



Kubernetes Topologies with vSAN

Storage options with Kubernetes for vSAN
in VMware Cloud Foundation 9.1.1

October 15, 2026

Table of Contents

- Introduction 3
 - Scope of Topics 3
- Data and Application Availability for Kubernetes 3
 - Kubernetes on Bare Metal 3
 - Kubernetes in VCF 4
- Operational Models for Zones in VCF 4
- Kubernetes Topology Design with VCF Storage 5
 - Single-Zone Topologies 6
 - Multi-Zone Topologies 7
 - Cross-Zone Topologies 10
- Container Native Storage (CNS) Implementation Priorities 13
 - Key Technical Considerations for Your Ideal Storage Topology Selection 13
- Recommendations..... 14
- Summary 14
 - Additional Resources 14
 - About the Author 15

Introduction

As enterprise environments look to accommodate containerized microservices and cloud-native applications, maintaining robust storage and compute topology designs is paramount. VMware Cloud Foundation (VCF) integrates core vSphere compute, [vSAN storage](#), and [vSphere Kubernetes Service](#) (VKS) / Supervisor infrastructure to provide scalable, highly available Kubernetes execution environments.

This document explores the flexibility of running Kubernetes on VCF 9.1.1, including:

- Data and application availability concepts for Kubernetes.
- Operational models for zones in VCF 9.1.1, including new support for more than one cluster in each zone for multi-zone deployment.
- Kubernetes topology design patterns when using vSAN storage clusters.
- General recommendations and guidance.

Scope of Topics

The information provided in this document will assume the use of vSAN 9.1.1 and/or VMware Cloud Foundation (VCF) 9.1.1. VCF deployments may have additional requirements and support limitations that fall outside of the scope of this document. The topologies discussed in this document refer to virtualized environments where vSphere clusters in VCF 9.1.1 are using disaggregated external storage, courtesy of vSAN. This external storage can be a [vSAN storage cluster, or a vSAN HCI cluster](#), where in either deployment model, the remote datastore provides data services to a vSphere cluster. For simplicity, this document will refer to the external cluster providing storage to the vSphere clusters as a “storage cluster.” But this cluster can also be an existing “vSAN HCI” cluster where the vSphere clusters have mounted its remote datastore. Any references to “vSAN HCI” in this document do not imply that the workload instances and the storage are aggregated on the same cluster.

Data and Application Availability for Kubernetes

Traditional applications virtualized in a data center rely on infrastructure-level availability technologies to make both the application and the data highly available using technologies within the hypervisor. For more information, see the post: [“Application Versus Infrastructure-Level High Availability with vSAN in VMware Cloud Foundation.”](#)

Persistent storage for Kubernetes workloads is typically presented as a “persistent volume” (PV) to the container. This is a stand-alone volume that can be used by existing or newly instantiated containers via a “persistent volume claim” (PVC). Persistent volumes in vSAN can be provided in one of two ways:

- **ReadWriteOnce (RWO).** This is backed by block-based storage solutions. The volume can only be mapped to a single container. vSAN provides RWO persistent volumes without the configuration of any additional services.
- **ReadWriteMany (RWX).** This is backed by file storage, such as NFS services. The volume can be mapped to multiple containers, much like you might see an NFS exported volume mapped to multiple Linux VMs. vSAN provides RWX persistent volumes when vSAN File Services is configured on the cluster.

Kubernetes on Bare Metal

While infrastructure level high availability mechanisms like vSphere HA and vSAN were a natural fit for traditional applications, Kubernetes workloads that generate and consume persistent data work under a different paradigm. Kubernetes has conceptual origins that assume it is installed directly on to an operating system of a physical server, or node. This meant that a Kubernetes application would store critical data on specific persistent volumes attached to its underlying node. Because a pod must remain attached to its specific volume, it cannot simply be restarted on any arbitrary node if a failure occurs. Application availability requirements could be met with running multiple instances, but high availability of persistent data needs to rely on storage orchestration engines (e.g. Ceph, OpenEBS, Longhorn, etc.) to provide data replication and availability.

To better account for high availability across physical or logical boundaries extending beyond a single node, **Kubernetes uses the concept of “zones”** - often referred to as “availability zones” by cloud providers. Zones provide a construct to represent a logical or physical failure domain. Zones could represent racks in a data center row, rooms or buildings of a campus, or a broader geographic region, such as a large data center infrastructure.

Kubernetes in VCF

VCF can run Kubernetes applications natively, as well as store its persistent data, all while accounting for the unique constructs of Kubernetes. The vSphere Kubernetes Service (VKS) is the built-in control plane component that automates the deployment and lifecycle management of upstream, CNCF-certified Kubernetes clusters within a vSphere environment. It allows developers to self-service provision enterprise workload clusters using standard Kubernetes APIs or declarative tools like Cluster API. Meanwhile, IT administrators maintain full governance and operational control by managing those automated clusters alongside traditional virtual machines through native vSphere tooling. **Running Kubernetes on VCF offers additional flexibility and efficiency for your environment by providing infrastructure-level resilience already used for your traditional applications.**

A key element of VMware VKS is the Supervisor. It is the platform engine in VCF that provides all the integration of Kubernetes with components of VCF, including vSAN storage, NSX networking, compute boundaries, security boundaries, and so on. A vSphere Supervisor transforms a vSphere cluster into an integrated Kubernetes control plane by embedding Kubernetes natively into vSphere. It allows IT teams to manage virtual machines, Kubernetes clusters, and containerized workloads side-by-side using unified Kubernetes APIs and standard vSphere tools. **The Supervisor can even account for zones, where it can map vSphere clusters or compute resources to underlying physical availability zones (like distinct racks or datacenters) to isolate infrastructure failures.** This allows Kubernetes workloads to distribute their pods and nodes across multiple independent zones, ensuring high availability and fault tolerance for enterprise applications.

Operational Models for Zones in VCF

In VCF, a zone using VKS maps physical infrastructure boundaries such as racks, power zones, server rooms, or distinct buildings into logical availability domains for Kubernetes workloads.

