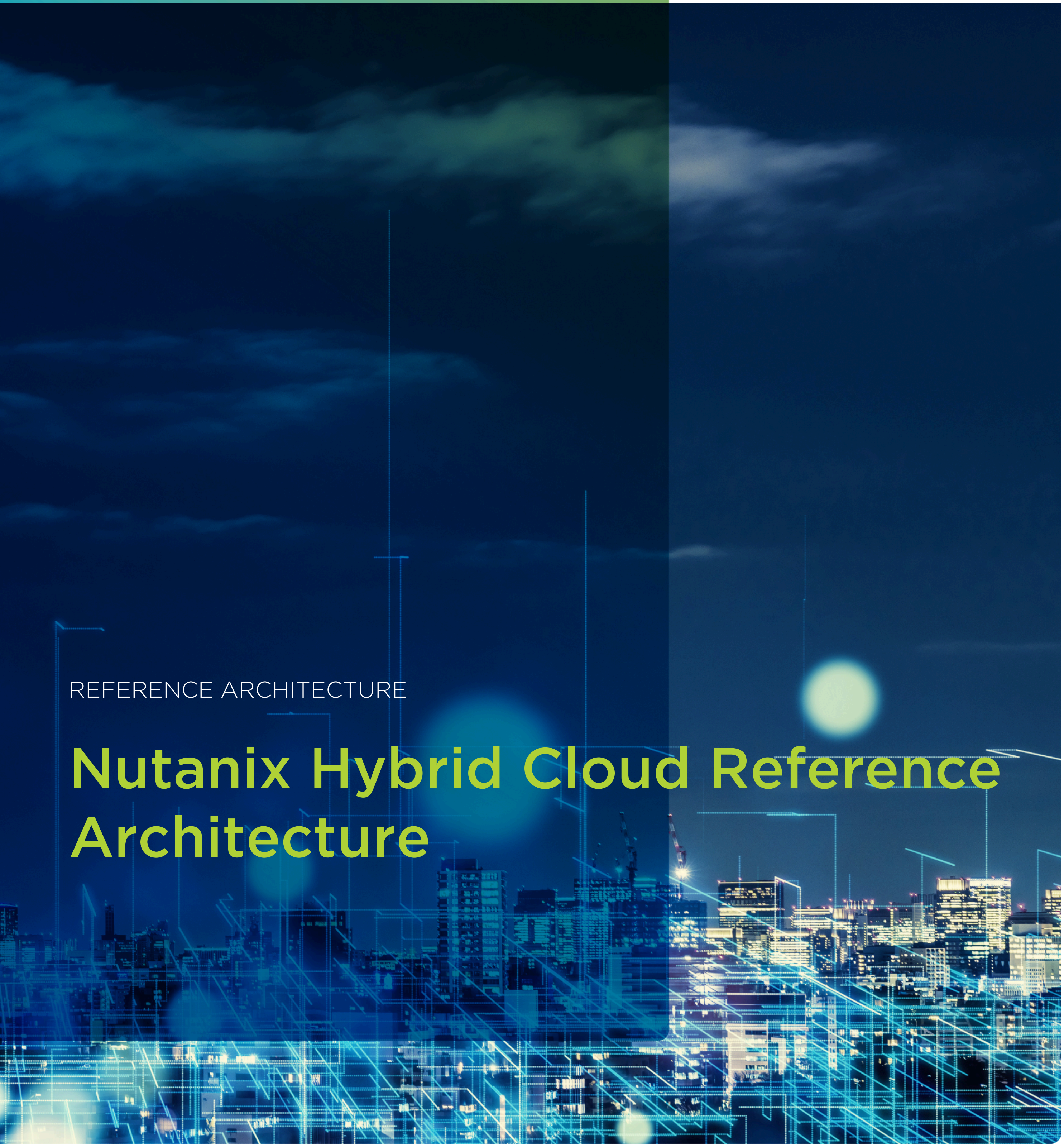
v3.0 | April 2021 | RA-PC21



REFERENCE ARCHITECTURE

Nutanix Hybrid Cloud Reference Architecture

NUTANIX™
YOUR ENTERPRISE CLOUD

Copyright

Nutanix, Inc.
1740 Technology Drive, Suite 150
San Jose, CA 95110

Contents

1. Introduction

Vision for Private, Hybrid and Multicloud

The Nutanix vision for cloud computing environments started in the datacenter with innovative solutions for private cloud, which greatly reduced the complexity and effort required to deploy and manage software-defined storage, compute, and networking infrastructure. In recent years, Nutanix's vision has expanded to include architectures for hybrid and multicloud that offer more alternatives to optimize costs.

Which type of cloud to use for a specific use case depends on a variety of characteristics of multi-tiered applications. These can be summarized in terms of:

- Use Case: web servers, application servers, databases, etc.
- Workload Type: virtual machines (VMs) and/or containers.
- Storage Format: block, object, file.
- Security Policies: requirements that could rule out the option to host workloads on particular public clouds.
- Unique Requirements: items like machine learning hardware or media transcoding that can only be met by the unique features of one specific Public Cloud offering (for example serverless computing, analytics, PaaS, and so on).

These characteristics ultimately shape an organization's cloud architectures and drive the decisions regarding where to host each application tier in a particular service.

Historically, the hosting architecture for multi-tiered applications didn't change without a labor-intensive migration. Nutanix innovation gives IT organizations the flexibility to dynamically place workloads. The Nutanix distributed architecture natively includes the ability to provision the same application blueprints in four different cloud configurations:

Table 1: Cloud Configurations

Cloud Configuration	Description
Private	Hosted on prem on AHV and/or VMware.
Hybrid	Some hosted on prem on AHV and/or vSphere and some hosted in AWS, Azure, or GCP.
Public (natively)	Hosted in AWS, Azure, or GCP as a native public cloud service offering.
Public (bare metal)	Hosted as bare metal in AWS running Nutanix Acropolis Operating System (AOS).

Nutanix further extends these cloud configurations by enabling the following four architectural principles:

Table 2: Architectural Principles

Architectural Principle	Description
Application Mobility	Multi-tiered applications can be aligned across all cloud providers to maximize architectural symmetry and promote application mobility.
Increased Standardization	Use of the same hardened OS gold image and business process orchestration (e.g. ITSM integration, approvals, emails, showback/ chargeback, etc.) across all cloud configurations.
Policy-Based Security Governance	Hosted in AWS, Azure, or GCP as a native public cloud service offering.
Policy-Based Cost Governance	Optimize cost by hosting workloads on the platform that meets the requirements of the service and has the lowest Total Cost of Ownership (TCO).

All of these principles contribute to an advanced hybrid and multicloud strategy that simplifies IT, stretches budgets, and accelerates time to value. This document provides the architecture and design-driven decisions to help our customers realize this strategic vision.

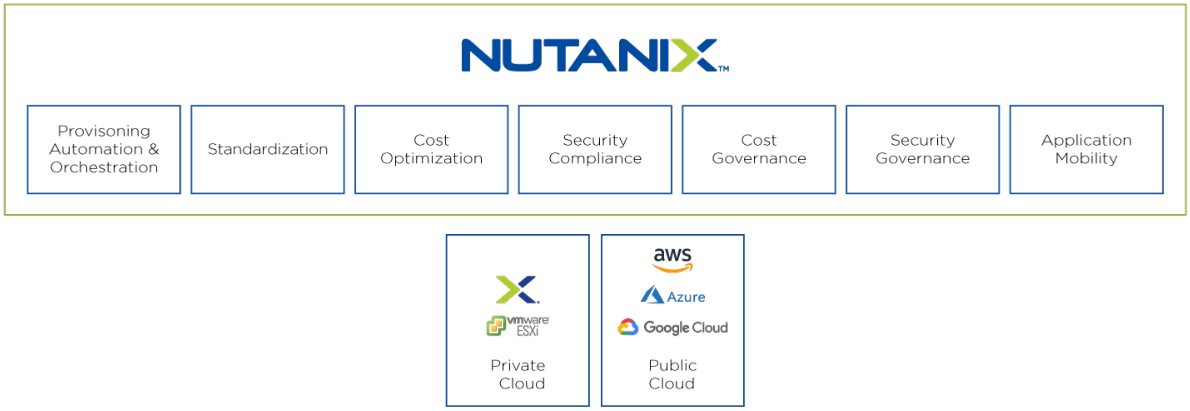


Figure 1: Nutanix Hybrid and Multicloud Strategy

Design Objectives

The objective of this document is to define, explore, and develop key design decisions required when implementing private, hybrid, or multicloud solutions based on the Nutanix platform. The objective can be further broken down as follows:

- To identify and enumerate the key design decisions that need to be documented in order to support a robust design methodology and practice.
- To explore each design decision, evaluating key viable options, tools, and methods for implementation and management so that organizations can make informed decisions relating to their specific design requirements.

Since simplicity is a key principle of all Nutanix products, some requirements may be met through native platform architecture without the need for superfluous design decisions, which are often required by competing platforms. The objective of this document is therefore not to educate readers about Nutanix features and functions, even though this may naturally occur as a side benefit. In these cases, this document will explain how these requirements are addressed by Nutanix-native features.

The design objectives of this document for the Hybrid Cloud implementations are as follows:

Table 3: Design Objectives of this Document for Private Cloud Implementations

Design Objectives	Description
Key objective	The platform is capable of hosting and provisioning workloads.
Workload Types	Virtual Machines and containers
Scope of deployment	Greenfield deployment that is adaptable to brownfield deployments with workload migrations.
Cloud type	Hybrid cloud
Number of regions	• Two regions. Region 1 provides two availability zones for resiliency. Region 2 provides long-term backup retention for recovery from disasters impacting all of Region 1. • Services must be available across two availability zones during normal production. Services must be available within one availability zone during disaster recovery. • A minimum distance of 300km between Region 1 (primary) and Region 2 (DR). • Provide running workloads and disaster recovery services in same region.

Design Objectives	Description
Availability	• SLA 24/7 • Uptime 99.999% across availability zones › Applies to certain production workloads › Scheduled downtime not included › 99% uptime required during disaster recovery scenario • RPO – 0 min across availability zones › Applies to certain production workloads › RPO = 0 can be achieved at the application layer for some services but the infrastructure layer must support RPO = 0 • RTO – 5 min across availability zones. › Applies to certain production workloads
Disaster Recovery	• DR RPO = 60 min between primary availability zones and DR availability zone • DR RTO = 24h between primary availability zones and DR availability zone • Best effort RTO in the region where production workloads are running and the disaster recovery capabilities fails.
Minimum number of nodes per cluster	3 nodes.
Maximum number of workloads	Unlimited number of workloads dependent on the pod based constructs described herein.
Types of Clusters	Management, Workload, Storage Heavy, Edge Clusters.
Virtualization	• Nutanix AHV and VMware ESXi hypervisors. • While Nutanix supports Microsoft Hyper-V, it is not considered in this design.

Design Objectives	Description
Management Plane	Nutanix Prism Element, Nutanix Prism Central, and VMware vCenter.
Scope	• Sizing recommendations and methodology for the amount of software-defined storage, compute and networking. • Physical implementation of storage, compute, and networking. • Logical configuration of clusters. • Scalability methods and recommendations. • Automation relating to cluster build, expansion, and lifecycle management. • Management and operations aspects such as capacity management, reporting, upgrades, backup and restore, monitoring and alerting, and logging.
Authentication, authorization, and access control	Microsoft Active Directory users and groups tied to Role Based Access Control.
Security Policy & Enforcement	• Least privilege access policies will be implemented so that end-users and administrators will need to be members of groups in order to be able to perform secure aspects of their job function. • Certificates are signed by a trusted certificate authority (CA). • Hardening of platform associated with hypervisor, control plane, and data plane. • Security policy definition and enforcement of drift away from defined policies and checksum verification. • Separation of traffic classes such as management and application.

Audience

This document for Multi Datacenter design is intended for infrastructure architects, infrastructure administrators, and infrastructure operators who want to deploy and manage datacenters based on Nutanix Enterprise Cloud, to address requirements for availability, capacity, performance, scalability, business continuity, and disaster recovery.

Design Decisions

This document makes recommendations and guides the reader to appropriate decisions where possible. In cases where a design decision is required, this document describes possible options.

Table 4: Design Decision Example

NET-001	Title Of Design Decision
Justification	Justification to support why the decision was made.
Implication	Additional implications as a result of the design decision.

Note: Appendix I: Table of Design Decisions includes a list of all the design decisions described throughout this document.

How To Use This Reference Architecture?

This document is subdivided into five major sections as follows:

Architectural Overview

Introduces key Architecture concepts that will be discussed throughout this design.

Design Considerations

Discusses key design considerations that will vary for each customer. Customers will be required to make decisions that will influence the design and build of their end-solution.

Detailed Design

Identifies key design decisions and in most cases determines the optimal configuration and decision that will be used for validating the design. For design decision, alternate options may be discussed along with pros and cons. For good reasons, Customers may choose to deviate from the decisions made in this document. The decisions made in this section are recommended by Nutanix, however they are by no means meant to be the only method, and Nutanix recognizes that alternate decisions may be appropriate depending on specific requirements.

Multi-site Design, Disaster Recovery and Business Continuity

Articulates considerations for designing across multiple data centers, and the related DR/BC architecture.

> Note: This section does not provide detailed operational guides or runbooks.

Incorporating Optional Nutanix Services

Describes the design considerations for additional services that can optionally be deployed in a Nutanix environment to address specific requirements. The list of capabilities covered includes: management automation, Nutanix deployment in public cloud, advanced security capabilities, file and object data services, integrated backup options, application orchestration, container support, and database-as-a-service (DBaaS).

2. Architecture Overview

This section describes the high-level Nutanix architecture, including major concepts and design elements that anyone designing a Nutanix deployment should understand. If you are already familiar with Nutanix hyperconverged infrastructure (HCI) and Nutanix software, you can skip this section.

The diagram below shows the high-level architecture covered in this document. This overview explains the elements in each layer. Later sections will explore the design decisions necessary for each layer.

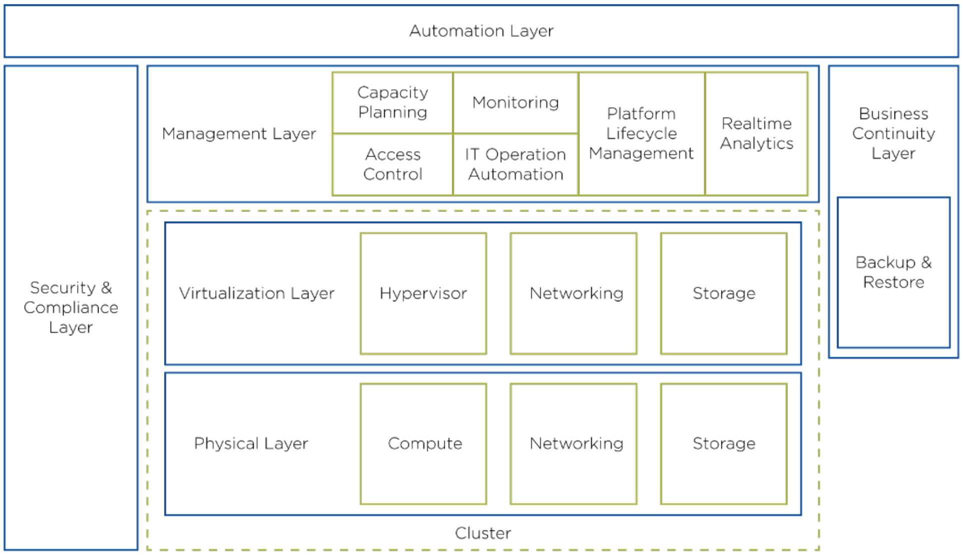


Figure 2: Architecture Overview

Physical Layer

Because the Nutanix Enterprise Cloud architecture is based on hyperconverged infrastructure, the physical layer is significantly different than it would be in a traditional datacenter architecture. Understanding the differences will allow you to make the best hardware choices for your Nutanix deployment.

Hyperconverged Infrastructure

Nutanix HCI converges the datacenter stack including compute, storage, storage networking, and virtualization, replacing the separate servers, storage systems, and storage area networks (SANs) found in conventional datacenter architectures and reducing complexity. Each node in a Nutanix cluster includes compute, memory, and storage, and nodes are pooled into a cluster. The Nutanix Acropolis Operating System (AOS) software running on each node pools storage across nodes and distributes operating functions across all nodes in the cluster for performance, scalability, and resilience.

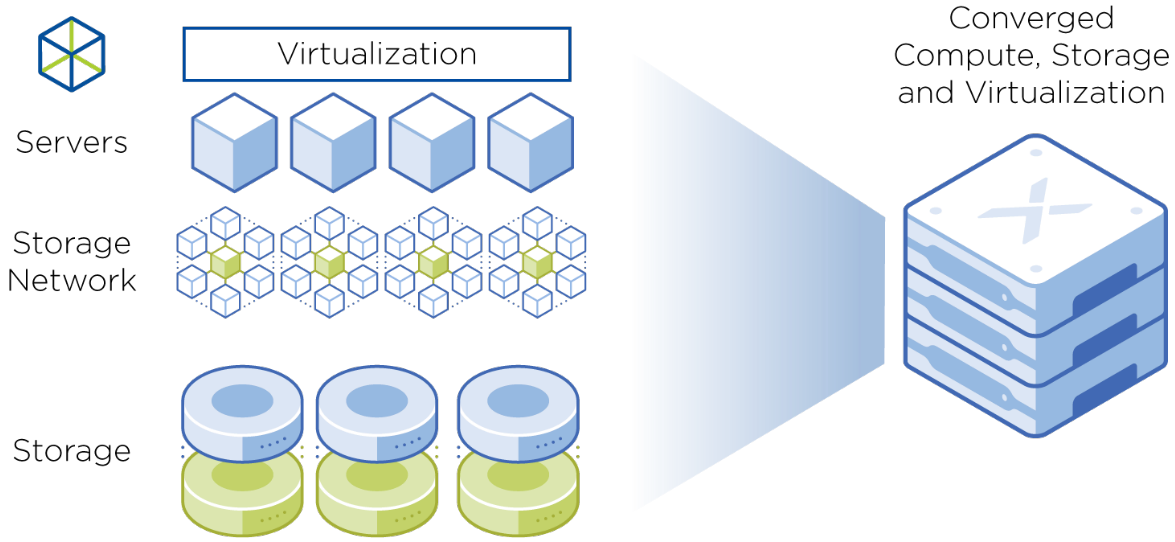


Figure 3: Hyperconverged Infrastructure

A Nutanix node runs an industry-standard hypervisor and the Nutanix Controller VM (CVM). The Nutanix CVM provides the software intelligence for the platform and is responsible for serving IO to running VMs.

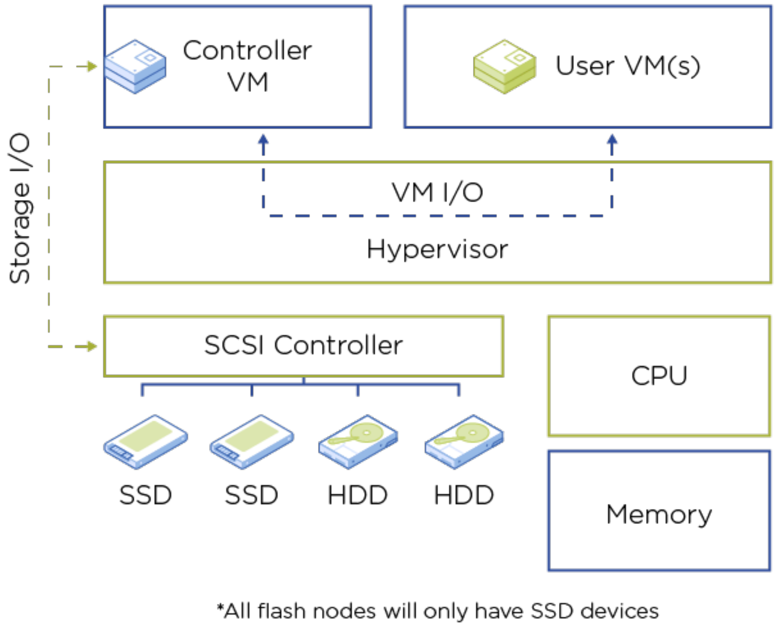


Figure 4: A Nutanix Node

Hardware Choice

Nutanix Enterprise Cloud provides significant choice when it comes to hardware platform selection. Available options include:

- Nutanix NX appliances.
- OEM appliances from leading vendors such as Dell, Lenovo, HPE, IBM, and Fujitsu.
- Other third-party servers from a wide range of vendors.

The Nutanix Support Portal contains the most up-to-date information on supported systems.

Hardware is available in a variety of chassis configurations from various vendors. Options range from multi-node chassis for high density to single-node rackmount chassis.

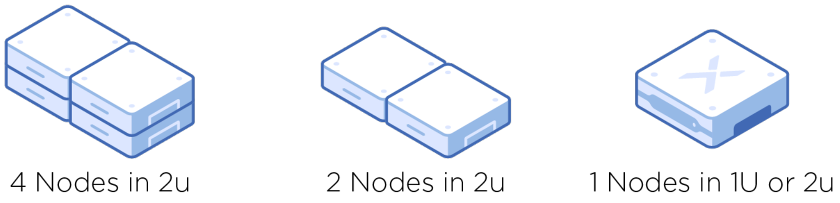


Figure 5: Chassis Configurations

Nutanix commonly refers to all of these chassis configurations as a block.

Compute

Sizing systems to meet compute needs in a Nutanix environment is similar to other architectures. However, it's important to ensure that your design provides enough compute (CPU/RAM) to support the CVM.

Storage

Nutanix nodes offer a range of storage configurations:

- Hybrid nodes combine flash SSDs for performance and HDDs for capacity.
- All-flash nodes utilize traditional flash SSDs.
- NVMe nodes utilize NVMe SSDs.

Different node types can be mixed in the same cluster. More information is provided in the document Product Mixing Restrictions.

For data resiliency, Nutanix uses replication factor (RF), maintaining 2 or 3 data copies. This approach enables a Nutanix cluster to be self-healing in the event of a drive, node, block, or rack failure. In a Nutanix cluster consisting of multiple blocks, RF can enable block awareness. Data copies are distributed across blocks to protect against the failure of an entire block. In configurations spanning multiple racks, RF can similarly provide rack awareness with resilience to a rack outage. For more information on Nutanix data resiliency, please refer to the Nutanix Bible.

Compression, deduplication and erasure coding (EC-X) can be enabled to increase data efficiency and save capacity.

Data locality and intelligent tiering ensure that the data associated with a VM is preferentially stored on that VM's local node. Active data is stored on the fastest media, delivering performance and eliminating the need for ongoing performance tuning.

Networking

Fast, low-latency and highly available networking is a key element of this document. The distributed storage architecture relies on the performance and resilience of the physical network. A good design provides high performance while maintaining simplicity.

In the detailed design section of this document we address common network topologies, selection of physical switches, and recommended connections between hosts and the physical network.

Cluster Design

A Nutanix cluster is the management boundary of the storage provided to a group of workloads. A Nutanix deployment can be architected to support either (a) mixed workloads in a single Nutanix cluster; or (b) dedicated clusters for each workload type in a block and pod design.

Designs that choose dedicated clusters may include any of the following cluster types:

Management clusters

Designed to run VMs that support datacenter management such as:

- Nutanix Prism Central.
- VMware vCenter.
- Active Directory Domain Controllers.
- Other management workloads, such as DNS, DHCP, NTP, Syslog.

Management clusters reside in the management workload domain. In this document, a management cluster occupies a separate rack or is distributed across multiple racks.

Workload clusters

Reside in a virtual infrastructure workload domain and run tenant virtual machines. You can mix different types of compute clusters and provide separate compute pools to address varying SLAs for availability and performance.

Storage clusters

Storage-only clusters provide dedicated data services to tenants. These are typically deployed for use cases focused on Object, File, or Block-level storage.

Edge/ROBO clusters

Reside at an edge and/or ROBO deployment to run virtual machines or ROBO workloads. These are typically distinguished from normal workload clusters by their small size and limited external bandwidth.

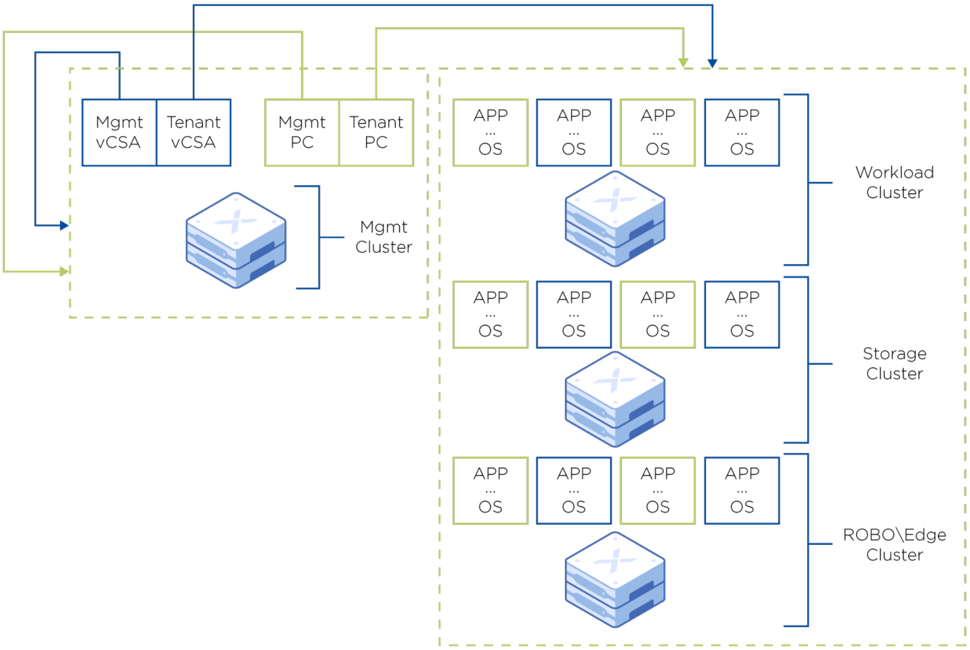


Figure 6: Cluster Design

Virtualization Layer

The virtualization layer sits logically above the physical layer, controlling access to compute, network, and storage resources. This document provides a choice between two hypervisors: Nutanix AHV and VMware ESXi. Both are enterprise-grade hypervisors, filling a similar set of requirements and use cases. Nutanix AHV is included at no additional cost with every Nutanix node.

