition to the obvious design decisions surrounding the server infrastructure itself, there are several less obvious considerations involving networking and 3rd party service integration.

This document is intended to highlight the most critical factors for SDDC design.

The key deliverables of SDDC design are as follows:

1. Determine the size requirements of the SDDC in terms of compute, storage, and network performance.
2. Develop an IP addressing scheme for the SDDC.
3. Determine a strategy for core services such as DNS and DHCP.
4. Determine a strategy for SDDC interconnectivity (connecting to on-premises, other cloud services, etc.).
5. Determine a strategy for service integration (e.g., cloud or other network services).
6. Develop a network security plan.

SDDC Sizing

A first step toward determining the overall footprint of an SDDC will be to gather data surrounding the expected CPU, memory, and storage consumption for workloads. For greenfield deployments, this may involve the manual process of interviewing the planned tenants of the SDDC. For brownfield (aka. migration) deployments, it is better to perform an audit of the current production environment and use this as the basis for sizing the new SDDC.

While the requirements for CPU and memory usage are relatively straight-forward to calculate, sizing exercises for storage are a bit trickier. Since the SDDC utilizes vSAN for storage, the storage policies used within the SDDC will impact the overall capacity available to workloads. It is important to understand the design implications of each option for vSAN storage policies and determine a strategy for storage policy assignment for workloads. See the document on vSAN architecture for more information.

When sizing the SDDC, keep the following points in mind:

- Remember to account for management overhead of the SDDC. This includes components such as vCenter, NSX, HCX, and any other appliances that will be required to operate the SDDC.
- Carefully plan your vSAN storage policies. The structure of these policies will impact the overall available storage capacity of the environment.
- Capture storage I/O requirements for workloads. High I/O workloads may require storage policies designed to optimize for performance.
- Map storage I/O requirements of workloads to host instance types that can meet those requirements. Some instance types offer only slower EBS-based storage that has lower performance than localized flash-based storage.

Compute

Selecting Host Instance Types

Once a basic sizing exercise has been performed, the next step will be to determine the host instance types which will be used within the SDDC. The available instance types vary not only in terms of the resources they contribute to SDDC, but also in terms of design considerations specific to a given instance type. You should understand the instance types available to you and the resources they contribute to the SDDC. This information will enable you to determine the total host requirements for the SDDC.

Designing a Cluster Layout

Clusters enable the vSphere High Availability (HA) and vSphere Distributed Resource Scheduler (DRS) features within an SDDC. As hosts are added to a Cluster, its total CPU, memory, and storage capacities are increased per the specifications for the instance type of that host. In addition to the standard design recommendations surrounding vSphere Clusters, there are a few additional considerations that are specific to cloud based SDDCs. The following items are of particular importance.

- Cluster limits - There are limits to both the hosts per Cluster and total Clusters per SDDC. Refer to the configmax document for details.
- Standard vs Stretched Cluster - You must understand the difference between the two available Cluster layouts for an SDDC. Refer to the architectural guide on stretched clusters for more information.

Determining an EDRS Strategy

EDRS is a feature which auto-scales an SDDC by dynamically adding and removing hosts based on resource contention within a given Cluster. While there are only a small number of configuration profiles available for EDRS, it is important to understand the differences between them and how a given profile will impact the design.

Storage

You must manage storage policies carefully, not only during the initial design and deploy, but as part of a long-term operations strategy. Consider the following recommendations.

- Deploy clusters with a minimum set of hosts to support your storage policies. For example, if you intend to use RAID-6 then deploy the cluster with the minimum hosts to support this policy.
- Larger deployments with multiple clusters must consider that a VM storage policy can be assigned to any cluster. This means considering everywhere a policy is being used before making any alterations. There are two strategies for managing and mitigating this complication: either standardize on any new clusters being at least 6-hosts or integrate the cluster name into the policies themselves. ***Do not use force provisioning to overcome compatibility issues.***
- Use erasure coding policies by default whenever possible to bias the cluster to capacity efficiency; however, do note that some applications with high I/O requirements perform better with a mirroring policy.
- Create policies based on the scope of management not based on the settings. For example, create a "VDI" policy, not a "raid-5" policy. Your policies are the scope of change, by preemptively sorting important workloads into differentiated policies you empower future change.
- Do not disable checksum unless data integrity is not important and silent data corruption is mitigated in the application layer.

Network

IP Administration

In some ways you can think of each SDDC you create as if it were just another remote data center. In that sense you should think about IP address management just as you would with any other traditional data center.