Supervisors support two primary operational models:

1. **Single-Zone Deployment:** One or more vSphere clusters mapped to a single availability zone.
2. **Multi-Zone Deployment:** One or more vSphere clusters mapped across exactly three zones.

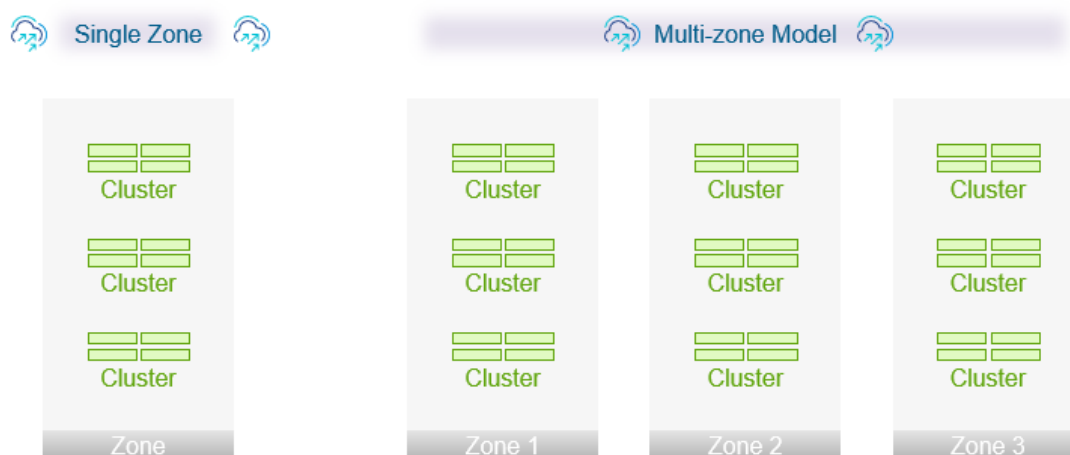


Figure. Single-Zone versus Multi-Zone deployment models in VCF.

Multi-zone deployments offer levels of application and data availability historically not possible with single-zone deployments. A multi-zone deployment can maintain full levels of availability in the event of failure or outage of a zone. However, there are some characteristics and requirements that may not make them ideal for all circumstances. For example.

- **Single vCenter Server.** A single vCenter Server must be used across the three zones participating in a multi-zone deployment.
- **Shared Network.** While the management network can use different port groups on the same virtual distributed switch (VDS), the workload networks must use the same network across zones. Supervisor deployment in VCF 9.1 will provide two options in the vSphere Client UI: NSX Virtual Private Cloud (VPC) and VDS.
- **Shared Storage Policy.** The vSAN datastores in each zone must be compatible with the shared storage policy.
- **Compliance Coherency.** Every rule in the policy must be satisfiable by the datastore in all respective zones.
- **Zone limits.** A multi-zone deployment should consist of precisely three zones. No more, and no less. Multi-zone deployments can only maintain availability if the zone failures do not exceed one zone.
- **Reliant on Application for cross-zone HA.** High availability of applications and data across zones is achieved through application-level HA and data replication respectively.

Support for more than one cluster in each zone for single zone and multi-zone deployments was introduced in VCF 9.1.1. This eliminates the historical limitation of a 1:1 mapping between a zone and a single vSphere cluster.

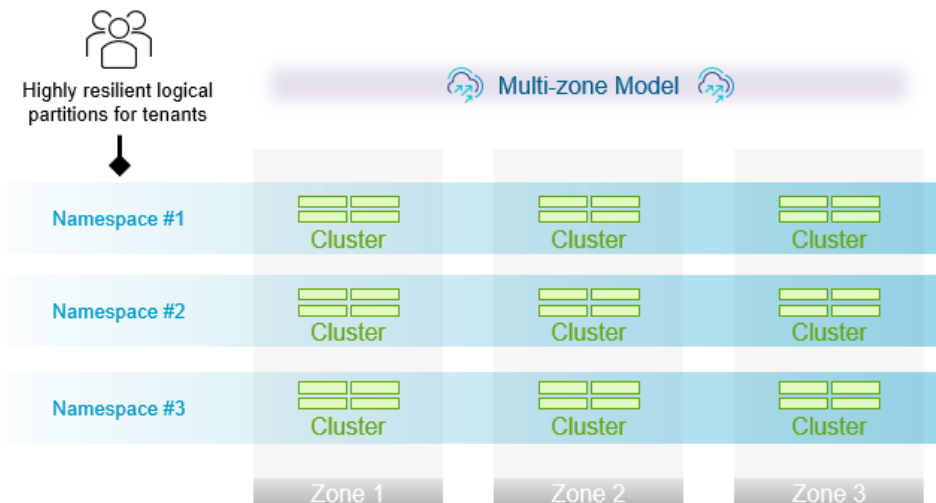


Figure. Support for more than one cluster in each zone for multi-zone deployments.

As shown in the figure above, each zone can contain multiple vSphere clusters. The namespace attaches to only one cluster with each zone at any single point in time. **These namespaces for a multi-cluster zone will select the cluster based on the available capacity of the cluster.** Cluster switching preserves the availability of data when the same datastore is mounted on another cluster within the zone. If the datastore requires maintenance to a level that prevents it from being available, the persistent volumes must be migrated to another datastore that supports the same storage policy. Note that even though namespaces create an abstraction layer to consume resources dispersed across zones, vSphere HA and vMotion activities are limited to within the specific cluster that is attached to the namespace. High availability is achieved through application-level techniques.

Kubernetes Topology Design with VCF Storage

The following eight topology design patterns summarize how Kubernetes can be deployed on VCF across a wide variety of topologies.

