v1.0 | April 2023 | NVD-2177

NUTANIX VALIDATED DESIGN

Cloud Native: AOS 6.5 with Red Hat OpenShift Design

NUTANIX
YOUR ENTERPRISE CLOUD

Copyright

Contents

1. Executive Summary

Nutanix and Red Hat's strategic partnership offers enterprise customers a best-in-class solution for building, scaling, and managing cloud-native applications on-premises and in a hybrid cloud. Red Hat OpenShift on Nutanix provides a resilient, scalable infrastructure and a cloud-native application platform, allowing organizations to deliver innovative solutions, drive competitive advantage, and meet customer expectations.

This Nutanix Validated Design (NVD) includes combinable components and recipes, including validation steps, to ensure that your deployment meets best practice standards and is ready to support your business.

Nutanix recommends using Red Hat OpenShift Platform Plus, part of the Red Hat OpenShift portfolio, which provides a single hybrid cloud platform for enterprises to build, deploy, run, manage, automate, and safeguard intelligent applications at scale.

Key Reasons to Use Red Hat OpenShift on Nutanix

- Best-in-class infrastructure stack for enterprise cloud-native applications. Nutanix offers better resilience for both OpenShift platform components and application data, and superior scalability to bare-metal infrastructure. Nutanix web-scale architecture self-heals and recovers proactively, delivering up to 85 percent less downtime than legacy architecture.
- Choice of both the underlying hardware and the hypervisor. The built-in life-cycle management performs software and firmware upgrades with comprehensive dependency management and one-click simplicity across hardware platforms.
- Built-in data services to address persistent storage needs for stateful containers. This solution includes a full featured Container Storage Interface (CSI) for file and block storage, native S3-compatible object storage for

unstructured data, and, with Nutanix Database Service (NDB), an integrated multivendor database as a service (DBaaS).

- Reduced risk through combined support. Customers in need of technical support benefit from the combined expertise of the award-winning Red Hat and Nutanix global support teams.
- Combined innovation for today and the future. Joint roadmap planning and delivery from both Red Hat and Nutanix ensures the delivery of new capabilities that provide business value for organizations.

Audience

This guide is part of the Nutanix Solutions Library and is intended for individuals responsible for designing, building, managing, and supporting Red Hat OpenShift on Nutanix infrastructures. Readers of this document should already be familiar with Red Hat OpenShift, Nutanix AOS, the Hybrid Cloud: AOS 6.5 with AHV On-Premises NVD, the Hybrid Cloud: AOS 6.5 with AHV Unified Storage NVD, and the Red Hat OpenShift on Nutanix tech note.

Purpose

This document describes Red Hat OpenShift components, integration, and configuration and covers the following topics:

- Core Nutanix infrastructure and related technology
- Red Hat OpenShift for running cloud-native applications
- Bill of materials

Software Versions

This NVD is validated and tested on the listed software versions, though you can use minor updates on each product.

Table: Software Versions Used in Validation Testing

Component	Software Version
Prism Central	2022.6.0.2
Nutanix AOS	6.5.1.8
Nutanix AHV	el7.nutanix.20220304.336
Nutanix CSI Operator	2.6.2
Red Hat OpenShift Container Platform (OCP)	4.12.4
Red Hat Advanced Cluster Management	2.7.2
Red Hat Quay Registry	3.8.4
Nutanix Files Manager	4.2.1.1
Nutanix Files	4.2.1.1
Nutanix Objects	3.6
Nutanix Objects Manager	3.6
Nutanix Life Cycle Manager (LCM)	2.5.0.4
Nutanix Cluster Check (NCC)	4.6.1
Global Load Balancer F5 BIG-IP	16.1.0 LTS

Document Version History

Version Number	Published	Notes
1.0	April 2023	Original publication.

2. General Platform Design

This solution uses Red Hat OpenShift as a single platform for cloud-native applications that lets the applications operate consistently and innovate continuously.

Red Hat Advanced Cluster Management (RHACM) hosted on the Nutanix management cluster manages the OpenShift clusters running in the Nutanix workload.

Based on business requirements, you can run a single, large footprint OpenShift cluster or multiple, smaller footprint clusters (test or dev) that differ in control plane size.

The size of the worker nodes always depends on application needs. See the Scalability and Performance guidelines for more information.

The following tables provide platform design requirements, assumptions, risks, and constraints.

Table: Platform Design Requirements

Platform	Component	Description
Nutanix	Management	Automate deployment of OCP clusters.
Nutanix	Storage	Provide CSI integration with underlying storage.
Nutanix	Storage	Support NFS, block storage, and S3.
Nutanix	Platform	Limit virtual CPU overcommitment to 2:1, or 2 vCPU per physical CPU core for OpenShift infrastructure and worker VMs.

Platform	Component	Description
Nutanix	Platform	Don't overcommit virtual CPU for OCP control plane VMs.
Nutanix	Monitoring	Enable platform fault monitoring and use email to send alerts.
Nutanix	Monitoring	Keep resource usage under 75 percent; usage over 75 percent generates an email alert.
Nutanix	Monitoring	Use email as the primary channel for event monitoring alerts.
Nutanix	Monitoring	Ensure that event monitoring is resilient. For example, when the management plane is the primary source of alerts, there must be a secondary method for monitoring the management plane itself. Then, if the management plane fails, an alert from the secondary source can trigger the action to recover the management plane.
Nutanix	Monitoring	Facilitate automated issue discovery and remote diagnostics.
Nutanix and OpenShift	Networking	Provide network segmentation and isolation on a container level.
Nutanix and OpenShift	Monitoring	Monitor performance metrics and store historical data for the past 12 months in Nutanix Objects.

Platform	Component	Description
OpenShift	Scalability	Deploy and manage applications across the whole OpenShift environment.
OpenShift	BCDR	Support a backup solution with an RPO and RTO of 15 min.
OpenShift	Platform	Support OVN-Kubernetes as the default Container Network Interface (CNI).
OpenShift	Platform	Support at least two control plane sizes: small footprint and large footprint.
OpenShift	Platform	Support disconnected installation.
OpenShift	Platform	Provide enterprise container registry for additional application security.
OpenShift	Platform	Use dedicated OCP infrastructure VMs to isolate infrastructure workloads from user workloads in OCP clusters.
OpenShift	Monitoring	Enable and configure the OpenShift Platform monitoring stack to store data in persistent volume in each OCP cluster.
OpenShift	Monitoring	Centralize monitoring of each OCP cluster with RHACM and store logs and metrics in Nutanix Objects.

Table: Platform Design Assumptions

Platform	Component	Description
Nutanix	Monitoring	IT operations teams can continuously staff the mailbox that receives monitoring alerts to address critical issues in a timely manner.
Nutanix	Monitoring	IT operations teams can provide email infrastructure with sufficient resilience to send, receive, and access emails even during critical outages.
Nutanix	Monitoring	Network security appliances allow the management plane to transmit telemetry data to Nutanix.
Nutanix	Platform	IT operations teams can deploy Active Directory, DNS, and a web load balancer (like F5 BIG-IP) in a highly available configuration in each Nutanix management cluster.

Table: Platform Design Risks

Platform	Risk	Mitigation
Nutanix	Datacenter outage	Fail over to disaster recovery site.
Nutanix	Nutanix cluster outage	Fail over to disaster recovery site.

Platform	Risk	Mitigation
Nutanix	Monitoring outage	If Prism Central becomes unavailable for any reason, the platform can no longer send alerts. To mitigate this risk, configure each Prism Element instance to send alerts as well. As this approach results in duplicate alerts during normal operations, send Prism Element alerts to a different mailbox that you can monitor when Prism Central is unavailable.
OpenShift	Scalability	Label OCP worker VMs with an underlying AHV host to mitigate the risk of multiple pods in a ReplicaSet running on the same AHV host.

Table: Platform Design Constraints

Platform	Component	Description
Nutanix	Platform	The number of VMs per Nutanix infrastructure pod doesn't exceed 7,500 (the limit of Flow Network Security policies per Prism Central instance).
OpenShift	Platform	The maximum number of OCP worker VMs (4 vCPU and 16 GB RAM) per Nutanix workload cluster is 360 (24 per usable AOS node).

Platform	Component	Description
OpenShift	Platform	The maximum number of OCP worker nodes per OCP control plane in default configuration is 500 (limited due to default /14 CIDR).
OpenShift	Platform	OCP control plane doesn't span across availability zones (AZs) or AOS clusters (not supported today for OpenShift IPI installer and Nutanix CSI).

3. Core Infrastructure Conceptual Design

This design incorporates three distinct Nutanix cluster types:

1. Nutanix management cluster: critical infrastructure and environment management workloads including OCP control planes with a large footprint
2. Nutanix workload cluster: AOS cluster running user workloads
3. Nutanix storage cluster: dedicated cluster to store application backup data locally

This section defines the overall high-level design, platform selection, capacity management, scaling, and resilience. This design follows the block-and-pod architecture defined in the Nutanix Hybrid Cloud Reference Architecture.

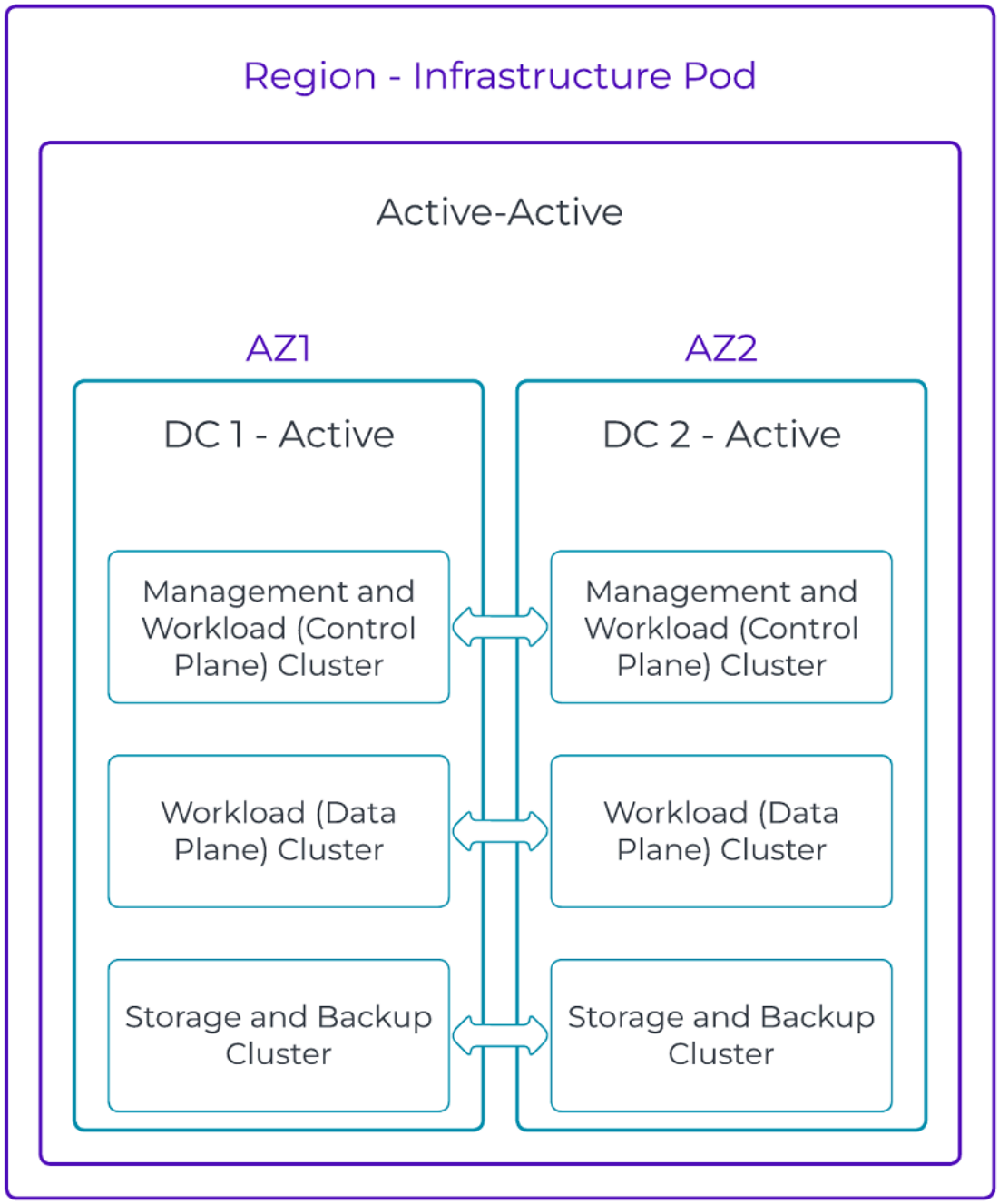


Figure 1: Cloud Native Conceptual Pod Design

The platform selection is aligned to our Hybrid Cloud NVD to enable a migration path to a full hybrid cloud or Nutanix Unified Storage validated design.

Table: Platform Design Decisions

Decision Name	Decision
Number of regions	Use 1 region.
Number of AZs	Use 2 AZs.
Number of datacenters	Use 2 datacenters: 1 per AZ.
Workload types per cluster	Use dedicated workloads per Nutanix cluster, as this design is for containerized applications on Red Hat OpenShift.
Minimum Nutanix workload cluster building block size	Use at least 4 AOS nodes as a minimum Nutanix cluster size.
Nutanix workload cluster building block expansion increments	Use 1 AOS node.
Maximum Nutanix workload cluster building block size	Use at most 16 AOS nodes.
Maximum Nutanix workload cluster building blocks per pod	Use at most 4 Nutanix workload cluster building blocks (1 per AZ) per pod.
Maximum number of running VMs per usable AOS node in the Nutanix workload cluster building block	Use at most 124 small VMs per usable AOS node.
Maximum number of VMs per Nutanix workload cluster building block	Use at most 1,860 small VMs per workload cluster building block.
Nutanix workload cluster building block node redundancy	Use n + 1 for redundancy.
Maximum usable AOS nodes per maximum Nutanix workload cluster building block	Configure at most 15 usable AOS nodes per maximum Nutanix workload cluster building block.
Nutanix workload cluster building blocks configuration	Use one rack per Nutanix workload cluster building block.
AOS cluster replication factor	Use replication factor 2.
AOS cluster high availability configuration	Guarantee high availability.