> Note: Nutanix HCI supports Microsoft Hyper-V, however this document does not describe deployment of this hypervisor.

Management Layer

The management layer is a key differentiator for Nutanix.

Nutanix Prism provides simplified end-to-end management for Nutanix environments. Prism combines multiple aspects of datacenter management into a single consumer-grade design that provides complete infrastructure and virtualization management, operational insights, and troubleshooting. Prism largely eliminates the need for separate management tools. All Prism functionality is also accessible via REST API.

The Prism family consists of three products that extend core capabilities:

Prism Element

The core Nutanix management platform enables management and monitoring at the cluster level for all infrastructure (compute, storage, networks) and virtualization. Key functionality of Prism includes:

- Full VM, storage, and hypervisor management.
- Network visualization.
- Role-based access control (RBAC).
- Nutanix 1-click upgrades. Prism orchestrates and streamlines platform upgrades, keeping track of the changes. Can upgrade all Nutanix software and firmware running in a Nutanix environment plus the ESXi hypervisor.

Prism Central

Enables management and monitoring of multiple Nutanix Prism Element clusters from a central interface.

Prism Pro and Ultimate

Adds advanced capabilities to the Prism platform, including performance anomaly detection, capacity planning, custom dashboards, reporting, advanced search capabilities, and task automation.

The Prism family of products are an integral part of a Nutanix cluster and do not require separate infrastructure. Prism Central runs as a separate VM, or as a cluster of 3 VMs for additional scale and resilience. More details on Nutanix management are provided later. Nutanix deployments that use the AHV hypervisor can be fully managed by Prism.

Nutanix deployments that use VMware vSphere should also include VMware vCenter Server. This is the centralized monitoring and resource management software for VMware virtual infrastructure. It performs a number of tasks, including resource provisioning and allocation, performance monitoring, workflow automation, and user privilege management.

Automated IT Operations

Prism Pro allows administrators to automate routine operational tasks, reducing administrator effort and time while increasing the quality of results. To provide this automation, Nutanix X-Play enables "if-this-then that" (IFTT) features that allow admins to create Playbooks that define automation actions that run when a particular trigger occurs.

The most common type of trigger is alert-based, where a system-defined or user- defined alert causes an action to occur. An alert could be something as simple as crossing a designated CPU or memory threshold. Other triggers can be manual; the associated playbook does not take action until an admin explicitly tells it to. With a manual trigger, an admin selects an entity such as a VM, and the specified playbook executes against it. Manual triggers allow the admin to control when and where the automation takes place. (See the section Prism Operations for more information on automation with Prism.)

Business Continuity Layer

Nutanix HCI is built with resiliency in mind including redundancy for power and other hardware components and the ability to architect resiliency for entire datacenters. Nutanix provides multiple ways to provide business continuity including backup/restore and disaster recovery:

There are five main categories where Nutanix provides native protection, via hypervisor, third-party software, or a combination:

- Hardware
  - Self-healing from disk failure
  - Redundancy of key components
- Node
  - Node failure resiliency
  - Non-disruptive upgrades
- Data
  - Tunable redundancy
  - Nutanix Mine provides secondary storage integration and works with leading backup vendors.
- VM
  - VM-centric Protection
  - High Availability (HA)
  - ADS/DRS
  - Eco-system integration such as Nutanix Mine
- Datacenter/Site
  - Multi-datacenter replication including ROBO
  - Multi-site replication including ROBO
  - Cloud Connect
  - Metro Availability / Sync Replication
  - Near-sync
  - Async
  - Nutanix Xi Leap provides cloud-based DR-as-a-service (DRaaS).

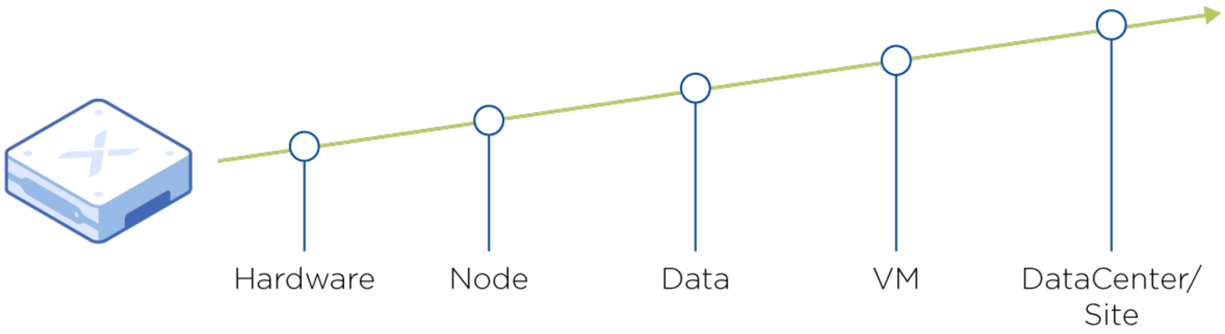


Figure 7: Business Continuity Layer

This document describes the configuration and use of Nutanix-native capabilities. Other solutions mentioned may be added to the completed design but are beyond the scope of this document.

For a list of supported 3rd party backup solutions see: Nutanix Technology Alliance Partners.

Automation Layer

Automation and orchestration are increasingly recognized as critical to IT success. By simplifying infrastructure management across the entire lifecycle, automating operations, and enabling self-service, Nutanix helps you deploy datacenter infrastructure that delivers a high degree of scalability, availability, and flexibility.

Nutanix improves efficiency with meaningful automation, self-service, and integration with development pipelines.

- Flexible task automation: Nutanix Prism Pro provides a low-code/no-code, visual approach to task automation, enabling any administrator to build, maintain, and troubleshoot automations. Common admin tasks, like adjusting resources allocated to a VM in response to a constraint, are easily automated. Even the most complex, multi-step procedures can be turned into one- click operations.
- Self-service with no loss of control: Enterprise teams want self-service access to infrastructure and services to accelerate time to market. With Nutanix Calm, you can create blueprints that model applications and tasks and

publish them to an internal marketplace or add them to a growing collection of pre-integrated, blueprints on the Nutanix Marketplace. Application owners and developers can request IT services from the marketplace whenever needed. (See the section Calm Application Orchestration for more information.)

- Simplified development: Nutanix eliminates the complexity of test and development automation, allowing developers and administrators to work more efficiently. Your team can deploy and maintain a fully automated CI/CD pipeline with continuous application deployment across on-premises and cloud locations.

DevOps is a way to standardize processes and improving communication and collaboration between development and operations teams in an enterprise organization. DevOps methodology is not discussed in this design, but the constructs and the logical components included are elements that will help you develop/improve your DevOps strategy.

Security and Compliance Layer

Designing for your security requirements is paramount to delivering robust solutions. This is another area where this document is well differentiated from conventional datacenter architectures.

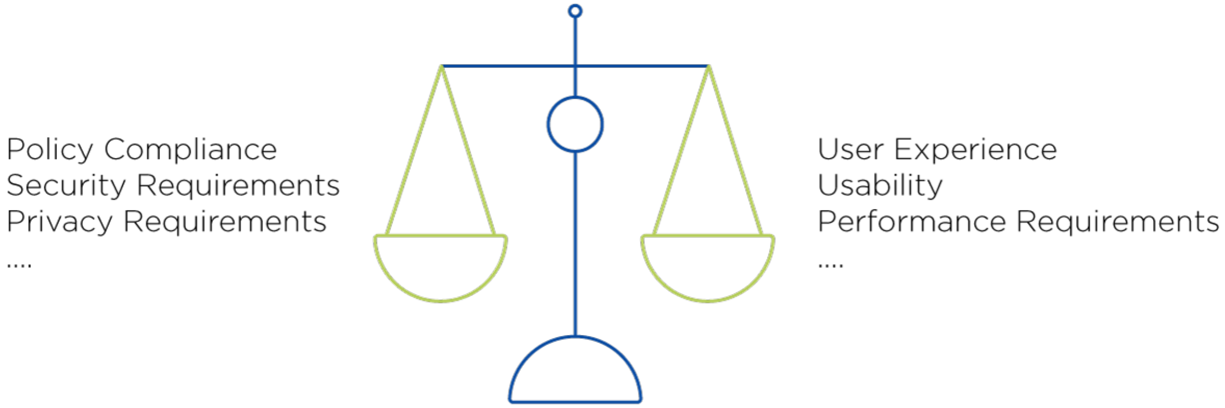


Figure 8: Designing for Security and Compliance

The Nutanix security approach is not predicated on having hundreds of different configuration options that you must set to achieve a secure environment. Nutanix takes a security-first approach including a secure platform, extensive automation, and a robust partner ecosystem. There are additional configuration options available if you need to add an extra layer of security based on business and or technical requirements.

Nutanix enables you to maintain a continuous security baseline to meet regulatory requirements more easily. Powerful security automation, called Security Configuration Management Automation (SCMA), monitors the health of storage and VMs, automatically healing any deviations from this baseline.

Nutanix provides customers with the ability to evolve from point-in-time security baseline checking to a continuous monitoring/self-remediating baseline to ensure all CVM/AHV hosts in a cluster remain baseline compliant throughout the deployment lifecycle. This new innovation checks all components of the documented security baselines (STIGs), and if found to be non-compliant, sets it back to the supported security settings without customer intervention.

In addition, data-at-rest encryption features and a built-in Key Management Server (KMS) further add our robust security capabilities.

These security features are discussed later in the Detailed Design section, and you can find more information online in the Nutanix Bible.

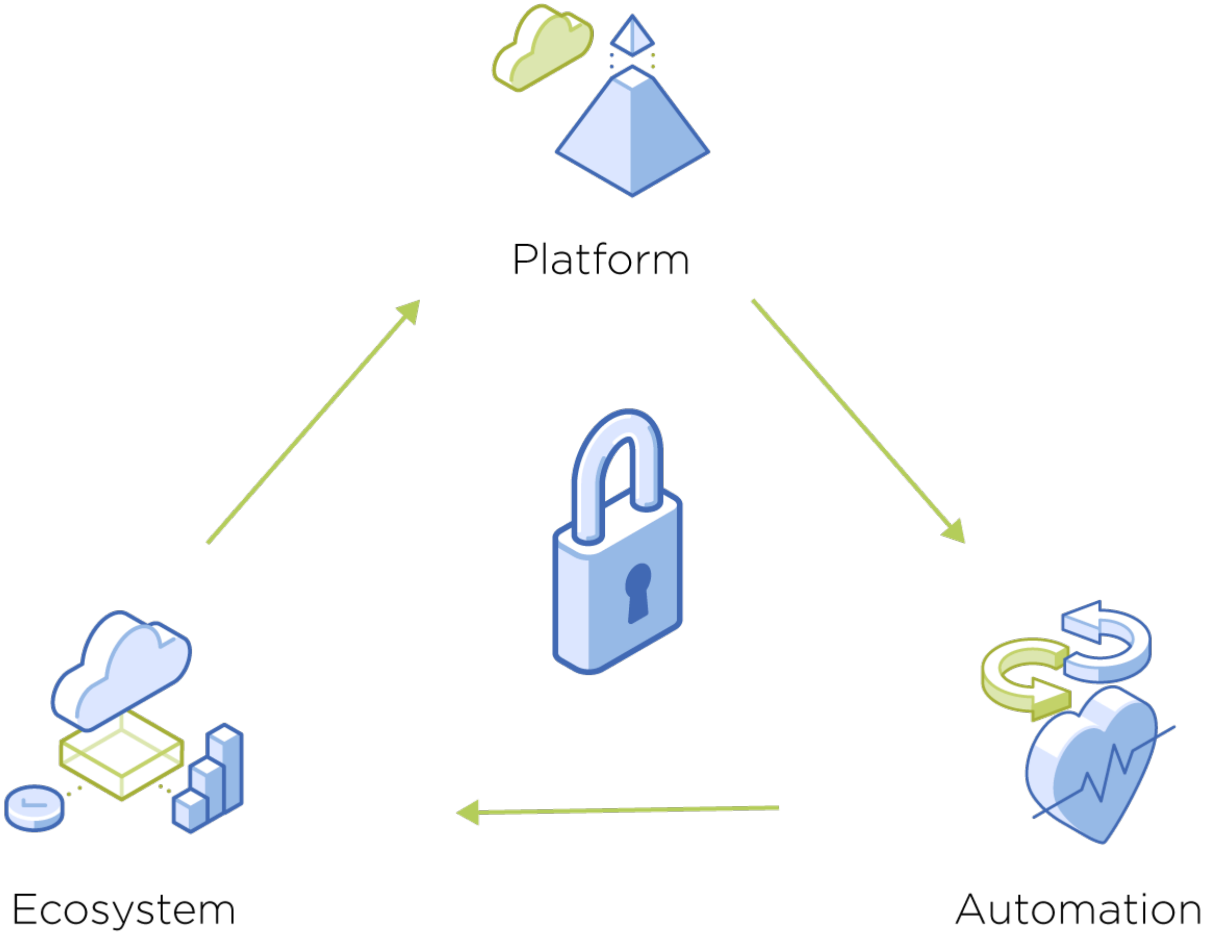


Figure 9: Nutanix Security

Nutanix incorporates security into every step of its software development process, from design and development to testing and hardening. Nutanix security significantly reduces zero-day risks. One-click automation and a self-healing security model ensure ongoing security maintenance requires much less effort.

Nutanix provides the following international security standards:

- 508 Compliant
- FIPS 140-2 Level 1
- National Institute of Standards and Technology (NIST) 800-53

- TAA Compliant

3. High-Level Design Considerations

A key feature of Nutanix Enterprise Cloud is choice. Nutanix customers have the flexibility to choose their preferred hardware vendor, CPU architecture, hypervisor, and more. This design is intended to help guide you to the best choices for your organizational needs.

There are a number of high-level design decisions that must be made before proceeding to a detailed technical design. This section provides the information necessary to help you make the following decisions:

- Choosing a datacenter architecture.
- Choosing a hypervisor.
- Choosing a cluster deployment model.
- Choosing a hardware platform.

Choosing A Datacenter Architecture

What availability must be delivered and what type of failures must the design protect against. These are key input requirements for a design since they will have an immediate impact on regions, availability zones, and data centers required plus budget.

Area and Availability Definitions

Region

A region is a geographical area with one or more availability zones (AZs). Regions are independent from each other so that failures in one region should not affect another region. Typical examples would be US east coast, US west coast, Europe north, Europe south, Asia east, Asia south.

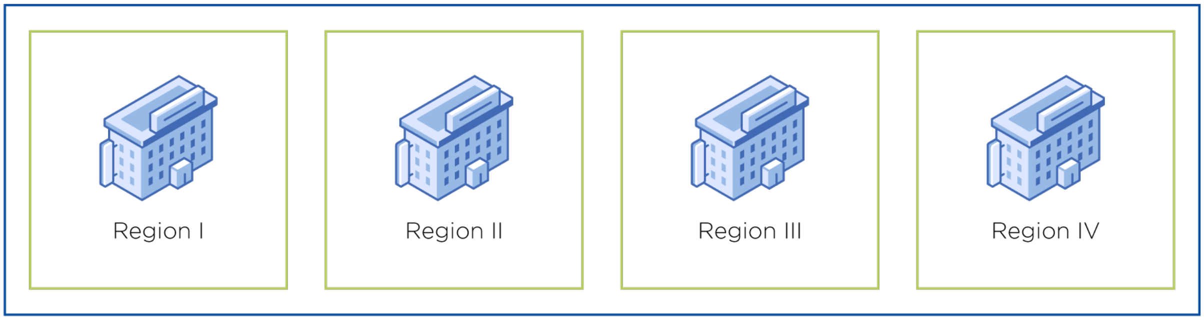


Figure 10: Regions

Availability Zone

A region holds one or more availability zones. Each availability zone contains one or more data centers.

Availability zones are implemented such that normal failures (such as a power plant failure) in one zone will not affect another. Natural and manmade disasters such as catastrophic earthquakes and nuclear strikes may disable more than one availability zone in a region.

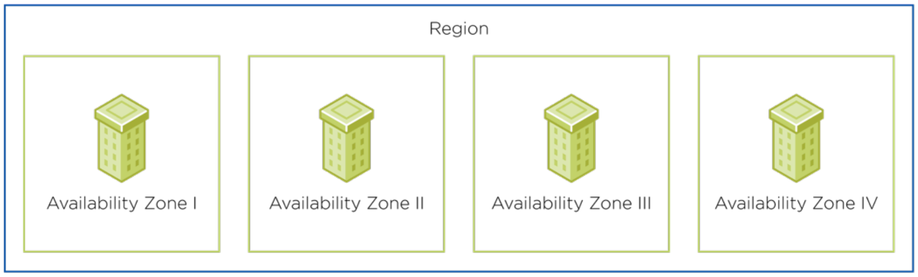


Figure 11: Availability Zones

Datacenter

Datacenters host hardware, management, and end user applications/services including network routers, network switches, firewalls, load balancers, physical servers running Nutanix software, and potentially third-party hypervisor(s).

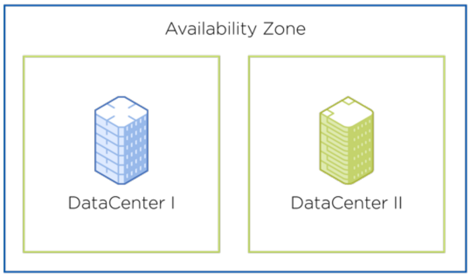


Figure 12: Datacenters

Datacenter Architectures

Nutanix provides built-in capabilities to provide support for a multi datacenter operating model, independent of the datacenters implementation architecture which usually can be divided into single location and multiple locations.

The different datacenter location models discussed can be leveraged using private or public implementation models, or a combination of the two typically referred to as a hybrid strategy.

Single Location

- Active
  - Applications are active within the datacenter
  - Backups are stored in the same datacenter as where the applications are running.
  - Long time archiving at offsite location
  - Limited protection against datacenter failure.

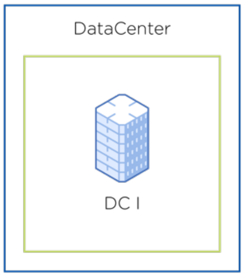


Figure 13: Single Location

Multiple Locations

- Active – Active
  - Applications are active in both datacenters during normal production.
  - Both sites can provide disaster avoidance and disaster recovery for each other.

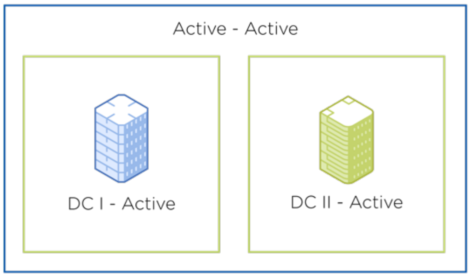


Figure 14: Multiple Locations

- Active – Passive
  - Applications are typically running from one datacenter only during normal production.
  - To make failover easier and more predictable the passive datacenter typically runs a limited number of infrastructure services during normal production.
  - The passive site provides disaster avoidance and disaster recovery functionality for the active site.

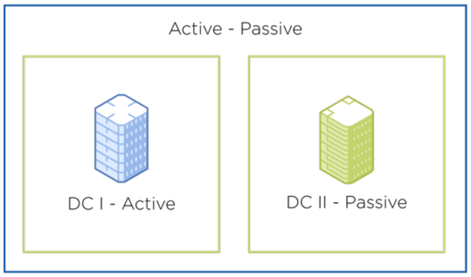


Figure 15: Active - Passive

- Active-Active-Passive
  - Applications are active in two datacenters during normal production.
  - The active datacenters provide disaster avoidance for each other.
  - To make failover easier and more predictable the passive datacenter typically runs a limited number of infrastructure services during normal production.
  - The passive site provides disaster recovery functionality.

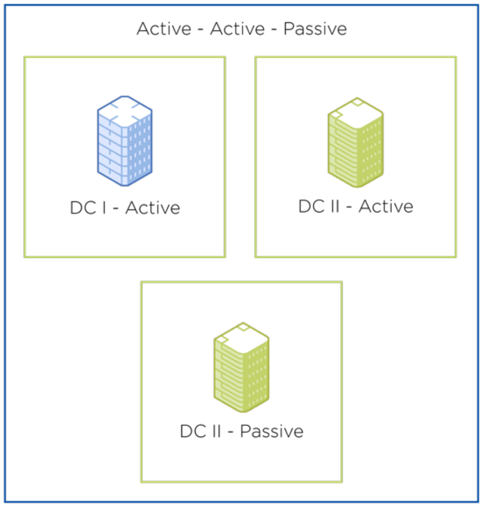


Figure 16: Active - Active - Passive

Remote Office/Branch Office Architectures

In addition to a multiple datacenter model Nutanix can be used to provide business continuity for remote office branch office (ROBO) implementation models with a datacenter as a fan-in/central starting point. including e.g.:

- Datacenter – ROBO
  - Multiple ROBO sites replicating data to the datacenters.
  - Datacenter acts as disaster recovery for the ROBO sites.

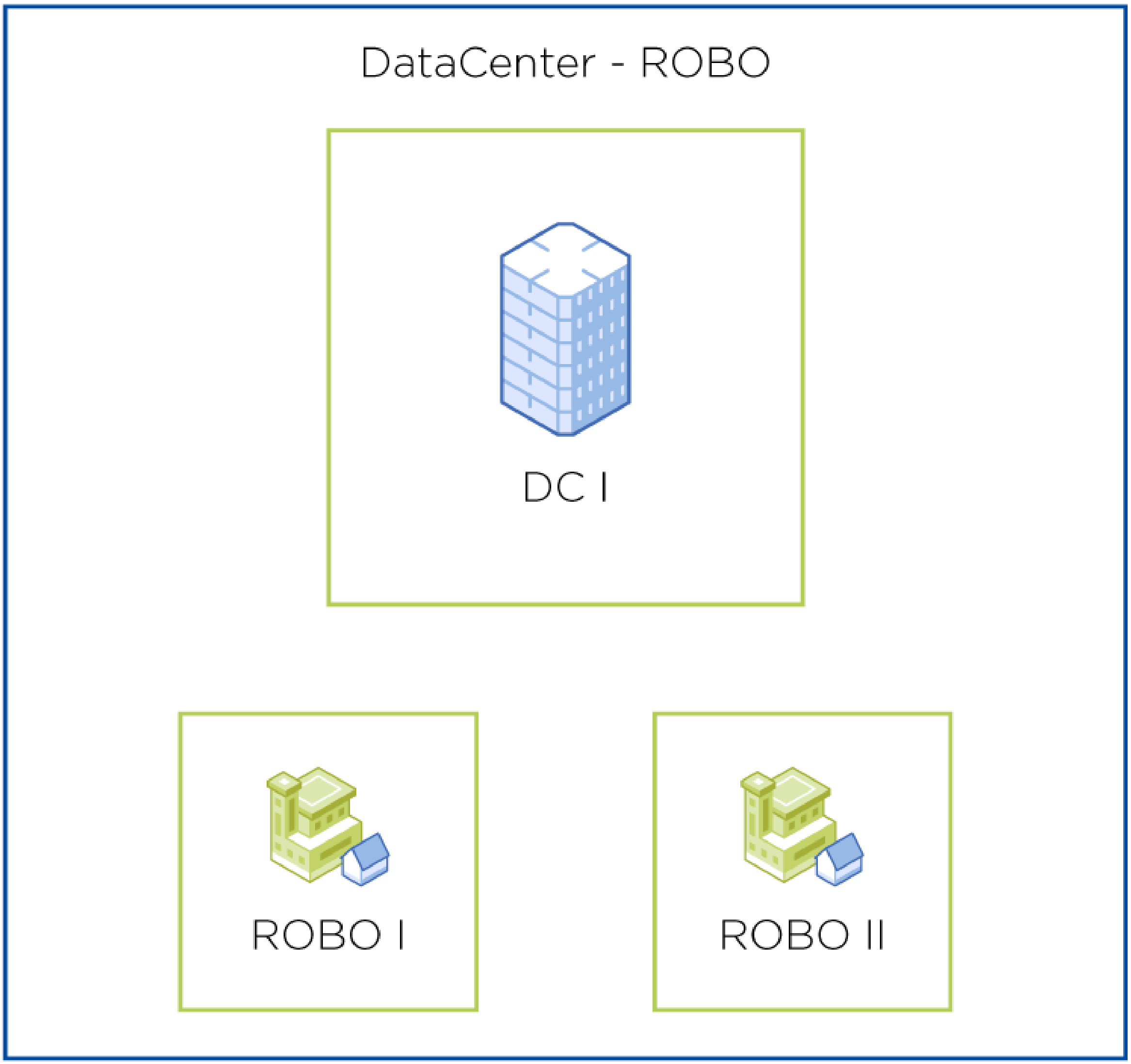


Figure 17: Datacenter ROBO

- Chain structure with Datacenter-ROBO-ROBO
  - Datacenter acts as disaster recovery for the closest ROBO site or sites
  - The first tier ROBO site/sites connected to the datacenter acts as disaster recovery sites for the next chain of ROBO sites. Some organizations use the term "regional data center" for these locations.