Zone Type	Design	Persistent Volume types validated	Stance
Single-Zone	Design 1: vSAN storage cluster with vSAN file services	RWO & RWX	Supported
Single-Zone	Design 2: vSAN stretched storage cluster with vSAN file services	RWO & RWX	Supported
Multi-Zone	Design 3: vSAN storage cluster isolated to each zone	RWO	Supported
Multi-Zone	Design 4: Multi-Zone vSAN storage cluster isolated to each zone, with per-zone vSAN file services	RWO & RWX	Supported
Multi-Zone	Design 5: Multi-Zone vSAN storage cluster isolated to each zone, with additional storage cluster for centralized shared vSAN file services	RWO & RWX	Supported with Caveats
Cross-Zone	Design 6: Cross-Zone vSAN stretched storage cluster	RWO	Recommended
Cross-Zone	Design 7: Cross-Zone vSAN stretched storage cluster with vSAN file services	RWX	Recommended
Cross-Zone	Design 8: Cross-Zone shared vSAN storage cluster	RWO	Recommended

Figure. Topology Design with VCF Storage Matrix

“Cross-zone” topologies described in this document are an adaptation of a multi-zone topology, with storage provided by a single vSAN cluster – stretched or non-stretched. This approach relies on vSAN for storage level resilience. Since the data is written synchronously across the datastore, it can be ideal for mission critical workloads, as a zone failure can quickly hand off compute responsibilities to pods in other zones, and does not need to wait for readiness of storage. As noted above, cross-zone topologies are generally recommended over other options.

Single-Zone Topologies

Single-zone deployments consist of one or more vSphere clusters mapped to a single availability zone.

Design 1: Single-Zone vSAN Storage Cluster with vSAN File Services

In this Single-Zone topology, compute and vSAN storage reside within the same zone.

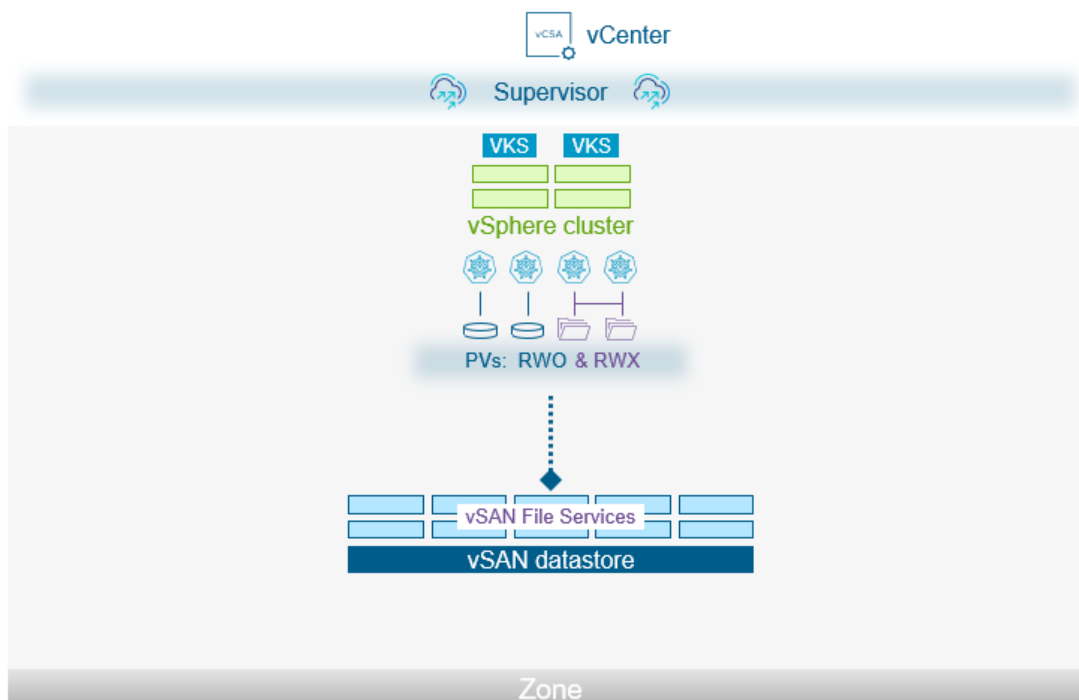


Figure. Single-zone vSAN storage cluster with vSAN file services.

A single-zone topology can be ideal for simpler environments and workloads that do not require beyond a single failure domain, such as a data center, campus building, room, or even a rack. This type of topology works well for environments

housing testing and development, as well as AI/ML. Persistent volumes can be block-based ReadWriteOnce (RWO), or file-based ReadWriteMany (RWX) courtesy of vSAN File Services. While it cannot tolerate an entire zone failure, the infrastructure-level resilience mechanisms within the zone (vSphere HA, vSAN as well as VKS container instantiation) make it highly tolerant of failures within the zone. This is a supported configuration.

Design 2: Single-Zone vSAN Stretched Storage Cluster with vSAN File Services

In this Single-Zone topology, compute and vSAN storage reside within the same zone, but the storage is stretched to maintain data availability in the event that a discrete vSAN fault domain fails.