Table: Platform Selection

Cluster Management	Workload	Storage
Node type	NX-3170N-G8	NX-3170N-G8
Node count	4–8 (increments of 2)	4–16 per building block (increments of 4, up to 16 maximum)
Processor	2 Intel Xeon Gold 5318Y 24-core 165 W 2.1 GHz (Ice Lake)	2 Intel Xeon Gold 5318Y 24-core 165 W 2.1 GHz (Ice Lake)
Memory	8 × 64 GB 2,933 MHz DDR4 RDIMM (512 GB total)	12 × 128 GB 2,933 MHz DDR4 RDIMM (1.5 TB total)
SSD	6 × 3.84 TB	6 × 3.84 TB
HDD	N/A	N/A
NIC	25 GbE Dual SFP+	25 GbE Dual SFP+
Form factor	1RU of single nodes	1RU of single nodes

OpenShift Infrastructure Conceptual Design

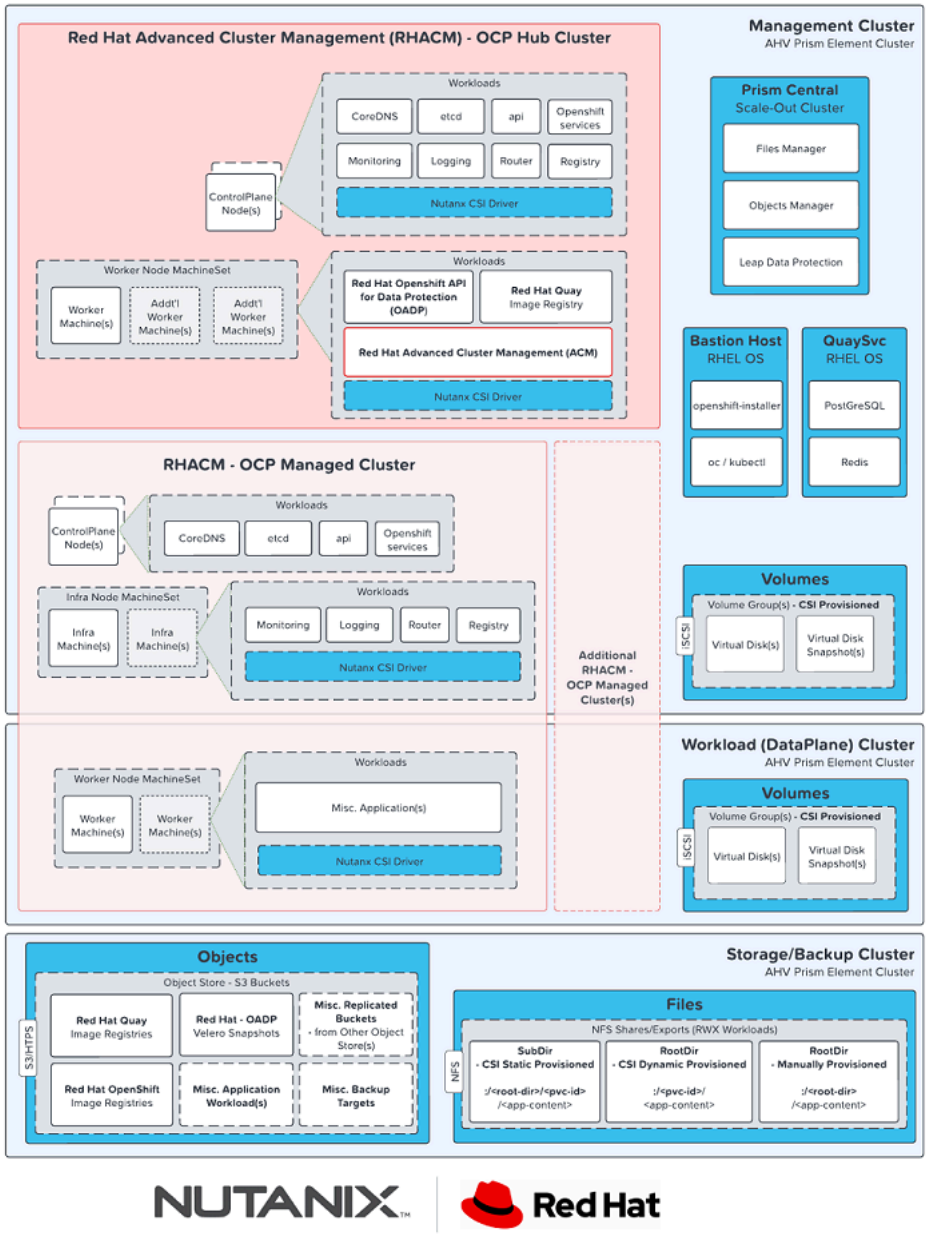


Figure 2: Red Hat OpenShift on Nutanix Cluster Architecture Overview

Table: OpenShift Cluster Sizing: OCP Hub (RHACM)

Component	Instances	Size
Control plane	3	8 CPU cores, 32 GB
Worker	6	8 CPU cores, 32 GB

Table: OpenShift Cluster Sizing: OCP Cluster (Workload) Small Footprint

Component	Instances	Size
Control plane	3	4 CPU cores, 16 GB
Infrastructure	3	4 CPU cores, 16 GB
Worker	2+ (max 25)	4 CPU cores, 16 GB

Table: OpenShift Cluster Sizing: OCP Cluster (Workload) Large Footprint

Component	Instances	Size
Control plane	3	16 CPU cores, 128 GB
Infrastructure	3	16 CPU cores, 128 GB
Worker	2+ (max 360)	Minimum 4 CPU cores, 16 GB

Quay Image Registry Conceptual Design

Red Hat Quay container registry platform provides secure storage, distribution, and governance of containers and cloud-native artifacts across OpenShift and other Kubernetes clusters. It allows you to keep OpenShift disconnected from the internet to provide higher security and reliability.

Quay is run as a geo-replicated setup in the management plane on OCP hub clusters with Nutanix Objects as backend storage. F5 BIG-IP provides global load balancing. The following image shows the Quay geo-replication process.

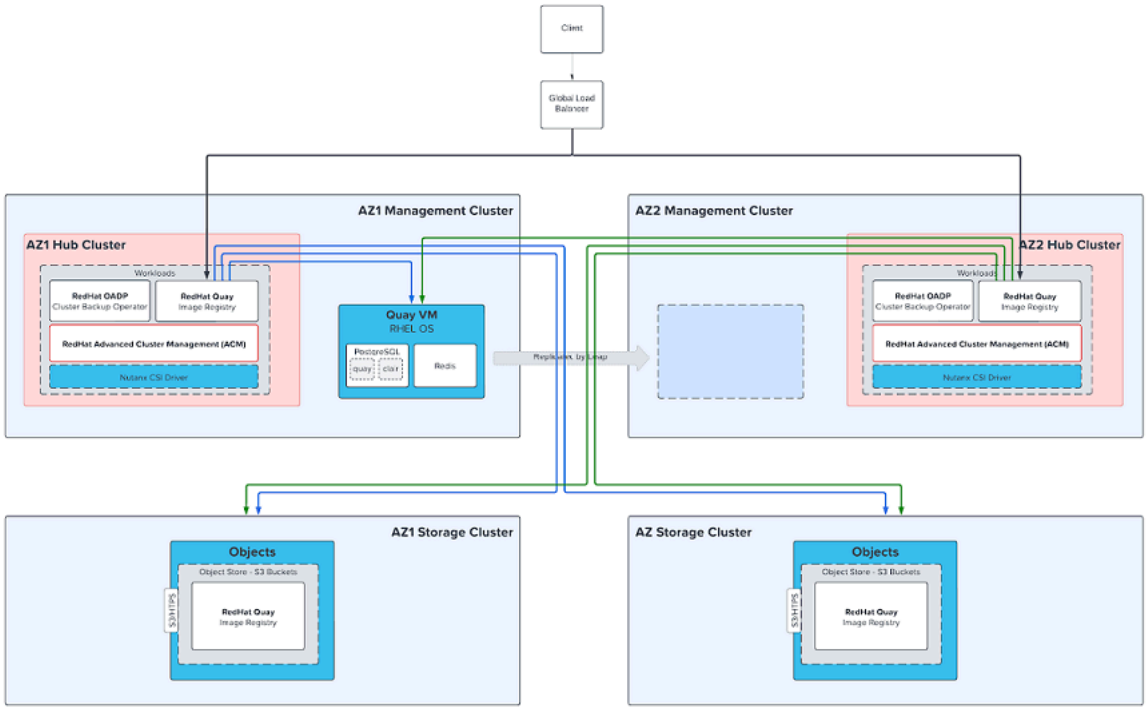


Figure 3: Quay Geo-Replication

OCP workload clusters use Quay Bridge Operator instead of an internal image registry, which simplifies the configuration. The Quay Container Security Operator fetches vulnerability scan results from Clair components and provides information in the OpenShift console of each OCP cluster.

Table: Image Registry Requirements

Component	Description
Security	Implement vulnerability scanning on images.
OpenShift	Integrate into OCP workload clusters.
Storage	Integrate into Nutanix Objects.
Manageability	Provide support for disconnected environments.

Table: Image Registry Risks

Risk Description	Impact	Likelihood	Mitigation
Datacenter outage	Small	Unlikely	If a database VM is impacted, protection domains recover the VM on the secondary side.
OCP hub cluster outage	Small	Unlikely	Quay can still provide services.
Object store outage	Small	Unlikely	Quay can still provide services.

Table: Image Registry Design Decisions

Component	Description
Manageability	Enable Quay Bridge Operator on OCP workload clusters.
Network	Use F5 BIG-IP as a global load balancer to access Quay.
Resilience	Install Quay in geo-replication mode.
Resilience	Use Red Hat Enterprise Linux (RHEL) VM for PostgreSQL and Redis; protect VM with Nutanix protection domains.
Security	Enable proxy cache.
Security	Configure mirroring of external repositories.
Security	Enable Quay Container Security Operator on all OCP clusters.

4. Scalability

Scalability is one of the core concepts of the Nutanix platform and refers to the ability to increase storage and compute capacity to meet both current and future workload demands. A well-designed cluster meets current requirements while providing a path to support future growth.

Scalability Conceptual Design

This NVD is based on a block-and-pod architecture. This section explains how to scale the solution with both OCP components and the underlying Nutanix infrastructure.

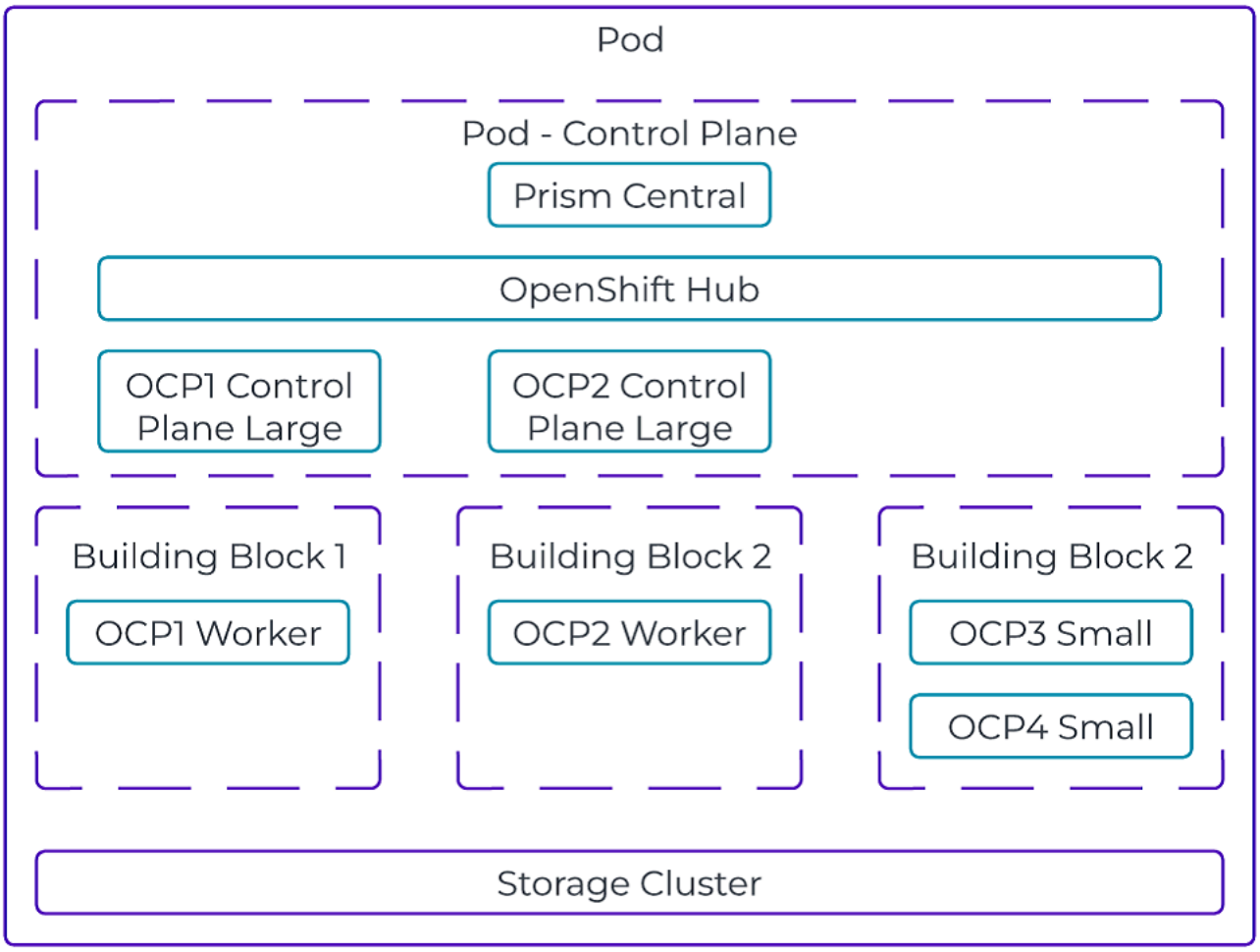


Figure 4: Cloud Native OpenShift Scalability

Core Infrastructure Scalability

This NVD permits horizontal and vertical scaling within the boundaries set by running workloads in a single rack per AOS cluster in a single AZ. If the workload grows, you can add nodes and storage capacity to the management or workload AOS clusters.

This design has a maximum of 16 AOS nodes per AOS cluster to allow you to use Nutanix LCM to complete nondisruptive Nutanix software, hardware, firmware, and driver maintenance in a 16-hour maintenance window (using Nutanix NX model hardware). You may use a smaller maximum size per Nutanix

workload building block to shorten maintenance windows and allow more small Nutanix clusters per Nutanix infrastructure pod without changing the maximum number of AOS nodes or VMs that each Nutanix infrastructure pod supports. For example, an 8-node Nutanix workload cluster building block reduces maintenance windows by half and allows twice the number of Nutanix clusters per Nutanix infrastructure pod without changing the number of AOS nodes supported. However, the number of usable AOS nodes decreases with the smaller cluster size, as 1 node per cluster is logically reserved for maintenance and failure.