As part of the SDDC deployment process you are required to specify an IP range that will be used for the Management Network of the SDDC. The choice of address space is extremely important since it cannot be changed without destroying the entire SDDC and rebuilding it from scratch. Here are some considerations when deciding upon the address space to use:

- Size - The range needs to be large enough to facilitate all hosts which will be deployed on day 1, but also must account for future growth.
- Uniqueness - You should ideally provision an IP range that is unique within your organization.
- Ability to summarize - Ideally this block should be a subnet of some larger space which is allocated to the SDDC as a whole. By subnetting a larger dedicated supernet you will gain the ability to simplify routing between your on-premises environment and the SDDC, and you will potentially simplify network security policies used to secure the SDDC.

DNS

By default, the SDDC is configured to use public DNS severs; however, these settings may be changed. Here are a few considerations for planning DNS services within the SDDC:

- The DNS servers must be reachable; either via public IPs or via one of the models described by the section on Shared Services in the document on solution design.
- The DNS servers must support recursive queries.
- The SDDC is pre-configured to internally resolve hosts within the vmc.local domain. All other domains require an external DNS server.
- Network segments within the Compute Network which have DHCP enabled will use the DNS servers configured on the SDDC.

DHCP

Network segments within the compute network may be configured to provide basic DHCP services. If your design requires a more advanced DHCP feature set, then the SDDC may be configured to provide DHCP relay services. Please refer to the section on Shared Services in the document on solution design for considerations on how to deploy a DHCP server for relay services.

SDDC Interconnectivity

Planning for SDDC interconnectivity involves integrating the SDDC with the external world and making it accessible both for management and workload traffic. When planning for SDDC interconnectivity, you must work within the options provided for edge uplinks. These are as follows:

- Internet – This uplink will be used for accessing the SDDC via public IP address. This may include management components such as vCenter, as well as individual workloads that have been assigned a public IP address.
- IPSec VPN – VTI uplinks for route-based IPSec VPN may be used to connect the SDDC to a remote network, another SDDC, or a local networking hub construct (eg., a Transit Gateway in AWS).
- Direct Connect (AWS) – This may be used to connect the SDDC to a remote facility, other SDDCs, or other VPCs within the region.
- VPC uplink (AWS) – Used to connect the SDDC to a cross-linked VPC for AWS service access.

Remember that traffic for all uplinks contribute to overall network utilization of the SDDC edge and that you must factor in the combined total of both ingress and egress traffic.

Considerations for IPSec VPN

IPSec VPN can easily become a bottleneck due to performance limitations on either end of the tunnel. One common strategy is to

create multiple, redundant VPN tunnels between the SDDC and a remote network. If the tunnels are terminated to separate devices, then this configuration provides additional availability as well as additional performance.

A configuration which utilizes multiple VPN tunnels will rely upon ECMP to load balance traffic between them. Prior to implementing a dual tunnel configuration, it is important to understand how EMCP functions. See the follow-on section more details.

Some key points to remember for IPSec VPN

- Performance is impacted by the crypto settings used for tunnels. Higher crypto typically means lower performance.
- Performance is typically measured as an aggregate of all tunnels on a device.
- Multiple tunnels + ECMP will provide additional availability and performance (on the remote end of the tunnel). SDDC edge will not gain additional performance with multiple tunnels.
- Performance numbers of remote network devices is vendor specific. You must gather and understand these numbers for your specific devices.

Considerations for Direct Connect (AWS)

Direct Connect (DX) integration comes with a few caveats which must be accounted for. Please keep the following in mind.

- The total number of routes learned over DX is limited. You must plan your design around these limits.
- There is no direct BGP exchange between DX and the SDDC edge. Instead, the SDDC relies on updates via AWS API. This integration introduces a certain amount of route update lag in situations where several routes are being exchanged over DX.

A Note on ECMP

Equal Cost Multipath Routing (ECMP) typically utilizes a 5-tuple hashing function (protocol, src/dst IP, src/dst port) for binding a given connection to a particular interface (and in some cases to a particular CPU core within the device). In the case of certain tunneling protocols, such as IPSec and GRE, a 3-tuple hashing function will be used (protocol, src/dst IP).

Note that nowhere in the hashing functions mentioned above is network load considered. This means that it is unreasonable to expect a perfect load balance between interfaces/cores, and that it is quite possible for 1 interface/core to be completely saturated while others are simultaneously underutilized. This is commonly referred to as link polarization and will result in some users experiencing degraded network connectivity that will appear to be random to the casual observer. The best resolution for this is to understand the flows that are being polarized and look for opportunities to change one or more factors in the 5 tuples in order to influence a different hashing decision and better utilize the available links. Note that in the case of tunneling protocols that use a 3-tuple hashing function, link polarization is even more exaggerated due to the fact that the tunneling effectively hides the various connections contained within the tunnel.

Due to the inherent limits of ECMP, it becomes extremely important to monitor the performance of each tunnel in an ECMP configuration in order to help diagnose link polarization.