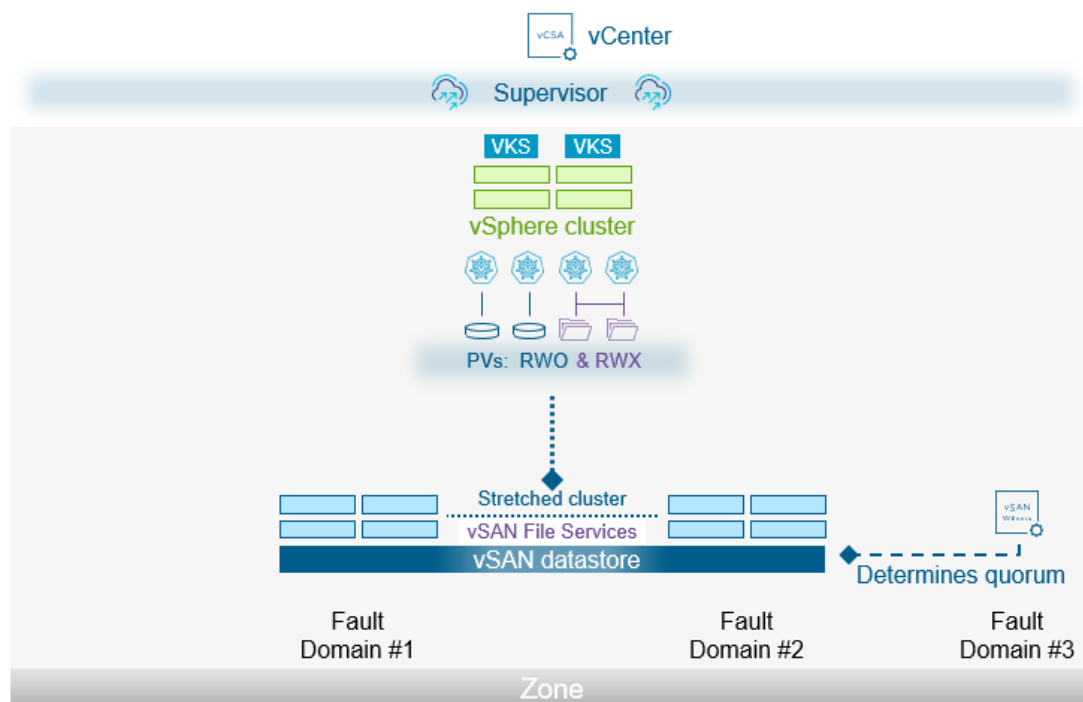


Figure. Single-zone vSAN stretched storage cluster with vSAN file services.

This variation of a single-zone topology uses a vSAN stretched cluster deployment to provide highly resilient storage across defined fault domains of “Preferred Site” “Non-Preferred Site” and “Witness Site.” Much like the previous example, persistent volumes can be block-based ReadWriteOnce (RWO), or file-based ReadWriteMany (RWX) courtesy of vSAN File Services. Validation testing demonstrated no disruption when the Non-Preferred Site or the Witness Site failed. For Preferred Site failures, Supervisor VMs took approximately 15-20 minutes to recover, and VKS VMs took about 10-15 minutes to recover – all courtesy of vSphere HA. This is a supported configuration.

Note that the use of a vSAN stretched cluster requires that the latency between sites must not exceed a round trip time (RTT) of 5ms. This topology will not be suitable for multi-zone topologies that cannot meet this requirement. For more information on the use of vSAN stretched clusters, see the document: “[vSAN Stretched Cluster Guide](#).”

Multi-Zone Topologies

Multi-Zone deployments distribute workloads across three distinct failure domains. Storage can be implemented locally per zone or disaggregated across zones.

Design 3: Multi-Zone vSAN Storage Cluster Isolated to Each Zone

This topology provides independent, dedicated vSAN storage clusters in each of the three zones.

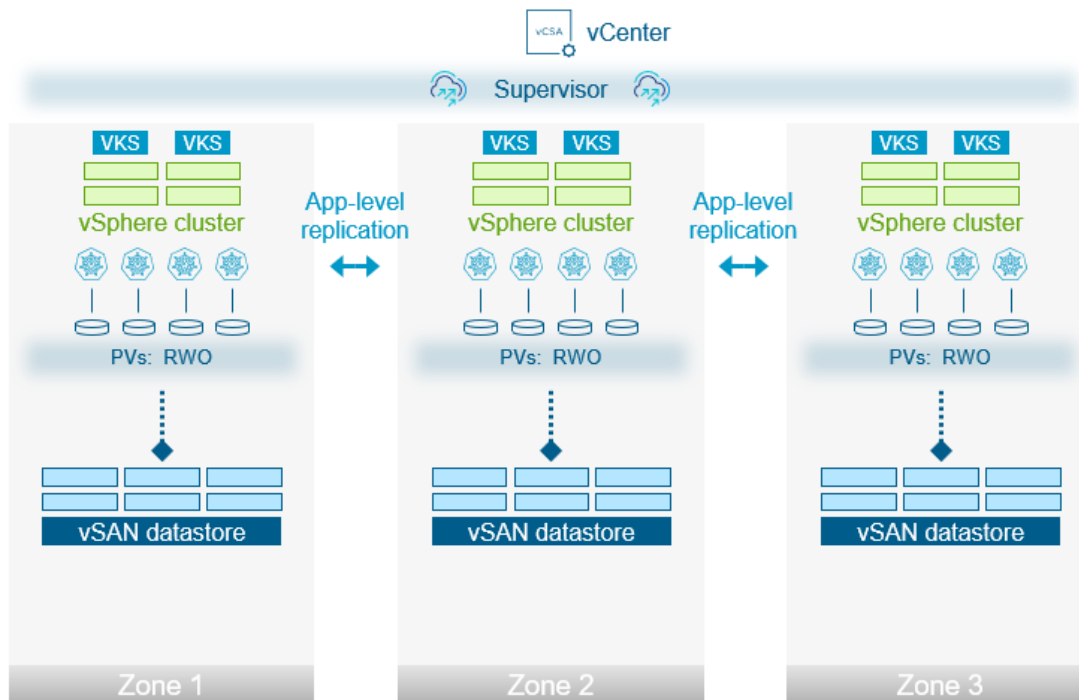


Figure. Multi-zone vSAN storage cluster isolated to each zone.

This multi-zone topology where each zone has its own storage works well with applications like Cassandra, PostgreSQL, and MongoDB, where built-in application-level replication ensures availability in the event of a zone failure. Persistent volumes (storage) and Pods (compute) are automatically co-located in each zone, enforced by the “WaitForFirstConsumer” storage class. In this configuration, block-based RWO volumes will be available for use by these applications. In addition to the infrastructure-level resilience mechanisms within the zone, this topology can tolerate up to one zone failure. This is a supported configuration.

Design 4: Multi-Zone vSAN Storage Cluster Isolated to Each Zone, with per-zone vSAN File Services

This topology provides independent, dedicated vSAN storage clusters in each of the three zones, with vSAN File Services running on each storage cluster.