Nutanix Management Cluster

The starting point is 4 AOS nodes and the maximum AOS cluster size is 16 nodes. The base AOS cluster can support infrastructure management plane workloads like Prism Central and OpenShift management (OCP hub).

If you want to use an OCP cluster with a large footprint, add the dedicated OCP control plane for up to 360 OCP worker nodes to the Nutanix management cluster. If you need to add more OCP clusters with a large footprint, add 2 more AOS nodes into the Nutanix management cluster to avoid CPU contention. A single management cluster can support up to seven OCP control planes with a large footprint.

If you use OCP clusters with a small footprint, the OCP control planes don't run in the Nutanix management cluster.

Nutanix Workload Cluster

The starting footprint is 4 AOS nodes, and you can scale up to 16 AOS nodes. If you need to scale above 16 AOS nodes, create a new Nutanix workload cluster.

As the size of OCP worker VMs might vary, each AOS node's memory is fully populated to accommodate the resulting mixed memory requirements. This approach also provides maximum memory performance, even if you don't need it. If memory pressure increases, add more nodes. The design uses all-flash disks to accommodate peak workload demands.

> Note: If workloads change or grow, you must account for that in the workload cluster and in any additional workload clusters in the resource locations.

Nutanix Storage Cluster

A Nutanix storage cluster can scale up to 8 AOS nodes, but the starting point is 4 nodes. A single cluster can support multiple Nutanix workload clusters (building blocks).

> Note: The number of building blocks supported by a single backup cluster depends on the application storage footprint and backup retention policies.

OpenShift Scalability

Red Hat OpenShift supports both scale-up and scale-out mechanisms. Follow these guidelines when scaling an OpenShift cluster:

- Add no more than 15 OCP workers at the same time per MachineSet.
- Scale up OCP workers (add CPU and RAM) until the Kubernetes Pod limit is reached (250 pods or OCP workers), up to a maximum of 16 CPU cores and 128 GB.
- Further scale out OCP workers up to 360 VMs; if you need to go beyond 360 OCP worker VMs, start a new OCP cluster.
- Always use proper OCP worker-node labeling and pod-placement rules to avoid running multiple pods on the same AHV host.

Increase limits for maxPods and number of worker nodes according to OpenShift documentation, which includes changes to Kubelet settings, hostPrefix, and CIDR ranges for host networks.

Table: OCP Scalability Summary

Component	Starting	Maximum	Increment	Comment
OCP hub control plane	3 VMs	3 VMs	N/A	OCP hub can support a maximum of 10 OCP clusters with a large footprint.

Component	Starting	Maximum	Increment	Comment
OCP hub worker	3 VMs	6 VMs	N/A	OCP hub can support a maximum of 10 OCP clusters with a large footprint. For additional workloads like ACS or GitOps, scale worker nodes to 6.
OCP cluster control plane	3 VMs	3 VMs	N/A	To scale about 25 OCP workers, use OCP control plane with a large footprint.
OCP cluster infrastructure	3 VMs	Based on infrastructure needs	1	To scale about 25 OCP workers, use OCP control plane with a large footprint.
OCP cluster worker	2 VMs	360 VMs	1	To scale about 25 OCP workers, use OCP control plane with a large footprint.

Table: Nutanix Scalability Summary

Component	Starting	Maximum	Increment	Comment
Nutanix management cluster	4 nodes	16 nodes	2	Nutanix management cluster can support a maximum of 7 OCP control planes for a large footprint.
Nutanix worker infrastructure	4 nodes	16 nodes	1	N/A

5. Resilience

Nutanix provides many resilience features, including storage replication, snapshots, block awareness, degraded node detection, and self-healing. These capabilities increase the resilience of all workloads, even if the application itself has limited resilience options. Nutanix layers these software features on hardware designed with resilience in mind (for example, with redundant physical components and power supplies, many of which are hot-swappable or otherwise easily serviceable). Running workloads in a virtualized environment adds another kind of resilience, as you can perform many maintenance operations without application downtime. A resilient network fabric that can sustain individual link, node, or block failures without significant impact completes the architecture.

Core Infrastructure Resilience Conceptual Design

All components are physically redundant. The physical components include the top-of-rack switches, the nodes and their internal parts, and the datacenter itself in case of a disaster.

To protect workloads to meet or exceed service-level agreements (SLAs), this NVD separates the Nutanix workload clusters from the management and backup clusters. The Nutanix management, workload, and backup cluster sizing allows n + 1 failure redundancy. Monitoring and alerting ensure that any issues result in an alert; consistently monitoring workload growth ensures that sufficient headroom is available at any time.

There is no ideal AOS cluster size for a generic workload. This NVD uses 16 single-chassis AOS node building blocks to take advantage of block awareness, a key platform resilience feature.

Table: Resilience Design Decisions

Decision Name	Decision
Component redundancy	Ensure the full redundancy of all components in the datacenter.
Availability resilience	Replicate application data to a secondary AZ.
Nutanix cluster resilience	Ensure that every Nutanix cluster is designed to support n + 1 redundancy.
Rack resilience	Back up application data out of the AOS cluster.

Replication factor 2 protects against the loss of a single component in case of failure or maintenance. During a failure or maintenance scenario, Nutanix rebuilds any data that falls out of compliance much faster than traditional RAID data protection methods. Rebuild performance increases linearly as the AOS cluster grows.

In the Nutanix architecture, rapid recovery in the event of failure is the standard, and there are no single points of failure. You can configure the AOS cluster to maintain three copies of data; however, for general server virtualization, Nutanix recommends that you distribute application and VM components across multiple AOS clusters to provide greater resilience at the application level.

Note: You can achieve rack-aware resilience when you split AOS clusters evenly across at least three racks, but this NVD doesn't use that approach because it adds configuration and operational complexity. Nutanix cluster replication factor 2 in this design is sufficient to exceed five nines of availability (99.999 percent).

Figure 5: Cloud Native OpenShift on Nutanix Availability

OpenShift and Application Resilience

OpenShift provides resilience for the container. If a container dies and is part of a Deployment, StatefulSet, ReplicaSet, or so on, OpenShift restarts it. This resilience is provided within the boundary of a single OCP cluster. You should also deploy a second OCP cluster in a different datacenter where the OCP clusters are completely independent and don't share any resources.

By using smaller and simpler OCP clusters, you can create multiple small failure domains.

Instead of using manual deployments, rely on GitOps features provided by RHACM to ensure reproducible deployments and consistency in any OCP cluster.

If possible, run your application in active-active mode across both sites, using a global load balancer on top. Otherwise, use active-passive mode, which always requires an automated or manual activity to activate the passive side.

In case of a stateful application, design the application to stay on the paradigm of shared nothing, even if running across multiple OCP clusters, using components that work well spread across them.

6. Storage Design

Nutanix uses a distributed, shared-nothing architecture for storage. For a discussion of Nutanix storage constructs, refer to the Storage Design section in the Nutanix Hybrid Cloud Reference Architecture. For information on node types, counts, and physical configurations, see the Platform Design section in this document.

Creating an AOS cluster automatically creates the following storage containers:

- NutanixManagementShare: Used for Nutanix features like Nutanix Files and Nutanix Objects and other internal storage needs. This storage container doesn't store workload vDisks.
- SelfServiceContainer: Used by the Nutanix Self-Service Portal and automation services.
- Default-Container-XXXX: Used by VMs to store vDisks for user VMs and applications.

> Note: You can delete the Default-Container and create a new one with your desired naming convention.

Because this NVD provisions workloads from images with NCM Self-Service, the SelfServiceContainer serves the workload and backup clusters. In both AZs, the management cluster uses the Default-Container to store VMs and their vDisks. This NVD enables inline compression on the Default-Container for all Nutanix management and workload clusters in the AZs. Because these AOS clusters have a fault tolerance level of 1, the replication factor for the containers is 2.

Storage Conceptual Design

Nutanix Files and Nutanix Objects are deployed on a dedicated Nutanix cluster in each AZ. The Nutanix Unified Storage Validated Design (NUS NVD) is the foundation for the Nutanix storage cluster; all configuration settings and design decisions are aligned with the current NUS NVD.

Data Reduction Options

To increase the effective capacity of the AOS cluster, this design enables inline compression with compression delay of zero. Nutanix disabled deduplication for this container and used the defaults for the other containers.

Table: Data Reduction Settings

Container	Compression	Deduplication	Erasure Coding
Default-Container-XX	On	Off	Off
NutanixManagementShare	On	Off	Off
SelfServiceContainer	On	Off	Off
PVC-OCP-XX	On	Off	Off

Table: Storage Design Requirements

Component	Description
Management	Integrate into an existing management plane.
Connectivity	Support the NFS and S3 protocols.
Application	Provide read-write once (RWO) and read-write many (RWX) persistent storage for containers.
Application	Provide S3-compatible storage for applications.
Business continuity and disaster recovery	Provide S3-compatible storage for application backup.
Resilience	Provide an application design that encourages a shared-nothing approach across the datacenters.

Table: Storage Design Risks

Component	Description
Management	If Prism Central becomes unavailable, the Nutanix Objects service continues to function, but object store creation, upgrade, and scale-out operations aren't available.
Business continuity and disaster recovery	Buckets are replicated to remote AZ. A global server load balancer routes traffic to the primary AZ until a planned or unplanned failover occurs.

Table: Storage Design Constraints

Component	Description
Nutanix Files service	The number of Nutanix Files file server VMs (FSVMs) per AOS cluster in a fully scaled Nutanix Files deployment doesn't exceed 16. The number of FSVMs doesn't exceed the number of physical nodes in the cluster.
Nutanix Objects clusters	The number of load balancer VMs doesn't exceed four. The number of worker VMs doesn't exceed the number of physical nodes in the cluster.

Table: Storage Design Decisions

Decision Name	Decision
Sizing an AOS cluster	All-flash configuration is mandatory for containerized workloads; Nutanix recommends write-optimized disks or NVMe.
Node type vendors	Don't mix node types from different vendors in the same AOS cluster.
Node and disk types	Use identical node types that have similar disks. Disks should be write-optimized.
Sizing for node redundancy for storage and compute	Size all AOS clusters for n + 1 failover capacity.

Decision Name	Decision
Fault tolerance and replication factor settings	Configure the AOS cluster for fault tolerance 1 and configure the container for replication factor 2.
Inline compression	Enable inline compression.
Deduplication	Disable deduplication.
Erasure coding	Disable erasure coding.
Availability domain for Nutanix workload cluster	Use node awareness.
Availability domain for combined Nutanix workload and management cluster	Use node awareness.

Nutanix Files Storage Design

Size and deploy FSVMs according to the number of connections and resources required. Each connection of a pod or container to a file share counts toward the total number of connections.

Nutanix Files has built-in high availability and resilience to recover from a range of service disruptions.

Table: Nutanix Files Configuration

Item	Detail
Version	4.1
Nutanix Files instance size	4 VMs
vCPUs per VM	4
Memory per VM	16 GB
Concurrent connections	3,000
NFS throughput	800 MBps
Share settings	Managed by CSI-Driver, use dynamic shares

Note: Although this NVD uses Nutanix Files 4.1, you can use newer versions. Scalability and sizing can change with newer versions of Nutanix Files.

To expand Nutanix Files, you can scale up the existing FSVMs with more compute resources or scale out by adding more FSVMs to the deployment. Scaling up compute resources and scaling out FSVMs is determined by the number of additional connections to Nutanix Files. Refer to Nutanix Files Sizing Guide (credentials required) for more information.

Nutanix Objects Storage Design

Size and deploy Nutanix Objects VMs based on the estimated required throughput and object store size.

Table: Nutanix Objects Configuration

Item	Detail
Version	3.6
Nutanix Objects store size	4 VMs

Note: Although this NVD uses Nutanix Objects 3.5.1, you can use newer versions. Scalability and sizing can change with newer versions of Nutanix Objects.

Nutanix Objects generates self-signed Secure Socket Layer (SSL) certificates by default. For control plane security in Nutanix Prism, the core Hybrid Cloud NVD describes replacing the default self-signed certificates with certificates signed by an internal certificate authority from a Microsoft public key infrastructure (PKI). You can choose alternative tools such as openssl for certificate generation and signing. These certificates secure communications between Nutanix Objects components and Nutanix Prism.

Generate additional sets of certificates in the same way, using the same certificate authority, and apply them to the object store in each AZ to provide strong security for S3 client connections using the HTTPS protocol. S3 clients that interact with Nutanix Objects should have the trusted certificate authority chain preloaded. To facilitate disaster recovery, configure each object store with the fully qualified domain name (FQDN) and certificate of the other object store in addition to its own. This configuration allows each object store to respond to client requests intended for the other store and ensures strong security for S3 client connections in a disaster recovery scenario.

Note: Certificate management is an ongoing activity, and certificates need to be rotated periodically. This NVD signs all certificates for one year of validity.

7. Network Design

A Nutanix cluster can tolerate multiple simultaneous failures because it maintains a set redundancy factor and offers features such as block awareness and rack awareness. However, this level of resilience requires a highly available network connecting an AOS cluster's nodes.

Nutanix Controller VMs (CVMs) send each write to another CVM in the AOS cluster. As a result, a fully populated AOS cluster sends storage replication traffic in a full mesh, using network bandwidth between all Nutanix nodes. Because storage write latency directly correlates to the network latency between Nutanix nodes, any increase in network latency adds to storage write latency. Protecting the AOS cluster's read and write storage capabilities requires highly available connectivity between nodes. Even with intelligent data placement, if network connectivity between multiple nodes is interrupted or becomes unstable, VMs on the AOS cluster can experience write failures and enter read-only mode.

A Nutanix environment should use datacenter-grade switches designed to handle high-bandwidth server and storage traffic at low latency. Refer to the Nutanix Physical Networking best practice guide for more information.

