VMware Cloud on AWS: Storage Architecture

VMware Architecture

Table of contents

VMware Cloud on AWS: Storage Architecture

Overview

A VMware Cloud on AWS Software Defined Datacenter (or briefly referred to as “SDDC”) uses storage capacity for 2 purposes:

1. Management Virtual Machines (VMs):
   This group contains the VMs of the VMware management stack (e.g. vCenter Server, NSX Manager), which are required for managing the SDDC.
2. Customer Workload VMs:
   This group contains the VMs deployed by the customer into the SDDC.

The primary choice of a storage option for SDDCs is VMware vSAN. vSAN aggregates local or direct-attached data storage devices to create a single logical storage pool shared across all ESXi hosts in a vSphere cluster.
In addition to vSAN, there are options to mount NFS-based external storage to the vSphere clusters of an SDDC. The availability of storage options is also dependent on which Amazon EC2 instance type (referred to as “SDDC host type”) you have chosen for the ESXi hosts of your SDDC.

SDDC Host Types and Storage Characteristics

Storage characteristics and options are dependent on which SDDC host type you have chosen for your SDDC. The table below provides an overview of SDDC Host types and their storage characteristics. Note: The listed storage options in the table will be explained further below in this document.

	SDDC Host Types			
	i3[1]	M7i	i3en	i4i
Total usable storage capacity	10.37TiB (~11TB), exclusively NVMe disks	No Local Storage NFS storage mount maximums apply according to VMware Configuration Maximums[2]	45.84TiB (~50TB), exclusively NVMe disks	vSAN OSA 20.46TiB (~22.5TB) vSAN ESA 27.28TiB (~30TB) exclusively NVMe disks
vSAN ESA	No	No	No	Yes[3]
vSAN OSA	Yes	No	Yes	Yes
VMware Cloud Flex Storage	Yes	Yes	Yes[4]	Yes[4]
Amazon FSx for NetApp ONTAP	Yes[4]	Yes	Yes[4]	Yes[4]

[1] There is an end of sale for the i3.metal instance, see Announcement of the end of sale, end of support and end of life timeline of the i3.metal instance type of VMware Cloud on AWS .

[2] This SDDC host type comes with NFS-only architecture and without vSAN. The VMware Management VMs (e.g. vCenter, NSX Manager) are stored on a fully managed external datastore provided by VMware.

[3] Currently the availability of this feature is gated. Customers need to partner with a VMware representative for its activation.

[4] For the specified host types NFS-type storage is only available as a supplemental option in addition to vSAN. Virtual machines dedicated to SDDC Management (like e.g. vCenter Server) are mandatorily stored on vSAN.

The above listing of SDDC host types is storage-centric. Please consult the document "Feature Brief: SDDC Host Types” for a more comprehensive discussion of SDDC host types.

Storage Options for the SDDC

vSAN

This type of storage is housed directly within the hosts of the SDDC and has a fixed per-host size depending on the SDDC host type being used. For each host that is added to a Cluster, the overall storage pool will increase. In the Amazon EC2 instances used for VMware Cloud on AWS, only high-performance NVMe storage is used as direct attached storage for vSAN.

vSAN Original Storage Architecture versus Express Storage Architecture - A Brief History

Within the vSAN storage option, two architectures are available – vSAN Original Storage Architecture (OSA) and vSAN Express Storage Architecture (ESA). The key differences between both architectures are briefly highlighted below. For more comprehensive details consult the documentation and other deep-dive materials on the VMware website.