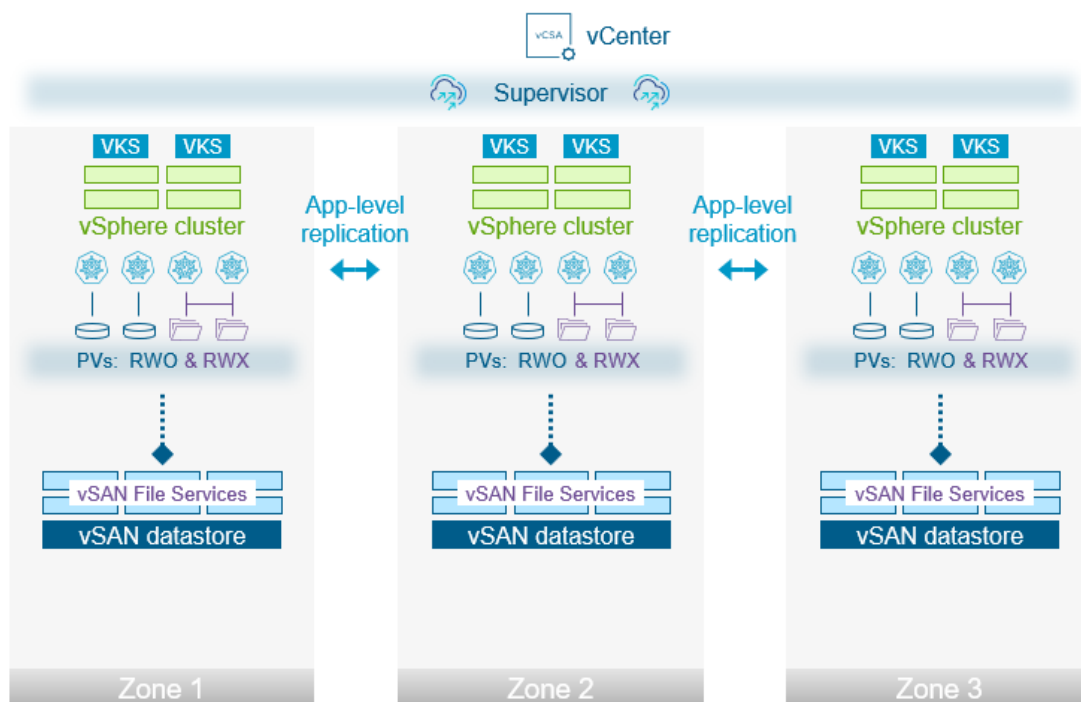


Figure. Multi-zone storage cluster isolated to each zone, With per-zone vSAN file services.

This is another multi-zone topology, where each vSphere cluster utilizes a dedicated vSAN storage cluster within the respective zone. This configuration has vSAN File Services enabled on each storage cluster, which provides the ability to not only serve up block-based RWO volumes, but NFS-based RWX volumes for highly flexible shared persistent volumes within the zone. This type of configuration is a more flexible variant of the topology previously discussed. This is a supported configuration.

Note that the use of a vSAN stretched cluster requires that the latency between sites must not exceed a round trip time (RTT) of 5ms. This topology will not be suitable for multi-zone topologies that cannot meet this requirement. For more information on the use of vSAN stretched clusters, see the document: “[vSAN Stretched Cluster Guide](#).”

Design 5: Multi-Zone vSAN Storage Cluster Isolated to Each Zone, with Additional Storage Cluster for Centralized Shared vSAN File Services

This multi-zone topology blends dedicated storage clusters in each of the three zones with a centralized storage cluster across zones.

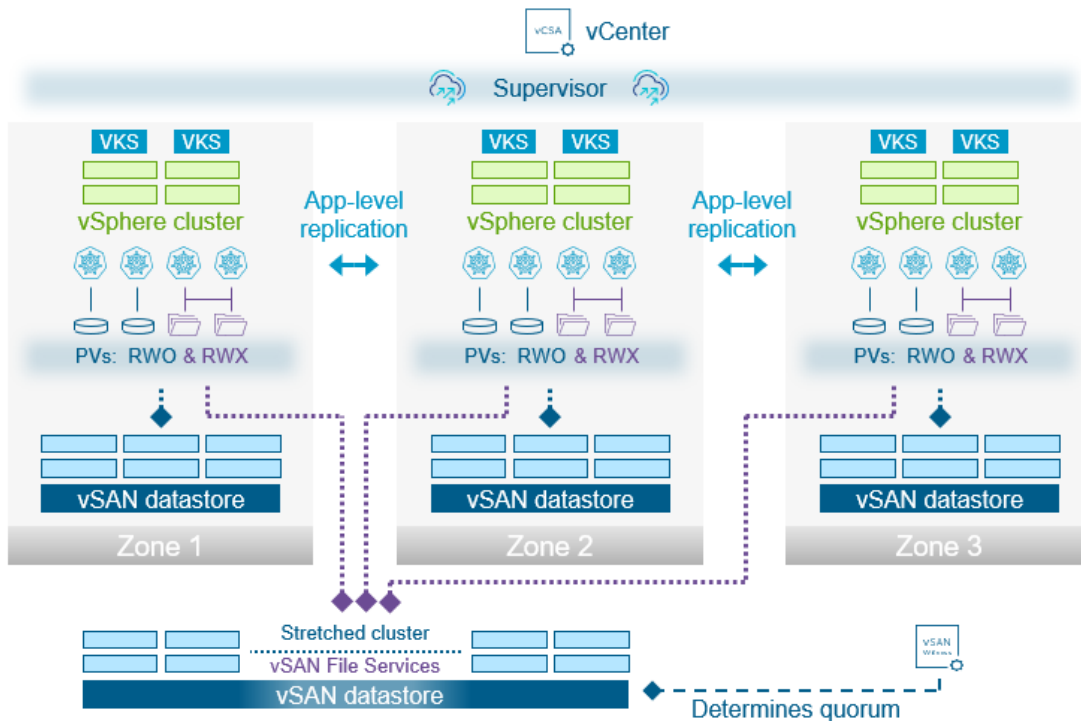


Figure. Multi-zone vSAN storage cluster isolated to each zone, with additional storage cluster for centralized shared vSAN file services.

