VMware Cloud on AWS: Stretched Clusters

VMware General

Table of contents

VMware Cloud on AWS: Stretched Clusters

Overview

What are Availability Zones?

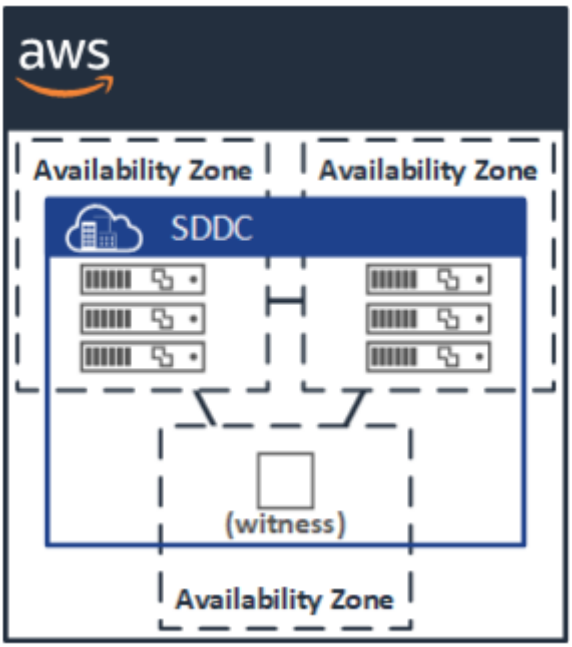
AWS regions are designed around the notion of fault domains, or Availability Zones (AZ). Within a Region, an AZ serves a couple of notable purposes:

1. It acts as a boundary for resource management. In other words, the compute (and other) resources available to an AZ are finite and may become exhausted by customer demands. AWS will remove an AZ from the list of available AZs should it become resource constrained.
2. It is built to be independently resilient, meaning failures in one AZ should not impact any other AZ.

The network infrastructure of an AWS Region is designed in such a way to support connectivity between Availability Zones. While network traffic within an AZ is not billable, network traffic which crosses between AZs will be billed. Data transfer charges and inter-AZ network performance are important considerations when designing a multi-AZ deployment within AWS.

What is a Stretched Cluster SDDC?

A standard (non-stretched) SDDC is one in which all hosts are deployed within a single AZ. On the other hand, a stretched cluster SDDC is one in which the hosts of the SDDC are evenly split between 2 AZs within an AWS Region (with a hidden “witness” host in a 3rd AZ).



The choice of a stretched cluster vs a standard SDDC is a decision which must be made at the time of SDDC deployment. Once it has been deployed, it is not possible to convert a standard SDDC to a stretched cluster SDDC (or vice-versa). Furthermore, all Clusters of the SDDC will be uniformly implemented as stretched or non-stretched.

Why Deploy A Stretched Cluster SDDC?

Simply put, stretched cluster SDDCs offer an availability strategy to customers. The solution is designed specifically to provide the SDDC with an extra layer of resiliency in the event of host-level failures within the cluster or with AZ-level failures within the region.

It is important to keep in mind that stretched cluster SDDCs only offer an additional layer of resiliency and do not address all failure scenarios. For instance, they do not protect against regional-level failures within AWS or data loss scenarios resulting from application issues or poorly planned storage policies. These types of scenarios fall outside of the scope of the protection offered by a stretched cluster SDDC and must be addressed by a detailed data recovery strategy.