- vSAN Original Storage Architecture (OSA)
  As the name already implies, vSAN OSA is the original architecture, which was introduced in the very first vSAN release in 2014. Especially at that time flash memory drives like SSDs were expensive and much lower in capacity compared to traditional HDD drives. This led to the design principle of using both drive types in 2 separate storage layers within a vSAN datastore: Flash drives were assigned as a smaller cache layer whereas the lower performance, but high-capacity HDD drives were aggregated in the 2nd layer, the capacity layer. Today this architecture still exists, but in nowadays implementations the HDDs have mostly been replaced by high-performance flash memory drives like SSD or NVMe, as these types of drives have become higher in capacity and lower in price during the past years.
- vSAN Express Storage Architecture (ESA)
  With dramatically dropping prices and at the same time increasing capacity of flash drives VMware decided to consider these new developments in the storage industry and introduced a new alternative vSAN ESA architecture in 2022 as part of the vSphere/vSAN 8.0 release. One of the fundamental differences compared to vSAN OSA is abandoning the dedicated caching layer in favor of a single, high-performance storage layer, where every single drive contributes to capacity. This fact leads to reduced costs per GB in vSAN clusters.
  For details on the new ESA architecture consult the vSAN documentation or various blogs and articles on the VMware website, e.g. on the vSAN Express Storage Architecture (ESA) Techzone page.

Availability of vSAN Architectures in VMware Cloud on AWS

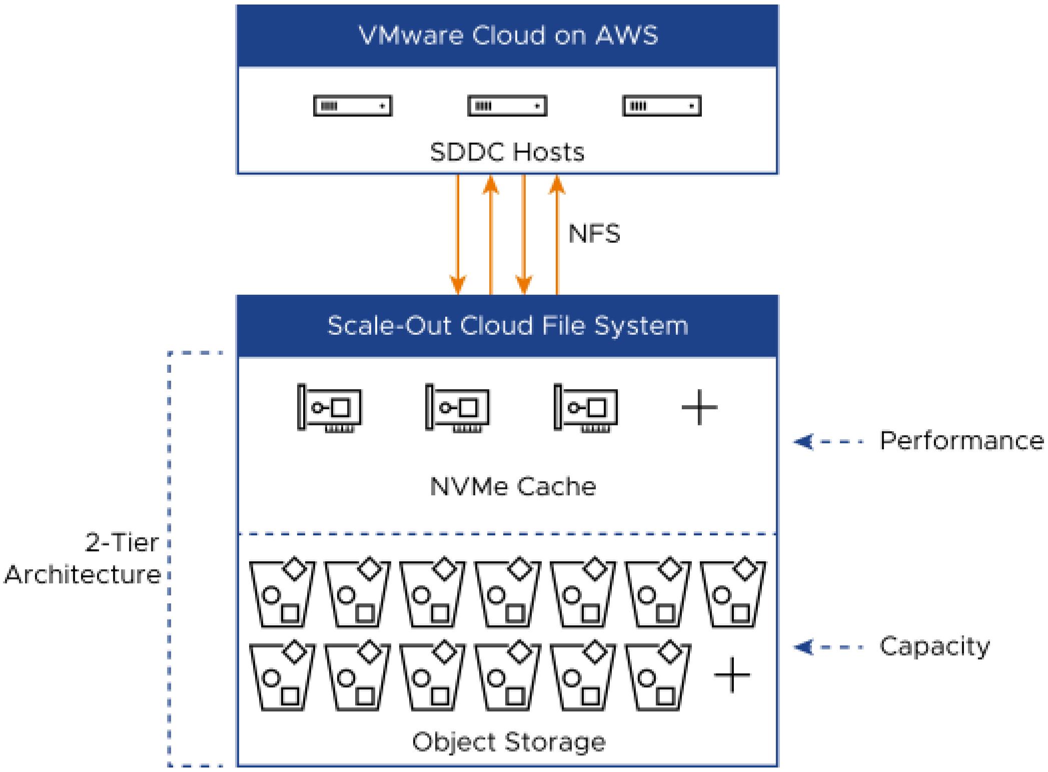
In VMware Cloud on AWS all SDDC host types except M7i support vSAN OSA. Additionally, vSAN ESA has been certified for the i4i SDDC Host Type. vSAN ESA in VMware Cloud on AWS is currently supported for Single-AZ-Clusters. It is not supported for Stretched Clusters yet. For specifics on vSAN ESA in the context of VMware Cloud on AWS see e.g. vSAN ESA with VMware Cloud on AWS: Technical Deep Dive and blog article vSAN ESA now available in VMware Cloud on AWS.