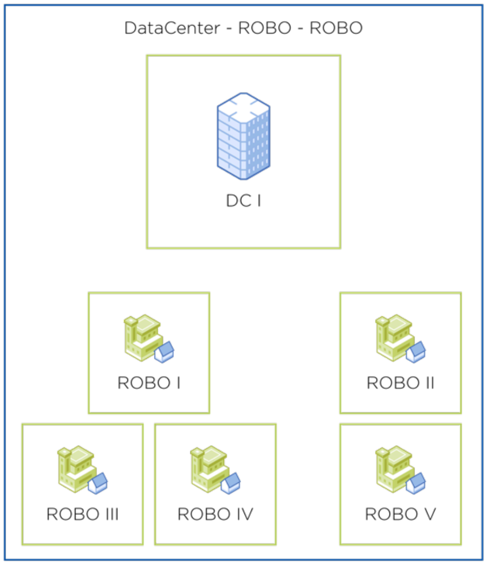


Figure 18: Chain structure with Datacenter-ROBO-ROBO

ROBO designated sites are typically a small footprint and differentiated from datacenters due to low bandwidth & high latency connections between datacenters and the ROBO sites. For larger customers, there will also be a proportionally high number of clusters/sites relative to the VM count.

Data can be replicated to more than one location for more advanced needs. An example of this policy might use NearSync between branches located in the same availability zone, and Asynchronous to a datacenter in another region

When choosing a ROBO architecture, keep in mind the restrictions that are most impactful:

- 100 millisecond or less latency between prism central and each cluster under management
- A scaled-out prism central can handle 300 clusters under management

- Accurate NTP access
- WAN restricted topology vs open internet access

Table 5: Number of Regions to be used

Region-001	NUMBER OF REGIONS TO BE USED
Justification	
Implication	

Table 6: Number of Availability Zones to be used

AZ-001	NUMBER OF AVAILABILITY ZONES TO BE USED
Justification	
Implication	

Table 7: Number of Datacenters to be used

DC-001	NUMBER OF DATACENTERS TO BE USED
Justification	
Implication	

This following table maps out the capabilities which can be applied to the different datacenter plus datacenter & ROBO architectures based on RPO and RTO requirements.

Table 8: RPO and RTO requirements

Nutanix Feature	RPO	RTO	Protect Against	Environment(s)
Time Stream / VM Snapshot	Minutes	Minutes	Minor Incident	Single Nutanix Cluster

Nutanix Feature	RPO	RTO	Protect Against	Environment(s)
Cloud Connect	Hours	Hours	Minor Incident	External Cloud
Asynchronous / Remote Replication	Minutes	Minutes	Major Incident	Multiple Nutanix Clusters
NearSync replication	Seconds	Minutes	Major Incident	Multiple Nutanix Clusters
Synchronous Replication	Zero	Minutes	Major Incident	Multiple Nutanix Clusters
Metro Availability	Zero	Near Zero	Major Incident	Multiple Nutanix Clusters

Typically, the multiple Nutanix clusters specified in the Environments column are placed in different datacenters but can be placed in same datacenter but in different fire zones.

Nutanix provides a variety of Node options and cluster sizes. Make sure to respect the current limitations:

- AOS Snapshot Frequency for Nutanix Nodes
- Single-Node Replication Target Requirements and Limitations

Choosing a Hypervisor

Nutanix supports a range of server virtualization options including: Nutanix AHV, VMware vSphere, and Microsoft Hyper-V. This design covers the deployment of either Nutanix AHV or VMware vSphere.

Nutanix AHV is included with AOS and delivers everything you'd expect from an enterprise virtualization solution: high performance, flexible migrations, integrated networking, security hardening, automated data protection and disaster recovery, and rich analytics. With robust, integrated management features, AHV is a lean virtualization solution.

A significant advantage of AHV as part of this validated design is the elimination of virtualization as a separate management silo and all of the complexity that entails. When designing a vSphere environment, for example, decisions must be made about the number of vCenters, the type of deployment mechanism, how to enable HA, and so on. Extensive ongoing training is often required for internal staff or consultants in order to effectively design, deploy, and upgrade the environment. AHV avoids these challenges.

AHV enables:

- Checkbox high availability configuration (vs. complex percentage or slot size config)
- No virtual SCSI devices required (vs. manually configured multiple SCSI devices for maximum performance).
- Distributed networking by default (vs. deciding between Standard or Distributed Switch).
- Automatic CPU Masking (vs. manual in vSphere).
- No need for Host Profiles.
- Fewer design choices.
- Simple control plane lifecycle.
- Simplified 1-click upgrades.
- Single support path.

VMware vSphere is a proven virtualization platform used by many organizations and has a robust ecosystem. It is a complex platform with many design choices and settings to tune; it often requires the purchase of additional licenses.

When deciding which hypervisor to use in your deployment, the choice comes down to which solution best meets your business and technical requirements and your budget. Key areas to consider when choosing a hypervisor include:

- Operating system support (including legacy and current OSes).
- Third-party virtual appliance support.
- Security/hardening baseline.

- Integration with third party products (backup, software-defined networking, anti-virus, security tools, etc.).
- Availability of automation tools.
- Staff skill set and training (architecture, administration, daily operations).
- Scalability of the solution.
- Licensing costs and model (AHV adds no licensing costs to a Nutanix deployment).
- Integration with the existing environment.
- Migration of existing VMs.
- Hypervisor and management plane technical features and performance.
- Simplicity or complexity of the solution.
- Features or products that support only one hypervisor or the other.
- Satisfaction with existing hypervisor platform(s).
- ROI/TCO of the full stack.
- Simplified support model (i.e. single support vendor vs multiple).
- Time to Deploy/Time to Value.

Freedom of choice is a key tenet of Nutanix. After considering these factors, it is likely that one hypervisor platform stands out as the best option for your deployment. No matter which hypervisor you choose, the solution is backed by world-class Nutanix support and a full ecosystem of node types. (Hardware selection is discussed in the section Platform Considerations).

Table 9: Choose a Hypervisor

VRT-001	CHOOSE NUTANIX AHV OR VMWARE ESXI AS THE HYPERVISOR FOR YOUR DEPLOYMENT
Justification	
Implication	

Choosing a Cluster Deployment Model

When designing Nutanix clusters for your deployment, there are several important design considerations:

- Choosing whether to deploy a separate management cluster?
- Choosing whether your cluster(s) will run mixed workloads or be dedicated to a single workload?

Separate Management Clusters

A separate management cluster is not a requirement of this design. If you are planning a smaller-scale deployment, it doesn't always make sense to design, purchase, and operate a separate cluster for a limited set of management VMs.

However, the infrastructure applications and services that run on a management cluster are critical. At a larger scale, there are several reasons to separate these management workloads:

- Availability: Separating management workloads on their own cluster(s) makes it easier to ensure they are always available.
- Security: you may wish to more strictly control access to management clusters and workloads, including additional controls such as firewalls, dedicated networks, more stringent role-based access (RBAC), and possibly others. While this can be accomplished on mixed clusters, it is easier to monitor and manage these additional security controls on a separate physical cluster.
- Operations: RBAC prevents those that are not authorized from interacting with important management services. Physical separation prevents any accidental or malicious actions that could compromise the availability or performance of important infrastructure services.
- Performance: Performance is just as important for infrastructure services as any other workload. Having a dedicated management cluster(s) simplifies troubleshooting of performance issues and reduces the potential for hard to diagnose workload conflicts. It eliminates the possibility that changes in the management space will affect application workloads and vice versa. In a highly elastic cloud environment, there may be workload expansions that

occur via self-service or automated events. Depending on whether resource control policies are in effect, this can cause resource contention in a shared cluster.

Note: You do not need a separate management cluster for each use case or environment unless you have strict security or compliance requirements that make it necessary. This means that a single management cluster can support multiple use cases such as: EUC deployments, private cloud, and general server virtualization environments.

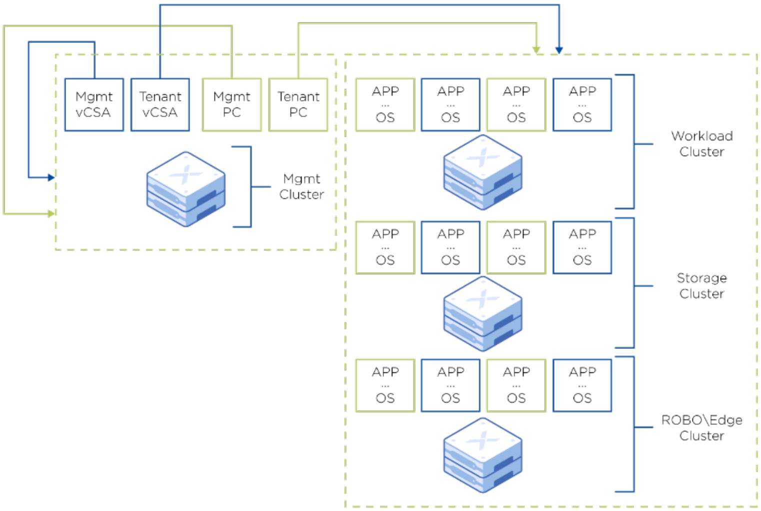


Figure 19: Separate Management Clusters

Should You Deploy Two Management Clusters?

In large-scale deployments, the management cluster can be split to create separate failure domains. With two management clusters, you can place redundant components into each cluster, enabling a higher level of availability. Should there be an issue with one of the management clusters, the other remains available to service requests.

Table 10: Management Cluster Architecture

PFM-001	MANAGEMENT CLUSTER ARCHITECTURE: DEPLOY A SEPARATE MANAGEMENT CLUSTER OR SHARE A CLUSTER WITH OTHER WORKLOADS. WHEN CHOOSING A SEPARATE MANAGEMENT CLUSTER, CONSIDER A REDUNDANT CONFIGURATION.
Justification	
Implication	

Will You Deploy Dedicated or Mixed Clusters?

The decision whether to mix workloads within a cluster or to dedicate a cluster for each type of workload is usually a question of scale. For example, if you have 200 general server VMs, a small Exchange deployment, and 10 average-sized database VMs, mixing the workloads in a single cluster is common and can be easily managed. However, if any or all of these workloads increase by 5-10x, the complexity of sizing and operating the mixed environment goes up dramatically.

Operating large-scale mixed environments creates a number of unique challenges. You have to decide whether you are willing to manage the challenges of mixed workloads or dedicate clusters for each workload. Here are the main factors to consider:

- Performance and capacity: The resource demands of different applications can vary widely. You need to understand the needs of each application when mixing workloads since the chance for conflicts to occur is increased. There may also be wildly different performance and capacity needs between applications which could require different node configurations within a single cluster. Unless you are going to isolate certain workloads to particular nodes, each node within a cluster needs to be able to handle the average daily mix of applications that might be running on it.
- VM resource-sizing requirements: The CPU and memory sizing for general server VMs, VDI VMs, business-critical application VMs, etc. can vary widely. While it’s fairly easy to account for memory sizing, CPU sizing is more complex. Each of these workloads consumes different amounts of CPU, and may require much different levels of CPU overcommit, if any.

If you have large groups of VMs with widely different resource requirements, it's typically better to build clusters to contain more uniformly sized VMs. From a hypervisor HA standpoint, you may require additional resources within a mixed cluster to ensure the ability to failover. This can also increase the day 2 operational support effort, since it may require manually tuning HA settings and increased monitoring to ensure HA resources remain in compliance.

- Software licensing: The most common reason for dedicated clusters is software licensing. There are a variety of reasons why using dedicated clusters make sense from a licensing standpoint. Here are two common examples:
  - Operating system licensing: Windows and Linux vendors may offer "all-you-can- eat" license models for licensing at the host level. Therefore, it makes more sense to have clusters dedicated to either Windows or Linux VMs, to minimize licensing costs.
  - Database licensing: Database licenses are frequently based on either CPU cores or sockets. These licenses can be expensive, and you often have to license all the nodes in a cluster to enable DB VMs to run on every node. Once again, you probably don't want to run other workloads on the cluster since that reduces the return on your license investment.

In addition, you may want to run nodes hosting database workloads on different hardware than nodes hosting general server VMs. For instance, having fewer CPU cores running at a higher clock speed may reduce your overall licensing costs while still providing the necessary compute power.

This is provided for informational purposes only, please refer to your licensing agreement with Microsoft for more information.

- Security: In many projects, security constraints are an overriding factor. When it comes to the security of mixed vs dedicated clusters, there are a few design considerations that are important to consider as you decide whether logical or physical separation is adequate to address your security requirements:
  - Operations: RBAC is the primary means of controlling access and management of infrastructure and VMs in a mixed cluster. Dedicated

clusters prevent non-ap proved parties from gaining any access whatsoever.

- Networking: A mixed cluster typically relies on separate VLANs and firewall rules per workload to control access. A dedicated cluster would only have the required networks presented to it for a single workload and would likely limit who and what has network access. Both approaches can control and limit access to cluster resources, but dedicated goes a step further by providing complete physical network isolation for those that require it.

If you are able to address the performance/capacity, resource-sizing, licensing, and security constraints discussed above, it is possible to successfully design and operate clusters of any size to run mixed workloads. The key is to thoroughly understand the requirements of each workload and size with that knowledge.

Recommendation

At scale, it is our recommendation to use dedicated clusters to the greatest extent possible for the reasons discussed above. Utilize mixed workload clusters where it fits.

Including Storage-Only Nodes in Clusters

Storage-only nodes contribute storage capacity and I/O performance within a cluster. Storage only nodes are available for clusters running either AHV or ESXi. Storage-only nodes can be any node type, but typically are configured with just enough CPU and memory resources to run the CVM since there aren't any application VMs running on these nodes. These nodes are a member of the AOS cluster but are not visible to the hypervisor cluster for non-storage functions, so the hypervisor won't schedule other VMs to run on them.

Table 11: Mixed or Dedicated Workloads Per Cluster

PFM-002	MIXED OR DEDICATED WORKLOAD PER CLUSTER
Justification	
Implication	

Choosing How You Will Scale

This document clearly lays out the Nutanix opinionated design for deploying HCI clusters and building a private or hybrid cloud for any type of workload. A challenge in any design, whether you're starting with one cluster or dozens, is how to scale the environment to reach your goal. To accomplish this, you need an overarching architectural plan that creates a repeatable process that can be followed by the organization. Defining this in advance removes confusion and simplifies future deployment. An overarching architectural plan allows you to track progress and ensure compliance with the design.

For this design, we recommend that an at-scale deployment should include:

- A control plane.
- A block and pod architecture.

Choosing a Control Plane

Within your design there should be a primary control plane where the majority of daily operational tasks are performed. In this design:

- The primary control plane is Prism Central.
- VMware vCenter is also required for those using VMware ESXi.

Each of these control planes has a maximum size that dictates the number of VMs, nodes, and clusters that can be managed by an instance.

Block and Pod Architecture

A repeatable architecture is needed to ensure safe and efficient scaling. This design uses the pod and block architecture because it's both familiar and easily consumed. We have adapted it to support Nutanix deployments. This section explains the different parts of the architecture and how they can be used to scale a deployment to any level.

What is a Pod?

In this design, a pod is a group of resources that are managed by a single Prism Central instance. The diagram below shows a single pod containing four building blocks. A pod is not bound by physical location limitations; all of its resources could be at a single site or at multiple sites. Examples of multiple sites include: a traditional multi-datacenter design, a hybrid cloud design,or a ROBO architecture.

Building Block

In this design, a build block is equivalent to a Nutanix cluster. Each of the clusters can run a single dedicated workload or mixed workloads. There is no need for building blocks to be uniform within your design.

Recommendation

For each workload:

- Establish whether it will have a dedicated cluster or share a mixed cluster with other applications.
- Define the maximum size of this type of building block. While building blocks for different workloads do not need to be the same, they certainly can be.
- Determine the maximum size of a building block based on:
  - The scale at which you will deploy workloads.
  - The need for failure domains.
  - Operational considerations such as upgrade timing.

These topics are discussed later in the Detailed Technical Design section of this document.

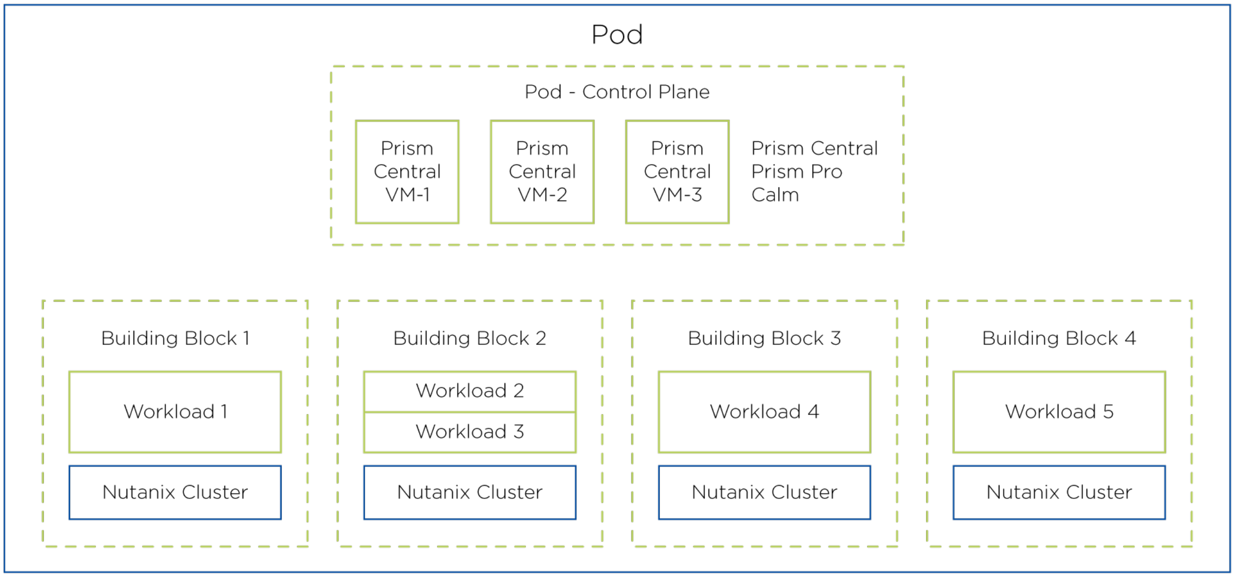


Figure 20: Block and Pod Architecture

Scaling the Block and Pod Architecture

You can scale an individual pod up to the maximums that Prism Central supports for the AOS version you are deploying. For the AOS version specified in this document, a scale out Prism Central deployment with large sized VMs can manage up to:

Prism Central Limits (assumes Scale-Out PC configuration):

- 25,000 VMs
- 200 clusters
- 1,000 nodes

If any of these limits are reached, a pod is considered full. For example, with an EUC deployment, the VM limit will likely be reached first since you typically have high VM density resulting in large numbers of VMs with fewer nodes and clusters. A large ROBO environment might hit the cluster count limit because you tend to have many sites each with a small cluster and a few VMs.

At very large scale (e.g. thousands of nodes), it can make sense to have pods dedicated to each workload, but for environments that have Nutanix deployed for multiple workloads, a pod will typically contain multiple applications.

Once a pod reaches a scaling limit, start a new pod with at least one building block. The new pod scales until it also reaches a scaling limit, and so on.

The building blocks within a pod scale in a similar fashion. A building block is started, and workloads are migrated onto it until it reaches its determined max size, and a new building block is started. New building blocks can be as small as 3 nodes, the minimum to start a Nutanix cluster, or any size up to the max size you've specified for that building block.

The starting size for each building block and the increments for scaling them are organizational decisions:

- For smaller or more agile organizations, starting small and scaling incrementally often makes sense.
- Larger organizations may prefer to deploy a new building block fully populated and migrate workloads onto it as schedules dictate.
- Although 3 nodes is the minimum size for a cluster, using 4 nodes provides a higher level of redundancy during maintenance and failure conditions.
- For ROBO and edge use cases, the starting size can be as small as one, two, or three nodes depending on requirements.

The diagram below illustrates a simple VDI building block example. With VDI it is easy to think in terms of number of users and the node count in a cluster. In this example, the building block is a 16-node cluster supporting 1,500 users. This works out to 100 users per node plus an additional node for HA. With the first building block full, a request for an additional 500 users requires a new building block to be started. This building block is then scaled up to its max size before starting a third.

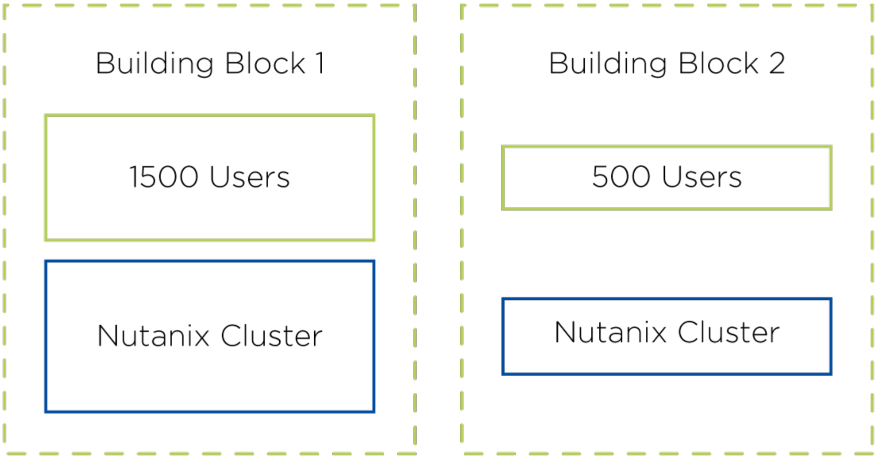


Figure 21: Example VDI Building Block

By having well established and documented design decisions for pod size and building block size, the architecture and operational teams are free to keep scaling without the need to revisit decisions to satisfy each resource expansion request.

Choosing the Right Licensing and Support

Now that you've chosen a hypervisor, decided on mixed or dedicated workloads, and established your block and pod architecture, you can make informed hardware decisions. When evaluating the different Nutanix platform options available, there a few key decision points:

- Choosing Your Software licensing and Support.
- Choosing Your Platform Vendor.

Software Licensing and Support Considerations

Nutanix nodes are available in two different purchasing/licensing options: appliances or software-only:

- Appliances are available directly from Nutanix or through our OEM relationships.
  - Appliance based licensing is referred to as “life of device licensing”, meaning it’s only applicable to the appliance it was purchased with.
  - The manufacturer of the appliance takes all support calls for software and hardware issues. For example, if you choose Dell appliances, Dell will take all support calls and escalate to Nutanix for software support as needed. With Nutanix NX appliance all support calls go directly to Nutanix.
- The Software Only option de-couples software and support licensing from the underlying hardware. This enables:
  - License portability. The same license can continue to be used when the underlying hardware is changed, such as a hardware vendor change or a node refresh. (note: licenses are portable for like-for-like hardware replacements. If the hardware specification of the nodes change, then there might be the need for additional licenses.).
  - Deployment on additional supported hardware platforms. Hardware is per our qualified list of platforms, which can be found on the Hardware Compatibility List (HCL).
  - Direct software support from Nutanix while the server vendor provides hardware support.
  - Another type of Software Only Licensing is the Core-Licensing option, which is best for those customers who prefer the benefits of the software-only model but want hardware from a specific OEM server vendor. Core licensing enables customers to purchase software only licenses and buy hardware from any of the appliance vendors. For example, XC-Core utilizes Dell XC OEM appliances but de-couples software and hardware support.

Table 12: Nutanix Software Licensing Level

PFM-003	SELECT NUTANIX SOFTWARE LICENSING LEVEL
Justification	
Implication	

Vendor Considerations

When it comes to selecting a hardware vendor, there are a number of factors to evaluate:

- Brand loyalty. This can be a strong factor in an IT buying decision. There may be purchasing commitments or discounts in place at an organizational level driving this loyalty, or you may simply have been happy with past experiences.
- Support Quality. The quality of the support experience can also be a factor in your hardware evaluation. For hardware failures, makes sure the hardware vendor responds quickly and can provide parts reliably within the contracted response time. The overall support experience is important. A vendor should be easy to contact, be responsive to requests, and provide resolution in a timely manner.
- Hardware Quality. The reliability and quality of hardware is of obvious importance. Today's servers are all very similar internally; they use many of the same components with just a few proprietary components for each vendor. The reliability of the leading server vendors is pretty similar so your decision may be contingent on your organization's past experiences.
- Operational Experience. When it comes to the ongoing operational experience, what does it take to carry out day 2 operations to support the lifecycle of the physical server and the vendor toolset (if any). This includes: monitoring server health, reporting and upgrading firmware, and monitoring for component issues and failures. Virtually all of the server vendors offer tools for these activities and when combined with the power of AOS and Prism, the experience is pretty similar. Nutanix Lifecycle Manager (LCM) offers firmware reporting and management for all Nutanix appliance options.

- Configuration Options. Physical configuration options may weigh strongly in choosing among server vendors, and in choosing a model type for each workload. Factors include:
  - Number of sockets.
  - CPU options.
  - Number of storage bays.
  - Network connectivity options.
  - Storage media options. While there is generally a level of parity between server vendors, a particular vendor may offer something the others do not, or one vendor may offer the latest options more quickly when new components are released. You may require or prefer a specific type of network card or need nodes with a large number of storage bays. Storage considerations, such as the number of SSD/NVMe options available, RDMA capabilities, and support for large-capacity media may be important for specific workload requirements.
- Physical form factor. This is another common decision point since it can affect the amount of space consumed in a rack, the power draw, the number of network connections, and the number of internal expansion slots. Availability of internal expansion slots may limit the number and type of network cards that can be deployed, and whether a node is capable of accepting GPU cards and the number of GPU cards it's capable of supporting. When it comes to different form factors, there are:
  - High-density chassis that offer either four or two physical nodes in 2U of rack space. These are popular options for a variety of workloads that do not require extensive internal expansion or a large number of storage bays.
  - Standard rackmount servers, typically with a 1U or 2U chassis and one physical server per chassis. These provide much wider capabilities in the number of storage bays available, the number of internal expansion slots available, and may also support more memory.