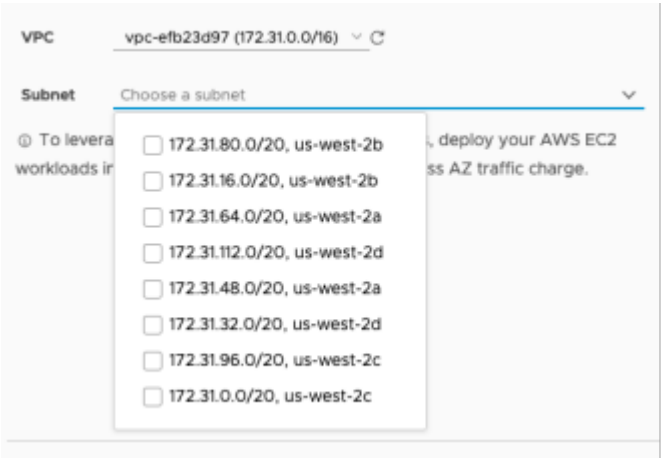
Architecture

Implementation Overview

Stretched cluster SDDCs are implemented using a vSAN feature of the same name. Per the requirements of vSAN, the SDDC provides 2 data sites and one witness host per cluster. The data sites are composed of 2 groups of hosts, which are evenly split between a pair of 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. This witness host is not reflected in the total host count of the SDDC.

Due to the requirement that stretched cluster SDDCS utilize a total of 3 AZs, they are only supported in AWS Regions that are able to provide at least 3 AZs.

When deploying a stretched cluster SDDC, AZ mapping for the data sites is controlled by the end user through the selection of cross-link Subnets for the SDDC. Additional details of the network design for stretched clusters are covered in the SDDC Network Architecture guide.