This topology augments the previous configuration with an additional vSAN storage cluster that spans across all three zones. Each zone has its own storage just as with previous examples, but another storage cluster with vSAN file services would span across all three zones, providing storage-level resilience for RWO and RWX persistent volumes. **In this example, all control-plane data (Supervisor CP, VKS node OS disks, etc.) reside on the local storage within each zone, while all user-based persistent volumes reside on the vSAN stretched storage cluster spanning across all zones.** If a full zone failure occurred, where vSphere HA restarts VKS worker VMs on another site, recovery may take approximately 10-15 minutes for VKS VMs and the pods it instantiates. This could conceptually be used for workloads already protected by application-level HA that require NFS-based RWX volumes for easy shared storage across zones. This type of configuration is currently supported in 9.1.1, but requires that the vSAN datastore is mounted, even if RWX persistent volumes are the only type of persistent volume used.

Note that the use of a vSAN stretched cluster requires that the latency between sites must not exceed a round trip time (RTT) of 5ms. This topology will not be suitable for multi-zone topologies that cannot meet this requirement. For more information on the use of vSAN stretched clusters, see the document: "[vSAN Stretched Cluster Guide](#)."

Cross-Zone Topologies

Cross-Zone topologies extend the storage access layer across physical availability zones, transferring high-availability responsibilities from the application layer to the vSAN storage infrastructure.

Design 6: Cross-Zone vSAN Stretched Storage Cluster

The cross-zone topology eliminates per-zone storage services in favor of cross-zone storage using vSAN's multi-site/zone construct.

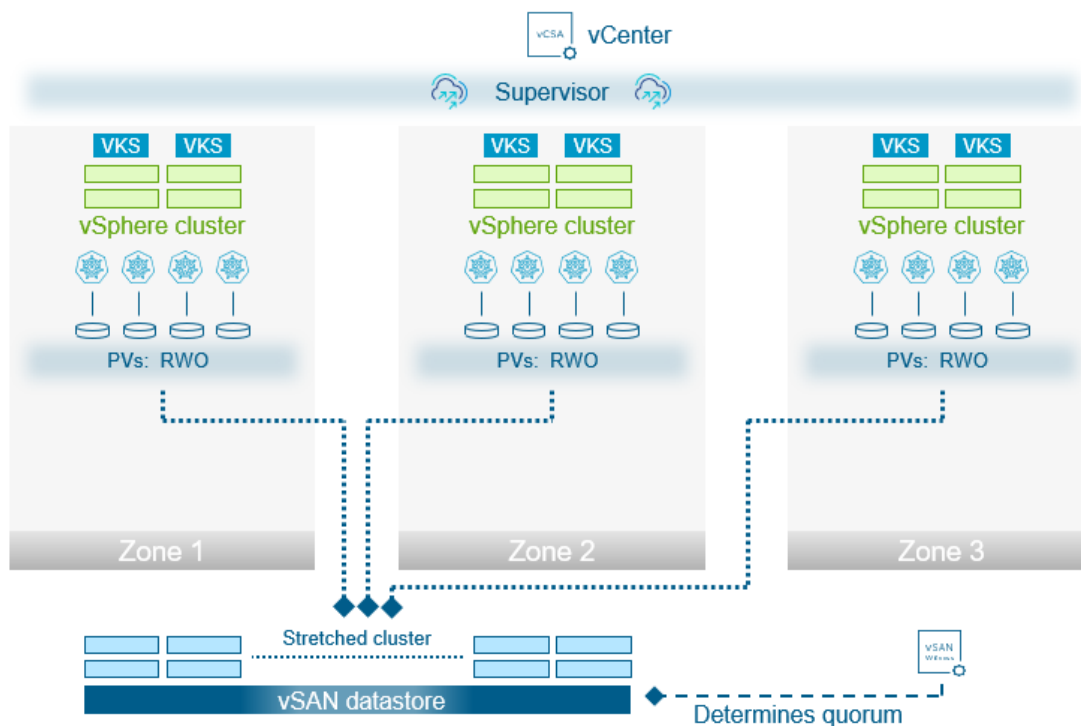


Figure. Cross-zone vSAN stretched storage cluster.

This cross-zone topology uses a similar approach to the previously mentioned design. Instead of a dedicated vSAN storage cluster within each zone, a vSAN stretched cluster is configured to provide highly resilient storage without using application-level storage replication technologies. vSAN provides full zone-level resilience with synchronous data mirroring across two of the three sites, and the third site is used for quorum arbitration. Just as with other configurations, VKS runs on the compute resources across all three instances but backed by a single vSAN datastore that can tolerate a zone outage and remain available. Upon a zone failure, worker nodes in the other zones can take over computational responsibilities with the same underlying persistent volumes. However, since these are RWO persistent volumes, they will be in an attached state to the failed zone. In VCF 9.1.1, a traditional manual detach and attach of a CNS volume attached to a disconnected VM may not complete correctly due to a lock on the volume. See [KB ID: 311105](#) and [“Recover Orphaned Virtual Machines”](#) in TechDocs for steps to ameliorate. This topology can be ideal for mission critical workloads needing only block based RWO persistent volumes that must remain resilient across zones. This is a **supported and recommended** configuration.

Note that the use of a vSAN stretched cluster requires that the latency between sites must not exceed a round trip time (RTT) of 5ms. This topology will not be suitable for multi-zone topologies that cannot meet this requirement. For more information on the use of vSAN stretched clusters, see the document: [“vSAN Stretched Cluster Guide.”](#)

Design 7: Cross-Zone vSAN Stretched Storage Cluster with vSAN File Services

This topology uses a shared cross zone storage infrastructure using vSAN, but adds file services capabilities.