Physical Network Architecture

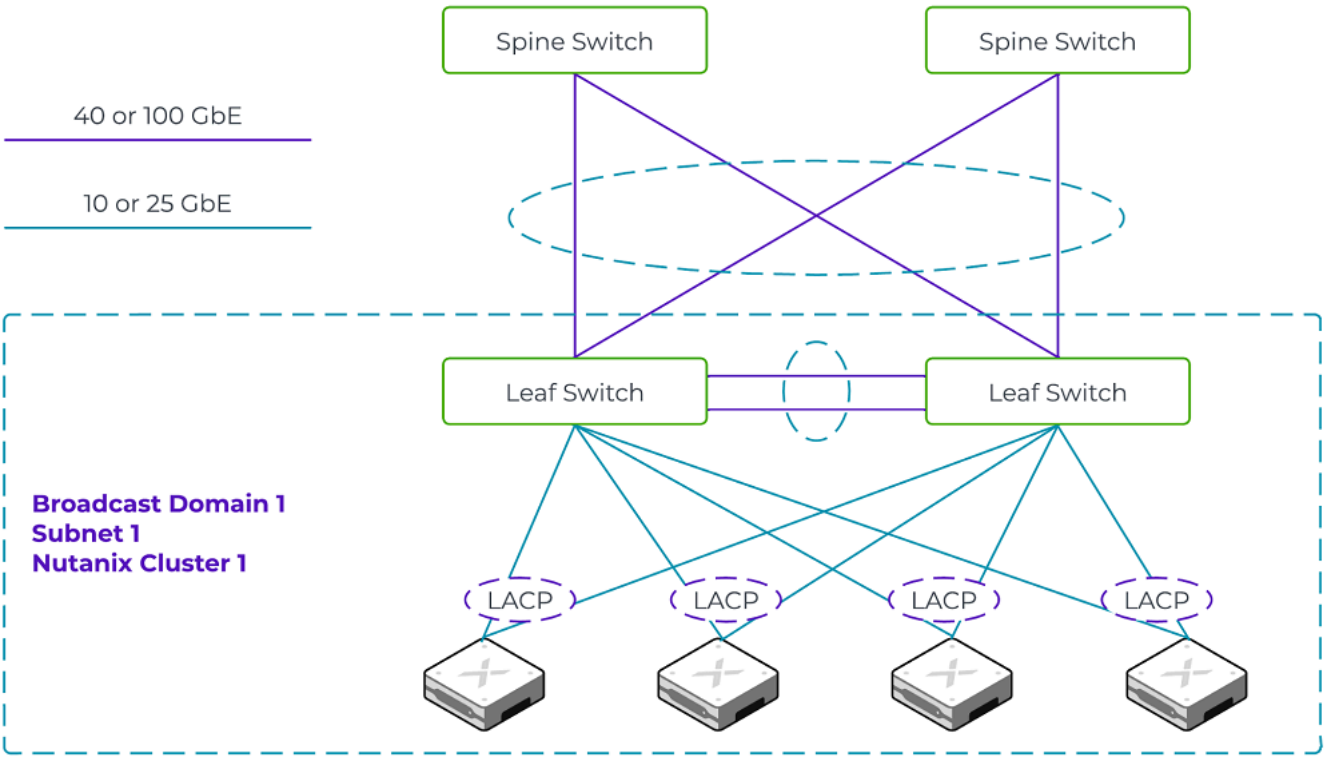


Figure 6: Physical Network Architecture

Table: Physical Network Design Decisions

Decision Name	Decision
Datacenter switch	Use large-buffer 25 Gbps switches for the datacenter.
Network topology for new environments	Use a leaf-spine network topology.
Top-of-rack switches	Populate each rack with two 25 Gbps top-of-rack switches.
Link aggregation group (LAG) type	Use MLAG configuration to avoid stacking and ensure network availability during individual device failure.

Decision Name	Decision
Number of switches between AOS nodes	Use at most three switches between any two Nutanix nodes in the same AOS cluster.
Network oversubscription	Reduce network oversubscription to achieve a 1:2 ratio.
Network design	Use a layer 2 network.

Table: AOS Node Connectivity Network Design Decisions

Decision Name	Decision
CVM and hypervisor VLAN	Configure the CVM and hypervisor VLAN as native, or untagged, on server-facing switch ports.
Switch ports for guest workloads	Use tagged VLANs on the switch ports for all guest workloads.
NICs for top-of-rack switches	Connect a 25 GbE NIC to each top-of-rack switch.
Virtual switches	Use one vs0 virtual switch with at least two of the fastest uplinks of the same speed.
NICs	Use NICs from the same vendor within a bond.
Logical network separation	Use VLANs to separate logical networks.
NIC load balancing mechanism	Use LACP (Link Aggregation Control Protocol).
MTU size	Set the MTU to 1,500 bytes.
Terminate L2/L3 networking	Have L2/L3 networking terminate on the spine.

Table: Nutanix Workload Cluster Networks

Decision Name	Decision
Shared infrastructure network subnet size	/23
VM network subnet size	/23

Decision Name	Decision
Number of addresses available per /23 network	490

Table: Nutanix Management Cluster Networks

Decision Name	Decision
Shared infrastructure network subnet size	/24
VM network subnet size	/24
Number of addresses available per /24 network	245
Number of VM networks	1

OpenShift Infrastructure Network Design

Red Hat OpenShift uses OVN-Kubernetes as the default CNI from Version 4.12. Each OCP cluster uses its own /23 machine network. This configuration provides segmentation between the different OCP clusters.

Red Hat recommends using nonoverlapping pod and service CIDRs between OCP clusters in case you use additional features like Submariner.

OCP worker VM IP addresses on all OCP clusters must be outside the pod and service CIDR ranges.

MachineNetwork needs access to Prism Central for MachineAPI features and CSI storage provision. MachineNetwork also needs access to the data services IP address of the hosting Prism Element cluster for iSCSI mount of persistent volume claims.

Table: OpenShift Infrastructure Network Design Decisions

Decision Name	Decision
Machine network	Configure as a /23 IPAM network on Nutanix workload clusters.
Pod and service CIDR	Pod and service CIDRs are unique across all OCP clusters.

Decision Name	Decision
CNI	Use OVN-Kubernetes.

OpenShift Application Network Architecture

You must direct external traffic to the OCP cluster and, in the case of disaster recovery, to the disaster recovery AZ as well.

This NVD uses F5 BIG-IP as a global load balancer for Red Hat Quay Registry and Nutanix Objects. F5 BIG-IP can also integrate with OpenShift in two key ways:

- F5 BIG-IP Local Traffic Manager (LTM) as an entry for the Core Services
- F5 Container Ingress Services (CIS): F5 BIG-IP Local Traffic Manager (LTM) as a replacement for the OpenShift Router

The NVD uses F5 CIS to provide access to applications. Because of CIS limitations, only one OpenShift cluster can be configured per F5 instance, and this NVD uses F5 CIS in the production app cluster.

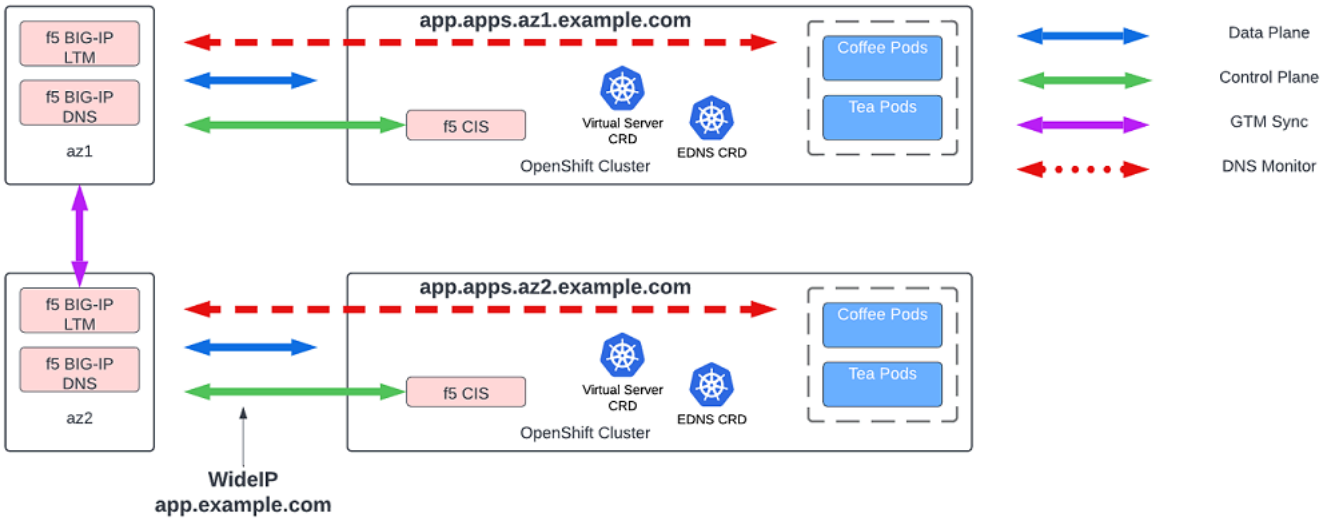


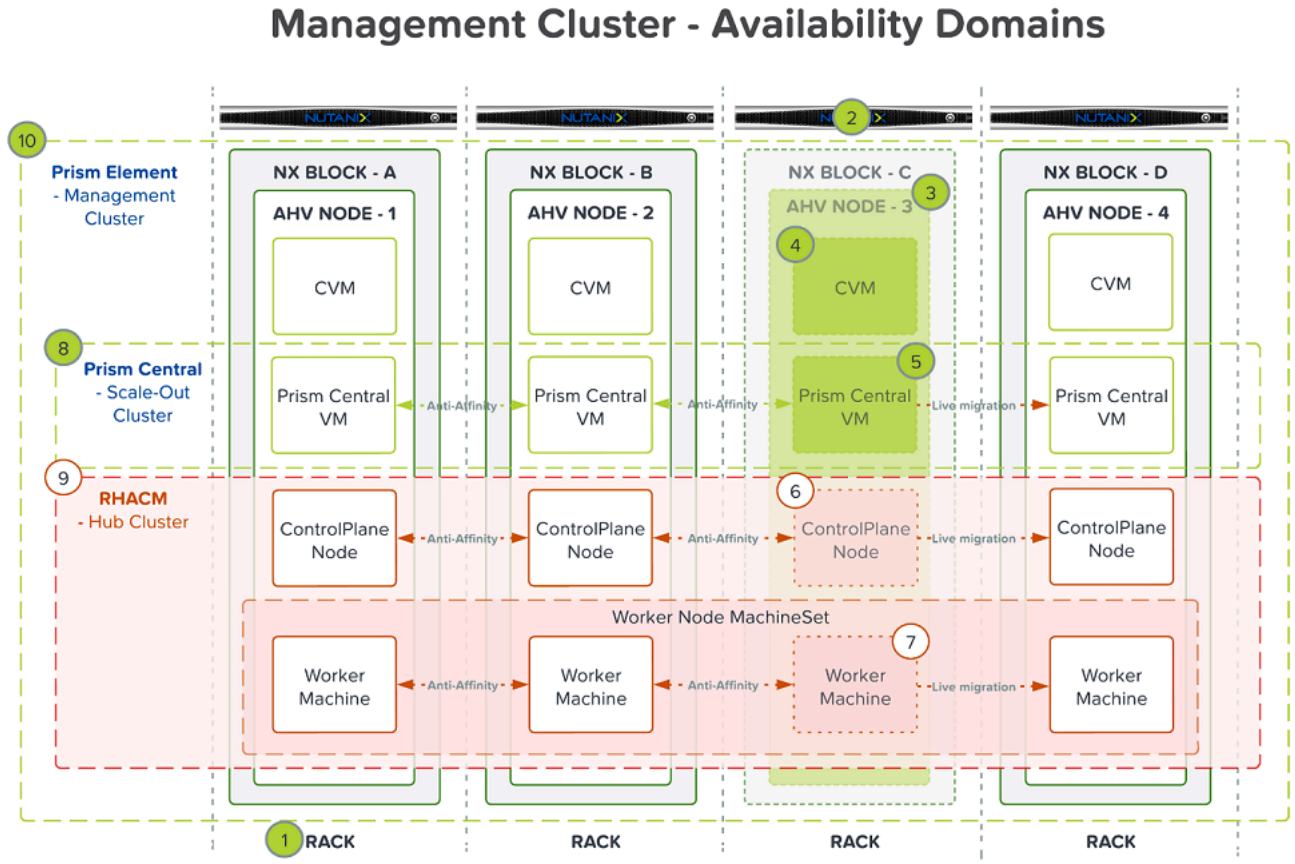
Figure 7: Application Load Balancing in OpenShift on Nutanix

This NVD uses virtual server custom resource definitions (CRDs) to configure necessary load balancing on each datacenter.

The application has a wide IP address host name that answers using round-robin for the different OpenShift environments. DNS has no layer seven path awareness; therefore, you need DNS monitors to determine the health of the application endpoints. Each ExternalDNS CRD specifies the DNS monitors on BIG-IP. If a monitor detects the http status failure, it removes the wide IP address from the DNS query.

8. Management Components

Management components such as Active Directory, DNS, and NTP are critical services that must be highly available. Together with Nutanix infrastructure management and OpenShift management, these components run in the dedicated Nutanix management cluster.



1. Rack (TOR Switch/PDU) Failure
2. Hardware/Block Failure
3. Hypervisor/Node Failure
4. Storage Controller VM (CVM) Failure
5. Prism Central VM Failure
6. Red Hat Openshift Control Plane (OCP) Control Plane/API Service Failure
7. Red Hat Openshift Control Plane (OCP) Worker Node Failure
8. Prism Central Cluster Outage
9. RedHat Advanced Cluster Management (RHACM) Hub Cluster Outage
10. Prism Element Cluster Outage

Figure 8: OpenShift on Nutanix Management Cluster Availability Domains

Table: Management Components Requirements

Component	Description
Prism Central	Deploy Prism Central in the Nutanix management cluster.
Nutanix Files	Enable Files Manager in Prism Central in the Nutanix management cluster. Use Files Manager to deploy, manage, and update file servers in the Nutanix storage cluster.
Nutanix Objects	Enable Nutanix Objects in Prism Central in the Nutanix management cluster. Use Nutanix Objects to manage object stores in the Nutanix storage cluster.
Active Directory	Integrate Active Directory for authentication for management components.
Virtual machine	Support at least small Prism Central size.
Virtual machine	Support at least six VMs (8 vCPU and 32 GB for an OCP hub cluster) for three OCP control plane VMs and three OCP worker VMs. Results may vary if you enable observability on the OCP cluster.
Virtual machine	Host each OCP control plane VM of the OCP hub cluster on a different AOS node.
OperatorHub	Give the OCP hub cluster access to the RHACM operator in the OperatorHub catalog.
Mirror registry	For disconnected network installs, the catalog registry.red hat.io/red hat/red hat-operator-index should be available in the mirror registry as it provides the RHACM operator. Include the packages advanced-cluster-manager and multicluster-engine from this catalog.

Component	Description
Mirror registry	For disconnected network installs, the catalog registry.red hat.io/red hat/ certified-operator-index should be available in the mirror registry as it provides the Nutanix CSI operator. Include the package nutanixcsioperator from this catalog.
Networking	Provide internet access for Prism Element.
Networking	Provide internet access for Prism Central.
Licensing	Use Red Hat OpenShift Platform Plus for RHACM.
Licensing	Use dedicated NUS Capacity Licensing for Nutanix Files and Nutanix Objects.
Monitoring	Use capacity forecast and planning functionality in Prism Central.
RHACM	Use RHACM on the OCP hub cluster to manage application life cycles across the managed OCP clusters.
RHACM	Use RHACM on the OCP hub cluster to import existing OCP clusters and manage life cycles.
RHACM	Use RHACM to enforce policies for the managed OCP clusters using Kubernetes-supported CRDs.