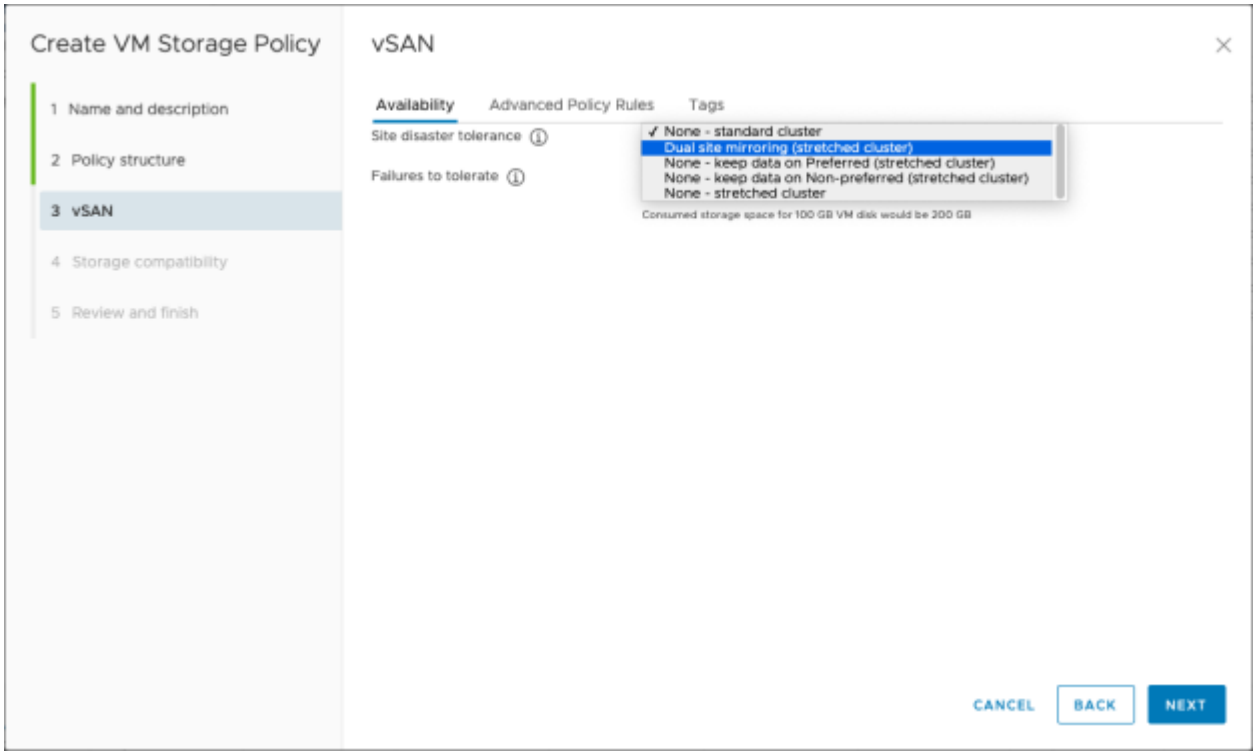


Exactly two subnets must be selected for cross-linking in a stretched cluster deployment. The first subnet “checked” will be used as the “preferred site” in vSAN, while the second subnet will be used as the “non-preferred site”. This designation is purely cosmetic since both “sites” will be active. However, it is worth documenting the preferred vs non-preferred site since this information is useful if you intend to not enable dual site mirroring for some workloads.

Site Disaster Tolerance

A stretched cluster SDDC offers additional resiliency for workloads; however, this feature is end-user controllable and is implemented through the vSAN policies applied to workloads. When constructing storage policy for stretched cluster SDDCs, there are 4 options to consider.

- Dual site mirroring
- None – keep data on Preferred
- None – keep data on Non-preferred
- None



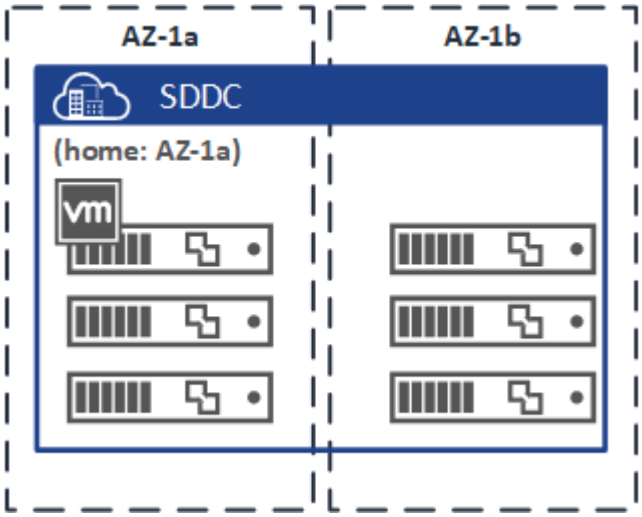
For workloads that require data resiliency between AZs, the “Dual site mirroring” option should be selected. This option will cause data to automatically be replicated between sites for all VMs/VMDKs that use this policy. It should be noted that, for stretched cluster SDDCs, the storage policy for the management appliances is automatically configured to implement dual site mirroring and cannot be changed.

For workloads that do not require data resiliency between AZs, one of the other options for stretched cluster may be selected. These options will allow you to either explicitly place data in one AZ or the other (the “None – Preferred” or “None - Non-preferred” options) or allow vSAN to place data based on available storage capacity (the “None” option). Note that, in general, you will want to run the workloads in the same AZ where their data is located (to avoid cross-AZ bandwidth charges). In order to ensure that this is the case, you must manage workload placement using affinity rules for the workloads.

Note that fault domains of vSAN are mapped to the AZs of the SDDC. This data is visible from the vSAN portion of the Cluster configuration within vCenter. The fault domains of individual hosts and VMs are visible from their summary pages, and you should use this information to assist you in creating tagging policies for any affinity rules that may need to be created.

AZ Affinity

AZ affinity refers to the AZ in which a VM “prefers” to run. This “preference” is used by DRS as a means of avoiding unnecessarily migrating a VM between AZs. The preferred (or “home”) AZ of a VM is based solely on the AZ in which it was most recently booted. For instance, if the SDDC is deployed within Availability Zones AZ-1a and AZ-1b, and a VM is booted on a host within AZ-1a, then that VM is considered to prefer AZ-1a as its “home” AZ. This means that the VM will tend to stay within that AZ (i.e., DRS will prefer to keep it within that AZ).