External NFS Storage

Complementary to vSAN, external NFS Storage offers an additional way to store your virtual machine data. It is particularly useful for workloads with high storage capacity demands. Scaling the capacity of vSAN storage cannot be done independently from your compute layer. In other words: when vSAN capacity needs to be added to a cluster you need to add one or more ESXi hosts. With external NFS storage, you can scale your storage capacity independently from your compute capacity and hence it is a viable choice for workloads that do not require vSAN performance. When connecting NFS storage to your SDDC, two options are available today.

VMware Cloud Flex Storage

VMware Cloud Flex Storage offers a scalable, elastic storage and data management service. As a VMware-managed service, it is natively integrated with the VMware Cloud Console and offers datastore-level storage access for storing virtual disks of your virtual machines. Note, that its purpose is solely to provide storage to your SDDC’s ESXi hosts. It is not supported for guest-level access by directly mounting it from your virtual machines.



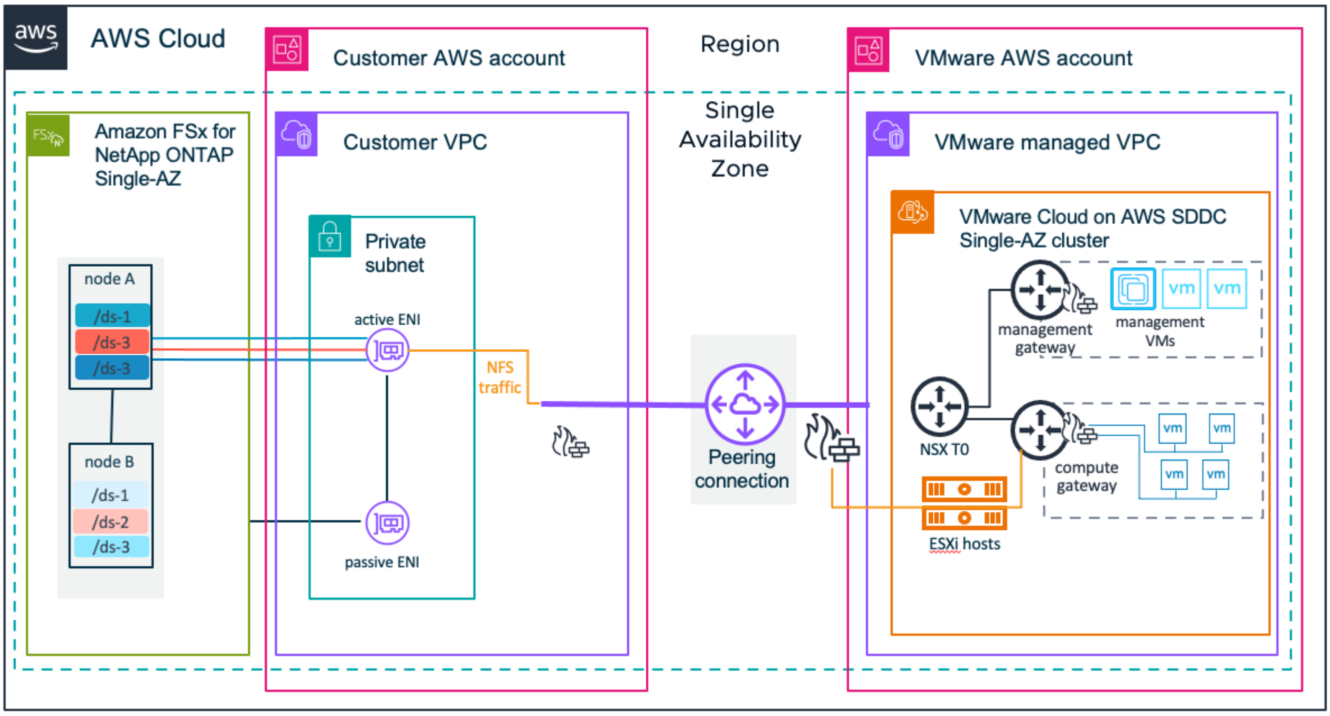
Amazon FSx for NetApp ONTAP and VMware Cloud on AWS

Amazon FSx for NetApp ONTAP integration with VMware Cloud on AWS is an AWS-managed external NFS datastore built on NetApp's ONTAP file system that can be attached to a cluster in your SDDC. It provides customers with a flexible, high-performance virtualized storage infrastructure that scales independently of compute resources.