Table 13: Physical Node Vendor

PFM-004	SELECT PHYSICAL NODE VENDOR
Justification	
Implication	

Mixed Configurations and Node Types

Nutanix clusters allow significant flexibility in terms of the node types and configurations you can utilize in a single cluster. This allows clusters to be operated and expanded over time without artificial constraints. A cluster can be expanded with different node configurations to accommodate new workloads or when previous nodes are no longer available.

Considerations include:

- Node models. Mixing node models within a cluster is a fairly regular occurrence. While it's possible to have the same CPU, memory and storage configuration in two different node types, it's not required in order to mix them in the same cluster.
- CPU configurations. Mixing nodes within a cluster with different CPU configurations such as core count, clock speed, or CPU generation is supported. This can be to address changing application requirements, inventory availability, financial constraints, time of purchase, or other factors.

While there is no limit to the drift between configurations, it's a commonly accepted best practice to keep the core counts and memory configuration of nodes within a cluster at similar levels. Using different CPU generations in the same cluster can limit the feature set / functionality of newer CPUs. The lowest common denominator is the level of the oldest CPU generation within a cluster. Having mostly uniformly configured nodes in a cluster makes it easier for humans to double check the HA and capacity planning recommendations of automated tools.

- Storage media. Having nodes with different storage configurations is also supported. Variations at the storage layer can include varying the size or number of SSDs or adding all-flash nodes to a hybrid cluster.

When introducing a storage configuration within a cluster that dramatically increases the amount of storage, you must ensure there is ample failover capacity to rebuild the largest node. For example, suppose the existing nodes in a cluster have 10TB of capacity each and you want to add node(s) with 40TB of capacity. Initially, it's best to add a pair of these 40TB nodes. Subsequent large-capacity-node additions can then be done one at a time.

Choosing the Right Hardware for Your Workload Types

When it comes to sizing Nutanix clusters for different application workloads there are many options. Most workloads can successfully run on any of the Nutanix and OEM models available, but some models and configurations may be a better fit than others.

Once you know the requirements of your workload(s) you can use the Nutanix Sizer to determine the best configuration for your cluster(s).

Nutanix Sizer is a web-based application that is available to Employees, partners and select customers. Sizer allows for the architect to input application and workload requirements and the node and cluster configurations are automatically calculated.

Different hardware platforms offer different characteristics to address differences in workloads. The following sections provide guidance on model selection, and performance considerations.

Model Types

The various Nutanix and OEM appliance models and hardware compatibility list (HCL) servers provide the flexibility to identify the right solutions to meet financial, space, and performance requirements for different projects. There may be differences in terms of the number and types of models available, but the same level of flexibility is generally available from all vendors. Some vendors offer fewer models but allow greater configuration flexibility.

Generally speaking, there are four different groups that servers fall into, translating to different use cases:

- General workloads and EUC: These are by far the most popular nodes deployed in Nutanix clusters. They can handle the vast majority of workloads including: general server virtualization, business-critical applications, VDI, and most others. There are a mix of form factors available.

- ROBO and Edge: These are similar to the general workload options, with the exception that they may offer few options for CPU and storage as they are optimized for these edge use cases.
- Storage dense: For workloads that require large amounts of storage capacity, storage-dense nodes offer a larger number of storage bays and dense media options, possibly with fewer CPU options. The physical configuration of these nodes is optimized for workloads such as Nutanix Files, Nutanix Objects, or to be utilized as storage-only nodes.
- High performance: The most demanding workloads and business-critical applications (BCA) may require additional CPU resources and storage performance. For these workloads there are nodes offering additional CPU configurations in terms of core count and clock speed, as well as quad-socket configurations. It’s common for these models to offer as many as 24 storage bays to allow for more flash devices or hard drives for workloads that can utilize the added performance characteristics.

Each of the above model alternatives offer one or several of the available physical form factors along with the density and performance characteristics discussed.

Nutanix software does not require any complex tuning or configuration to support the different workloads, but there are plenty of hardware options to tailor your selections to different use cases.

Performance Considerations

Nutanix and AOS meet the performance demands of different workloads without continuous performance tuning. The Nutanix HCI storage fabric is powerful and intelligent enough to handle nearly any type of workload.

However, different cluster design and configuration options still yield performance benefits. Selecting the appropriate node model and configuration to meet application and solutions requirements is an important design decision. The primary design considerations for performance are:

- Number of drives. The number of hard disk drives (HDDs) or flash devices in a node can dramatically affect its performance characteristics. However, simply picking the node with the most device bays won't improve the performance of every workload.
  - Write-heavy workloads benefit from additional storage devices to provide performance and consistency. Other workload characteristics such as read/write ratio and I/O size should also be considered.
  - Workloads such as VDI typically have minimal capacity requirements but higher IOPS demands. It's common to utilize nodes with partially populated storage bays and as few as 2 flash devices per node, providing the right amount of storage capacity while still exceeding performance demands.
- All Flash. All-flash configurations are available from Nutanix, OEMs, and supported third-party server vendors. All-flash clusters utilize only SSDs, and these configurations provide higher IOPS and a more consistent I/O profile. While all flash configurations have become common, they are not absolutely necessary for every workload and use case.
- NVMe. There are a number of new technologies available now and coming soon that offer additional performance capabilities. NVMe is the first of these to be widely available and offers a number of benefits over SSD. New flash technology allows NVMe devices to deliver higher levels of I/O with lower latencies.
- RDMA. To realize the full benefits of NVMe, nodes are typically configured with remote direct memory access (RDMA). RDMA allows one node to write directly to the memory of another node. This is done by allowing a VM running in the user space to directly access a NIC, which avoids TCP and kernel overhead resulting in CPU savings and performance gains.
- Size of flash tier. In hybrid configurations containing SSD and HDD devices, the bulk of the performance comes from the flash tier. Therefore, it's important to understand the workload being deployed on a hybrid cluster. The data an application accesses frequently is typically referred to as the working set. The flash tier in hybrid clusters should be sized to meet or exceed the size of the working set for all of the applications that will run on

the cluster. There is no penalty for having too much flash in a cluster but not having enough can result in inconsistent performance.

Table 14: Node Models

PFM-005	**SELECT NODE MODEL(S) PER USE CASE**
Justification	
Implication	

Note: As your organization works through the Detailed Technical Design elements in the following section, be prepared to revisit your model decisions to fine-tune CPU, memory, and storage configurations.

4. Detailed Technical Design

With the necessary high-level design decisions made, including hypervisor, deployment model, and hardware platform and models, you can now plan the technical design for your Nutanix deployment.

This section provides technical guidelines for each layer of the design stack. Disaster recovery and business continuity guidelines are provided in chapter 5. Where possible, we've organized the sections so that you don't have to spend time reading material that doesn't apply. For instance, if you are not deploying VMware, you can skip all sections that are applicable only to it.

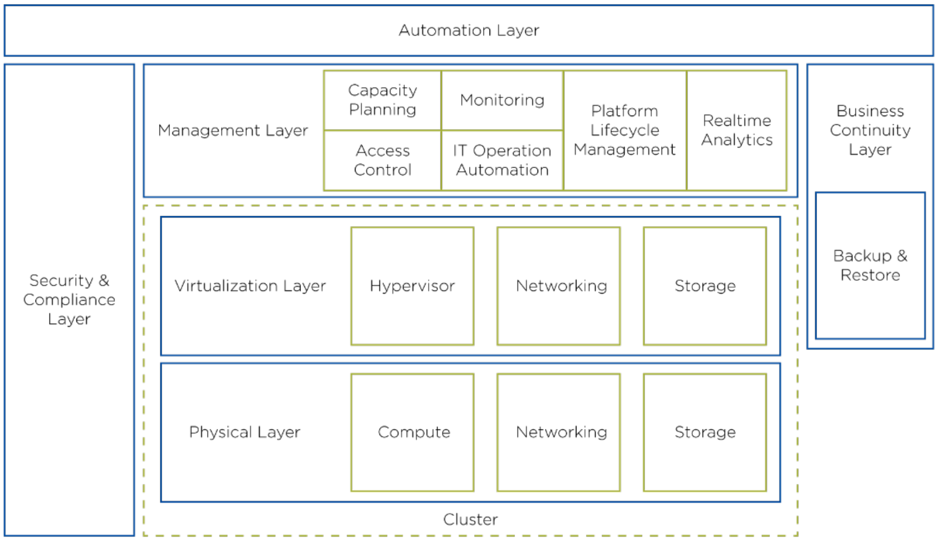


Figure 22: Planning the Technical Design

Required Software Versions

This design assume that you will be running the following software versions:

Table 15: Software Versions

Name	Version	Description
Prism Central	Latest version	Use latest version available (pc.2020.7 at the time of writing).
Nutanix AOS	Latest LTS	Use latest LTS available (5.15 at the time of writing).
Hypervisor Options		
Nutanix AHV	Latest LTS	Use latest LTS available (5.15 at the time of writing).
VMware vSphere	6.7 or 7.0	
VMware vCenter	6.7 or 7.0	Recommended for vSphere deployments.

Physical Layer Design

This section guides you through the process of designing all physical aspects of your Nutanix deployment, including:

- Compute and Storage Design.
- Networking.
- Choosing cluster size.
- Failure domain considerations.
- Designing workload domains and cluster layout.

Choosing the Optimum Cluster Size

This section provides guidance for cluster sizing. When designing a Nutanix cluster it’s important to consider more than just technical limits and recommendations. There are other important considerations including: hardware vendor recommendations, security, and operational requirements that may dictate the optimal cluster size.

The following table provides a comprehensive list of the design considerations that pertain to cluster size:

> Note: These considerations apply to cluster sizing. Additional information for each area is provided in later sections.

Table 16: Cluster Size Design Considerations

AREA	LIMITING FACTOR(S)	CONSIDERATIONS
Operations/ manageability	Maintenance window	Define/validate your maintenance window and make sure the cluster upgrade process fits in the window. Example: • Maintenance window: 12h. • Full single node upgrade: at least 45m (Hardware, FW, AOS, and hypervisor). Time depends on hardware vendor. • Maximum cluster size: 12 -15 nodes.

AREA	LIMITING FACTOR(S)	CONSIDERATIONS
Security & compliance	Security zones within an organization or business unit	Collect all relevant security and compliance requirements. Example: • Multiple security zones: Internet DMZ, PROD, test & dev, inner-DMZ • Every security zone may dictate different cluster sizes: › Internet-facing DMZ clusters usually have a smaller number of nodes and smaller number of workloads to minimize impact of a security breach or DDoS attack. › For example, Dev/Test clusters that have lower criticality may also have less strict change management policies and can therefore have a large numbers of hosts.
Vendor recommendations and limitations	• Hypervisor limitations • Management plane limitations • Vendor recommendations	Each product has limitations and vendor recommendations. Make sure you do not cross boundaries set by the vendor. See the vendor limitations/ recommendations table.

AREA	LIMITING FACTOR(S)	CONSIDERATIONS
Business continuity	• RPO • RTO • Backup window	Collect BC/DR requirements, RPO, RTO, backup and restore time window, backup system performance statistics. Example: • RPO 24h • RTO 48h Ensure you can recover/restore from backup and restart workloads within 48h. A cluster where total storage capacity exceeds technical capabilities of the backup system could fail to meet desired RTO.

AREA	LIMITING FACTOR(S)	CONSIDERATIONS
Workload considerations	• Application architecture • Application licensing • Application criticality	Verify application architecture with the application team/vendor, including HA, DR, scale-in vs scale-out, and performance requirements. Consider licensing model for each application and its implications. Example #1: Oracle or MS SQL licensing Licensing models are based on physical core count. Design clusters for database performance and capacity requirements to avoid cluster oversizing and minimize license costs. Example #2: Application has its own HA or DR If application can provide native HA and/or DR, RPO/RTO considerations described under Business Continuity (above) may not apply.

AREA	LIMITING FACTOR(S)	CONSIDERATIONS
Networking	• Total available network switch ports. • Available network switch ports per rack.	Available physical ports per rack and rack row is important when choosing cluster size and number of clusters. Example: • 96 ports (10GbE) available per rack. • 48 ports (1Gbps) available per rack. • 2 x 10GbE uplinks per Nutanix host. • 1 x 1Gbps uplink for Out-of-Band management. Maximum nodes per rack is 48 (total capacity of the TOR switches).
Datacenter Facility	• Available server rooms. • Total power and cooling. • Power and cooling per rack. • Available total rack units. • Available rack units per rack. • Floor weight capacity.	Power and cooling is one of the most important factors limiting Nutanix cluster size and node density. You have to ensure you do not exceed any hard limits when designing your cluster layout. When calculating power consumption and thermal dissipation use maximum values provided by vendor. Typical datacenter rack is 42U. Some datacenters have racks up to 58U.

Maximums and Minimums: AHV Deployments

The following table shows the maximum limits for management plane software components in AHV deployments:

Table 17: Maximum Limits for Management Plane Software Components in AHV Deployments

Management Software	Max. # of Hosts	Max # of VMs	Notes
Nutanix Prism Central	Up to 200 clusters or 25000 VMs.	1000 nodes or	Assumes Prism Central scale out is deployed.

The following table provides guidance regarding minimum and maximum number of nodes supported in a single Nutanix AHV cluster:

Table 18: Min and Max Number of Nodes in a Single AHV Cluster

Min # of Nodes	Max # of Hosts	Hypervisor	Recommendation
1	No Limit. Cluster size based on a variety of factors.	Nutanix Acropolis Hypervisor (AHV).	Single and dual-node clusters are for ROBO only.

Maximums and Minimums: VMware vSphere Deployments

The following table shows the maximum limits for management plane software components in VMware deployments:

Table 19: Maximum Limits for Management Plane Software Components in VMware Deployments

Management Software	Max. # of Hosts	Max # of VMs	Notes
Nutanix Prism Central	Up to 200 clusters or 25000 VMs.	1000 nodes or	Assumes Prism Central Scale-Out.
VMware vSphere vCenter	2000 hosts	25000 powered on	

The following table provides guidance regarding minimum and maximum number of nodes supported in single VMware cluster:

Table 20: Min and Max Number of Nodes in a Single VMware Cluster

Min # of Nodes	Max # of Nodes	Hypervisor	Recommendation
2	64	VMware vSphere ESXi	Dual-node clusters are for ROBO workloads only. Maximum number of nodes in single Nutanix cluster with VMware ESXi is limited by hypervisor version. For more details refer to official VMware documentation.

Table 21: Number of Clusters

PFM-006	NUMBER <<TYPE AND SIZE>> OF CLUSTERS
Justification	
Implication	

Failure Domain Considerations

Failure domains are physical or logical parts of a computing environment or location that is adversely affected when a device or service experiences an issue or outage.

The device or services that are affected can greatly affect the size of the failure domain and its potential impact. For example, a router generally has a bigger failure domain than a wireless access point since more endpoints rely on a single router than a single access point. Identifying possible failure domains and keeping the size of failure domains small or manageable where possible, reduces the chance of widespread disruption.

Building redundancy within and/or across failure domains is an important method to help mitigate the risks of failure.

When designing a Nutanix deployment, you can take steps to mitigate risk for each of the following failure domains:

- Drives

- Nutanix node
- Nutanix block
- Management plane
- Nutanix cluster.
- Datacenter rack and server room
- Datacenter

Nutanix clusters are resilient to a drive, node, block, or rack failure, which is enabled by Redundancy Factor 2, the default. Redundancy Factor 3 can enable simultaneous drive, node, block, or rack failures with the right architecture. After drive, node, block or rack failure, a Nutanix cluster self-heals to reach the desired redundancy factor and rebuilds resilience for additional subsequent failures.

> Note: You can configure your Nutanix environment to be fault tolerant to node, block, and rack failures. This is described later in the section: Data Redundancy and Resiliency. Mitigating the risks of network failure domains is described in the section: Networking.

The Management Plane

One of the most important failure domains, and one that is often overlooked by architects, is the management plane. The more workload domains managed by a single management plane, the bigger the impact of a failure. When deploying the management plane, consider the following risk mitigations to reduce the impact of a failure.

Table 22: Management Plane Risk Mitigations

AREA	RISK MITIGATION
Availability	• Design and deploy management plane to be highly available. • At a minimum, design to meet the availability. requirements of the managed workload or service with the highest uptime requirement.

AREA	RISK MITIGATION
Limit the impact	• Confine workload domain to a single datacenter or site. • Confine workloads domain to a defined security zone. • Ensure the API gateway is always available because other 3rd party integrations may rely on it. (e.g. 3rd party backup vendor integration.).
Access Control	Configure built-in RBAC to restrict access to management platform resources.

> Note: The section Management Layer provides more details on management plane configuration.

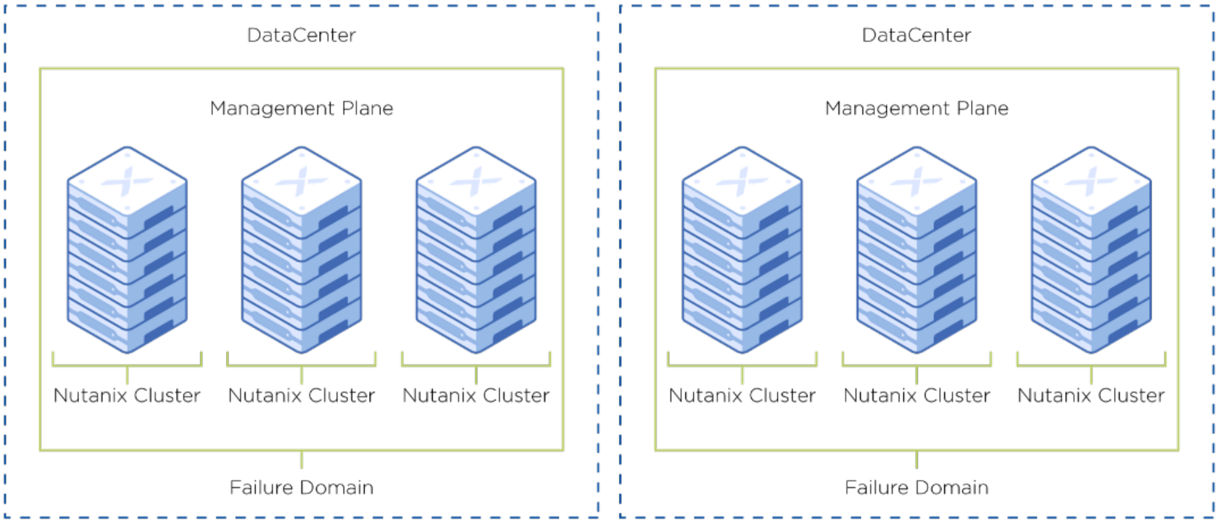


Figure 23: Management Plane Failure Domain

The Nutanix Cluster as a Failure Domain

When evaluating each Nutanix cluster as a failure domain, you have to consider the risks and potential effects in terms of both the size of the cluster (as described above in Choosing the Optimum Cluster Size) and the workloads running in the cluster, including whether the cluster is running mixed or dedicated workloads.

Large clusters result in larger failure domains and potentially higher business impacts, since they typically host considerably more workloads. To mitigate the risk of data unavailability or service disruption, design for redundancy at a cluster level to protect data and services, as described in the following table:

Table 23: Designing for Redundancy at the Cluster Level

AREA	RISK MITIGATION
Power	Redundant power from two different power supplies.
Networking	Redundant TOR switches. Redundant upstream connectivity to TOR switches from each Nutanix node.
Cluster	Leverage Nutanix scale-out architecture and data protection capabilities to replicate data to a second Nutanix cluster in the event that one cluster fails.
Application	Deploy application across multiple clusters.

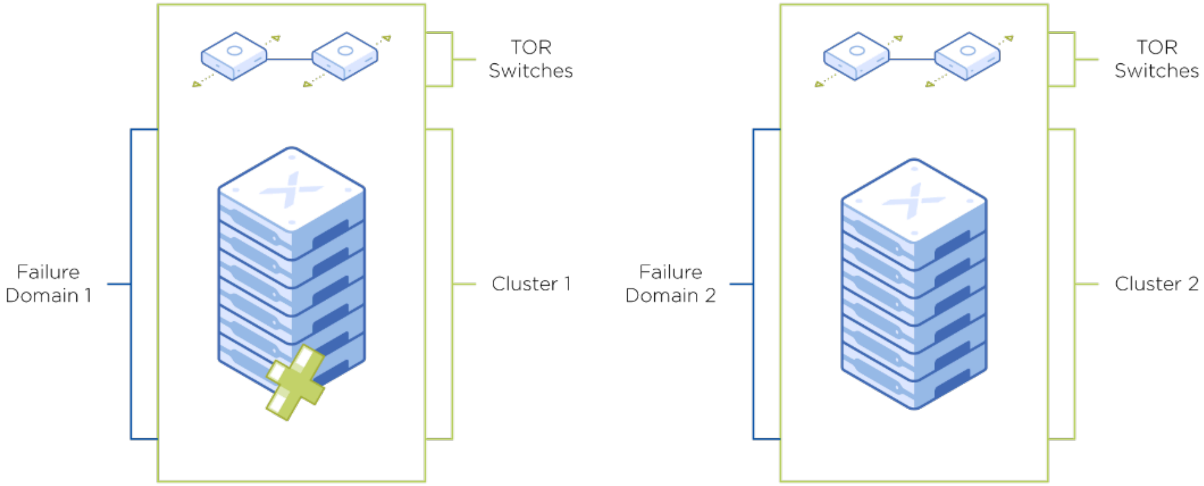


Figure 24: The Nutanix Cluster as a Failure Domain

Datacenter Rack and Server Room Failure Domains

When considering the datacenter rack failure domain, the primary mitigations are redundant power from two different power supplies to each rack, redundant TOR switches, and redundant network uplinks.

When considering the datacenter server room failure domain, it is critical to examine all datacenter components to ensure they are not shared among multiple server rooms, as described in the following table:

Table 24: Datacenter Rack and Server Room Risk Mitigation

AREA	RISK MITIGATION
Power	Redundant power
Cooling	Independent cooling
Server Room	Provide redundant server room within separate firezone (in the same datacenter)
Application	Place application in multiple server rooms. For example, active directory domain controllers in separate server rooms.

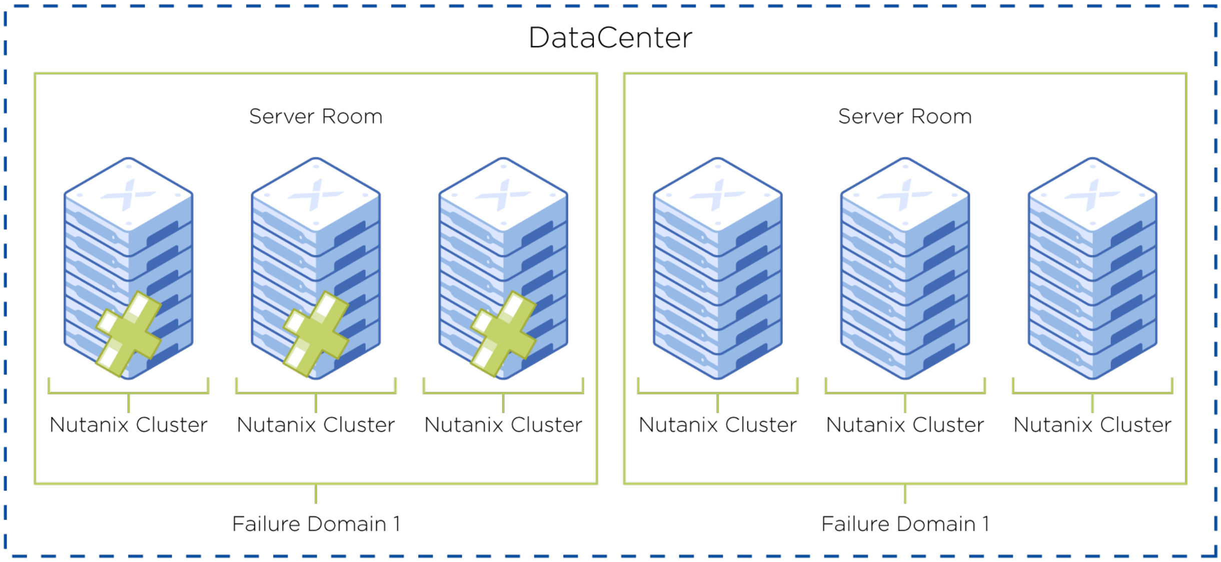


Figure 25: Datacenter Rack and Server Room Failure Domains

Datacenter Building

When considering the datacenter failure domain, it is critical to examine the redundancy of all connections to the outside to ensure they are not shared, as described in the following table:

Table 25: Datacenter Building Risk Mitigation

AREA	RISK MITIGATION
Power	Redundant power supply from different suppliers.
Cooling	Independent/redundant cooling system for each datacenter room.
Network Connectivity	Redundant networking connectivity between datacenter buildings from different providers Redundant internet connectivity from different ISP
Datacenter buildings/server room	Where possible, utilize multiple datacenter buildings and/or server rooms in separate datacenter buildings so a single server room does not affect all of production. Distribute multi-component services equally across datacenter server rooms.