Should the AZ suffer a host failure, vSphere HA will attempt to recover the VM on another host within the same AZ. If there are insufficient resources to recover the VM in that AZ, then it will attempt to recover the VM on a host within the other AZ.

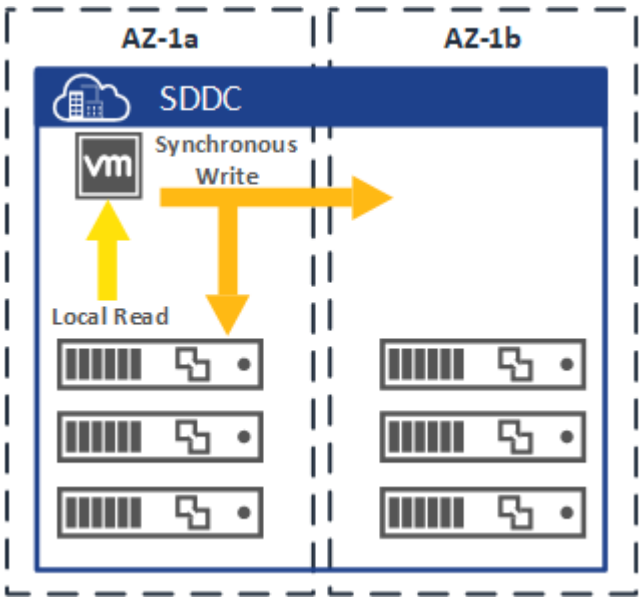
A VM that is recovered in a different AZ will now consider that AZ as its "home".

Workload Balancing

There is no mechanism today to automatically balance workload placement between AZs. This process must be managed by the customer.

Data Reads and Writes

It is important to understand the read/write behavior of workloads which utilize a Dual Site Mirroring enabled vSAN policy. For reads, the data is read from the local copy (i.e., the "home" AZ of the workload). For writes, the data is written locally as well as synchronously replicated to the hosts of the twin AZ. These synchronous writes will result in cross-AZ network traffic, the amount of which is highly dependent on the I/O profile of the workloads.



Design Considerations

Adding and Removing Hosts

One notable difference between a standard SDDC and a stretched cluster SDDC is the way hosts are added and removed from the SDDC. Since hosts of a stretched cluster SDDC must be equally balanced between AZs, add/remove operations will be done in pairs. In other words, you may only increment or decrement the host count of a Cluster two at a time.

Recovering from an AZ failure

If it a full or partial AZ failure occurs, Elastic DRS scales out the cluster in the remaining AZ. It adds non-billable hosts in the remaining AZ until the cluster reaches its original host count. This scale out is dependent on available capacity and is not guaranteed. When the failed AZ is restored, Elastic DRS redeploy the hosts in the restored AZ and scales in the cluster to remove the extra hosts from the original AZ.

Note: entry-level stretched clusters (1-1-1 and 2-2-1) can be scaled in to the original host count after an AZ failure.

Maximum Cluster Size

A common misconception is that, since a stretched cluster SDDC is managing host groups that are split across AZs, that it offers twice the total host count per Cluster. In reality, the maximum host per Cluster for a stretched cluster SDDC is identical to that of a standard SDDC; the hosts just happen to be split across two AZs.

The witness host of the Cluster does not count toward the maximum.

Minimum Host Counts

Stretched clusters can be deployed with as few as two hosts - one host per Availability Zone plus a witness in a third AZ. These entry level 2-host and 4-host clusters are fully supported and are covered by a 99.9% SLA. They can be expanded to 6+ hosts to qualify for the 99.99% SLA. Once a stretched cluster increases to 6 or more hosts, the minimum size for that cluster becomes 6 hosts - larger clusters cannot be shrunk to 2 or 4 hosts.

Network Architecture

The network architecture for stretched cluster SDDCs is slightly different than that of a standard SDDC. The details of this are covered in the SDDC Network Architecture guide.

Authors and Contributors

Author: Dustin Spinhirne