Availability and Connectivity Options for Amazon FSx for NetApp ONTAP

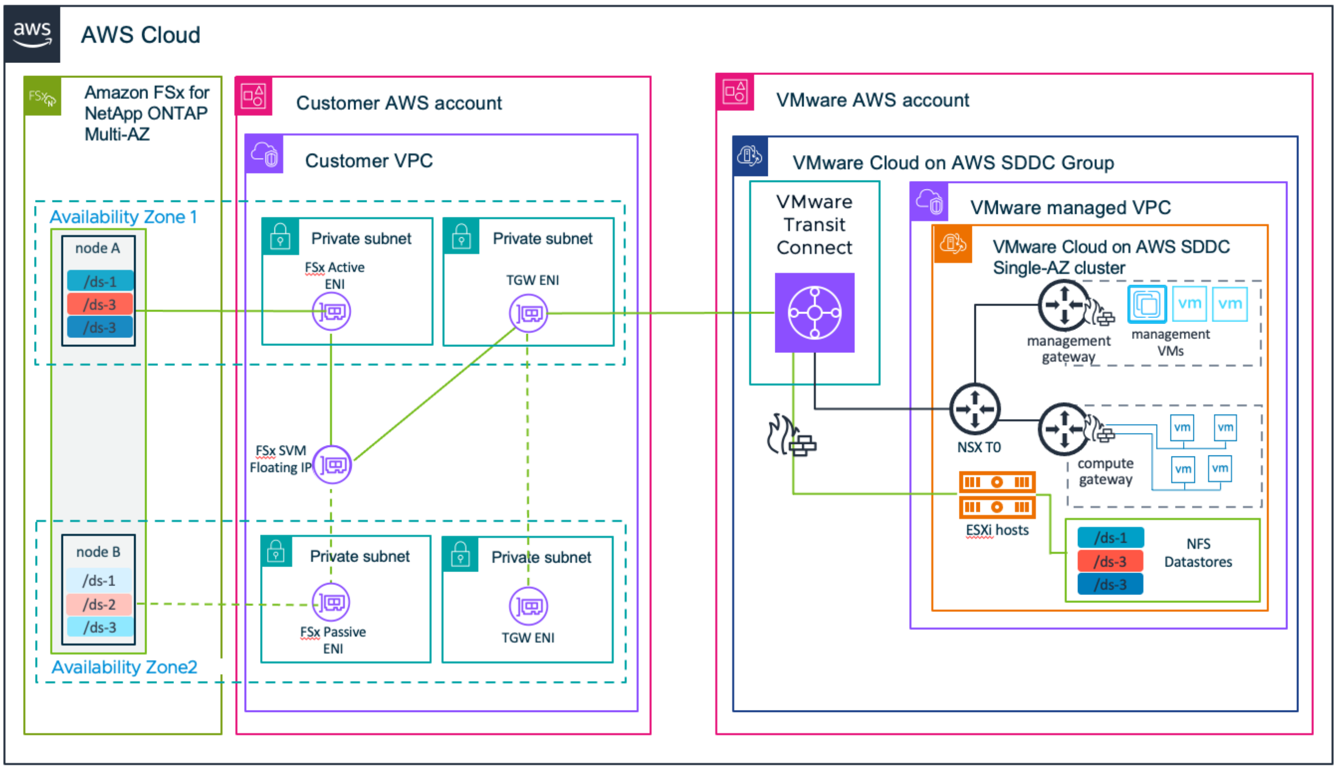
Amazon FSx for NetApp ONTAP can be configured either in a standard single AWS Availability Zone (AZ) or a Multi-AZ setup[5]. A Multi-AZ configuration can meet extended availability SLAs compared with a Single-AZ setup. VMware Cloud on AWS SDDCs supports both Single-AZ and Multi-AZ configurations of Amazon FSx for NetApp ONTAP. Depending on the chosen configuration (Single-AZ versus Multi-AZ) you have to implement different connectivity options. Basically, you differentiate between 2 connectivity options for attaching Amazon FSx for NetApp ONTAP storage to an SDDC:

- VPC Peering for Single-AZ Designs
  VPC Peering is the preferred model to connect an SDDC with Amazon FSx for NetApp ONTAP storage if both the SDDC and the VPC hosting the Amazon FSx for NetApp ONTAP storage are in the same Availability Zone.
  The following figure illustrates the architectural design:



For implementation details on VPC peering for external storage please consult Feature Brief: VPC Peering for External Storage.

- VMware Transit Connect for Multi-AZ Designs
  While VMware Transit Connect can be selected for Single-AZ designs as well to connect your SDDC to Amazon FSx for NetApp ONTAP Storage, it really enables Multi-AZ designs, where SDDC and Amazon FSx for NetApp ONTAP storage are hosted in different availability zones or where Amazon FSx for NetApp ONTAP for itself is already deployed in a Multi-AZ fashion. The following figure illustrates the architectural design of a Multi-AZ connectivity between SDDC and Amazon FSX for NetApp ONTAP:



The VMware TechZone document "VMware Cloud on AWS integration with Amazon FSx for NetApp ONTAP Deployment Guide" discusses the connectivity- and other integration aspects in detail.

[5]Note, that in the context of discussing NFS storage options the term “Multi-AZ” only applies to the implementation of the NFS storage infrastructure on AWS. It does not apply to Stretched Cluster SDDCs, as this type of SDDCs cannot have NFS datastores mounted.

Differences between Amazon FSx for NetApp ONTAP and VMware Cloud FlexStorage

	VMware Cloud Flex Storage	Amazon FSx for NetApp ONTAP
Datastore-level SDDC access	Yes	Yes
Guest-level access	No	Yes
Stretched Cluster Support	No	No
Storage Resiliency	Single-AZ	Single-AZ Multi-AZ
Storage Efficiency	Data Compression Data Deduplication	Data Compression Data Deduplication
Data Security	Data-at-Rest Encryption	Data-at-Rest Encryption Data-in-Transit Encryption
Integrated in VMware Cloud Console	Yes	No
Managed by	VMware	AWS
Use Cases	For workloads demanding high storage capacity, which needs to be independently scalable from the compute layer (Valid for both options).	
	For customers seeking a seamless operational experience fully integrated with the VMware Cloud Console	Suitable especially for customers, who already use NetApp ONTAP in their storage landscape and benefit from its features - either on-premises or as a cloud-based offering like Amazon FSx for NetApp ONTAP.

Summary

vSAN is still considered the first and easiest choice in VMware Cloud on AWS SDDCs when deploying a VMware Cloud on AWS SDDC. This is especially true for workloads with high storage performance requirements and low latencies, but also for scenarios, where certain features like stretched cluster support are required. However, there are certain scenarios and use cases, where customers may opt to go with external NFS storage. For example, when workloads with extraordinarily high storage capacity demand benefit from scaling storage capacity independent from the compute layer. Also, it is worth highlighting, that the M7i SDDC host type does not support vSAN, as it has no direct attached storage built in. When selecting the M7i instance type, you have to choose an external NFS storage option, which is either VMware Cloud Flex Storage or Amazon FSx for NetApp ONTAP.

Storage for the Management Cluster of an SDDC

The management cluster of an SDDC has a special role, as it is rolled out during the initial deployment of the SDDC and hence has to provide the storage for SDDC management VMs (e.g. vCenter, NSX) running in this cluster. For all SDDC host types except M7i, these management VMs are stored on the vSAN datastore. The M7i host type does not have vSAN and management VMs are stored on a fully managed external datastore provided by VMware.

With vSAN being used within the primary cluster of the SDDC, it has been modified compared to its standard version to present two logical datastores from the same underlying physical storage: one for management appliances and the other for end-user workloads. It is important to point out that this separation exists purely as a means of enforcing permissions on the storage for management appliances.