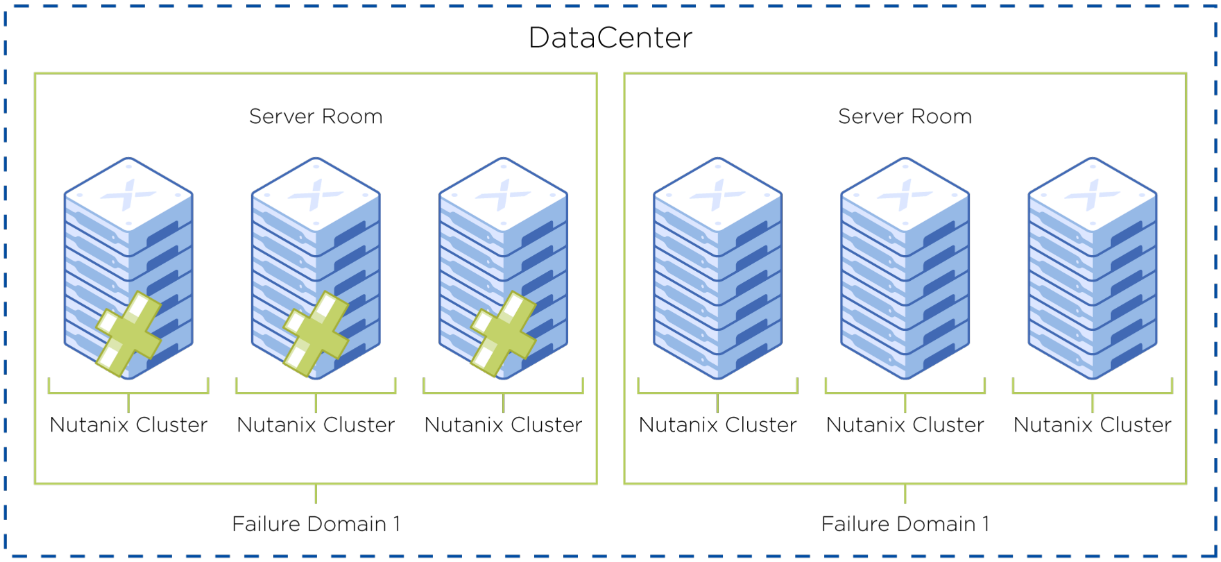


Figure 26: Datacenter Building Failure Domain

Designing Datacenter Locations

Based on typical enterprise requirements the logical implementation for DR across regions, availability zones and datacenters will use the following layout.

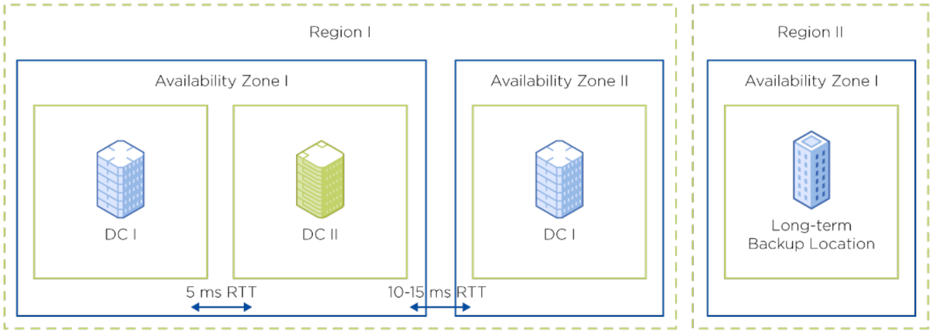


Figure 27: Designing Datacenter Locations

Regions

Table 26: Two region implementation for DR

REGION-001	TWO REGIONS ARE USED, SUCH AS: REGION 1 (PRIMARY REGION) RUNS ACTIVE WORKLOADS AND PROVIDES TWO AVAILABILITY ZONES FOR DISASTER RECOVERY AND REGION 2 (SECONDARY REGION) PROVIDES A LOCATION FOR LONG-TERM BACKUP RETENTION.
Justification	Regional separation of running workloads and disaster recovery capabilities plus long-term backup retention.
Implication	It may take a long time to restore services in or from Region 2 when Region 1 fails but the likelihood of an entire region failing is low.

Table 27: Single region implementation for DR

REGION-002	ONE REGION HOSTS PRIMARY AND DISASTER RECOVERY AVAILABILITY ZONES.
Justification	• Addresses the requirement to have active workloads and disaster recovery capabilities in a region separate from where the long-term backups are stored. • Services must be available in one region only based on end-user location • Staff only available in one region • Budget constraints • One region can host multiple datacenters within 100 km and provide RTT of maximum 5ms which will be required to meet infrastructure RPO 0. • One region can host datacenter separation of more than 200 KM which is required for disaster recovery purposes.

REGION-002	ONE REGION HOSTS PRIMARY AND DISASTER RECOVERY AVAILABILITY ZONES.
Implication	• If Region 1 fails, the services has to be rebuilt from backups from outside Region 1. • Staff has to travel to Region 2 to build up the services from backups, or backups must be relocated to Region 1 when Region 1 becomes available after the failure, and the services must be rebuilt/restored.

Table 28: Long term retention backups are placed in Region 2

REGION-003	LONG TERM RETENTION BACKUPS ARE PLACED IN REGION 2.
Justification	Addresses the requirement to have long-term backups stored in a separate region from active workloads and provides disaster recovery capabilities in case of regional disasters.
Implication	• If Region 1 fails, the services has to be rebuilt from backups from outside Region 1. • Staff has to travel to Region 2 to build up the services from backups, or backups must be relocated to Region 1 when Region 1 becomes available after the failure, and the services must be rebuilt/restored.

Availability Zones

Table 29: Three availability zones in two regions

AZ-001	THREE AVAILABILITY ZONES IN TWO REGIONS WILL BE USED, SUCH AS: DC1 AND DC2 IN REGION 1 RUNNING WORKLOADS, AZ2 IN REGION 1 FOR DISASTER RECOVERY PURPOSES, AND AZ1 IN REGION 2 WHERE LONG-TERM BACKUPS ARE STORED.
Justification	Required to provide the separation required

AZ-001	THREE AVAILABILITY ZONES IN TWO REGIONS WILL BE USED, SUCH AS: DC1 AND DC2 IN REGION 1 RUNNING WORKLOADS, AZ2 IN REGION 1 FOR DISASTER RECOVERY PURPOSES, AND AZ1 IN REGION 2 WHERE LONG-TERM BACKUPS ARE STORED.
Implication	• Many availability zones to manage. • Multiple network connections to manage between availability zones.

Table 30: Two availability zones in Region 1

AZ-002	TWO AVAILABILITY ZONES IN REGION 1 WILL HOST PRIMARY WORKLOADS.REGION 2 DR SITE WILL HOST ANY WITNESS FUNCTIONALITY REQUIRED FOR SERVICE AVAILABILITY.
Justification	• Addresses the requirement to have applications available across multiple availability zones during normal production • Distance is more than 100km and provides RTT of less than 5ms meaning it is technically possible to meet an RPO = 0 requirement.
Implication	Network connections between availability zones are critical for service availability and are also exposed to more risk compared to intra-availability-zone connections.

Table 31: One availability zone in Region 1 to provide disaster recovery

AZ-003	ONE AVAILABILITY ZONE IN REGION 1 WILL BE USED TO PROVIDE DISASTER RECOVERY CAPABILITIES.
Justification	Addresses the requirement to provide disaster recovery across availability zones within the same region.
Implication	No availability zone redundancy during a disaster recovery scenario.

Table 32: One availability zone in Region 2 hosts backups.

AZ-004	ONE AVAILABILITY ZONE IN REGION 2 HOSTS LONG-TERM BACKUPS.
Justification	Addresses the requirement to have long-term backups stored in a separate region from running workloads, and disaster recovery capabilities are provided.
Implication	Staff has to travel to Region 2 to build up the services from backups, or backups must be relocated to Region 1 when Region 1 becomes available after the failure, and the services must be rebuilt/restored.

Designing Workload Domains

This document uses workload domains as building blocks. Each workload domain consists of a set of Nutanix nodes that are managed by the same management instance (Nutanix Prism Central/VMware vCenter) and connected to the same network domain.

Workload Domain Architecture

A single workload domain can include different hardware and software combinations, have homogeneous ESXi or AHV nodes, or mix ESXi (running VMs) and AHV (not running VMs) in a single cluster. These can be configured to satisfy redundancy, performance, and capacity requirements.

Normally, single workload domains occupy a single rack. However, you can aggregate multiple workload domains in a single rack or span a single workload domain across multiple racks.

Workload Domain Rack Mapping

Mapping workload domains to datacenter racks is not one to one. While the workload domain is a repeatable building block, a rack is a unit of size. Workload domains and datacenter racks can have different characteristics; you map workload domains to datacenter racks according to the use case and rack specifications.

Nutanix Cluster Layout

Single Rack, Single Workload Domain

One workload domain can occupy a single datacenter rack. All nodes from a workload domain are connected to a single pair of TOR switches.

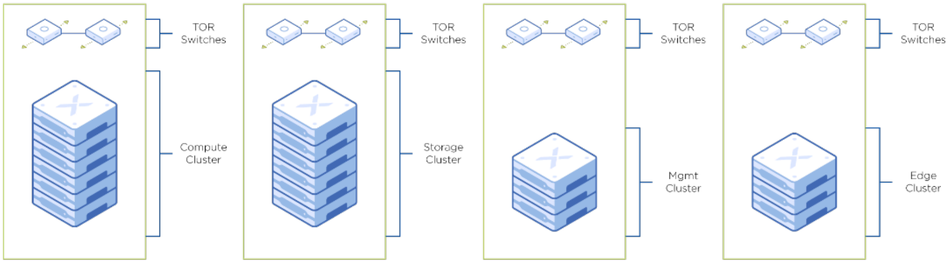


Figure 28: Single Rack Single Workload Domain

Single Rack, Multiple Workload Domains

One workload domain can occupy a single datacenter rack. All nodes from a workload domain are connected to a single pair of TOR switches.

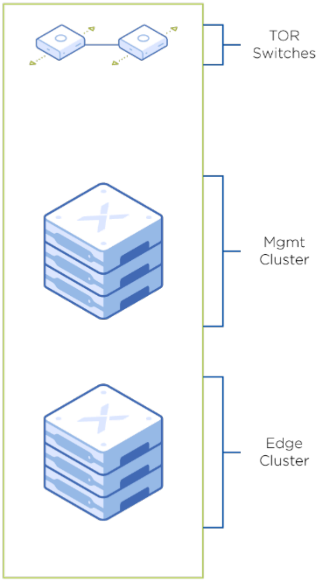


Figure 29: Single Rack, Multiple Workload Domains

Multiple Racks, Single Workload Domain

A single workload domain can span multiple racks. For example, to provide an additional level of data protection (using rack awareness/fault tolerance, see the section Data Redundancy and Resiliency) or if a single workload domain is bigger than a single rack can contain.

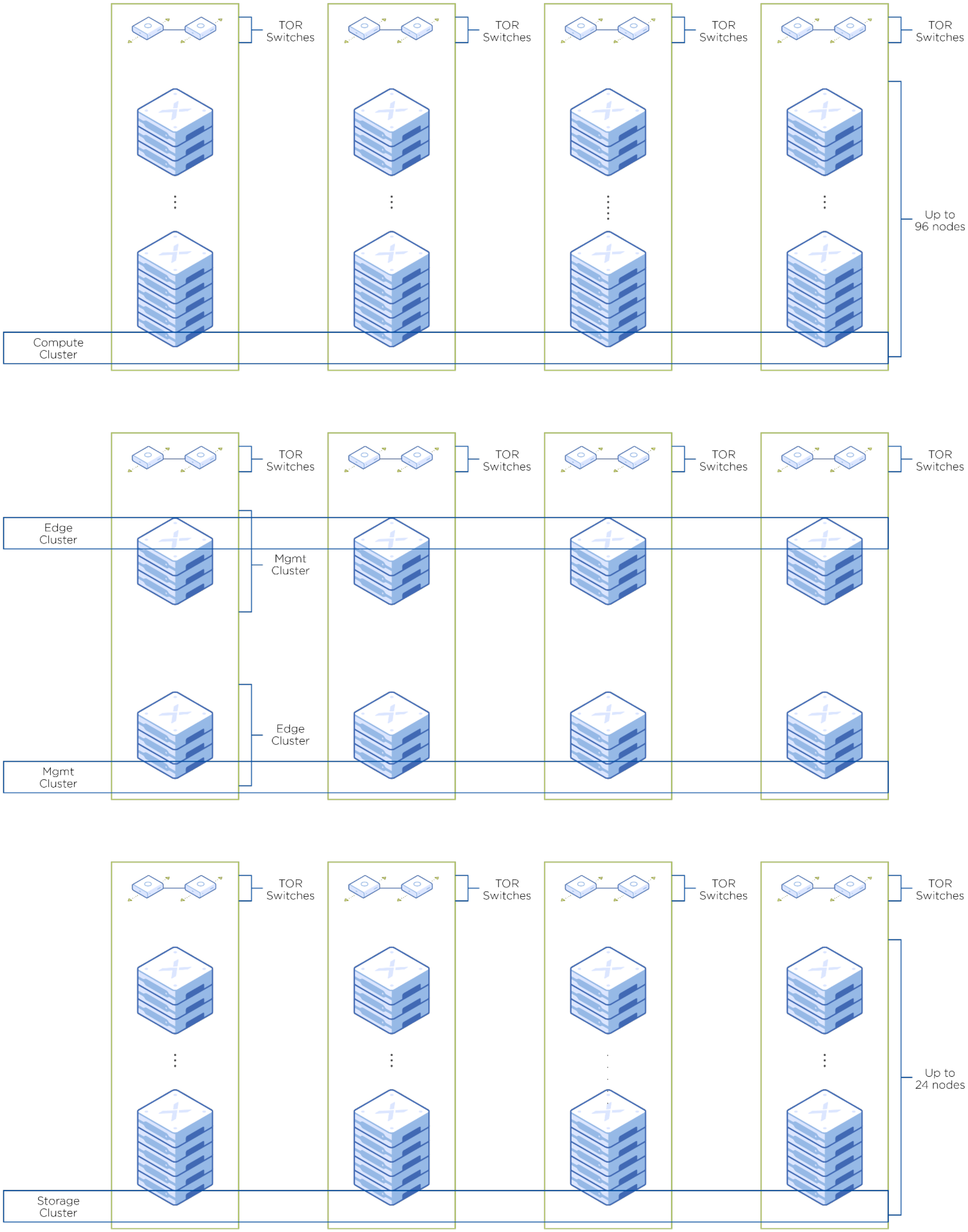


Figure 30: Multiple Racks, Single Workload Domain

Pros and Cons

A single workload domain can span multiple racks. For example, to provide an additional level of data protection (using rack awareness/fault tolerance, see the section:

Data Redundancy and Resiliency) or if a single workload domain is bigger than a single rack can contain.

Table 33: Pros and cons

	PROS	CONS
Single Rack/	• Default redundancy level. • Simple to consume.	May use rack space inefficiently.
Single Workload Domain	• Efficient space usage. • Default redundancy level. • Simple to consume.	Increases impact of rack failure.
Single Rack/ Multiple Workload Domains	• Efficient space usage. • Increases resiliency for each workload domain. • Decreases impact of rack failure.	Complexity.

Table 34: Workload Domains Spanning Single or Multiple Racks

PFM-007	DECIDE WHICH WORKLOAD DOMAINS WILL SPAN A SINGLE OR MULTIPLE RACKS
Justification	
Implication	

Networking

Well-designed networks are critical to a Nutanix deployment's resilience and performance.

A Nutanix cluster can tolerate multiple simultaneous failures because it maintains a set redundancy factor and offers features such as block and rack awareness. However, this level of resilience requires a highly available, redundant network connecting a cluster's nodes. Protecting the cluster's read and write storage capabilities also requires highly available connectivity between nodes. Even with intelligent data placement, if network connectivity between more than the allowed number of nodes breaks down, VMs on the cluster could experience write failures and enter read-only mode.

To optimize I/O speed, Nutanix clusters choose to send each write to another node in the cluster. As a result, a fully populated 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.

Physical Switches

A Nutanix environment should use datacenter switches designed to handle high- bandwidth server and storage traffic at low latency. Do not use switches meant for deployment at the campus access layer. Campus access switches may have 10 Gbps ports like datacenter switches, but they are not usually designed to transport a large amount of bidirectional storage replication traffic. Refer to the Nutanix physical networking best practices guide for more information.

The deployment size and purpose also influence physical switch choice. Datacenter switches with large buffers are critical in a large AOS cluster that is going to grow beyond eight nodes or host storage-intensive applications. In smaller clusters or ROBO deployments that have fewer than eight nodes or do not host write-intensive applications, the switch may not experience buffer contention, and you can relax these switch restrictions. There are also some switch types you should never use for any Nutanix deployment because of oversubscription or other architecture choices; we list examples of these as well.

Datacenter switches should have the following characteristics:

- Line rate: Ensures that all ports can simultaneously achieve advertised throughput.
- Low latency: Minimizes port-to-port latency as measured in microseconds or nanoseconds.
- Large per-port buffers: Accommodates speed mismatches from uplinks without dropping frames.
- Nonblocking, with low or no oversubscription: Reduces chance of drops during peak traffic periods.
- 10 Gbps or faster links for Nutanix CVM traffic: Only use 1 Gbps links for additional user VM traffic or when 10 Gbps connections are not available, such as in a ROBO deployment. Limit Nutanix clusters using 1 Gbps links to eight nodes maximum.

Switch manufacturers' datasheets, specifications, and white papers can help identify these characteristics. For example, a common datacenter switch datasheet may show a per-port buffer of 1 MB, while an access layer or fabric extension device has a per-port buffer of around 150 KB. During periods of high traffic, or when using links with a speed mismatch (such as 40 Gbps uplinks to 10 Gbps edge ports), a smaller buffer can lead to frame drops, increasing storage latency.

The following table is not exhaustive, but it includes examples of model lines that meet the above requirements for high-performance or large clusters. Models similar to the ones shown are also generally good choices.

Table 35: Examples of Recommended Switch Models

EXAMPLES OF HIGH PERFORMANCE SWITCH MODELS		
Arista 7050X3	Arista 7160	Arista 7170
Arista 7280	Cisco Nexus 9000	Cisco Nexus 7000

Cisco Nexus 7000	Dell S5200-ON	HPE FM3810
HPE FM3132Q	Juniper QFX-5100	Lenovo NE2580O
Mellanox SN2010	Mellanox SN2100	Mellanox SN2410

The following are examples of switches that do not meet high-performance datacenter switch requirements but are acceptable for ROBO clusters and clusters with fewer than eight nodes or low performance requirements:

Table 36: Examples of Robo Or Smb Switch Models

EXAMPLES OF ROBO OR SMB SWITCH MODELS		
Arista 7050	Arista 7150S	Cisco Nexus 3000
Cisco Catalyst 9000	Cisco Catalyst 3000	HPE FM2072

The following are examples of switches that are never acceptable for any Nutanix deployment:

Table 37: Switches that are not acceptable

SWITCH	REASON NOT RECOMMENDED
Cisco Nexus 2000 (Fabric Extender)	Highly oversubscribed with small per-port buffers.
10 Gbps expansion cards in a 1 Gbps access switch	10 Gbps expansion cards provide uplink bandwidth for the switch, not server connectivity.

Each Nutanix node also has an out-of-band connection for IPMI, iLO, iDRAC, or similar management. Because out-of-band connections do not have the same latency or throughput requirements of VM or storage networking, they can use an access layer switch.

> Note: Nutanix recommends an out-of-band management switch network separate from the primary network to ensure management availability. Configure server-facing ports in the management network as access ports and do not use VLAN trunking for these ports. Access to this critical management network should be restricted.

Table 38: Using a Large Buffer Datacenter Switch at 10GBPS or Faster

NET-001	USE A LARGE BUFFER DATACENTER SWITCH AT 10GBPS OR FASTER
Justification	Achieves high performance for critical storage and VM traffic converged on the same network fabric.
Implication	Requires data center switches that may be more expensive than campus or access layer switches.

Network Topology

In a greenfield environment, Nutanix recommends a leaf-spine network topology because it is easy to scale, achieves high performance with low latency, and provides resilience. A leaf-spine topology requires at least two spine switches and two leaf switches. Every leaf connects to every spine using uplink ports.

There are no connections between the spine switches or between the leaf switches in a conventional leaf-spine design. To form a Nutanix cluster it's critical that all nodes are in the same broadcast domain, thus in any leaf-spine design all leaf switches connecting nodes in a cluster should carry the Nutanix VLAN. This can be accomplished with physical connections between switches, using an overlay network, or using a pure layer 2 design. The following example shows a pure layer 2 design or overlay.

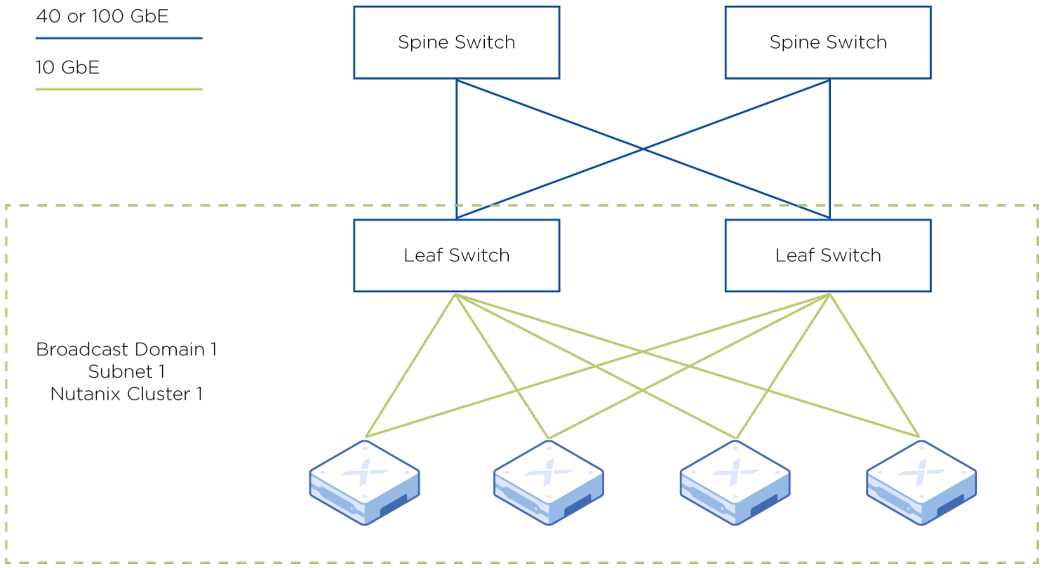


Figure 31: Pure Layer 2 Design

You may also choose a leaf-spine topology that uses links between switches to guarantee layer 2 connectivity between Nutanix nodes.

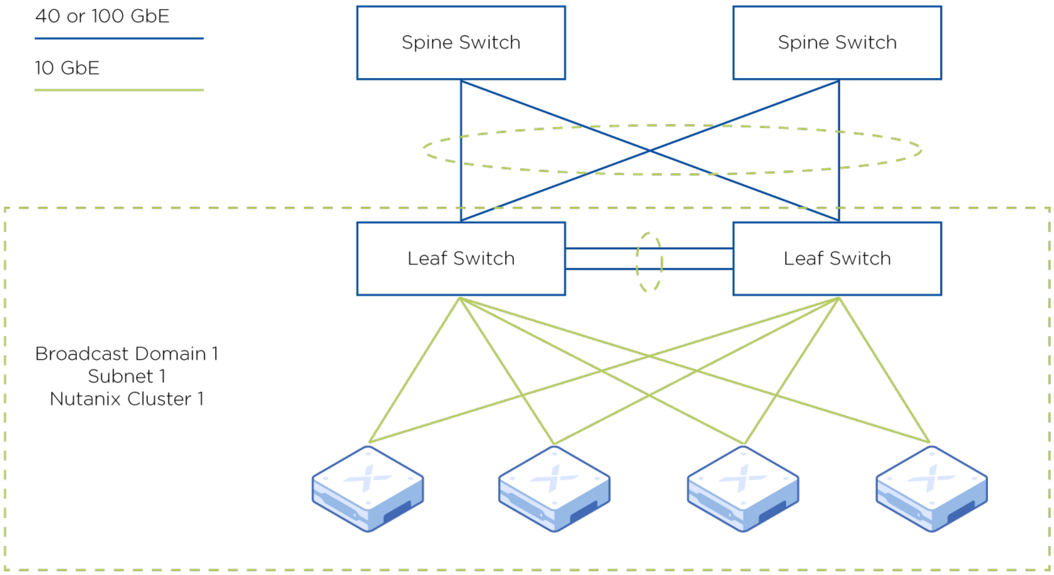


Figure 32: Leaf Spine Topology

Use uplinks that are a higher speed than the edge ports to reduce uplink oversubscription. To increase uplink capacity, add spine switches or uplink ports as needed.

The core-aggregation-access (or multi-tier) network design is a modular layout that allows you to upgrade and scale layers independently. Nutanix clusters perform well in the core-aggregation-access topology, but extra caution should be taken around scaling the Nutanix cluster.

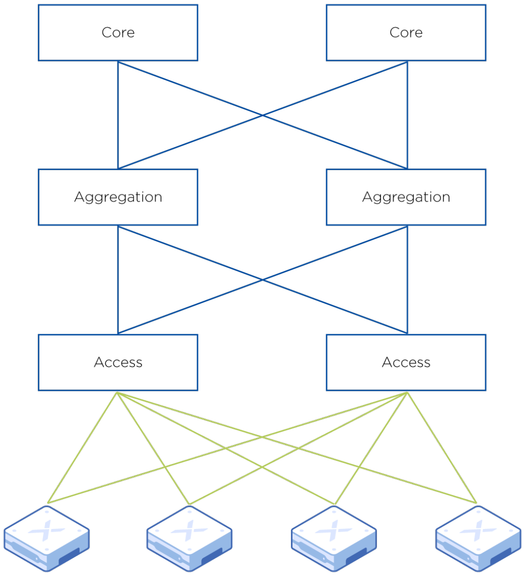


Figure 33: Core Aggregation Access Design

In pre-existing environments, you may not have full control over the network topology, but your design should meet the following requirements:

Guidelines:

- Networks must be highly available and tolerate individual device failures.
- Ensure that each layer of the network topology can tolerate device failure.
- Avoid configurations or technologies that do not maintain system availability during single device outages or upgrades such as stacked switches.
- Ensure that there are no more than three switches between any two Nutanix nodes in the same cluster. Nutanix nodes send storage replication traffic to

each other in a distributed fashion over the top-of-rack network. One Nutanix node can therefore send replication traffic to any other Nutanix node in the cluster.