Table: Management Components Assumptions

Component	Description
OCP cluster	The maximum number of OCP worker VMs per Nutanix cluster is 360 if you use an OCP control plane with a large footprint.
Nutanix cluster	The maximum number of VMs managed by a Prism Central instance is 7,500.

Component	Description
Nutanix cluster	All management components reside in the same AZ.
Network	There is network connectivity between the Nutanix management, workload, and storage clusters at all times and to the internet.
Network	Files Manager on Nutanix management cluster can reach file servers in the Nutanix storage cluster at all times.
Network	Objects Manager on Nutanix management cluster can reach object stores in the Nutanix storage cluster at all times.
Infrastructure	Active Directory, DNS, NTP, and load balancer infrastructure services are available.

Table: Management Components Risks

Risk Description	Impact	Likelihood	Mitigation
Datacenter outage	Large	Unlikely	Restore management components to disaster recovery AZ.
Nutanix management cluster outage	Large	Likely	Restore management components to disaster recovery AZ.
Prism Central outage	Large	Likely	Restore Prism Central component in the same AZ.
OCP hub cluster outage	Large	Likely	Restore components to disaster recovery AZ.

Risk Description	Impact	Likelihood	Mitigation
OCP etcd cluster outage	Large	Unlikely	Rebuild etcd cluster and restore etcd backup from snapshot.
OCP control plane VM loss or failure	Medium	Unlikely	Restore control plane VM from protection domain snapshot if deployed through UPI or Assisted Installer; otherwise allow IPI to redeploy.
OCP workload cluster outage	Large	Likely	Restore control plane VM consistency group from protection domain snapshot.
OCP infra VM or worker VM loss or failure	Medium	Likely	Delete machine from respective MachineSet; MAPI controller reconciles.
OCP workload cluster pods in pending state	Medium	Likely	Increase the number of respective node MachineSet replicas to allow pods to be scheduled.

Nutanix Infrastructure Management Design

Nutanix recommends that you have a dedicated Nutanix management cluster in the datacenter AZ for both Nutanix and non-Nutanix environment management and control plane instances. For this validated design, the Nutanix management clusters contain at least 4 nodes. The design hosts critical infrastructure and environment management workloads, including OCP workload cluster control planes for large footprints. You can expand the Nutanix management cluster to a maximum of 16 AOS nodes in increments of 2 AOS nodes.

The Nutanix management cluster includes the following components:

Prism Central

Nutanix Prism is a comprehensive interface and unified control plane that simplifies and streamlines management of virtualized datacenter environments. Manage everything from storage and compute to VMs and deploy clusters for storage and virtualization within minutes—with one-click cluster management, one-click VM setup, one-click storage management, and self-service provisioning. This NVD recommends a scale-out Prism Central installation (a set of three VMs) because this installation increases the capacity and resilience of Prism Central.

Files Manager

Files Manager lets you view and control all your file servers from a single control plane. Files Manager is enabled from Prism Central, where you can deploy, manage, and update new file servers. Files Manager provides the Smart DR feature, which lets you protect file servers at the share level. In the event of a planned or unplanned loss of service, you can restore write access to protected shares by failing over to a recovery site file server.

Nutanix Objects

Nutanix Objects is a simple, scale-out cloud object storage solution that offers secure S3-compatible storage at massive scale. It helps simplify storage operations while offering high performance for cloud-native, big data analytics, and deep archive workloads. Nutanix Objects is flexible and easy to use, with policy-driven data tiering to any S3-compatible cloud provider.

Nutanix Disaster Recovery

Nutanix Disaster Recovery helps customers design disaster recovery plans that meet their specific needs. With options that include on-prem, public cloud (NC2), MSPs, and disaster recovery-as-a-service (DRaaS), Nutanix Disaster Recovery offers one-click failover, failback, and automated recovery—so organizations can meet performance SLAs while eliminating data silos, protecting business-critical applications, and reducing TCO. This NVD uses it to protect the database and Redis store used by the Quay container registry.

Active Directory

Active Directory authenticates Prism Central in Nutanix and the RHACM and OCP clusters in OpenShift.

DNS

The internal infrastructure services use DNS, and the global load balancer also integrates with DNS to direct client traffic to the application. The domain names must be appropriately configured so that traffic flows seamlessly to the disaster recovery AZ in the event of the primary AZ failure.

Syslog

A third-party syslog server that runs in the Nutanix management cluster in each AZ provides syslog services.

OpenShift Infrastructure Management

RHACM offers end-to-end management, visibility, and control of an OCP cluster. It provides management capabilities for application life cycles as well as improved security and compliance for the entire Kubernetes domain—across multiple datacenters.

RHACM is installed as an operator through Operator Lifecycle Manager (OLM) in a dedicated OCP cluster, called the OCP hub cluster. The OCP hub cluster resides in the Nutanix management cluster of each AZ. RHACM runs active-passive in this environment. Find more details about RHACM backup and recovery in the Business Continuity and Disaster Recovery section.

The additional OCP clusters that are managed by the OCP hub cluster are called managed OCP workload clusters. An agent installed in a managed OCP cluster receives requests from the OCP hub cluster and administers and services life cycle, application life cycle, governance, and observability on these managed OCP clusters.

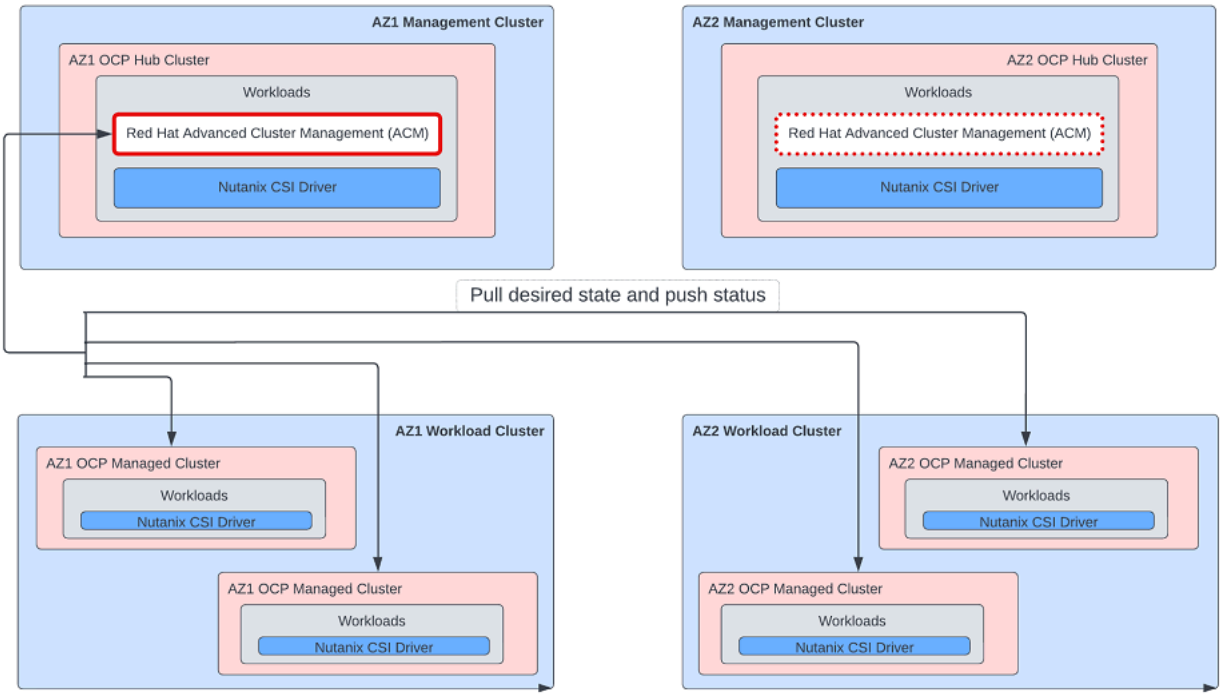


Figure 9: RHACM Overview

Application Management

RHACM manages application resources on your managed OCP clusters. The application management functionality provides simplified options to create and deploy multicluster applications through channel and subscription constructs. A channel is a source repository for the application manifest, which can be a Git repository, Helm release registry, or an object storage repository. Subscriptions allow OCP clusters to subscribe to a channel. You can also set up a GitOps environment to automate application consistency across OCP clusters.

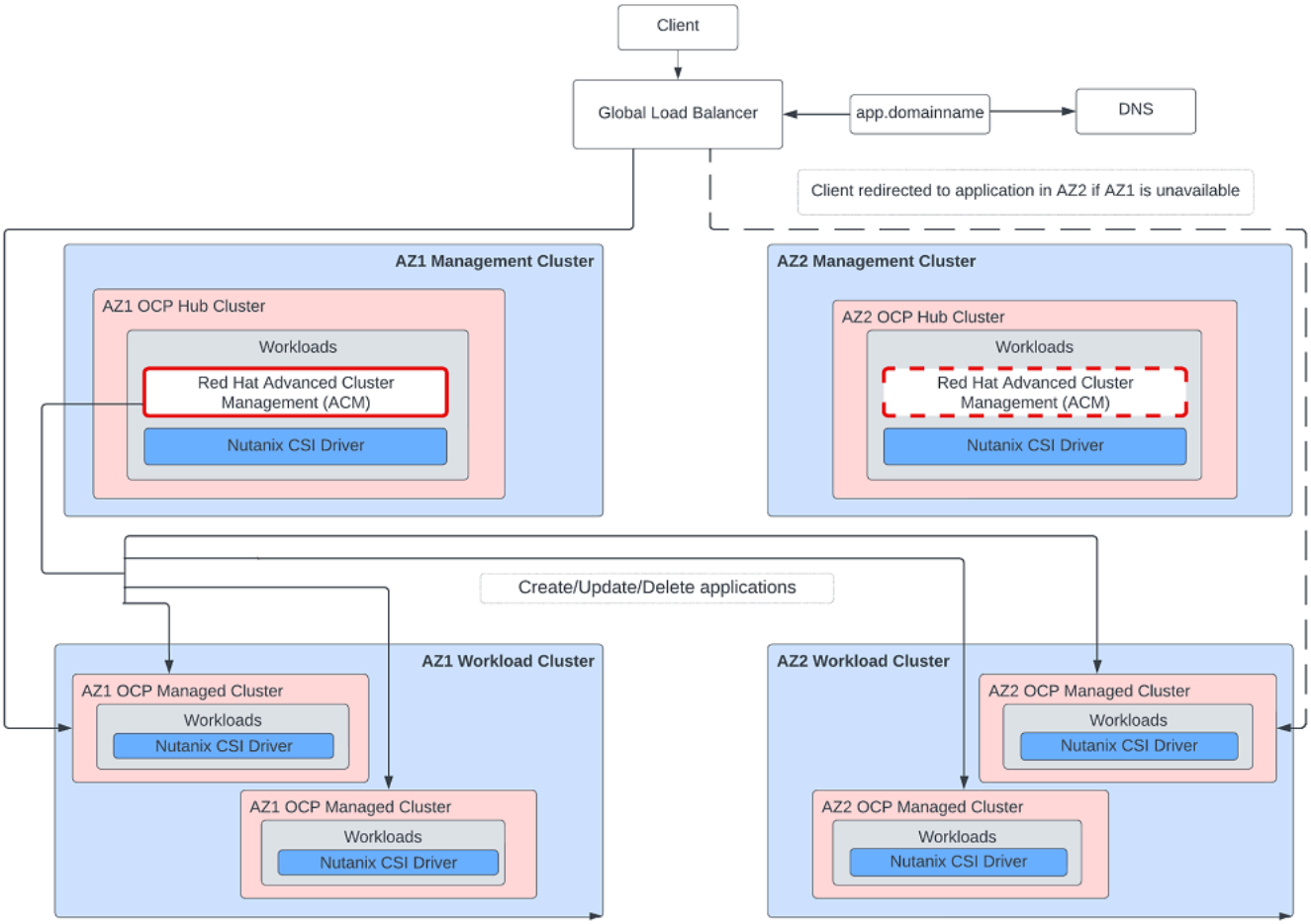


Figure 10: RHACM Application Management

Policy Management

RHACM lets you define policies to enforce security compliance or inform you of changes that differ from the compliance requirements in the managed OCP clusters. The central interface lets you manage policies for all the connected managed OCP clusters.

Use the configuration policy controller to create any Kubernetes resource and apply security policies across your OCP clusters. When a policy defines a Kubernetes object, the configuration policy controller compares it with objects

on the managed OCP clusters. If the policy is noncompliant, the controller can enforce the configuration on the target managed OCP cluster.

You can deploy and manage the Nutanix CSI Operator on individually managed OCP clusters using a configuration policy.

This NVD uses policies for the following tasks:

- Deploying Nutanix CSI Operator
- Deploying OpenShift Quay Bridge Operator
- Deploying OpenShift Container Security Operator
- Verifying that there is at least an OADP backup custom resource for each user project
- Configuring ClusterLogging and ClusterLogForwarder custom resources

9. Observability Design

Observability is provided by different OpenShift components as operators. This solution collects metrics from all managed OCP clusters and displays them in a centralized dashboard and collects log files from the entire OpenShift environment.

Monitoring

Use RHACM to collect metrics and monitor the status of all managed OCP clusters by creating a custom resource called MultiClusterObservability.

The RHACM Operator must be available to install the custom resource. Create a StorageClass in the OCP hub cluster, as the observability components require persistent storage. An S3-compatible object store must be available to store metrics. You can use Nutanix Objects for the object store.

When the service is enabled, a controller deploys to each managed OCP cluster that collects data from that OCP cluster and sends it to the OCP hub cluster. A Grafana dashboard instance is available in the OCP hub cluster for data visualization.

You can also define custom rules with Prometheus to create alert conditions and send notifications to an external messaging service.