Key Takeaways

- SDDC management VMs are running in the primary cluster and stored on the cluster’s connected storage (vSAN or NFS)
- Primary storage for an SDDC is vSAN for all host types except M7i. For M7i the only supported storage type is NFS.
- The logical vSAN datastores reflect the same underlying pool of capacity. Do not mistake them for independent sets of storage.
- The free space, used space, and total capacity numbers will be identical between the two logical Datastores, as they relate to a single, “physical” vSAN datastore.
- You must consider management appliance storage footprints within the primary cluster when sizing an SDDC.

vSAN Slack Space Requirements

vSAN requires that a certain percentage of raw storage within a Datastore be reserved as “slack” space. This reserved capacity is used for operations such as deduplication, object re-balancing, and recovery from hardware outages within the underlying pool of storage capacity. The current requirement is to maintain a 20% buffer for slack space.

As part of its service level agreement with customers, VMware will ensure the health of the SDDC by enforcing this slack space requirement using the default EDRS policy whenever necessary. A notification will be sent when you approach the 20% threshold and EDRS will automatically scale up the SDDC if available slack space comes close to 20%.

Deduplication and Compression

Deduplication and Compression are designed to optimize storage by reducing the amount of space required to store data within the datastore.
The following table provides an overview of these features for the various VMware Cloud on AWS host types:

	i3	M7i	i3en	i4i
Compression	Enabled	Available within the NFS storage provider	Enabled	Enabled
Deduplication	Enabled	Available within the NFS storage provider	Disabled	Disabled

Note: For vSAN, these settings cannot be changed on VMC on AWS.

Storage Capacity Reclaim Using TRIM/UNMAP

The TRIM/UNMAP commands for the respective ATA and SCSI protocols provide a means for guest operating systems to reclaim unused storage space as free space. VMC SDDCs configured to use vSAN ESA on i4i instances have TRIM/UNMAP enabled by default. This configuration is unique to vSAN ESA. For vSAN OSA configurations the feature can be enabled on request.

For more information about this feature see document vSAN Space Efficiency Technologies.

Encryption at Rest and in Transit

Encryption at rest

vSAN implements encryption at rest using the AWS KMS service. Although much of the functionality behind vSAN encryption is automated, it is worth understanding the key components and which of them may be customized.
The components of encryption are as follows:

- Customer Master Keys (CMK) - This is the master key that is used to encrypt all other keys used by vSAN. It is controlled and managed by VMware/AWS and may not be updated. One CMK is required per Cluster.
- Key Encryption Key (KEK) - This key is used by vSAN to encrypt DEKs and may be updated through the vCenter UI (referred to as a shallow rekey). One KEK is required per Cluster.
- Disk Encryption Key (DEK) - This key is used to encrypt disk data and may not be updated. One DEK is required per disk in vSAN.

In addition to vSAN encryption, the SDDC host types use self-encrypting NVMe drives.

Encryption in transit

For all SDDC host types except i3, we added in-transit hardware-level encryption between instances within the SDDC boundaries.

Storage Policy Based Management

Storage Policy Based Management (SPBM) is a declarative system for the management of storage within the SDDC. SPBM provides a framework to define policies with specific data protection and performance criteria and permits it to be applied granularly to objects within vCenter. SPBM allows multiple policies to coexist and permits them to be applied at both the VM and VMDK levels. Once defined, these policies may be extended across clusters. The Policy configuration and assignment of management appliances is controlled by VMware and may not be modified. The default policy configuration for any workloadDatastores is also maintained by the service but customers may create and assign custom policies if they wish to override this behavior. Note, that whenever custom policies are created a customer would have to make sure that these self-created and managed custom policies still adhere to the requirements of the service level agreement to stay eligible to receive any SLA credits in case of SLA events.

The first consideration when designing policy is availability, which defines the level of redundancy desired for objects within the datastore. The second consideration for policy design is performance, which defines both the processing overhead required to implement the policy and the overall end-user impact in terms of IOPS for a given workload. Both of these considerations are discussed below.