- The network should provide low and predictable latency for this traffic. Leaf-spine networks meet this requirement by design. For the core-aggregation-access model, ensure that all nodes in a Nutanix cluster share the same aggregation layer to meet the three-switch-hop rule.

Oversubscription occurs when an intermediate network device or link does not have enough capacity to allow line rate communication between the systems connected to it. For example, if a 10 Gbps link connects two switches and four hosts connect to each switch at 10 Gbps, the connecting link is oversubscribed. Oversubscription is often expressed as a ratio—in this case 4:1, as the environment could potentially attempt to transmit 40 Gbps between the switches with only 10 Gbps available. Achieving a ratio of 1:1 is not always feasible.

Recommendation 1

- Keep the oversubscription ratio as small as possible based on budget and available capacity.

In a typical deployment where Nutanix nodes connect to redundant top-of-rack switches, storage replication traffic between CVMs traverses multiple devices.

Recommendation 2

- To avoid packet loss due to link oversubscription, ensure that the switch uplinks consist of multiple interfaces operating at a faster speed than the Nutanix host interfaces. For example, for nodes connected at 10 Gbps, interswitch connections should consist of multiple 10, 40, or 100 Gbps links.
- Connect all Nutanix nodes that form a cluster to the same switch fabric. Do not stretch a single Nutanix cluster across multiple, disconnected switch fabrics. A switch fabric is a single leaf-spine topology or all switches connected to the same switch aggregation layer. Every Nutanix node in a cluster should therefore be in the same L2 broadcast domain and share the same IP subnet.

See section Security > Network Segmentation for service placement information and design decisions.

Recommendation 3

- Use native, or untagged, VLANs for the hypervisor host and CVM for ease of initial configuration. Ensure that this untagged traffic is mapped into the CVM and hypervisor VLAN only on the required switch ports to reduce risk.
- Use tagged VLANs for all guest VM traffic and add the required guest VM VLANs to all connected switch ports for hosts in the Nutanix cluster.
- Limit guest VLANs for guest VM traffic to the smallest number of physical switches and switch ports possible to reduce broadcast network traffic load.

Table 39: Using a Leaf-Spine Network Topology for New Environments

NET-002	USE A LEAF-SPINE NETWORK TOPOLOGY FOR NEW ENVIRONMENTS.
Justification	Achieves high performance for critical storage and VM traffic and is easy to scale.
Implication	Requires more network connections between switches and may require new network design.

Table 40: Populate Each Rack with Two 10GBE or Faster TOR Switches

NET-003	POPULATE EACH RACK WITH TWO 10GBE OR FASTER TOR SWITCHES
Justification	Simplifies the design, follows the leaf/spine model, and provides high performance and high availability to the network.
Implication	Increases rack space requirements, costs, and the number of leafs.

Table 41: Avoid Switch Stacking to Ensure Network Availability During Individual Device Failure

NET-004	AVOID SWITCH STACKING TO ENSURE NETWORK AVAILABILITY DURING INDIVIDUAL DEVICE FAILURE.

Justification	Critical network infrastructure must be redundant between all CVMs in the same Nutanix cluster.
Implication	More network devices may be required to provide high availability and reduce single points of failure.

Table 42: Use No More Than Three Switches Between Two Nutanix Nodes in the Same Cluster

NET-005	ENSURE THAT THERE ARE NO MORE THAN THREE SWITCHES BETWEEN ANY TWO NUTANIX NODES IN THE SAME CLUSTER.
Justification	Storage latency is directly related to network latency, so the network distance between nodes in the same cluster must be reduced.
Implication	A Nutanix cluster cannot span multiple sites or switch fabrics, instead use data replication technologies instead to provide high availability between sites.

Table 43: Reduce Network Oversubscription to Achieve a Near-1:1 Ratio

NET-006	REDUCE NETWORK OVERSUBSCRIPTION TO ACHIEVE AS CLOSE TO A 1:1 RATIO AS POSSIBLE.
Justification	Dropped network packets or a congested network will immediately impact storage performance and must be avoided in the design phase.
Implication	Higher bandwidth paths must be created using more uplink paths of faster speed between switches.

Table 44: Configure CVM and Hypervisor VLAN on Server-facing Switch Ports

NET-007	CONFIGURE THE CVM AND HYPERVISOR VLAN AS NATIVE, OR UNTAGGED ON SERVER FACING SWITCH PORTS.
Justification	Newly added nodes use untagged traffic for discovery and will work out of the box, reducing manual server configuration.
Implication	Network switches must be configured to accept untagged frames on ports facing Nutanix servers and place the traffic into the CVM and hypervisor VLAN.

Table 45: Use Tagged VLANs on Switch Ports for All Guest Workloads

NET-008	USE TAGGED VLANS ON THE SWITCH PORTS FOR ALL GUEST WORKLOADS.
Justification	Workloads should be separated from each other and from the CVM and hypervisor network using VLANs.
Implication	Multiple VLANs and IP subnets are required.

Broadcast Domains

Performing layer 3 routing at the top of rack, creating a layer 3 and layer 2 domain boundary within a single rack, is a growing network trend. Each rack is a different IP subnet and a different layer 2 broadcast domain. This layer 3 design decreases the size of the layer 2 broadcast domain to remove some common problems of sharing a large broadcast domain among many racks of servers, but it can add complexity for applications that require layer 2 connectivity.

In contrast, the traditional layer 2 design shares a single broadcast domain or VLAN among many racks of switches. In the layer 2 design, a switch and server in one rack share the same VLANs as a switch and server in another rack. Routing between IP subnets is performed either in the spine or in the aggregation layer. The endpoints in the same switch fabric have layer 2 connectivity without going through a router, but this can increase the number of endpoints that share a noisy broadcast domain.

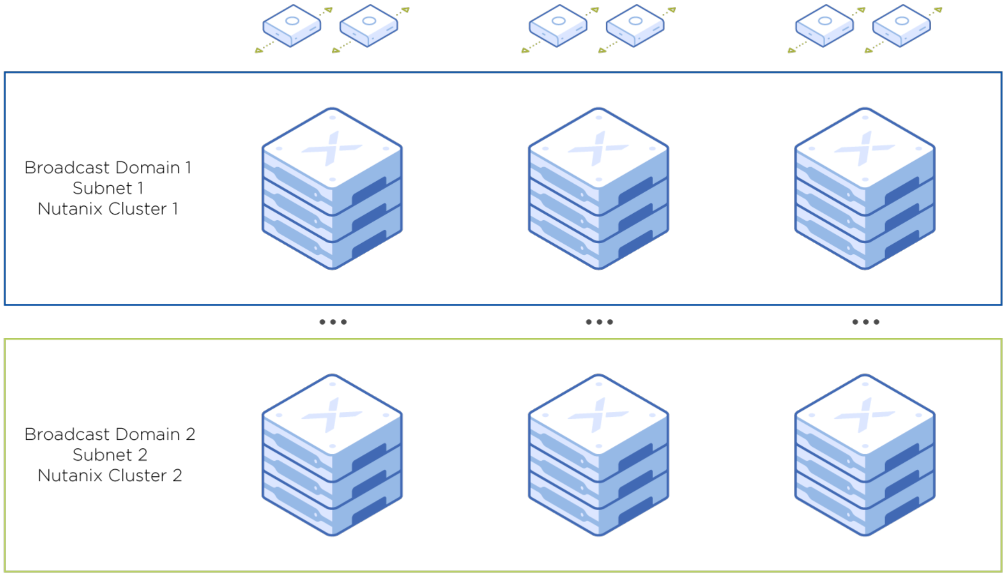


Figure 34: Layer 2 Network Design

Nutanix recommends a traditional layer 2 network design to ensure that CVMs and hosts can communicate in the same broadcast domain even if they are in separate racks. The CVM and host must be in the same broadcast domain and IP subnet.

If a layer 3 network design is chosen, there are two possible ways to make this work with a Nutanix deployment:

- Keep all Nutanix nodes in a cluster inside the same rack.
- Create an overlay network in the switch fabric that creates a virtual broadcast domain that is shared across racks for the Nutanix CVMs and hosts.

Keeping all of the Nutanix nodes in a single cluster in a single rack has the limitation of losing resilience against rack failure. Creating an overlay in the network fabric:

- Requires special configuration.
- May require specific network hardware to achieve.
- Increases configuration complexity.

Spreading a broadcast domain across racks virtually may also raise concerns that the original intent of the layer 3 design, reducing broadcast domain scope, is now violated.

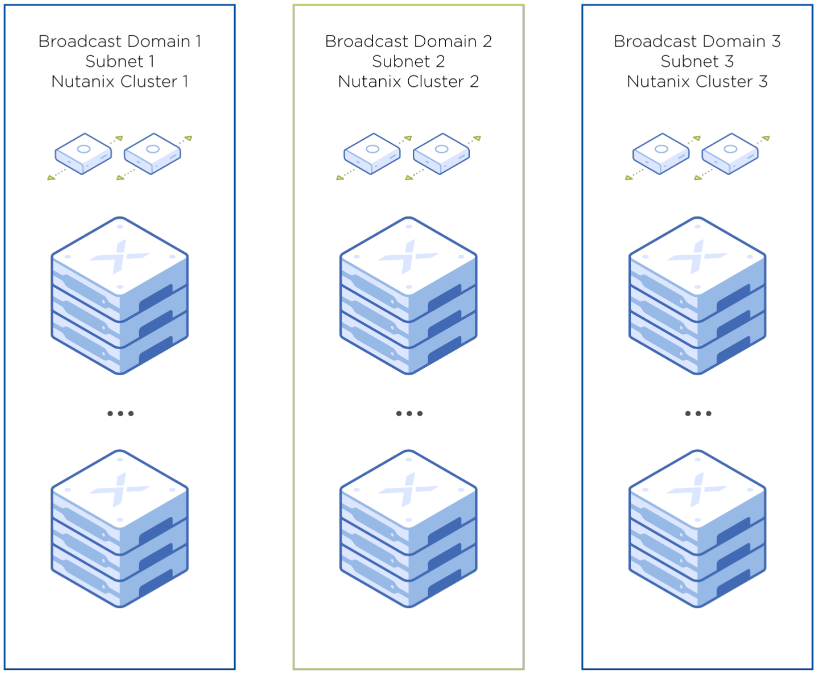


Figure 35: Layer 3 Network Design

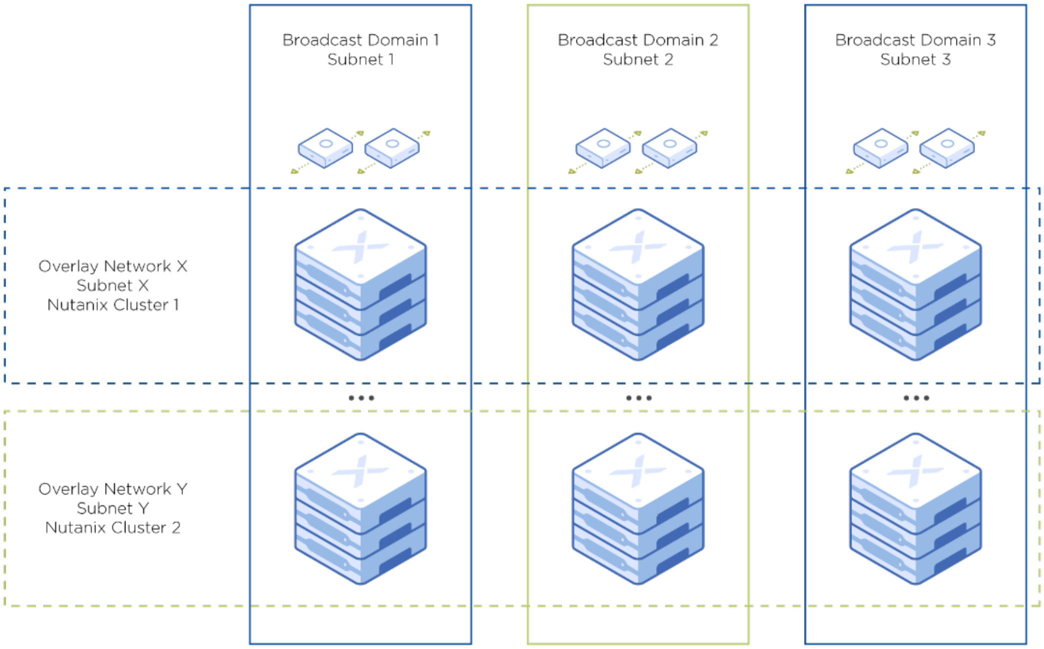


Figure 36: Layer 3 Network Design with Overlay

Table 46: Use a Layer 2 Network Design

NET-009	USE A LAYER 2 NETWORK DESIGN
Justification	Nutanix CVMs and hypervisor hosts must be in the same broadcast domain to function properly after a CVM becomes unavailable.
Implication	A broadcast domain failure or storm, can span multiple racks.

Scaling the Network

To scale the leaf-spine network:

- Add two leaf switches and a management switch for every rack.
- Add spine switches as needed.

Because there are no more than three switch hops between any two Nutanix nodes in this design, a Nutanix cluster can easily span multiple racks and still use the same switch fabric. Use a network design that brings the layer 2 broadcast domain to every switch that connects Nutanix nodes in the same cluster.

The following example shows a leaf-spine network using either an overlay or a pure layer 2 design, with no connections are required between leaf switches.

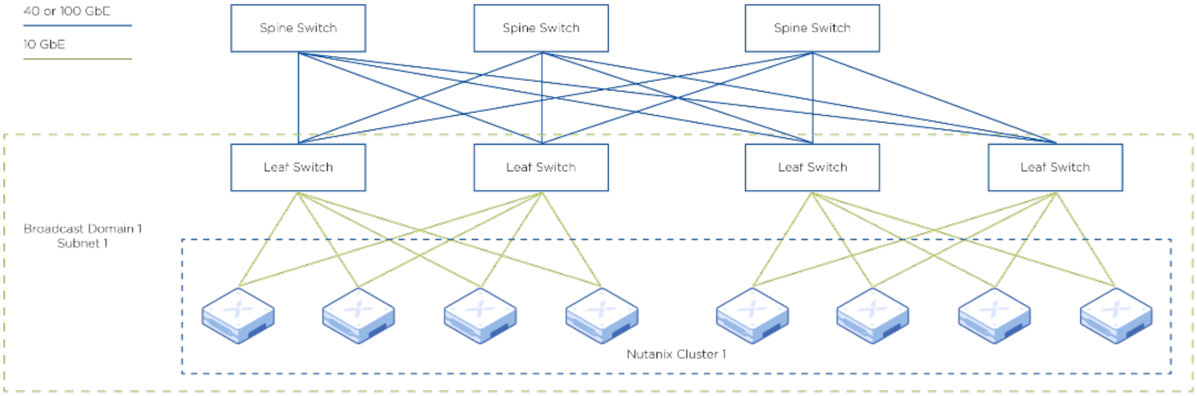


Figure 37: Logical View of Leaf Spine Scale

You may choose a leaf-spine design where a pair of leaf switches are connected using link aggregation. Discuss scale plans with the switch vendor when scaling beyond two spine switches in this design.

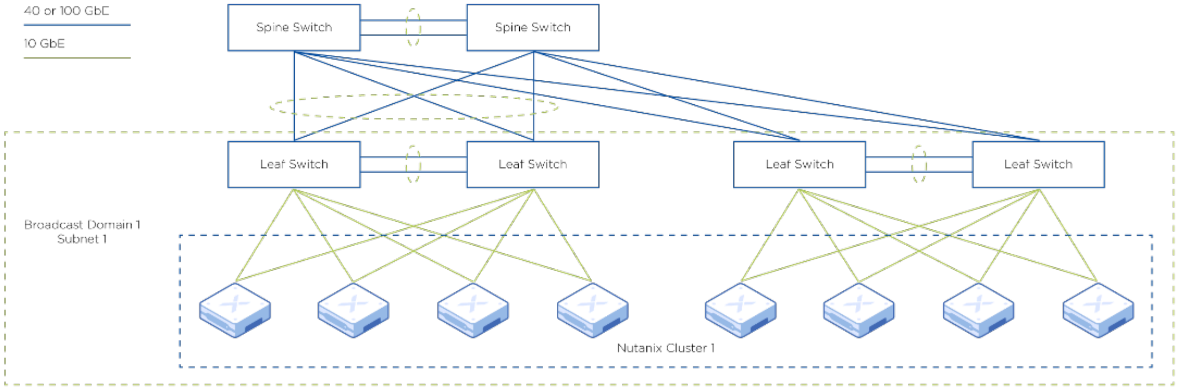


Figure 38: Leaf Switches Connected Using Link Aggregation

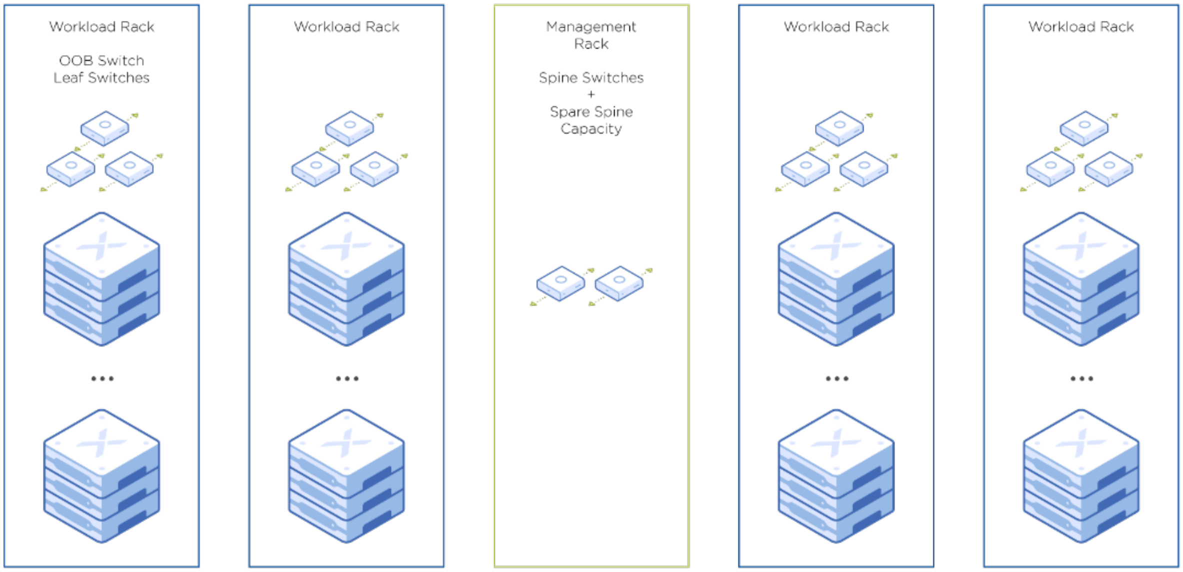


Figure 39: Rack Level View of Leaf Spine

Determining when to add new spine switches is a function of:

- How many ports the spine switch has.
- How many leaf switches (or racks) there are.
- How many connections are formed between each leaf and spine.

It's also important to take into account the throughput capacity of the switch and the throughput required between each rack.

In the previous example with 4 workload racks, assume you have 2 x 32-port spine switches with a single 100 Gbps connection to every leaf. Also assume each TOR leaf switch has 48 x 10 Gbps ports and 6 x 100 Gbps ports.

Spine switch port utilization is at 8 out of 32 ports on each spine, leaving capacity to grow up to 16 racks at each spine (32 spine ports divided by 2 ports per rack).

Each leaf would be using 1 out of the 6 available uplinks and each leaf can support up to 48 connected servers. However, at this point these two spine switches will be processing a lot of traffic and may be overloaded depending on their specifications.

Calculating oversubscription ratios, if you assume that each rack has 24 x 10Gbps dual-connected servers with dual leaf switches, for a total of 240 Gbps bandwidth capacity at each leaf, oversubscription is 2.4:1. This design is oversubscribed and would not be recommended.

One way to reduce oversubscription is to add another 100Gbps uplink from each leaf to the existing spine switches. However, that reduces the total number of supported racks by half to 8, (32 spine ports divided by 4 ports per rack), so you must carefully consider which leaf and spine switches are selected and what scale you would like to achieve.

Another way to grow spine capacity is to add a spine switch, which would then require another uplink from every leaf, and greatly increase the total throughput of the switch fabric without reducing the number of racks supported. If you added another spine switch to bring the total to 3 spines, each leaf switch would add another 100Gbps uplink (reducing oversubscription), but each spine could still support 16 racks (32 spine ports divided by 2 ports per rack). With 6 uplinks on every leaf switch, you can support designs with up to 6 spines in this example.

Scaling the multi-tier network design may require adding another aggregation and access layer to the core. In this case, there would be more than three switch hops between the two access layers.

> Note: The guideline here is to ensure that you add Nutanix nodes in separate aggregation and access layers to separate clusters to keep the number of switch hops between nodes in the same cluster to three or fewer.

In the following figure, Cluster 1 connects to one aggregation layer and Cluster 2 connects to another.

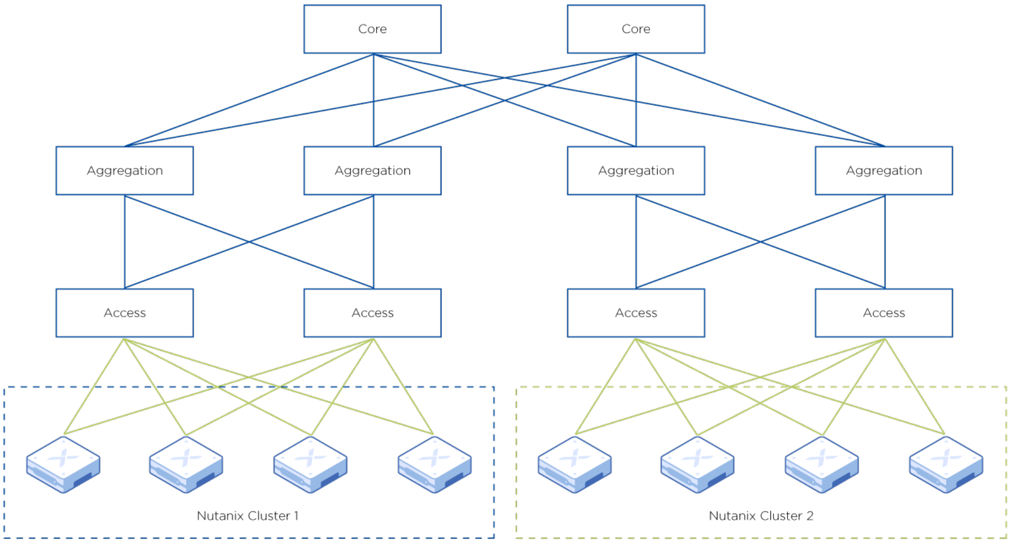


Figure 40: Clusters Connecting to Different Aggregation Layers

Host Networking

AHV Networking

AHV leverages Open vSwitch (OVS) for all VM networking. The virtual switch is referred to as a bridge and uses a br prefix in the name. Read the AHV Networking section on the Nutanix Support Portal for in depth guidance on any settings not covered here.

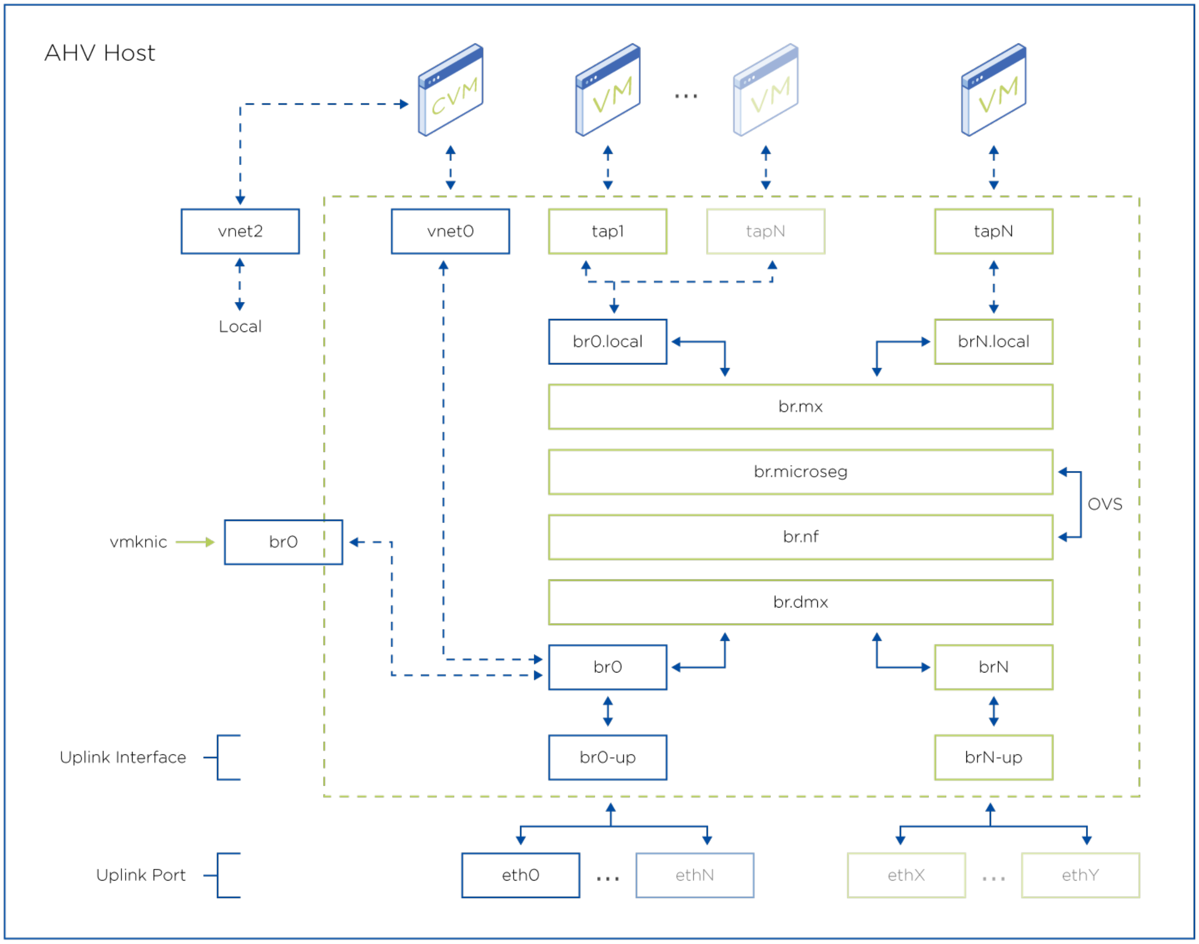


Figure 41: AHV Networking

Bridge

The default bridge, br0, behaves like a layer 2 learning switch that maintains a MAC address table. Additional uplink interfaces for separate physical networks are added as brN, where N is the new bridge number. To ensure that all VM traffic goes through the same set of firewall rules and the same set of network functions, a bridge chain is created with all microsegmentation and network function rules being placed in br. microseg and br.nf respectively.