Network Capacity Planning

A critical but often overlooked step in planning is to gather data on application network flows and overall network utilization within the environment. This data is critical to the planning process and is useful in determining:

- Application interdependencies which were previously unknown.
- End-user network usage patterns.
- Periodic spikes in network utilization due to backups or other events.

- Network capacity requirements for the target SDDC.

The gathering of network utilization data can take 2 forms: basic or in-depth. The quality of the overall planning will be greatly impacted by the approach taken.

Basic Utilization Data

At minimum you should make some attempt to gather network utilization data from the existing environment. Network utilization data will ideally include historical graphing for ingress/egress bit-per-second and packets-per-second on key points within the infrastructure. Key points for data gathering would include:

- VM utilization – Gather network statistics for individual VMs. This will give you a rough idea of how much network utilization a given VM generates. Aggregating this data for all VMs of an application will give you a rough ideal of total utilization for that application.
- Aggregate router utilization – Gather the statistics for router aggregation points in the network. This will give you a rough idea of east-west utilization levels for the environment.
- Edge router utilization – Gather statistics at the network edge. This will give you a rough idea of the north-south utilization of the environment.

Basic utilization data is exactly that; basic. At this level, it is impossible to effectively profile application traffic within the existing environment. However, this level of data is better than nothing and should be considered the minimum level of effort for data gathering.

In-Depth Utilization Data

An in-depth analysis of network utilization and flow patterns yields much better data than a basic analysis. This type of data gathering exercise relies on collecting and analyzing network flow data from key endpoints within the network. The resulting data will help you to profile application traffic in the environment and will help you to make projections on performance expectations for the edge router of the production SDDC. Use the following guidance for performing in-depth analysis:

- Your analysis should be based on network flow data collected from network endpoints (hypervisors). Network Insight is an ideal solution for performing this type data collection.
- Collect data for as long as possible. A minimum 24 hours, but a full month is better.
- Tailor reports to provide information on inter-application flows, intra-application flows, traffic flows between applications and end-users, and traffic flows to the internet.
- Capture data on traffic types (TCP/UDP) and average packet sizes.
- Capture utilization rates in terms of ingress/egress bits-per-second and packets-per-second.

Other Considerations

No matter the data gathering approach chosen, the key will be to account for ***peak*** utilization. Peak utilization rates will help you to determine minimum performance requirements of the SDDC while it is under stress.

An additional consideration for network capacity planning is factoring in how traffic patterns may change once workloads have been deployed within the SDDC. Consider the following points:

- How will network utilization be impacted by end users access the SDDC from a remote environment?
- How will network utilization be impacted by inter-application traffic between fault domains of a stretched-cluster SDDC? Between SDDCs? Between an SDDC and other cloud services?

- How will traffic bursts from periodic backups impact the SDDC?
- How will SDDC management and vSAN replication traffic impact the SDDC?

Considerations for SDDC Edge

The SDDC edge router serves as the north-south border device for the SDDC. This means that all ingress/egress traffic through all uplinks will transit this device. This includes all traffic destined for internet, VPN, Direct Connect, and to the cross-linked VPC. When sizing an SDDC, it is vital to plan for traffic loads on the SDDC edge.

You should keep the following points in mind for the SDDC edge:

1. It is a virtual appliance and is thus subject not only to limitations inherent to the appliance itself, but also to those of the underlying host.
2. 100% of all north-south traffic originating from VMs must pass through the edge. VMs exist within an NSX logical overlay network. Any traffic which leaves this overlay must pass through the edge. There is no getting around this.
3. Within VMware Cloud on AWS, there is one exception to the above rule. For vSphere replication traffic originating from the ESXi hosts themselves, and only in situations where Direct Connect private VIF is attached to the SDDC, replication traffic from the hosts will bypass the edge.
4. As mentioned previously, network appliances tend to be limited on the total packets-per-second they can process (typically denoted as a combined total of ingress/egress). This is true of both the edge and the underlying hosts, which have performance limits in terms of packets-per-second they may process. This is an aggregate number of ingress and egress passing through the edge and host.
5. VPN traffic will incur an extra performance penalty on the edge. The exact amount depends on the crypto settings applied to the VPN.
6. The rated performance of the underlying infrastructure itself. For example, there are per-flow performance limitations within the AWS infrastructure. These rates vary depending on whether the flow is local to an Availability Zone (AZ), crossing between AZs, destined for another Region, the internet, or Direct Connect, and also the Placement Group configuration of the SDDC.