SPBM Availability - Conceptual Overview

Availability is broken down into 2 levels: the ability to survive a site-level disaster and the ability to survive failures within a given site. Within the context of VMware Cloud on AWS, a site falls within the boundary of an AWS Availability Zone (AZ). The ability of a VM or VMDK to survive an AZ-level outage applies only to stretched cluster designs. Specifically, the “dual site mirror” option for stretched clusters will cause data to be replicated across AZs and will enable an object to survive a complete failure of either AZ.

See the document on stretched cluster SDDCs for more information.

Within an AZ, there are two settings to consider for data resiliency:

- Failures to Tolerate (FTT) - This defines the number of host or disk failures to tolerate. In other words, it defines the number of devices that can fail before data loss occurs.
- Fault Tolerance Method (FTM) - This defines the type of data replication used: mirroring (RAID1) or erasure coding (RAID5/RAID6)

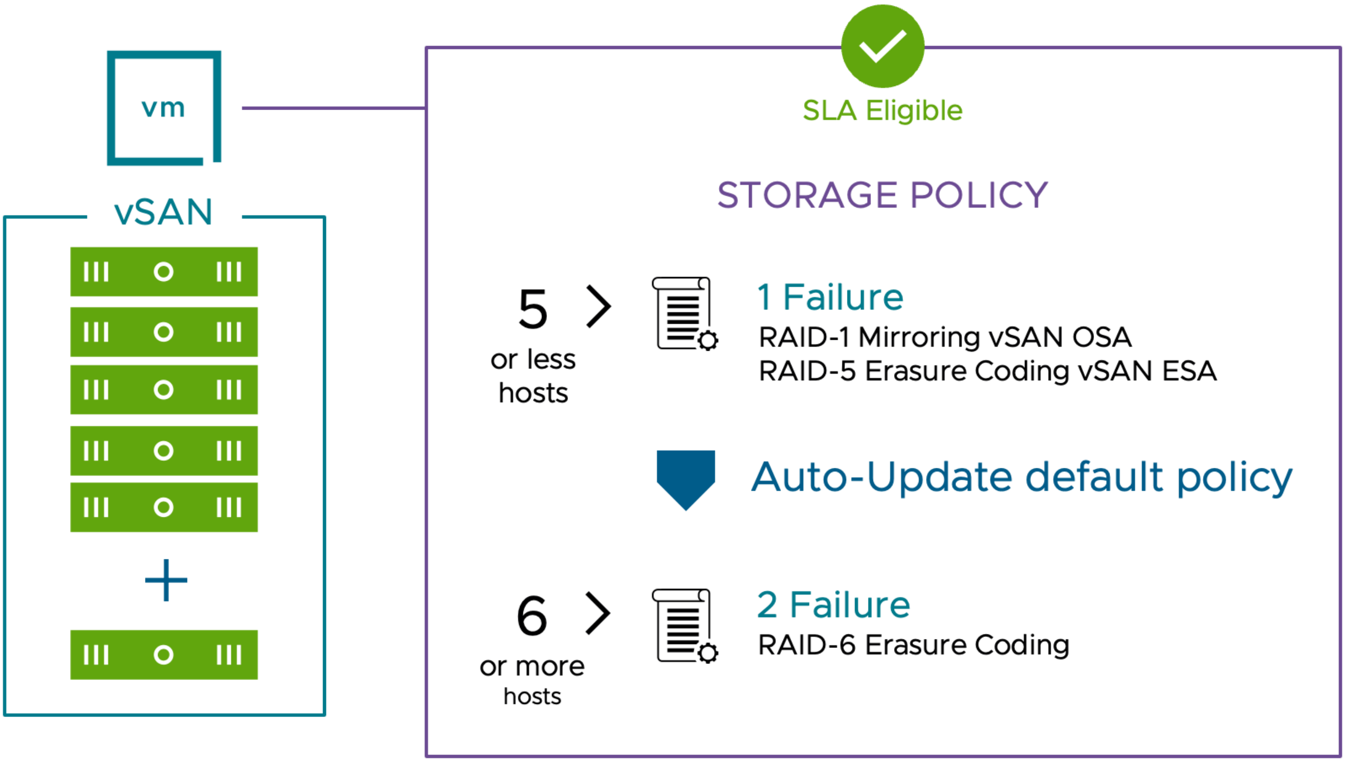
The options available for FTT/FTM will be determined by the number of hosts within a cluster, the vSAN architecture being used (vSAN OSA or vSAN ESA), and the vSAN cluster type (stretched vs non-stretched/standard). The table in the following chapter provides an overview of the available FTT/FTM options. The options chosen will impact the total amount of usable capacity of the datastore.