Traffic from VMs enters the bridge at brN.local. Next, traffic is multiplexed onto the bridge chain, goes through the firewall and network function ruleset, and is then demultiplexed on the correct brN for physical network forwarding. The reverse path is followed for traffic flowing from the physical network to VMs.

All traffic between VMs must flow through the entire bridge chain, since brN is the only bridge that performs switching. All other bridges simply apply rules or pass traffic up and down the bridge chain.

Recommendation

- Use only the default bridge, br0, with the two fastest network adapters in the host.
- Converge the management, storage, and workload traffic on this single pair of uplink adapters.
- Additional brN bridges should only be added when connection to a separate physical network is required. For example, if the top-of-rack has two pairs of switches, one pair for storage and management, and another pair for workload traffic, it makes sense to create another bridge, br1 and place the workloads on this bridge.
- Other valid use cases include separate physical networks for iSCSI Volumes workloads, backplane intracluster replication, or workloads like virtual firewalls that require access to a physically separate network. In these cases, the additional bridge is used to connect to the separate top-of-rack network. If there is only a single top-of-rack physical network, then a single bridge with VLAN separation is sufficient.

AHV also includes a Linux bridge called virbr0. The virbr0 Linux bridge carries management traffic between the CVM and AHV host. All other storage, host, and workload network traffic flows through the br0 OVS bridge, or additional brN bridges if configured.

> Note: Do not modify the configuration of any bridges inside the AHV host unless following an official Nutanix guide.

Bond

Bonded ports aggregate the physical interfaces on the AHV host. By default, a bond named br0-up is created in bridge br0. After the node imaging process, all interfaces are placed within a single bond, which is a requirement for the foundation imaging process. A bridge can have only a single bond.

Bonds allow for several load-balancing modes, including active-backup, balance-slb and balance-tcp. Link Aggregation Control Protocol (LACP) can also be activated for a bond for link aggregation. The “bond_mode” setting is not specified during installation and therefore defaults to active-backup, which is the default configuration.

Recommendation

Keep the following recommendations in mind for all bond scenarios to prevent undesired behavior and maintain NIC compatibility.

- Ensure that each bond has at least two uplinks for redundancy.
- Do not mix NIC models from different vendors in the same bond.
- Do not mix NICs of different speeds in the same bond.

Uplink Load Balancing

The following bond modes are available:

- active-backup: Default configuration which transmits all traffic over a single active adapter. If the active adapter becomes unavailable, another adapter in the bond will become active. Limits host throughput to the bandwidth of a single network adapter. No additional switch configuration required.
- balance-slb: Distributes VM NICs across adapters in the bond and periodically rebalances for even uplink utilization. Limits VM per-nic throughput to a single network adapter, but allows utilization of multiple host physical interfaces for multiple VM NICs. NOTE: Not recommended due to negative interaction with default IGMP snooping and pruning network settings for multicast.
- balance-tcp with LACP: Distributes VM NIC TCP or UDP session across adapters in the bond. Limits per-nic throughput to the maximum bond bandwidth (number of physical uplink adapters * speed). Requires physical switch link aggregation. Used when LACP negotiation is required by the datacenter team or throughput requirements exceed a single NIC.

Recommendation

- To keep network configuration simple, use the standard 1,500 byte MTU in the hosts, CVMs, and workload VMs. Nutanix does not recommend jumbo frames unless specifically required by high-performance Nutanix Volumes iSCSI workloads or specific workload requirements.

- When switching from 1,500-byte frames to 9,000-byte frames, performance improvements are generally not significant unless the workload uses the maximum network bandwidth for read traffic.

For more information on when to use jumbo frames, see the Nutanix Volumes Guide and AHV Networking guides.

Table 47: Connecting NICs to TOR Switches

NET-010	CONNECT AT LEAST ONE 10 GBE OR FASTER NIC TO EACH TOR SWITCH
Justification	Maintains high availability in the event of the loss of one switch.
Implication	Requires at least two TOR switches, which will impact cost and rack space. Increases total bandwidth to each host to 20 Gbps.

Table 48: Using Bridges with Uplinks

NET-011	USE A SINGLE BR0 BRIDGE WITH AT LEAST TWO OF THE FASTEST UPLINKS OF THE SAME SPEED UNLESS MULTIPLE PHYSICAL NETWORKS ARE REQUIRED.
Justification	Simplifies the design, requiring fewer physical switches and switch ports, while relying on high bandwidth of 10Gbps and faster adapters.
Implication	All traffic is sent over a single virtual switch to a single pair of TOR switches.

Table 49: Using NICs of the same vendor within a bond

NET-012	USE NICs OF THE SAME VENDOR WITHIN A BOND
Justification	Ensures NIC compatibility within the bond and prevents undesired failover behavior.
Implication	NIC vendor choice must be considered when initially purchasing or adding NICs.

Table 50: Using VLANs to separate logical networks

NET-013	USE VLANS TO SEPARATE LOGICAL NETWORKS
Justification	Physical hosts have a limited number of network ports and each port adds complexity. Traffic separation can be logically accomplished without numerous physical ports.
Implication	Enables multiple networks to be used at once without additional hardware, saving on costs, and simplifying the design.

Table 51: Load Balancing

NET-014	USE ACTIVE-BACKUP UPLINK LOAD BALANCING
Justification	Simplifies the design and does not need any additional configuration.
Implication	All traffic is transmitted over one adapter, limiting bandwidth to at least 10 Gbps.

Table 52: Using Standard 1500 Byte MTU

NET-015	USE STANDARD 1,500 BYTE MTU AND DO NOT USE JUMBO FRAMES.
Justification	Simplifies the design and does not need any additional configuration. Reduces risk and complexity from non-standard configuration that only increases performance in certain use cases.
Implication	All network frames are limited to the default 1,500-byte maximum size for interoperability, potentially creating more network overhead for high throughput write workloads.

vSphere Networking

VMware vSphere networking follows many of the same design decisions as AHV networking. The critical design choices for vSphere networking are covered here. See the VMware vSphere Networking Guide for more details.

Nutanix hosts with vSphere ESXi use two virtual switches (vSwitches), named vSwitchNutanix and vSwitch0. vSwitchNutanix is the internal, standard vSwitch used for management and storage traffic between the CVM and the hypervisor.

Recommendations

- Do not modify vSwitchNutanix. vSwitch0 is also a standard vSwitch by default, used for communication between CVMs as well as workload traffic.
- Convert vSwitch0 to the distributed vSwitch following KB 4552 for Converting to the distributed vSwitch allows central management of networking for all hosts, instead of host by host networking configuration. The distributed vSwitch also adds capability for advanced networking functions such as load based teaming, LACP, and traffic shaping.
- Connect at least two of the fastest adapters of the same speed into vSwitch0 and use the "Route Based on Physical NIC Load" load balancing method to ensure traffic is balanced between uplink adapters.
- Connect these adapters to two separate top-of-rack switches to ensure redundancy.
- Do not add more vSwitches unless a connection to another physical network is needed to meet security or workload requirements.
- All CVM storage, hypervisor host, and workload traffic should flow through vSwitch0 using VLAN separate between the workload traffic and all other traffic.
- Use the default 1,500 byte frame size on all uplinks unless there is a specific performance or application requirement that would justify 9,000 byte jumbo frames.

Table 53: Use Virtual Distributed Switch (VDS)

NET-016	USE VIRTUAL DISTRIBUTED SWITCH (VDS)

Justification	Reduces the amount of configuration required in the environment. Makes port group provisioning faster and less error prone. Provides additional features such as LACP and bandwidth shaping. Important: Leave the internal vNetwork Standard Switch used by ESXi host and CVM for internal communication in place, including security settings.
Implication	Ability to configure port groups and other features at the cluster switch level vs. individual nodes. Reliance on vCenter for host and VM network configuration management.

Note: Leave the internal vNetwork Standard Switch used by ESXi host and CVM for internal communication in place, including security settings.

Table 54: Connecting NICs to the TOR Switch

NET-017	CONNECT AT LEAST ONE 10 GBE OR FASTER NIC TO EACH TOP-OF-RACK SWITCH
Justification	Maintains high availability in the event of the loss of one switch.
Implication	Requires at least two ToR switches, which will impact cost and rack space. Increases total bandwidth to each host to 20 Gbps.

Table 55: Using Virtual Switches and Uplinks

NET-018	USE A SINGLE VSWITCH0 WITH AT LEAST TWO OF THE FASTEST UPLINKS OF THE SAME SPEED
Justification	Simplifies the design, requiring fewer physical switches and switch ports, while relying on high bandwidth of 10Gbps and faster adapters.
Implication	All traffic is sent over a single virtual switch to a single pair of top-of-rack switches.

Table 56: Load Balancing

NET-019	USE ROUTE BASED ON PHYSICAL NIC LOAD UPLINK LOAD BALANCING
Justification	Achieves load balancing and use of both adapters with no switch side configuration required.
Implication	Requires the vSphere Distributed Switch to spread VM NICs among uplinks to reduce the most utilized physical network adapter.

Table 57: Using Standard 1500 Byte MTU

NET-020	USE STANDARD 1,500 BYTE MTU AND DO NOT USE JUMBO FRAMES.
Justification	Simplifies the design and does not need any additional configuration. Reduces risk and complexity from non-standard configuration that only increases performance in certain use cases.
Implication	All network frames are limited to the default 1,500-byte maximum size for interoperability, potentially creating more network overhead for high throughput write workloads.

Compute and Storage Design

Nutanix HCI is a converged storage and compute solution which leverages local hardware components (CPU, memory, and storage) and creates a distributed platform for running workloads.

Each node runs an industry-standard hypervisor and the Nutanix CVM. The Nutanix CVM provides the software intelligence for the platform. It is responsible for serving all IO to VMs that run on the platform.

The following logical diagram illustrates the relationship of the components of a Nutanix node:

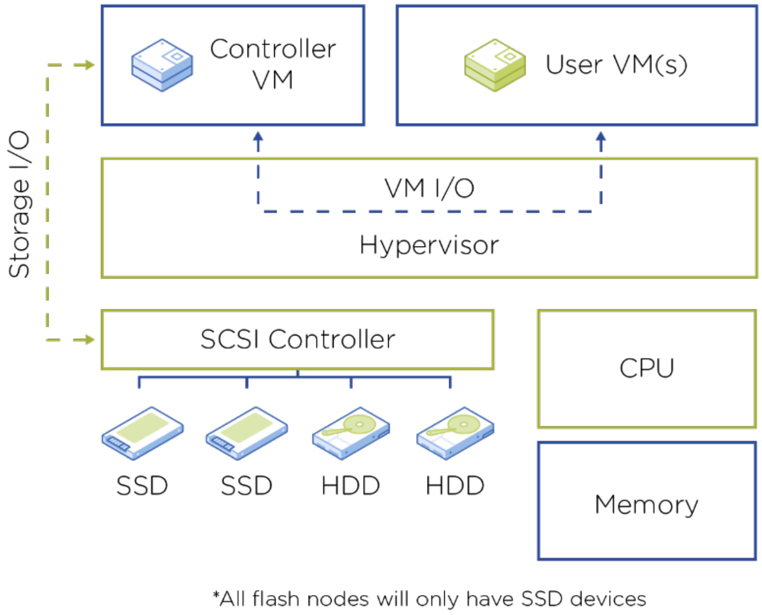


Figure 42: Relationship between the Components of the Nutanix Node

The CVM controls the storage devices directly and creates a logical construct called a storage pool of all disks from all nodes in the cluster. For AHV it uses PCI passthrough and for ESXi it uses VMDirectPathIO.

Compute Design

AHV CPU and Memory Planning

AHV lets you configure vCPUs (similar to sockets) and cores per vCPU (similar to CPU cores) for every VM. Except for the memory used by the CVM and the AHV software, all physical memory can be allocated for use by VMs. AHV guarantees memory allocated to each VM and doesn't overcommit or swap which can impact performance.

The amount of memory required by an AHV host varies based on a number of factors. The host uses from 2GB to 10GB of memory on hosts with 512GB memory.

NUMA and VMs

Processor sockets have integrated memory controllers. A VM has faster access to memory attached to the physical CPU where it is running versus memory attached to another CPU in the same host. This is called Non-Uniform Memory Access (NUMA). As a result, the best practice is to create VMs that fit within the memory available to the CPU where the VM is running whenever possible.

Table 58: Running Non-NUMA-aware Applications on a VM

CMP-001	IF RUNNING A NON-NUMA-AWARE APPLICATION ON A VM, CONFIGURE THE VM'S MEMORY AND VCPU TO FIT WITHIN A NUMA NODE ON AN AHV HOST.
Justification	This prevents the VM's vCPU and memory access from crossing NUMA boundaries.
Implication	Avoiding crossing NUMA boundaries results in better, more predictable performance.

For example, if a host has 2 CPU sockets with 8 cores each and 256GB memory, the host has 2 NUMA nodes, each with:

- 8 CPU cores
- 128GB memory

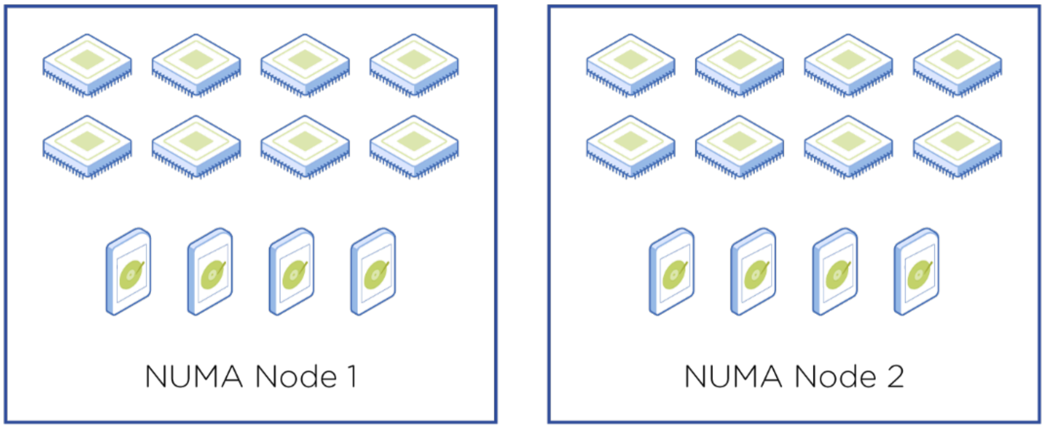


Figure 43: Example Host with 2 NUMA Nodes

For example, a VM that has 4 vCPUs and 64GB memory, will fit within a single NUMA node on the host and achieve the best performance.

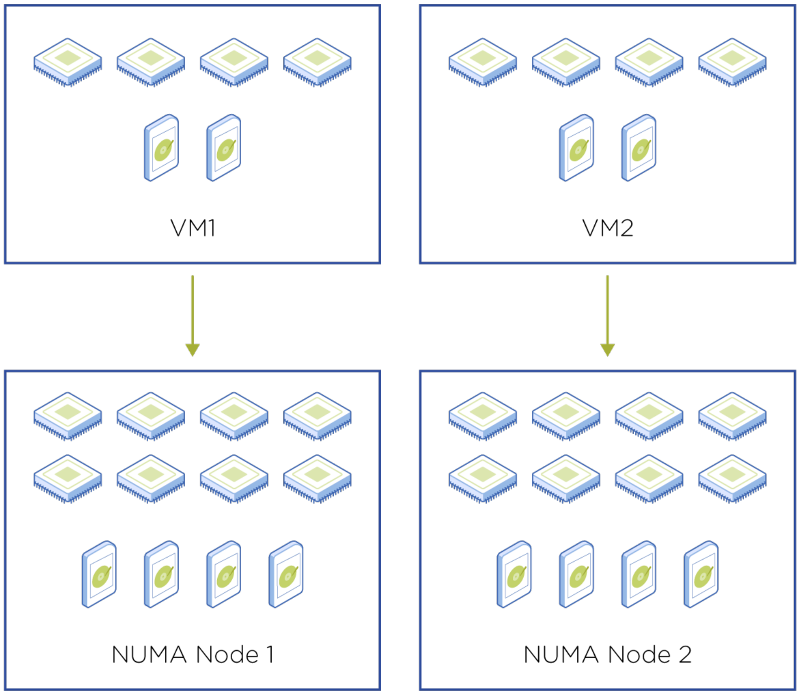


Figure 44: Example of a Single NUMA Node

Virtual Non Uniform Memory Access (vNUMA) and User VMs

The primary purpose of vNUMA is to give large virtual machines or wide VMs (VMs requiring more CPU or memory capacity than is available on a single NUMA node) the best possible performance. vNUMA helps wide VMs create multiple vNUMA nodes.

Nutanix supports vNUMA with both AHV and ESXi. vNUMA requires NUMA aware applications.

Each vNUMA node has virtual CPUs and virtual RAM. Pinning a vNUMA node to a physical NUMA node ensures that virtual CPUs accessing virtual memory see the expected NUMA behavior. Low-latency memory access in virtual hardware (within vNUMA) matches low-latency access in physical hardware (within physical NUMA), and high-latency accesses in virtual hardware (across vNUMA boundaries) match high-latency accesses on physical hardware (across physical NUMA boundaries).

In AHV, administrators can configure vNUMA for a VM via the aCLI (Acropolis CLI) or REST API; this configuration is VM-specific and defines the number of vNUMA nodes. Memory and compute are divided in equal parts across each vNUMA node. Refer to AHV admin guide for steps on how to do it.

In ESXi, vNUMA is automatically enabled for a VM with >8 vCPU.

Building on the previous example, if a host with 16 CPU cores and 256GB memory has a VM with 12vCPUs and 192GB memory that is not vNUMA configured, the vCPU and memory assignment will span NUMA boundaries.

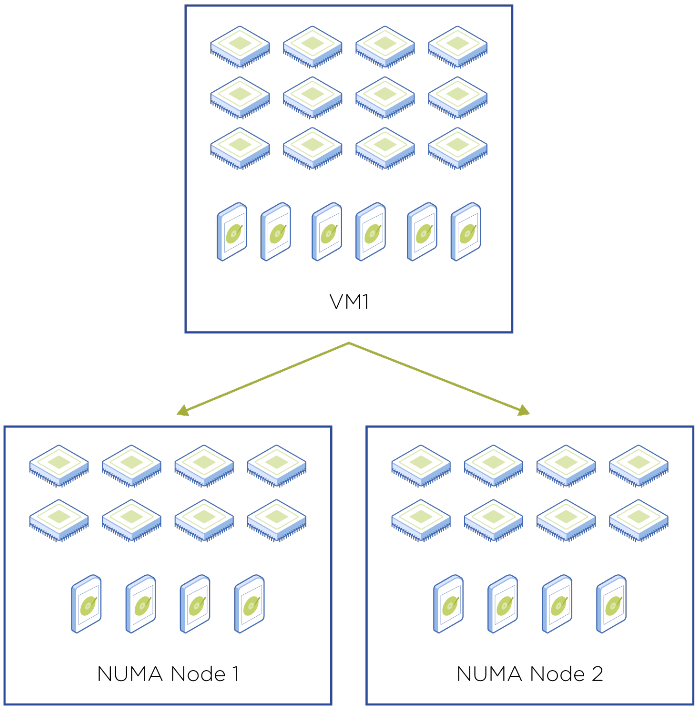


Figure 45: vCPU and Memory Assignment Spanning NUMA Boundaries

To ensure the best performance for wide VMs like this, vNUMA must be configured.

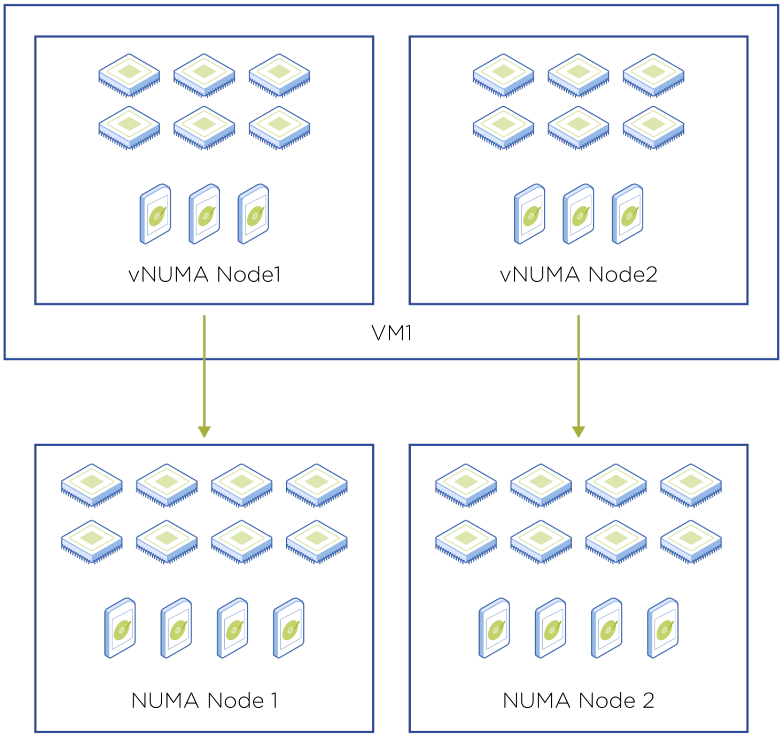


Figure 46: Best Performance vNUMA Configuration for Wide VMs

vNUMA VM configurations require strict fit. With strict fit, for each VM virtual node, memory must fit inside a physical NUMA node. Each physical NUMA node can provide memory for any number of vNUMA nodes. If there is not enough memory within a NUMA node, the VM does not power on. Strict fit is not required for CPU. To determine how many vNUMA nodes to use for a user VM, follow application-specific configuration recommendations provided by Nutanix.

CVM CPU and Memory Considerations

Optimal CVM vCPU values depend on the type of system and how many cores per NUMA node are present in the system. The CVM is pinned to the first physical CPU of the node. The following table gives an overview of minimum vCPU and memory values for the CVM (as of AOS 5.15):

Table 59: Minimum CPU and Memory Values as of AOS 5.15

Platform Storage or Type	Minimum Memory (GB)	Minimum vCPU
Platforms up to 80TB per node	32	8
Platforms up to 120TB per node	36	12
Platforms up to 120TB per node deploying Nutanix Objects	36	12

For specific recommendations on CVM sizing relating to specific workloads (e.g. Microsoft, Oracle, etc) please refer to Nutanix Solution Briefs and Best Practice Guides for those workloads.

Recommendations

Use Nutanix Foundation to set up a new cluster or node and configure the CVM(s). Nutanix Foundation software automatically identifies the platform and model type. Based on that information, Nutanix Foundation configures the appropriate default values for CVM memory and vCPU.

Storage Design

Nutanix HCI combines highly dense storage and compute into a single platform with a scale-out, shared-nothing architecture and no single points of failure. The AOS Distributed Storage Fabric (DSF) appears to the hypervisor like any centralized storage array, however all of the I/Os are handled locally to provide the highest performance.

Storage planning for this document, requires that you understand some of the high-level concepts for DSF:

Storage Pool

A group of physical storage devices within the cluster. This can include HDDs and both NAND and NVMe SSDs. The storage pool spans multiple Nutanix nodes and expands as the cluster expands. A default storage pool is created when a cluster is created.

Container

Logical segmentation of the Storage Pool is provided by containers, which contain a group of VMs or files (vDisks). Configuration options and data management services are configured at the container level. When a cluster is created, a default storage container is configured for the cluster.

> Note: When VMware ESXi is used, containers correspond to NFS datastores.

Container Guidelines:

A key design question is how many containers to configure in a cluster:

- Nutanix typically recommends minimizing the number of containers created. For most use cases only 1 container is enough.
- Exceptions can be made if two or more applications running on the same cluster require different data reduction and replication factor values. In this case, additional containers can be created for those workloads.

Data reduction, including compression, dedupe, erasure coding and replication factor (RF) values, is configured on a container basis and multiple containers need to be created when different values are desired.

All Nutanix containers are thin provisioned by default. Thin provisioning is a widely accepted technology that has been proven over time by multiple storage vendors, including VMware. As the DSF presents containers to VMware vSphere hosts as NFS datastores by default, all VMs are also thin provisioned by default. vDisks created within AHV VMs are thinly provisioned by default. Nutanix also supports thick provisioned vDisks for VMs in ESXi by using space reservations. Using thick provisioned vDisks might impact data reduction values on containers since Nutanix reserves the thick provisioned space and it cannot be oversubscribed.

Table 60: Creating vDisks in ESXi

STR-01	WHEN CREATING VDISKS IN ESXI, ALWAYS USE THIN-PROVISIONED VDISKS.
Justification	There is no performance difference between thin-provisioned and thick-provisioned disks. Thin provisioning is more space efficient.
Implication	All Nutanix data reduction technologies can be used on the vDisk.