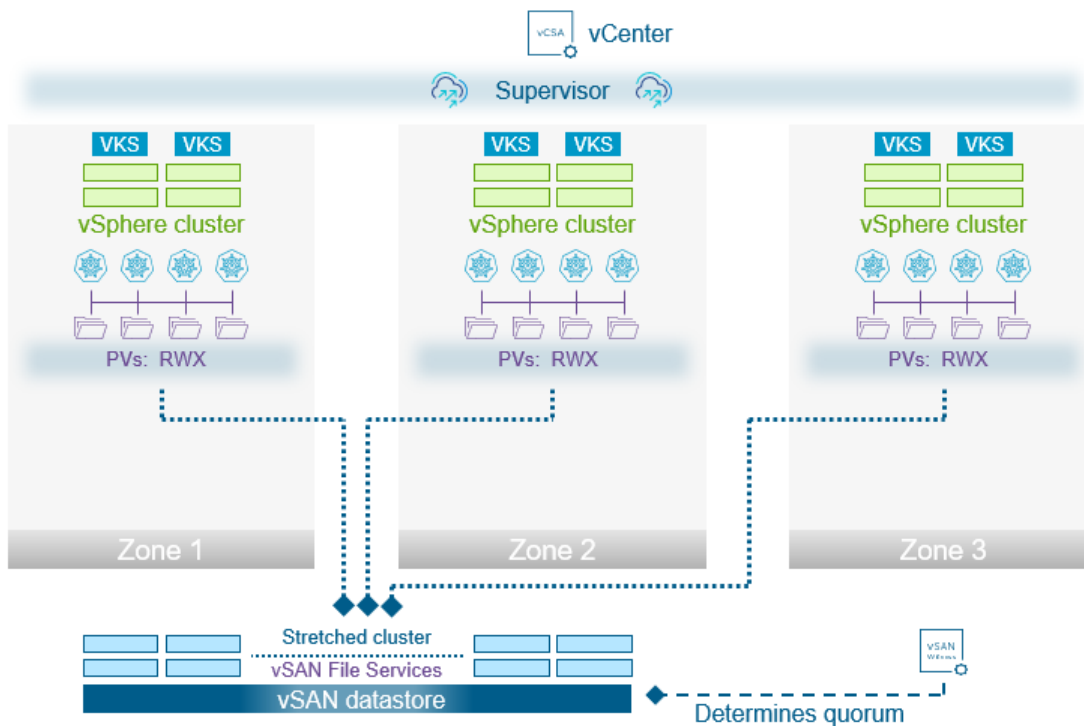


Figure. Cross-zone vSAN stretched storage cluster with vSAN file services.

This topology is like the previously mentioned cross-zone topology where a vSAN stretched cluster is configured to provide highly resilient storage across all three zones. However, this design runs vSAN file services to provide RWX volumes. By running vSAN file services on the vSAN stretched cluster, this storage topology can provide file-based RWX persistent volumes across zones. Since pods in each zone can mount to the same RWX volume, this provides continuous availability to the data. This topology can be ideal for mission critical workloads that also require shared persistent volumes. This is a **supported and recommended** configuration.

Note that the use of a vSAN stretched cluster requires that the latency between sites must not exceed a round trip time (RTT) of 5ms. This topology will not be suitable for multi-zone topologies that cannot meet this requirement. For more information on the use of vSAN stretched clusters, see the document: [“vSAN Stretched Cluster Guide.”](#)

Design 8: Cross-Zone Shared vSAN Storage Cluster

This topology uses a shared cross zone storage infrastructure without using vSAN’s site/zone construct.

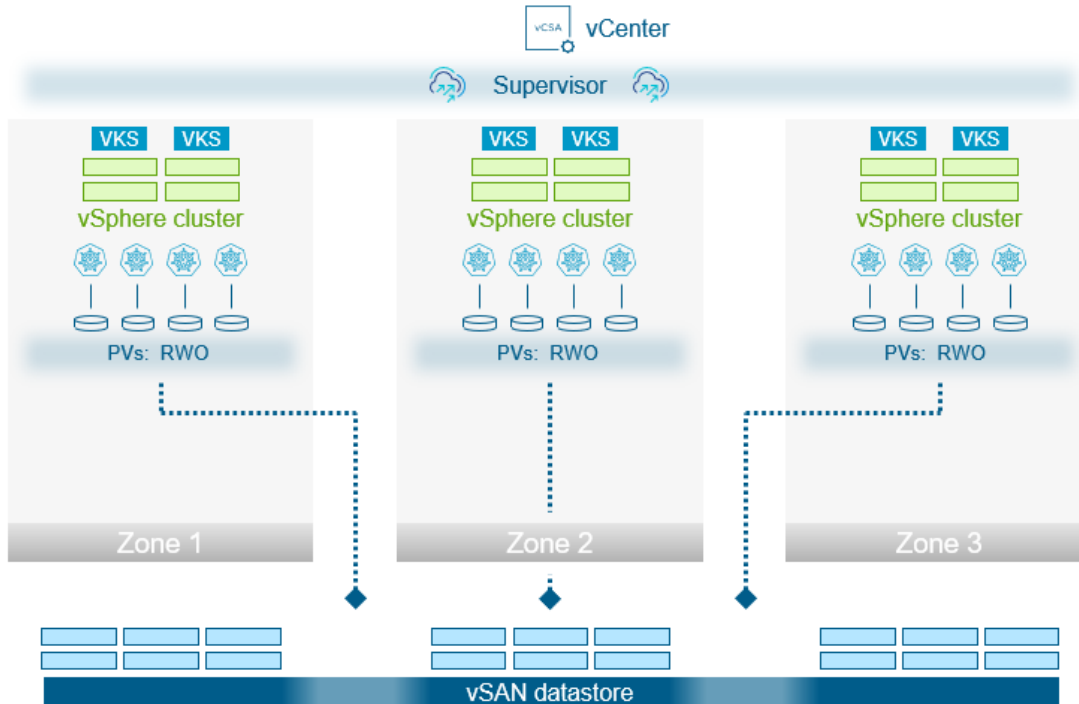


Figure. Cross-zone shared vSAN storage cluster.