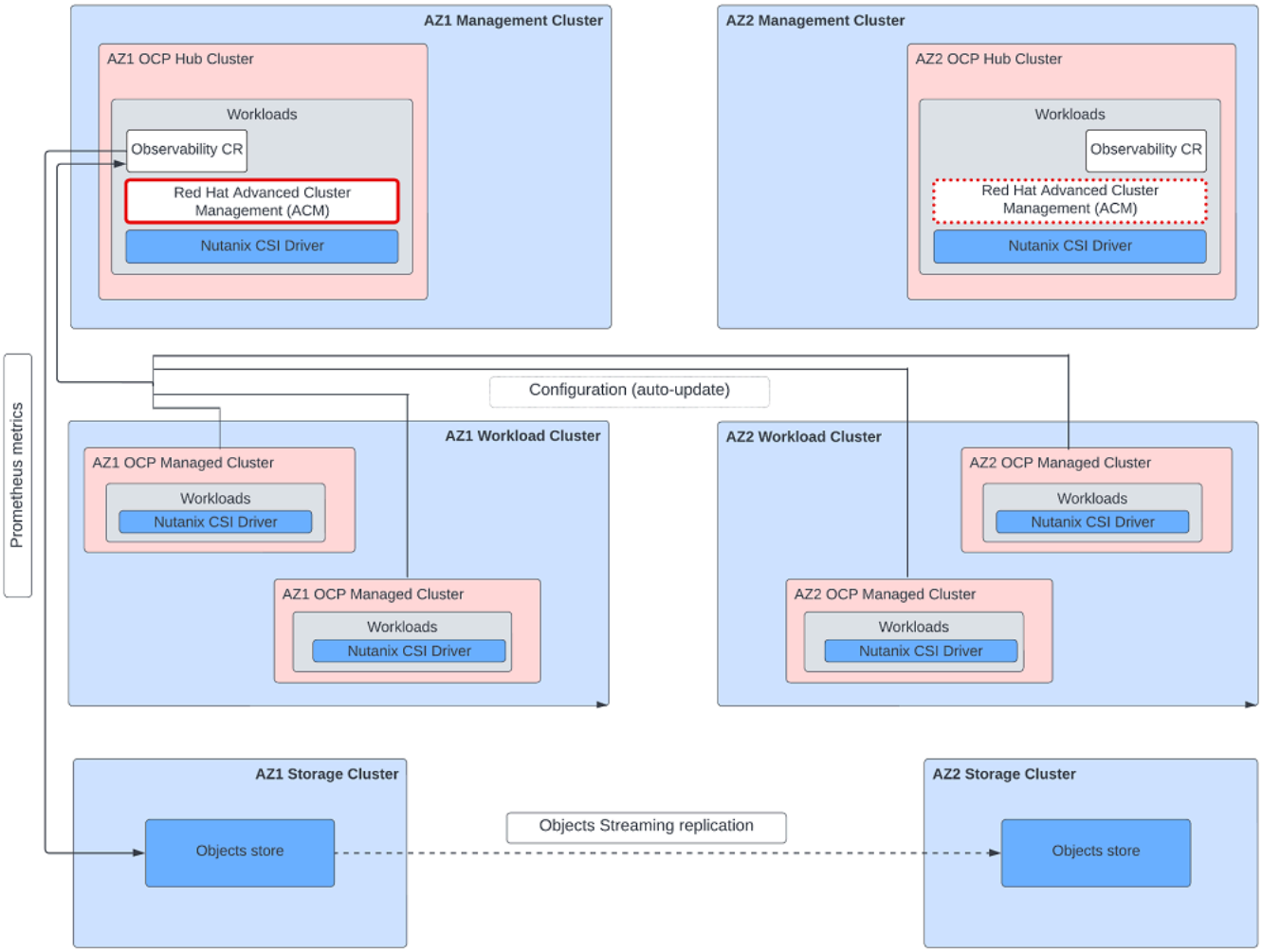


Figure 11: Cloud Native OpenShift Monitoring

Logging

The OpenShift Logging Operator manages all log files that are created in an OCP cluster.

The log types are one of the following:

- Application: Container logs generated by user applications running in the OCP cluster, except infrastructure container applications.

- Infrastructure: Container logs from pods that run in the openshift*, kube*, or default projects and journal logs sourced from the node file system.
- Audit: Audit logs generated by the node audit system, auditd, Kubernetes API server, OpenShift API server, and OVN network.

This NVD uses GrayLog as a central log management application so log files aren't stored directly in the managed OCP cluster or in the OCP hub cluster.

Use RHACM to push policies into every OCP cluster to forward all log entries into the central syslog appliance.

If needed, you can also use different targets for the log types to store infrastructure-related messages in syslog, forward application logs on a namespace-level into an external Elasticsearch instance, and process audit messages in your SIEM application.

10. Business Continuity and Disaster Recovery

Stateful Kubernetes applications use persistent volumes to store their data. A persistent volume has a life cycle independent from the underlying pods that consume it to preserve its data in case of failure.

To protect a stateful application, two pieces of data are most important:

- Persistent volume data: PersistentVolume API allows users to dynamically provision persistent volumes for user workloads. Persistent volume data is managed and stored with Nutanix Volumes for RWO access through iSCSI and Nutanix Files for RWX access through NFS.
- Kubernetes resources: Application configurations are commonly stored as Kubernetes resources in the etcd distributed database and accessed using the Kubernetes API server. The system typically exports these Kubernetes resources as JSON or YAML files.

This NVD uses Red Hat OpenShift API for Data Protection (OADP) to manage protection operations in both PersistentVolumes and Kubernetes resources against data loss and disaster scenarios. Red Hat OADP backs up Kubernetes objects and internal images by saving them as an archive file on Nutanix S3 object storage. Red Hat OADP backs up persistent volumes by either creating snapshots with the CSI or performing a generic volume backup of the resources and persistent volumes data with Restic.

Red Hat OADP currently supports the following capabilities:

- Backup: Back up all resources in a target cluster and alternatively filter the resources by type, namespace, or label.
- Restore: Restore backed up resources and persistent volumes from point-in-time scheduled (or on-demand) backup and alternatively restore all objects in a backup or filter the restored objects by namespace, persistent volume, or label.

- Hooks: Use hooks to run commands in a container on a pod either before or after a backup or restore. Restore hooks can run as an init container or directly in an application container.

Business continuity and disaster recovery requirements:

- Support the following disaster recovery events:
  - Datacenter outage
  - Single Nutanix cluster outage
  - Top-of-rack switch outage
  - Single VLAN outage
  - Accidental data loss (for example, human error)
  - Malicious data loss (for example, security breach or ransomware attack)
  - Application misconfiguration or software defect
  - Performance degradation caused by infrastructure (Nutanix cluster or network) or hardware components
- Schedule a backup job that supports a minimum RPO of 15 minutes for a given application.
- Filter (exclude or include) OCP cluster- and application-scoped Kubernetes resources as part of backup or restore policies.
- Support generic file system volume backups and restores of all Kubernetes persistent volume types to provide cross-cluster portability.
- Archive backups to S3-compatible storage (for example, Nutanix Objects and other major cloud providers such as AWS S3).
- Restore an application to an alternative namespace in the same OCP cluster (or entirely separate OCP cluster) for testing and QA purposes.
- Use life-cycle hooks as part of a backup workflow to easily coordinate or orchestrate integration with internal and external application or database services to ensure that you can properly flush all pending I/O resources prior to performing any backups or snapshots.

- Use life-cycle hooks as part of a restore workflow to easily orchestrate or transform resources on workflows on secondary OCP clusters.

Business continuity and disaster recovery assumptions:

- Disaster recovery avoidance causes minimal application and VM downtime.
- Customer provides redundant WAN connectivity between AZs.
- Customer provides WAN connectivity with sufficient bandwidth and latency (round-trip time (RTT) below 5 ms) to meet RPO requirements.
- Supporting infrastructure elements like GTM, LTM, DNS, Active Directory, and IPAM are available in both AZs.
- OCP clusters in both AZs have been properly configured with Nutanix CSI Operator and the Kubernetes cluster resources required for successfully recovering each application.
- The infrastructure administrator is responsible for setting up backup and recovery locations and monitoring any respective policy; the application administrator provides application backup policies.
- Failure domains associated with the loss of standard Nutanix management components (for example, disk, node, block, rack, zone, region, hypervisor, CVM, Nutanix Volumes, Nutanix Files, or Nutanix Objects) aren't accounted for in this design as they're already accounted for in the NUS NVD design overall.
- Network communication directly between Kubernetes pods and service networks managed by CNI doesn't require any level of advanced routing or tunneling to achieve connectivity.
- Because OCP clusters are in a warm state between each AZ, Nutanix data protection and recovery policies for individual VMs aren't necessarily required, given a cloud-native backup and disaster recovery solution.

Table: Business Continuity and Disaster Recovery Risks

Risk Description	Impact	Likelihood	Mitigation
Accidental data loss (for example, human error)	Large	Likely	Restore application to last known good state following data outage.
Malicious data loss (for example, security breach or ransomware attack)	Large	Likely	Restore application to last known good state following data outage.
Application misconfiguration	Large	Likely	Repair badly misconfigured applications by restoring application artifacts and configurations to a known good state.

Table: Business Continuity and Disaster Recovery Design Constraints

Constraint Description	Comment
Use Red Hat OADP Operator for cloud-native data protection solution for OCP clusters.	The Red Hat OADP operator seems to be targeted at independent software vendors (ISVs) who want to integrate easily with OpenShift APIs for backup and data protection; therefore, capabilities around policy management, scheduling, and overall UI or UX are currently limited. While this NVD does recommend the Red Hat OADP operator (Velero), Kasten K10 by Veaam is an alternative solution that includes more advanced policy management and scheduling capabilities.
Use Restic volume backup and recovery for RWX PersistentVolumes using Nutanix Files.	The Nutanix CSI Volume Driver doesn't currently support Snapshot API for Nutanix Files and requires an alternative method for performing volume snapshots, backups, synchronization, and recovery. Velero (with Restic) has native integration that supports this capability.

Constraint Description	Comment
Use identical names for source and target buckets.	Bucket names must be identical to simplify overall streaming and replication and S3 client redirection during a disaster recovery failover event. This naming consistency, combined with the configuration of the source object store's FQDN as a secondary FQDN on the disaster recovery object store, ensures that the file server's tiering profile can locate the disaster recovery bucket.

Table: Business Continuity and Disaster Recovery Design Decisions

Decision Name	Decision
Disaster recovery orchestration product	Use OADP with Nutanix Objects to support end-to-end cloud-native backup and disaster recovery workflows for both stateless and stateful workloads across all Kubernetes clusters on-premises or in the cloud.
Disaster recovery failover and testing	Use a combination of OADP restore custom resources and inherent hook capabilities to automatically restore persistent volumes (blocks and files) and transform Kubernetes resources during recovery to support planned and unplanned disaster scenarios.
Supported RPOs	Use Nutanix Objects streaming replication with replication rules to support RPOs of 15 min, 1 h, and 24 h.
Maximum number of buckets for streaming replication (Nutanix Objects)	There are no limits for Nutanix Objects replication.
Back up workloads within AZs or across AZs	To optimize the backup window and save WAN bandwidth, back up VMs, Nutanix Volume Groups, and Nutanix Files shares that are in the local AZ first, then replicate.
Backup policy RPO	Set 24-h RPO on backup policies.

Decision Name	Decision
Backup product	Use RHACM policies and OADP to synchronize backup and restore policies between clusters.
Storage solution for backup repository	Use the OADP DataProtectionApplication custom resource to define Nutanix Objects as the primary BackupStorageLocation for both Kubernetes resources and metadata and persistent volumes.
Number of S3 buckets for backup repository	Use one object store with one bucket as the backup repository.
Replication method for backups between AZs	Use Nutanix Objects bucket replication between object stores to manage backup replication.

Backup Conceptual Design

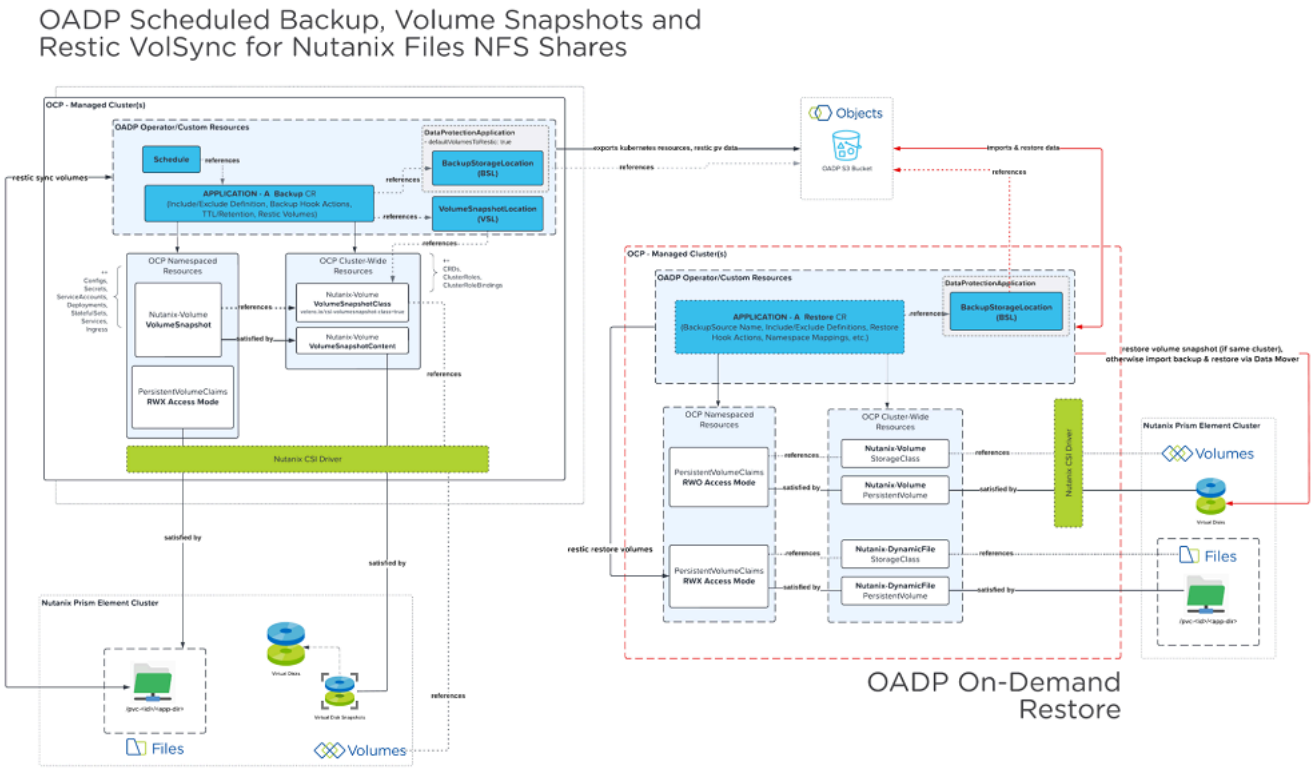


Figure 12: Cloud Native OpenShift Backup Conceptual Design

OCP Management Backup

RHACM is protected by the OADP-based cluster backup and restore operator.

The active or primary OCP hub cluster that manages the OCP clusters backs up resources at defined time intervals.

The passive OCP hub cluster continuously retrieves the latest backups and restores the passive data posted by the primary hub. The passive hub is on stand-by to become a primary OCP hub cluster when the primary hub becomes unavailable. It's connected to the same storage location where the primary hub backs up data, so it can access the primary hub backups.