As mentioned previously, network appliances tend to be limited on the total packets-per-second they can process (typically denoted as a combined total of ingress/egress). Due to the fact that the SDDC edge is a virtual appliance, and it is resident within a single host within the SDDC, there are 3 limits which come into play:

1. The rated performance of the SDDC edge in terms of raw packet processing ability. Since the SDDC edge may serve as both a router and a VPN endpoint, it is also important to consider how IPSec VPN (if applicable) will impact overall edge performance.
2. The rated performance of the Elastic Network Adapter (ENA) of the underlying host. This number will vary depending on the specific instance type hosting the edge.
3. The rated performance of the AWS infrastructure itself. There are per-flow performance limitations within the AWS infrastructure. These rates vary depending on whether the flow is local to an Availability Zone (AZ), crossing between AZs, destined for another Region, the internet, or Direct Connect, and also the Placement Group configuration of the SDDC.

There are two options available for increasing the overall packet processing capacity of the edge:

1. Edge scale-up: This strategy will provide the edge appliance with additional CPU and memory resources. Larger edge profiles will provide improved processing ability of the edge.
2. Edge isolation: If the edge is co-resident with other workloads on a given host, then there may be contention for resources on that host. Isolating the edge to a dedicated host will eliminate this contention.

In some cases, a single SDDC is simply insufficient to meet the network capacity requirements of the deployment. In these situations, you may consider splitting the deployment into multiple SDDCs.

Common Traffic Flows

When designing an SDDC, pay special attention to a few of the most notable types of traffic flows and account for them as part of your capacity planning efforts.

User and Application Traffic

This is the normal traffic which you expect to see in the environment. The difference is that you must now account for how these traffic patterns may have changed as the result of the migration. Consider the following:

- What may have previously been east-west user traffic in the on-premises environment is now north-south traffic.
- If applications have been split between SDDCs, then you must factor in how this impacts north-south traffic.
- Consider how connectivity to shared services may add to north-south traffic.

vSAN Replication Traffic

vSAN replication traffic is an important consideration for high utilization hosts (such as those that house the SDDC edge). Consider how vSAN replication for high I/O workloads might contribute to network contention for a given host.

Backup Traffic

Backups are an example of a periodic activity which results in a temporary spike in network utilization. Consider scheduling backups during off-hours or non-peak hours to avoid network contention.

HCX Replication Traffic

Keep in mind that replication traffic will contribute greatly to the north-south traffic utilization of the SDDC for as long as replications are being performed. You must account for this during your initial planning and monitor for it during migrations.

Note that HCX IX appliances create their own dedicated VPN tunnels. They will not utilize SDDC edge VPN tunnels.

Network Extension Traffic

Network extension traffic will contribute to the north-south utilization of the SDDC edge, and the exact amount of traffic resulting from network extension will vary greatly between projects. Due to its sub-optimal and unpredictable nature, it is vital to minimize network extension traffic wherever possible.

Note that HCX L2C appliances create their own dedicated VPN tunnels. They will not utilize SDDC edge VPN tunnels.

Cloud Service Integration (AWS)

There are a wide variety of services available from AWS which may be utilized by the SDDC. In all cases, accessing these services will result in additional network utilization on the edge as well as the underlying ENA of its host. Consider the following points when planning for AWS service integration:

- Traffic flow – Which edge uplink will be used to access the service? Some services, such as S3, may be accessed either by public IP (internet uplink) or via private endpoints in the cross-linked VPC (VPC uplink). Other services, such as Transit Gateway, may be accessed via VPN which will cross the internet uplink of the edge.
- Traffic rates – You must consider the both the amount of traffic which will be generated on the edge as well as performance expectations of the service. Remember that the AWS infrastructure provides different per-flow performance depending on how a given network flow transits their infrastructure.
- Bandwidth charges – Keep in mind that north-south traffic from the edge will incur bandwidth charges. The exception to this rule is for traffic to the cross-linked VPC which remains within the same Availability Zone as the edge.

Network Security

By default, the gateway firewalls of the SDDC are configured to deny all traffic. Firewall rules must be specifically created to permit access. This applies to all uplink interfaces of the edge. Due to the locked-down nature of the SDDC, an important part of the planning process is to determine a basic security policy for use within the SDDC.

Here are some things to consider:

- Determine who within your organization is required to review and approve security policy decisions.
- Determine how the SDDC will be accessed remotely, from where, and what source/destination IP addresses and TCP/UDP ports are required to facilitate the connectivity.
- Determine what services must be accessed within the on-premises and other cloud-native environments.
- Understand that the gateway firewalls filter traffic in both directions. This means that security policy must be explicitly defined for inbound requests as well as for outbound requests initiated from the SDDC.
- For AWS, remember that Security Groups may be applied within the cross-linked VPC. These may impact connectivity between that VPC and the SDDC.

Authors and Contributors

Author: Dustin Spinhirne