Like the other cross-zone topologies described, this topology does not have any dedicated vSAN storage clusters within each zone. Storage is provided by a single vSAN cluster that spans across all three zones, but unlike the previous two cross-zone topologies, vSAN is not deployed as a stretched cluster. The vSAN cluster is configured using its “Fault Domains” capability that allows the hosts to be group within a respective fault domain that aligns with its physical zone. This distributes the persistent volume data across zones that will remain available in the event of a single, full zone outage. Application-state replication is still required to ensure the application instance can pick up operations upon a single zone failure. **This is a supported and recommended configuration.**

Note that the use of fault domains in vSAN requires that the latency between the fault domains must not exceed a round trip time (RTT) of 1ms. This topology will not be suitable for multi-zone topologies that cannot meet this requirement. For more information on the use of vSAN fault domains, see the post: “[Using Fault Domains in vSAN ESA.](#)”

Container Native Storage (CNS) Implementation Priorities

The Container Native Storage (CNS) plugin in vSphere acts as the bridge between Kubernetes StorageClass primitives and underlying vSAN object-based storage policies. Note that in vSAN for VCF 9.1, discrete storage policies will relinquish responsibility of resilience settings to Auto-RAID: a new feature in 9.1 that automatically assigns the maximum level of resilience to data based on the characteristics of the cluster. For more information, see the post: “Auto-RAID in VMware vSAN for VCF 9.1 – Comprehensive System-Managed Data Resilience.”

Key Technical Considerations for Your Ideal Storage Topology Selection

Volume Binding Modes

For Multi-Zone Local vSAN deployments, StorageClasses must enforce volumeBindingMode: WaitForFirstConsumer. This ensures volume creation is deferred until the Kubernetes Scheduler selects an execution node, preventing volume placement in a zone different from where the Pod is scheduled.

Application-Level vs. Storage-Level High Availability

If an application natively manages quorum, state, and intra-application synchronization (e.g. Cassandra, CockroachDB, Kafka, etc.), use multi-zone local storage. This avoids doubling up on data replication (vSAN synchronous replication and application replication).

If the application is legacy, monolithic, or lacks internal multi-node state replication, use a cross-zone vSAN stretched cluster configuration. This will provide highly durable synchronous replication, providing an RPO=0.

Network Latency Constraints

Multi-zone topologies (both isolated and cross-zone) require low latency and high-bandwidth interconnects between zones to support synchronous commit semantics for storage write operations. Higher latencies or insufficient bandwidth between zones will result in higher latencies for the application. vSAN stretched clusters require less than 5ms round trip time (RTT) between zones. For more information, see the document: “[vSAN Stretched Cluster Bandwidth Sizing](#).”

Multi-zone topologies also require contiguous L2 broadcast domains or NSX overlay networks that span all participating vSphere host clusters.

Recommendations

Designing Kubernetes infrastructure on VMware Cloud Foundation 9.1.1 requires matching the underlying vSAN storage architecture to the resilience capabilities of the hosted applications:

- **For Cloud-Native / Distributed Applications:** Implement Multi-Zone Local vSAN Topologies. Use application-layer replication while benefiting from failure isolation per zone.
- **For Mission-Critical / Legacy Applications:** Implement Cross-Zone vSAN Stretched Storage Clusters. This shifts high-availability responsibilities to the infrastructure layer, ensuring automatic recovery and zero data loss.
- **For Flexible Scaling:** Transition to Multi-Cluster Zones in VCF 9.1.1 to decouple compute maintenance from consumer-facing zones and streamline capacity expansion.

Some of the topologies described in this document imply that physical infrastructure may have some geographic distances between failure domains, such as zones and vSAN fault domains. Verify that your environment can meet the minimum requirements for the topology of interest. Common requirements include network bandwidth, network latency, physical host counts, etc.

Summary

VMware Cloud Foundation can provide the substrate for your cloud native applications that use Kubernetes. Your topology, availability requirements and applications used will dictate what configuration that works best for your environment.

Additional Resources

The following are a collection of useful links that relate to Kubernetes Topologies on vSAN.

[vSAN Stretched Cluster Guide](#). This guide describes the concepts and requirements around vSAN stretched clusters.

[vSAN Network Design Guide](#). This network design guide applies to environments running vSAN 9 and later.

[Migrating vSphere-Based Kubernetes Workloads to VKS Using Zero-Copy FCD Adoption](#). A guide for migration from vSphere-based Kubernetes to VKS on VCF 9.

[Migrating Workloads Between VKS Clusters](#). Migrating stateful and stateless workloads between VKS clusters.

[Migrating Non-vSphere Kubernetes Sources to VKS Using Filesystem Replication](#). Storage agnostic migration from any non-vSphere K8's distribution to VKS on VCF 9.

[Migrating Kubernetes Workloads to VKS Using Velero and S3 Compatible Object Storage](#). Migrating application objects and data from any Kubernetes source.

[vSAN technical blogs](#). Stay up to date on the most recently published technical information about vSAN. These posts are created by the vSAN Technical Marketing team.

[VMware Resource Center](#). The location for design guides, operations guides and other technical white papers on vSAN. These assets are created by the vSAN Technical Marketing and Product Enablement teams.

[Official vSAN documentation](#). The location for all “how to” documentation on vSAN.

About the Author

Pete Koehler is a Product Marketing Engineer in the VCF division at Broadcom. With a primary focus on vSAN, Pete covers topics such as design and sizing, operations, performance, troubleshooting, and integration with other products and platforms.



Copyright ©2025 Broadcom. All rights reserved.

The term "Broadcom" refers to Broadcom Inc. and/or its subsidiaries. For more information, go to www.broadcom.com. All trademarks, trade names, service marks, and logos referenced herein belong to their respective companies. Broadcom reserves the right to make changes without further notice to any products or data herein to improve reliability, function, or design. Information furnished by Broadcom is believed to be accurate and reliable. However, Broadcom does not assume any liability arising out of the application or use of this information, nor the application or use of any product or circuit described herein, neither does it convey any license under its patent rights nor the rights of others.