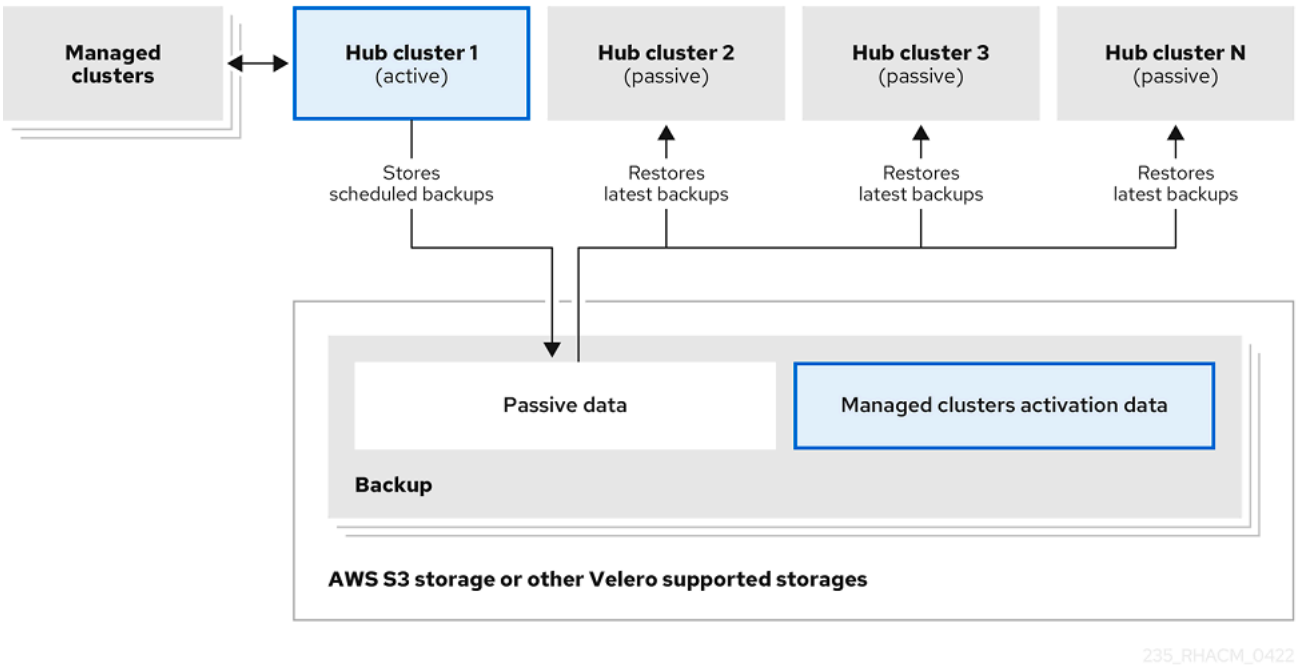


Figure 13: Cloud Native OpenShift Management Backup

When the primary OCP hub cluster is unavailable, the administrator can choose the passive OCP hub cluster to take over the managed OCP clusters and restore the managed OCP cluster activation data. Because the administrator must decide if the unavailable primary OCP hub cluster needs to be replaced or if there is some network communication error between the OCP hub cluster and managed OCP clusters, this process isn't automated.

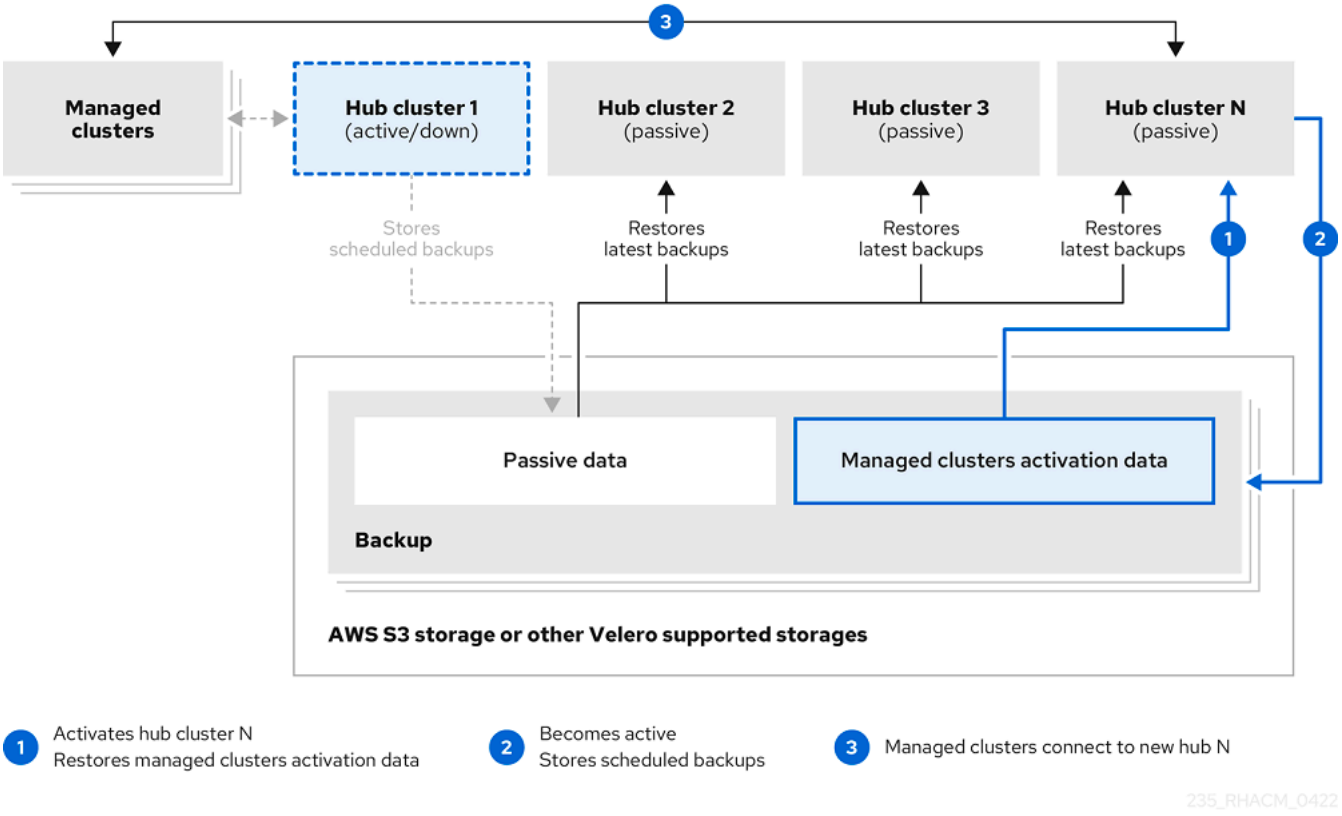


Figure 14: Cloud Native OpenShift Management Restore

OCP Control Plane Backup and Disaster Recovery

As a cluster administrator, you might experience situations where the OCP cluster itself doesn't work as expected:

- You have an OCP cluster that isn't functional after the restart because of unexpected conditions, such as node failure or network connectivity issues.
- You deleted something critical in the OCP cluster by mistake.
- You lost the majority of your control plane hosts, leading to etcd quorum loss.

The OCP and Kubernetes in general persist the state of all resource objects in etcd, a highly resilient and distributed key-value store. Nutanix recommends that you snapshot and back up your OCP cluster's etcd data regularly and store it in a secure location external to the OCP environment, such as your S3-

compatible Nutanix Objects bucket. Additionally, because the etcd snapshot results in a high I/O cost, Nutanix recommends that you schedule etcd backups outside peak usage hours.

It's critical to take an etcd backup immediately before and after you restart or upgrade your underlying OCP. This step is especially important during an upgrade, as you must use an etcd backup from the same z-stream release when you attempt to restore your OCP cluster. For example, an OCP 4.12.z cluster must use an etcd backup that was taken from 4.12.z.

From a security perspective, Nutanix strongly recommends that you encrypt all secrets stored in the etcd database. Therefore, you must account for the workflow used to both back up and restore the encrypted database. It might also make sense to use Nutanix buckets with versioning to ensure that backups and snapshots are considered immutable and explicitly guarantee integrity.

To recover from a disaster situation, the cluster administrator should restore the OCP cluster to its previous state by using one of the previously saved etcd snapshots that's been stored in the S3-compatible Nutanix Objects bucket.

Application Backup and Disaster Recovery

As a cluster administrator, you can back up, restore, and migrate applications running on OCP and Nutanix by using a combination of the OADP operator (available via OperatorHub Marketplace) and Nutanix Objects as the underlying S3-compatible storage.

The following image depicts a high-level overview of the backup and restore workflow when using OADP.

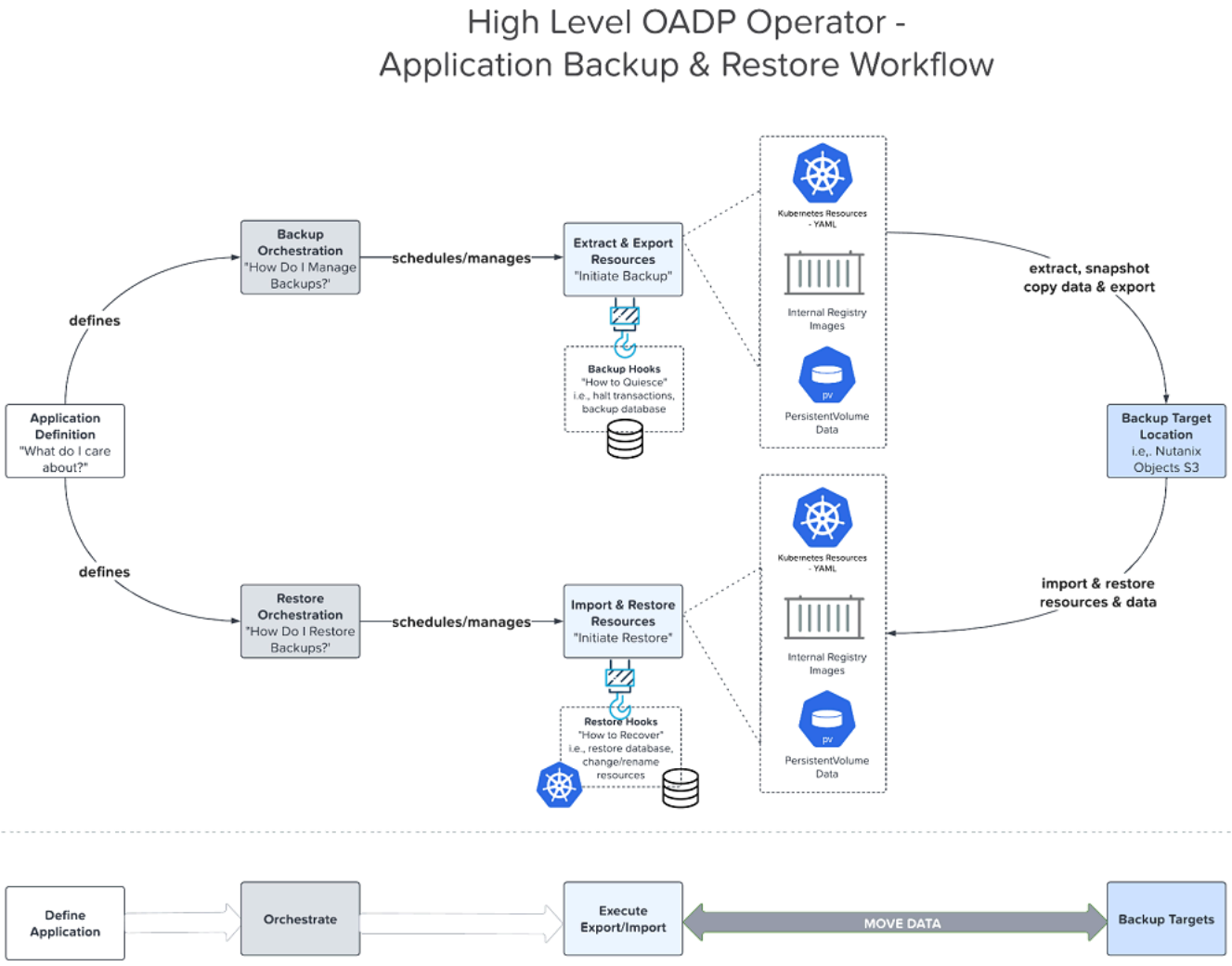


Figure 15: OADP Backup and Restore Workflow

You can use OADP to back up applications by first defining a DataProtectionApplication custom resource that references the target BackupStorageLocation (Nutanix Objects) that serves as the underlying Restic repository needed for uploading the backup content (and subsequently synchronize content downstream to secondary managed OCP and OADP clusters).

Once you deploy an application, initiate an on-demand or scheduled backup by creating a Velero backup or scheduling a custom resource. The OADP and

Velero backup controllers ensure that you can use the appropriate backup custom resource to define backup hooks (used during the life cycle of backup operations), backup schedules, custom filters (include or exclude specific namespaces or resources) and, most importantly, custom parameters such as whether the backup workflow should use file system backups (Restic) versus VolumeSnapshots with Nutanix CSI by default.

NFS shares that are dynamically provisioned by the Nutanix Files CSI provisioner don't currently support the Kubernetes CSI Snapshot API. Typically, OADP backs up and restores the underlying persistent volumes by using the native CSI Snapshot API to perform a volume snapshot against the target storage provider's APIs.

As an alternative solution, OADP and Velero support generic file system backups and restores of Kubernetes volumes attached to pods from the volumes' file system through a feature called File System Backup or Pod Volume Backup. A free, open-source backup tool called Restic provides this capability. OADP and Velero's native Restic integration provides an out-of-the-box solution for backing up and restoring almost any type of Kubernetes volume, including Nutanix Files CSI.