Details regarding the erasure code concepts and data placement schemes of the various vSAN fault tolerance methods are described in the VMware Cloud Platform Tech Zone document vSAN Space Efficiency Technologies.

Automated Storage Policy Management in VMware Cloud on AWS

Whenever a vSAN-based cluster is created in VMware Cloud on AWS, a default storage policy is created and its configuration is automatically maintained and adjusted by the service. This mechanism ensures that the customer workloads, for which the default storage policy is automatically applied, stay within the SLA specification, meaning they are “SLA eligible”.
The adjustment of the default policy during the lifetime of a cluster becomes necessary, when certain events occur, specifically the addition(“scale-out”) and removal(“scale-in”) of hosts to/from the cluster either by the customer or by Elastic DRS.

The following figure illustrates the automated storage policy configuration for a Single-AZ (“non-stretched”) cluster:



The following table provides a complete overview of the default storage policy configuration parameters depending on the chosen vSAN architecture, the number of hosts in the cluster and the cluster type (Single-AZ vs. stretched cluster):

	Single-AZ Cluster			**Stretched Cluster**	
	2 hosts	**3 to 5 hosts**	**>= 6 hosts**	**2 or 4 hosts**	**>= 6 hosts**
vSAN OSA	FTT=1 FTM = RAID-1 Site Disaster Tolerance (SDT) = None	FTT = 1 FTM= RAID-1 SDT=None	FTT = 2 FTM = RAID-6 SDT=None	FTT = 0 FTM = None SDT = Dual Site Mirroring	FTT = 1 FTM = RAID-1 SDT = Dual Site Mirroring
vSAN ESA	Currently not supported	FTT = 1 FTM= RAID-5 SDT=None	FTT = 2 FTM = RAID-6 SDT=None	Stretched Cluster is not yet supported for vSAN ESA.	

Performance

Data that is stored using an erasure coding FTM will consist of the actual data itself, which is broken into segments, along with parity copies. While erasure coding provides better efficiency in terms of storage overhead, this practice of segmenting data with added parity comes at a performance cost. This performance penalty is especially evident during failure scenarios when data must be calculated from parity.

In general, when performance is a concern, the order from most to least performant is: RAID1, RAID5, RAID6. This statement holds true for vSAN OSA configurations. For vSAN ESA configurations the claim is that we can achieve RAID-1 performance even with RAID-5/RAID-6 configurations due to significant improvements achieved with the ESA architecture. Please consult the blog post RAID-5/6 with the Performance of RAID-1 using the vSAN Express Storage Architecture for details.

NFS network connectivity could be a bottleneck for external storage, and VMware is consistently enhancing performance in this area.
Two features are worth highlighting in this regard:

- nConnect:
  nConnect is an enhancement for the NFSv3 protocol which was introduced in VMware Cloud on AWS v1.22. Without nConnect, a single TCP/IP connection from each host to each datastore was created. With nConnect support, multiple sessions are created in parallel from each host to each NFS datastore. VMware Cloud on AWS supports two parallel sessions to each NFS datastore. This will significantly enhance the available throughput of each datastore. This feature is enabled by default for each new SDDC and supports both VMware Cloud Flex Storage and Integration with Amazon FSx for Netapp ONTAP.
- Jumbo Frames/Large MTU Size:
  Using jumbo frames NFS storage traffic throughput can be significantly increased by reducing the overhead associated with transmitting NFS requests and responses. This can improve NFS performance, but all network devices in the path must support jumbo frames for it to be effective. Starting from VMware Cloud on AWS v1.24 SDDC supports increased MTU to provide better performance for NFS storage connections.