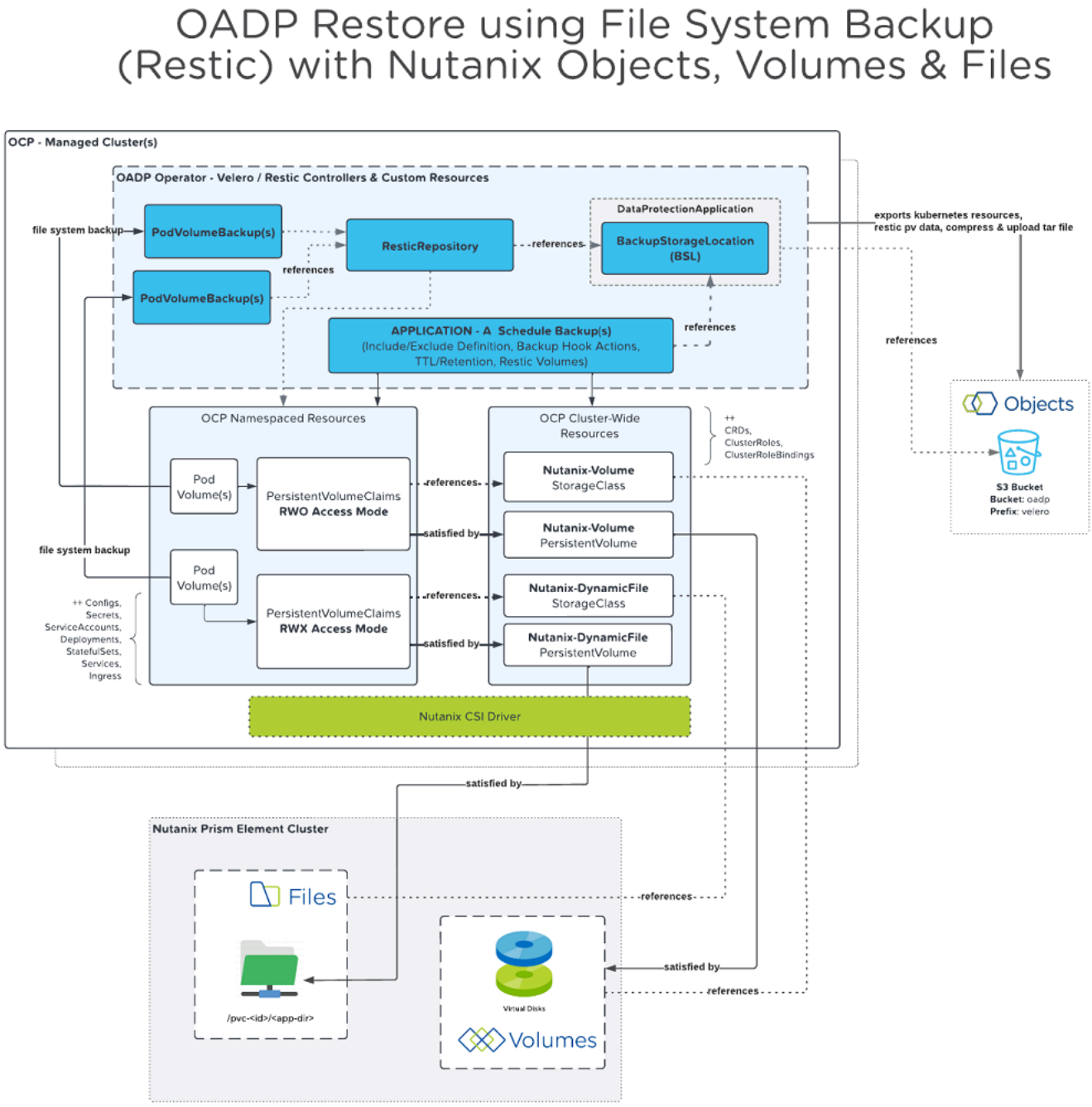


Figure 16: OADP File System Backup Resources

During a backup restore (to the same OCP cluster) or disaster recovery event (to a secondary OCP cluster), a cluster administrator can restore applications by

creating an OADP restore custom resource. Much like that of the OADP backup custom resource, you can configure restore hooks to manage the recovery life cycle (that is, running commands in init containers or directly in the respective application container during the restore operation).

Image Registry Backup and Disaster Recovery

To protect Quay image registry, use Nutanix protection domains to replicate data between Nutanix management clusters in two different AZs. Create and configure protection domains according to the following table.

> Note: If you have a license for Nutanix Disaster Recovery software, use protection policies instead of protection domains.

Table: Protection Domain Settings

Setting	Value
Protection domain name	Mgmt_Quay_01
Repeat every	15 minutes
Retention policy	Local: 1 day; remote: 1 day

11. Security and Compliance

Nutanix recommends a defense-in-depth strategy for layering security throughout any enterprise datacenter solution. This design section focuses on validating the layers that Nutanix can directly oversee at the control and data plane levels. Refer to the Network Design section of the Hybrid Cloud: AOS 6.5 with AHV On-Premises Design for more information on network-based security, as this NVD doesn't include Flow Network Security or Prism Central. See the Security and Compliance Layer section of the Nutanix Hybrid Cloud Reference Architecture for additional details.

Authentication and Authorization

Refer to the Hybrid Cloud: AOS 6.5 with AHV On-Premises Design for guidance on configuring the AOS and network components and integrating them with Active Directory authentication.

Red Hat OpenShift can also use the same LDAP sources for user authentication.

Red Hat OpenShift needs a user account with access to Nutanix Prism Central for managing VM resources, which you can configure with least privileges.

AOS Hardening

This NVD enables additional nondefault hardening options in each AOS cluster:

- Advanced Intrusion Detection Environment (AIDE)
- Hourly security configuration management automation (SCMA)

Both features are trivial to enable, introduce little to no discernible system overhead, and help detect and prevent internal system configuration changes that might otherwise compromise service availability. These features add to the intrinsic hardening built into AOS.

Syslog

For each control plane endpoint, system-level internal logging goes to a centralized third-party syslog server that runs in the Nutanix management cluster in each AZ. The Nutanix infrastructure is configured to send logs for all available modules when they reach the syslog error severity level. Use TCP transport via TLS where available.

The same syslog server collects all log messages from OpenShift infrastructure. However, a customer can choose to use different pipelines and targets for application or audit messages.

Certificates

SSL endpoints serve all Nutanix control plane web pages. This NVD replaces the default self-signed certificates on Prism Central, Prism Element, and Nutanix Objects with certificates signed by an internal certificate authority from a Microsoft public key infrastructure (PKI). Any client endpoints that interact with the control plane should have the trusted certificate authority chain preloaded, preventing browser security errors.

Valid and trusted certificates on Prism Central endpoints are mandatory for Red Hat OpenShift integration. Each OCP cluster needs to be preloaded with a trusted certificate authority chain.

Note: Certificate management is an ongoing activity, and certificates need to be rotated periodically. The NVD signs all certificates for one year of validity.

Data-at-Rest Encryption

Nutanix AOS can perform data-at-rest encryption (DaRE) at the AOS cluster level; however, as the NVD doesn't have a stated requirement that warrants enabling it, this design doesn't use it. If requirements change, you can enable DaRE nondisruptively after you create the AOS cluster and populate data. Once you enable DaRE, existing data is encrypted in place and all new data is written in an encrypted format.

Note: To enable DaRE, you must also deploy an encryption key management solution.

Our decision to not use DaRE doesn't preclude the use of in-guest encryption techniques such as system-level encryption, database encryption (for example, Microsoft SQL Transparent Data Encryption (TDE)), or the storage of encrypted files; however, in-guest encrypted data can't be compressed at the storage layer in most cases. Although this design enables compression, in-guest encrypted data isn't likely to be compressible, so using in-guest encryption might affect the amount of available storage.

Table: Security Design Decisions

Decision Name	Description
DaRE	Disable DaRE; don't deploy a key management server.
SSL endpoints	Sign control plane SSL endpoints with an internal certificate authority (Microsoft PKI).
Certificates	Provision certificates with a yearly expiration date and rotate accordingly.
Authentication	Use Active Directory LDAPS authentication (port 636).
Control plane endpoint administration	Use a common administrative Active Directory group for all control plane endpoints.
AOS cluster lockdown mode	Don't enable AOS cluster lockdown mode (allow password-driven SSH).
Nondefault hardening options	Enable AIDE and hourly SCMA.
System-level internal logging	Enable error-level logging to external syslog server for all available modules.
Syslog delivery	Use TCP transport for syslog delivery.

12. Datacenter Infrastructure

This design assumes that datacenters in the hosting region can sustain two AZs without intraregional fate-sharing—in other words, failures in one datacenter's physical plant or supporting utilities don't affect the other datacenter. This NVD addresses points where the Nutanix gear touches the datacenter equipment to ensure that all your needs are met.

Rack Design

The following figure shows the initial density for this design, with the designated requirements, assumptions, and constraints. Rack 1 runs the Nutanix management and storage clusters; Rack 2 runs the Nutanix workload cluster.

You can add more racks as needed, depending on top-of-rack network switch density and the datacenter's power, weight, and cooling density capabilities per square foot. Refer to the Platform Selection table for the specific node models selected for this NVD.

Figure 17: Cloud Native OpenShift Rack Layout

When you scale the environment, consider physical rack space, network port availability, and the datacenter's power and cooling capacity. In most

environments the Nutanix workload clusters are the most likely to grow, followed by the Nutanix storage clusters.

In this design's physical rack space, one generic 42RU rack contains 12RU for management and storage with 3RU reserved for two data switches and one out-of-band switch. The second rack runs the Nutanix workload cluster, using 16RU and again 3RU reserved for networking and out-of-band management.

For network ports, the 24 nodes in this NVD consume 12 ports on each of the two data switches for Rack 1 and 16 ports each for Rack 2, leaving free ports for additional growth.

For power, cooling, and weight, you need the minimums specified in the previous figure, and you should assume at least double these values for a fully loaded rack including network switches. Datacenter selection is beyond the scope of this design; have a conversation about fully loaded racks with datacenter management prior to initial deployment. Planning to properly support the environment's long-term growth might change where in the facility you want to set up the equipment.

13. Ordering

This bill of materials reflects the validated and tested hardware, software, and services that Nutanix recommends to achieve the outcomes described in this document. Consider the following points when you build your orders:

- All software is based on core licensing whenever possible.
- Nutanix Xpert Services or an affiliated partner selected by Nutanix provides all services.
- Nutanix based the functional testing described in this document on NX series models with similar configurations to validate the interoperability of software and services.

Substitutions

- Nutanix recommends that you purchase the exact hardware configuration reflected in the bill of materials whenever possible. If a specific hardware configuration is unavailable, choose a similar option that meets or exceeds the recommended specification.
- You can make hardware substitutions to suit your preferences; however, such changes may result in a solution that doesn't follow the recommended Nutanix configuration.
- Avoid software product code substitutions except when:
  - You need different quantities to maintain software licensing compliance.
  - You prefer a higher license tier or support level for the same software product code.
- Adding any software or workloads that aren't specified in this design to the environment (including additional Nutanix products) might affect the validated density calculations and result in a solution that doesn't follow the recommended Nutanix configuration.

- Professional Services substitutions to accommodate customer preferences aren't possible.

Bill of Materials

The following sections show the bill of materials for the primary and secondary datacenter management and workload clusters and the primary and secondary datacenter storage clusters.

Primary and Secondary Datacenter Management Clusters

Hardware

- Product code: NX-3070N G8
  - Quantity: 4
  - NX-3070N-G8, 1-node configuration
  - Hardware support:
    - Support level: Production
    - NRDK support: No
    - NR node support: No
  - Per-node hardware configuration:
    - Processor: 2 × Intel Xeon-Gold 5318Y processor (2.1 GHz, 24 cores, 165 W) (48 cores)
    - Memory: 8 × 64 GB (3,200 MHz DDR4 RDIMM)
    - HDD: No HDD included
    - SSD: 6 x 3.84 TB
    - Network adapter: 1 x 100, 40, or 25 GbE 2-port Mellanox CX-6

Software

- Nutanix Cloud Platform Pro
  - Product code: SW-NCP-PRO-PR
  - Quantity: From hardware

Cluster Install Services

- Xpert Services, cluster installation is part of CNS-INF-A-SVC-DEP-ULT

Primary and Secondary Datacenter Workload Clusters

Hardware

- Product code: NX-3070N G8
  - Quantity: 4
  - NX-3070N-G8, 1-node configuration
  - Hardware support:
    - Support level: Production
    - NRDK support: No
    - NR node support: No
  - Per-node hardware configuration:
    - Processor: 2 × Intel Xeon-Gold 5318Y processor (2.1 GHz, 24 cores, 165 W) (48 cores)
    - Memory: 12 × 128 GB (3,200 MHz DDR4 RDIMM)
    - HDD: No HDD included
    - SSD: 6 × 3.84 TB
    - Network adapter: 1 × 100, 40, or 25 GbE 2-port Mellanox CX-6

Software

- Nutanix Cloud Platform Pro
  - Product code: SW-NCP-PRO-PR
  - Quantity: From hardware

Cluster Install Services

- Xpert Services, cluster installation is part of CNS-INF-A-SVC-DEP-ULT

Primary and Secondary Datacenter Storage Clusters

Hardware

- Product code: NX-8055N-G8
  - Quantity: 4
  - NX-8055N-G8, 1-node configuration
  - Hardware support:
    - Support level: Production
    - NRDK support: No
    - NR node support: No
  - Per-node hardware configuration:
    - Processor: 2 × Intel Xeon-Silver 4310 processors (2.1 GHz, 12 cores, 120 W) (24 cores)
    - Memory: 8 × 32GB Memory Module (3200MHz DDR4 RDIMM)
    - HDD: 8 × 18 TB 3.5″ HDD
    - SSD: 4 × 7.68 TB SSD
    - Network adapter: 1 × 100, 40, or 25 GbE 2-port Mellanox CX-6

Software

- Subscription, contains Nutanix Files and Nutanix Objects
  - Product code: SW-NUS-PRO-PR
  - Quantity: 10 TiB

Cluster Install Services

- Xpert Services, cluster installation is part of CNS-INF-A-SVC-DEP-ULT

Professional Services

With the following professional services, Nutanix can implement this NVD as designed, built, and tested. These services are outcome-based, with fixed prices for the scope described by the services SKUs included in the bill of materials. See the Nutanix Xpert Services information available on Nutanix.com for more details on each of the SKUs included.

Table: Professional Services for Platform

Product Code	Description	Quantity
CNS-INF-A-WRK-DES-ULT	HCI Design Workshop Ultimate	1
CNS-INF-A-SVC-DEP-ULT	HCI Cluster Deployment Ultimate	2
Check with your account team	Red Hat OpenShift Design Workshop	2
Check with your account team	Red Hat OpenShift Deployment	4
Custom scope	OpenShift DR with OADP	Custom scope based on the number of BCDR policies required by the customer. Check with your account team for a quote.

14. Appendix

References

1. Hybrid Cloud: AOS 6.5 with AHV On-Premises Design
2. Hybrid Cloud: AOS 6.5 with AHV Unified Storage Design
3. Red Hat OpenShift on Nutanix tech note
4. Nutanix Hybrid Cloud Reference Architecture
5. Physical Networking best practice guide
6. OpenShift Container Platform 4.12 Documentation

About Nutanix

Nutanix offers a single platform to run all your apps and data across multiple clouds while simplifying operations and reducing complexity. Trusted by companies worldwide, Nutanix powers hybrid multicloud environments efficiently and cost effectively. This enables companies to focus on successful business outcomes and new innovations. Learn more at Nutanix.com.

List of Figures