Note: Multi-writer VMDK requires thick provisioned vDisks.

vDisk

All storage management is VM-centric, and I/O is optimized at the vDisk level.

The software solution runs on nodes from a variety of manufacturers. A vDisk is any file over 512KB stored in DSF, including vmdks and VM hard disks. vDisks are logically composed of vBlocks.

Recommendations

Use multiple vDisks for an application versus a single large vDisk.

Using multiple vDisks allows for better overall performance due to the ability to leverage multiple OS threads to do more parallel processing versus a single vDisk configuration. Refer to the appropriate application solution guide for application-specific configuration recommendations for how many vDisks to configure.

Nutanix Volumes

In addition to providing storage through the hypervisor, Nutanix also allows supported operating systems, to access DSF storage capabilities directly using Nutanix Volumes. Nutanix designed Volumes as a scale-out storage solution where every CVM in a cluster can present storage volumes via iSCSI. This solution allows an individual application to access the entire cluster, if needed, to scale out for performance.

Volumes automatically manages high availability to ensure upgrades or failures are non- disruptive to applications. Storage allocation and assignment for Volumes is done with volume groups (VGs). A VG is a collection of one or more disks (vDisks) in a Nutanix storage container. These Volumes disks inherit the properties (replication factor, compression, erasure coding, and so on) of the container they reside in. With Volumes, vdisks in a VG are load balanced across all CVMs in the cluster by default.

In addition to connecting volume groups through iSCSI, AHV also supports direct attachment of VGs to VMs.In this, the vdisks are presented to guest OS over the virtual SCSI controller. The virtual SCSI controller leverages AHV Turbo and iSCSI under the covers to connect to the Nutanix DSF. By default, the vdisks in a VG directly attached to a VM are hosted by the local CVM. Load balancing of vdisks on direct attached VGs can be enabled via acli.

Load balancing of vDisks in a VG enables I/O-intensive VMs to use the storage capabilities of multiple CVMs. If load balancing is enabled on a VG, AHV communicates directly with each CVM hosting a vDisk.

Each vDisk is served by a single CVM. Therefore, to use the storage capabilities of multiple CVMs, create more than one vDisk for a file system and use OS-level striped volumes to spread the workload. This improves performance and prevents storage bottlenecks, but can impact your network bandwidth.

Guideline: Core use-cases for Nutanix Volumes:

- Shared Disks
  - Oracle RAC, Microsoft Failover Clustering, etc.
- Where execution contexts are ephemeral, and data is critical
  - Containers, OpenStack, etc.
  - Guest-initiated iSCSI which is required for MS Exchange on vSphere (for Microsoft support).
- Bare-metal consumers
  - Stand alone or clustered physical workload.

Hybrid and All-Flash Nodes

A Nutanix cluster can consist of hybrid nodes, all-flash nodes, or a combination. Irrespective of the node type, each node has an OpLog that serves as a persistent write buffer for bursty, random write traffic. The OpLog is similar to a filesystem journal.

An extent store provides persistent bulk storage for DSF. A hybrid node has a combination of SSDs and HDDs with OpLog stored on the SSDs. All-Flash nodes are composed of SSDs. If there is a combination of SATA and NVMe SSDs, OpLog is stored on the fastest medium.

DSF automatically tiers data across the cluster to different classes of storage devices using intelligent lifecycle management (ILM). For best performance, ILM makes sure the most frequently used data is available in memory or in flash on the node local to the VM. The default threshold where the algorithm will move data from the hot tier to the cold tier is 75%. Data for down-migration is chosen based on last access time. In All-Flash node configurations, the Extent Store consists only of SSDs and no tiering occurs.

Recommendations:

For hybrid configurations, Nutanix recommends:

- Sizing the SSD tier so that the active data set of application fits in 75% of usable SSD capacity.
- Nutanix has a collector tool that can be run to determine the approximate active data set size.
- Alternatively, use application-specific tools such as: MAP toolkit for MSSQL; AWR reports and Nutanix scripts for Oracle.
- A general rule is to check how much data has been backed up during a month. Assuming a 70%/30% R/W pattern (if the R/W patterns are unknown), multiply the data you get from backups by 4, which would give an approximate value for hot data. That amount should fit within 75% of the usable cluster capacity.

For All-Flash configurations there is no ILM, so there is no equivalent recommendation.

Table 61: SSD Capacity Considerations When Sizing a Cluster

STR-02	WHEN SIZING A HYBRID CLUSTER, MAKE SURE TO HAVE ENOUGH USABLE SSD CAPACITY TO CONTAIN THE ACTIVE DATA SET OF RUNNING APPLICATIONS.
Justification	Keeping active data on SSD ensures the application achieves the best performance. ILM moves data between tiers. The best-case scenario is active data in the SSD tier all the time.
Implication	More SSD capacity may need to be sized to keep active data on the SSD tier.

Recommendations

For hybrid configurations, Nutanix recommends:

- Use All-Flash node configurations for business-critical applications for consistent performance. As mentioned before,
- All-Flash configurations can consist of all SATA SSDs or a combination of SATA and NVMe.
- If the application requires very high IOPS with consistent low latency, there is an option to use RDMA with NVMe. RDMA requires RoCEv2 to be configured on the network fabric; Nutanix makes this setup very easy through Prism.

Table 62: Mixing Nodes from Different Vendors in the Same Cluster

STR-03	DO NOT MIX NODE TYPES FROM DIFFERENT VENDORS IN THE SAME CLUSTER.
Justification	Mixing node types from different vendors is unsupported.
Implication	Nodes from different vendors can be setup in separate clusters and all managed centrally from Prism Central.

Table 63: Mixing NVMe and Hybrid SSD/HDD Nodes in the Same Cluster

STR-04	DO NOT MIX NODES THAT CONTAIN NVME SSDS IN THE SAME CLUSTER WITH HYBRID SSD/HDD NODES.
Justification	The performance of NVMe SSD nodes is significantly higher compared to hybrid nodes and mixing them can potentially create performance issues.
Implication	NVMe Nodes can be setup in their own clusters or added to clusters with All SSD nodes and managed centrally from Prism Central.

Node Selection Recommendations

- If your high-level design decision is to use hybrid nodes, you need to decide the number of HDDs and SSDs per node.

- For applications that require high performance like DBs and Tier-1 applications, more drives per node is better.
    - OpLog will be spread across up to 8 SSDs, so more SSDs results in overall better and more consistent performance. If NVMe devices are present, OpLog is placed on them upto 8 NVMe SSDs.
    - More SSDs provides more usable space for tiering data.
- For All-Flash nodes, more SSDs provide better performance by making read and write access more parallel, especially for non-optimal workload patterns.
- Use Nutanix Sizer to size usable capacity and obtain node recommendations based on your workload use case.
- Also refer to specific application solution guides and/or best practice guides.

Table 64: HDD to SSD Ratio

STR-05	A MINIMUM 2:1 HDD TO SSD RATIO IS REQUIRED FOR HYBRID CLUSTERS.
Justification	This ensures there is enough bandwidth in the slower storage tier to absorb down migrations triggered by ILM.
Implication	If this rule is not followed, it may result in performance degradation during down migrations.

Note: 4 drive slot and 10 drive slot systems from Dell, Lenovo, and Fujitsu can have 2+2 and 4+6 SSD+HDD configurations respectively.

Table 65: Sizing for Redundancy

STR-06	SIZE FOR N+1 NODE REDUNDANCY FOR STORAGE AND COMPUTE WHEN SIZING. FOR MISSION CRITICAL WORKLOADS THAT NEED HIGHER SLAS, USE N+2 NODE REDUNDANCY.
Justification	Having the storage and compute capacity of an additional node (or two) ensures there is enough storage and compute capacity available for VMs to re-start when a node failure occurs.

Implication	Additional storage and compute capacity is required for N+1 or N+2 redundancy.

Availability Domains and Fault Tolerance

Availability domains are used to determine component and data placement. Nutanix availability domains are:

- Disk Awareness (always).
- Node Awareness (always).
- Block Awareness (optional).Requires a minimum of 3 blocks for FT1 and 5 for FT2, where a block contains either 1,2 or 4 nodes. With Erasure Coding, minimums are 4 blocks for FT1 and 6 blocks for FT2.
- Rack Awareness(optional).Requires a minimum of 3 racks for FT1 and 5 for FT2, and you must define what constitutes a rack and where your blocks are placed. With Erasure Coding, minimums are 4 racks for FT1 and 6 blocks for FT2.

Closely tied to the concepts of awareness and RF is cluster resilience, defined by the redundancy factor or Fault Tolerance (FT). FT is measured for different entities, including:

- Data
- Metadata
- Configuration data
- Free space

FT=1: A cluster can lose any one component and operate normally. FT=2: A cluster can lose any two components and operate normally.

Depending on the defined awareness level and FT value, data and metadata are replicated to appropriate locations within a cluster to maintain availability.

The following table gives an overview of Data awareness and FT levels supported.

Table 66: Data Awareness and Fault Tolerance Levels Supported

DESIRED AWARENESS TYPE	FT LEVEL	MIN. UNITS*	SIMULTANEOUS FAILURE TOLERANCE
Node	1	3 nodes	1 node
Node	2	5 nodes	2 nodes
Block	1	3 blocks	1 block
Block	2	5 blocks	2 blocks
Rack	1	3 racks	1 rack
Rack	2	4 racks	2 rack

*The minimum units increase by 1 if EC-X is enabled.

FT Guidelines

- FT=2 implies RF=3 for data and RF=5 for metadata by default.
  - Use this setting when you need to withstand two simultaneous failures and have cluster and VM data still be available.
- FT=1 implies RF=2 for data and RF=3 for metadata by default.
  - Use this setting when the application is fine with having only 2 copies of data. This will keep the cluster and VM data available after one failure.
- Another option is to set FT=2 for a cluster with specific containers set to RF=2.
  - With this configuration, data is RF=2, but metadata is RF=5.
  - For containers that are RF=2, if there are 2 simultaneous failures there is a possibility that VM data becomes unavailable, but the cluster remains up.
  - This is different than the FT=1, RF=2 case where two simultaneous failures will result in a cluster being down.

Table 67: Fault Tolerance for Workloads/Clusters with Higher SLAs

STR-07	USE FT=2 AND RF=3 FOR WORKLOADS AND CLUSTERS THAT NEED HIGHER SLAS OR FOR CLUSTER SIZES >32.
Justification	Using FT=2 and RF=3 provides additional resilience and capability to survive 2 simultaneous failures.
Implication	Additional storage space is required for FT=2 and RF=3, reducing usable capacity.

Capacity Optimization

Nutanix provides different ways to optimize storage capacity that are intelligent and adaptive to workloads characteristics. All optimizations are performed at the container level, so different containers can use different settings.

Erasure Coding (EC-X)

To provide a balance between availability and the amount of storage required, DSF provides the ability to encode data using erasure codes (EC). Like RAID (levels 4, 5, 6, etc.) where parity is calculated, EC encodes a strip of data blocks across different nodes and calculates parity. In the event of a host and/or disk failure, the parity data is used to calculate any missing data blocks (decoding). In the case of DSF, the data block must be on a different node and belong to a different vDisk. EC is a post-process operation and is done on write cold data (Data that hasn't been overwritten in more than 7 days). The number of data and parity blocks in a strip is chosen by the system based on number of nodes and configured failures to tolerate.

The following table summarizes the expected overhead for EC-X vs standard RF2/RF3 overhead:

Table 68: Expected Overhead for EC-X vs Standard RF2/RF3

	FT1 (RF2 Equivalent)		FT2 (RF3 Equivalent) 3X Storage Overhead
# of nodes	EC strip size	3 nodes	1 node
(data/parity)	EC overhead	EC strip size	2 nodes

(data/parity)	EC overhead	3 blocks	1 block	
4	2/1	1.5X	N/A	N/A
5	3/1	1.33X	N/A	N/A
6	4/1	1.25X	2/2	2X
7	4/1	1.25X	3/2	1.6X
8+	4/1	1.25X	4/2	1.5X

EC-X Guidelines

EC-X provides the same level of data protection while increasing usable storage capacity.

- Turn on EC-X for non-mission-critical workloads and workloads that have a significant amount of write cold data, since erasure coding works on write cold data and provides more usable storage.
- For more information refer to specific application guides.

Compression

DSF provides both inline and post-process data compression. Irrespective of inline or post-process compression, write data coming into oplog that is >4k and shows good compression, will be written compressed in OpLog.

- Inline: Compresses sequential streams of data or large size I/Os (>64K) when writing to the Extent Store.
- Post-Process: Compresses data after it is drained from OpLog to the Extent Store after compression delay is met during the next curator scan.

Nutanix uses the LZ4 and LZ4HC compression algorithms. LZ4 compression is used to compress normal data, providing a balance between compression ratio and performance. LZ4HC compression is used to compress cold data to improve the compression ratio. Cold data is characterized as:

- Regular data: No access for 3 days.
- Immutable data: No access for 1 day.

Compression Guidelines

Compression provides on-disk space savings for applications such as databases, and results in a lower number of writes being written to storage. Turn ON inline compression for all containers.

Table 69: Enabling Inline Compression for All Containers

STR-08	ENABLE INLINE COMPRESSION FOR ALL CONTAINERS.
Justification	Since inline compression operates only on sequential data when writing to extent store, enabling it provides immediate space efficiency. Write data coming into oplog is already compressed.
Implication	Slightly higher CPU consumption for compressing/decompressing data.

Deduplication

When enabled, DSF does capacity-tier and performance-tier deduplication. Data is fingerprinted on ingest using a SHA-1 hash that is stored as metadata. When duplicate data is detected based on multiple copies with the same fingerprint, a background process removes the duplicates. When deduplicated data is read, it is placed in a unified cache, and any subsequent requests for data with the same fingerprint are satisfied directly from cache.

Deduplication Guidelines

Deduplication is recommended to be used for full clones, P2V migrations and persistent desktops.

Table 70: Enabling/Disabling Deduplication

STR-09	ENABLE/DISABLE DEDUPLICATION
Justification	Increases effective capacity of caching layer for dedup-able data like VDI clones.
Implication	Slightly higher CPU consumption for running deduplication processes.

Refer to specific application Solution Guides for determining whether to enable deduplication or not.

Virtualization Layer Design

This section describes technical design considerations for design using either Nutanix AHV or VMware ESXi.

Nutanix AHV

Introduction

The Nutanix hypervisor, AHV, is an attractive alternative hypervisor that streamlines operations and lowers overall costs. Built-in to Nutanix Enterprise Cloud and included at no additional cost in the AOS license, AHV delivers virtualization capabilities for the most demanding workloads. It provides an open platform for server virtualization, network virtualization, security, and application mobility. When combined with comprehensive operational insights and virtualization management from Nutanix Prism and Prism Central, AHV provides a complete datacenter solution.

Control Plane

The control plane for managing the Nutanix core infrastructure and AHV are provided by:

- Prism Element (PE)
  - Localized cluster manager responsible for local cluster management and operations. Every Nutanix Cluster includes Prism Element built-in.
  - 1-to-1 cluster manager
- Prism Central
  - Multi-cluster manager responsible for managing multiple Nutanix clusters to provide a single, centralized management interface. Prism Central is an optional software appliance (VM).
  - 1-to-many cluster manager
  - A dedicated PC is required at the source and target sites for any regions or availability zones using DR runbook automation for failing over workloads.

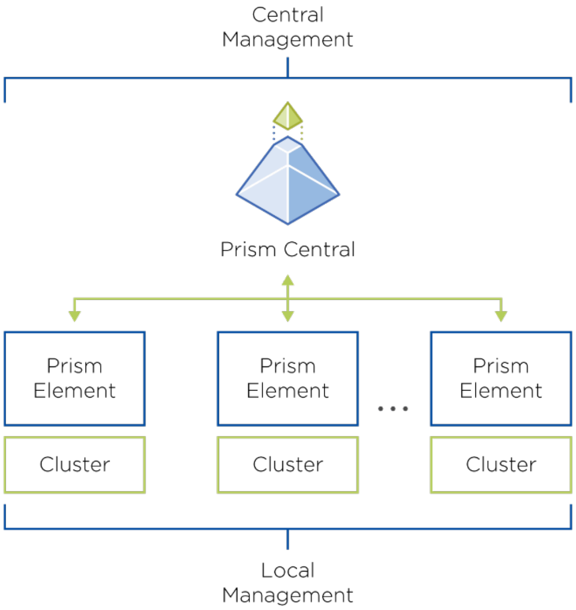


Figure 47: The Nutanix Control Plane

Table 71: Deploying Prism Central for Enhanced Cluster Management

VRT-001	DEPLOY SCALE-OUT PRISM CENTRAL FOR ENHANCED CLUSTER MANAGEMENT.
Justification	Increases the total number of managed objects, increases management plane availability.
Implication	Requires some additional cluster resources to run the necessary VMs.

Table 72: Deploy Prism Central in Each Region or Az Using Runbook Dr Automation.

VRT-002	DEPLOY PRISM CENTRAL IN EACH REGION OR AZ USING RUNBOOK DR AUTOMATION.
Justification	Required for DR runbook automation feature that the source and target sites have separate PC instances controlling them.
Implication	Requires some additional cluster resources to run the necessary VMs.

HA/ADS

Nutanix AHV has built-in VM high availability (VM-HA) and a resource contention avoidance engine called Acropolis Dynamic Scheduling (ADS). ADS is always on and does not require any manual tuning. VM-HA can be enabled and disabled via a simple Prism checkbox. There are two main levels of VM-HA for AHV:

- Best effort. In this HA configuration, no node or memory reservations are required in the cluster. In case of a failure, VMs are moved to other nodes based on the resources/ memory available on each node. This is not a preferred method of HA, since if no resources are available in the cluster after a failure some VMs may not be powered-on. Best effort is the default configuration. Most applicable to non-production environments.
- Guarantee. In this HA configuration, some memory is reserved on each node in the cluster to enable failover of virtual machines from a failed node. The Acropolis service in the cluster calculates the memory to be reserved based on virtual machine memory configuration. All nodes are marked as schedulable with resources available for running VMs. Recommended for production environments. Enable this mode by checking the Enable HA Reservation box in Prism settings.

VM-HA considers memory when calculating available resources throughout the cluster for starting VMs. VM-HA respects Acropolis Distributed Scheduler (ADS) VM-host affinity rules, but it may not honor VM-VM anti-affinity rules, depending on the scenario.

Table 73: Using VM-HA Guarantee

VRT-003	USE VM-HA GUARANTEE
Justification	Ensuring VM high availability is important in production environments. Guarantee should be used in production environments. “Best effort” (default) is optimal in test/ dev environments.
Implication	Reserves enough cluster resources such that an entire host’s worth of VMs can be successfully restarted (RF=2), or two hosts (RF=3). RF=3 must be manually configured, since it is not the default.

For additional information, see: Virtual Machine High Availability.

AHV CPU Generation Compatibility

Similar to VMware's Enhanced vMotion Capability (EVC) which allows VMs to move between different processor generations, AHV will determine the lowest processor generation in the cluster and constrain all QEMU domains to that level. This allows mixing of processor generations within an AHV cluster and ensures the ability to live migrate between hosts.

VMware vSphere

Introduction

VMware vSphere is fully supported by Nutanix with many points of integration. vSphere major and minor releases go through extensive integration and performance testing, ensuring that both products work well together in even the largest enterprises. Rest assured that if you deploy vSphere, the solution is fully supported.

Control Plane

When using vSphere as a part of this document, the control plane is vCenter to manage the vSphere components and Prism Central for managing Nutanix components.

Table 74: Managing ESXi-based Nutanix Clusters

VRT-004	DEPLOY AN HA VCENTER INSTANCE WITH EMBEDDED PSC TO MANAGE ESXI-BASED NUTANIX CLUSTERS.
Justification	Nutanix upgrade automation features, such as 1-click upgrades and LCM, require the advanced control features that vCenter provides. In addition, vCenter Server is required to create and manage the VMware vSphere cluster responsible for DRS and HA.
Implication	Requires additional vCenter licenses and added installation time to fully configure each vCenter instance.

Table 75: Deploy a vCenter in Region or Az Using Runbook Dr Automation

VRT-005	DEPLOY A vCENTER IN EACH REGION OR AZ USING RUNBOOK DR AUTOMATION.
Justification	Required for DR runbook automation feature that the source and target sites have separate PC and local vCenter instances controlling them.
Implication	Requires some additional cluster resources to run the necessary VMs.

EVC

VMware's Enhanced vMotion Capability (EVC) allows VMs to move between different processor generations within a cluster. EVC mode is a manually configured option, unlike the corresponding feature in AHV. EVC should be enabled at cluster installation time to negate the possible requirement to reboot VMs in the future when it is enabled.

Table 76: Enabling EVC Mode

VRT-006	ENABLE EVC MODE AND SET IT TO THE HIGHEST COMPATIBILITY LEVEL THE PROCESSORS IN THE CLUSTER WILL SUPPORT.
Justification	Even if a cluster is homogenous at the outset, it is likely new nodes will be added over time. Enabling EVC mode when a cluster is built helps ensure future node additions are seamless.
Implication	Requires manual configuration to determine the highest compatible EVC mode.

HA/DRS

VMware HA and DRS are core features which should be utilized as a part of this document. Nutanix best practices dictate that a few HA/DRS configuration settings be changed from the default. These changes are outlined in the design decisions below.

Table 77: Enabling HA

VRT-007	ENABLE HA
Justification	In production environments VM availability is extremely important. Automatically restarting VMs in case of node failure helps ensure workload availability.
Implication	Requires additional cluster resources to ensure VMs on the failed node(s) can be restarted elsewhere in the cluster.

Table 78: Enabling DRS

VRT-008	ENABLE DRS WITH DEFAULT AUTOMATION LEVEL
Justification	Hot spots can develop in a cluster. DRS moves VMs as needed to ensure optimal performance.
Implication	Initiates vMotion for optimal VM performance in case of resource contention. Moving VMs may temporarily impact Nutanix data locality.

Table 79: Disabling VM PDL and APD Component Protection

VRT-009	DISABLE VM PDL AND APD COMPONENT PROTECTION
Justification	These settings are designed for non-Nutanix storage. They should not be enabled in Nutanix clusters because they can cause storage unavailability. APD monitoring is managed via a Nutanix component.
Implication	Permanent device loss (PDL) and all paths down (APD) monitoring is disabled, and vSphere will not monitor for these conditions.

Table 80: Configuring DAS.IGNOREINSUFFICIENTHBDATASTORE

VRT-010	CONFIGURE DAS.IGNOREINSUFFICIENTHBDATASTORE IF ONE NUTANIX CONTAINER IS PRESENTED TO THE ESXI HOSTS
Justification	In many Nutanix clusters a single datastore is used. The warning message is not applicable in Nutanix environments.
Implication	Eliminates false positive warnings when a cluster uses a single datastore.

Table 81: Disabling Automation Level for All CVMs

VRT-011	DISABLE AUTOMATION LEVEL FOR ALL CVMS
Justification	CVMs are stored on local storage and cannot be vMotioned to another host. HA does not need to reserve resources for CVMs since they are bound to a single node and cannot be restarted elsewhere in the cluster.
Implication	DRS will not try and move a CVM to another host. Nor will CVM resources be taken into account for HA calculations.

Table 82: Setting Amount of Reserved Resources in the Cluster

VRT-012	SET HOST FAILURES CLUSTERS TOLERATE TO 1 FOR RF2 AND TO 2 FOR RF3.
Justification	Starting in vSphere 6.5, cluster resource percentage is now tied to “host failures cluster tolerates” and automatically adjusts the percentage based on the number of nodes in the cluster.
Implication	Automatically sets the proper amount of reserved resources in the cluster.

Table 83: Setting Host Isolation Response for VMs

VRT-013	SET HOST ISOLATION RESPONSE TO “POWER OFF AND RESTART VMS”
Justification	In an HCI environment all communications, including storage, are via the ethernet network. If a host becomes isolated on the network, VMs need to be moved to a healthy host to continue to function.
Implication	If a host becomes isolated on the network, the VMs on that host will be powered off and restarted on another node in the cluster.

Table 84: Setting Host Isolation Response for the CVM

VRT-014	SET HOST ISOLATION RESPONSE TO “LEAVE POWERED ON” FOR CVMS

Justification	If there is a transient network disruption, you do not want the CVMs to be powered off.
Implication	If a host becomes isolated on the network, the CVMs will remain on.

Table 85: Disabling HA/DRS for Each Nutanix CVM

VRT-015	DISABLE HA/DRS FOR EACH NUTANIX CVM
Justification	CVM uses local storage and no attempt to restart on other hosts should be performed.
Implication	vSphere will not try and restart the CVM on another host, which is impossible since the CVM uses local storage.

Table 86: Disabling SIOC

VRT-016	DISABLE SIOC
Justification	If SIOC is enabled, it can have negative side effects. These include storage unavailability, creating unnecessary lock files, and complications with Metro Availability.
Implication	Enhanced storage statistics are not available.

Management Layer Design

Control Plane

This document provides a primary control plane where the majority of daily operational tasks are performed. In this design:

- The primary control plane is Prism Central.
- VMware vCenter is also required if you are using VMware ESXi.

Each of these has a maximum size that limits the total number of VMs, nodes, and clusters that can be managed by an instance.

Prism Central

Prism Central (PC) is a global control plane for Nutanix, which includes managing VMs, replication, monitoring, and value-add products such as:

- Nutanix Calm for application-level orchestration. (See the section Calm Application Orchestration for more information.)
- Nutanix Flow for microsegmentation of east-west traffic. (See the section Nutanix Flow for Microsegmentation for more information.)
- Prism Pro for advanced task automation and capacity planning. (See the section Prism Operations for more information.)

These products are installed from Prism Central and managed centrally for all Prism Element clusters associated to the Prism Central instance.

Prism Central plays an important part in a deployment regardless of the scale. There are two deployment architectures for Prism Central that can be utilized and scaled depending on the size and goals of the design.

Single-VM Prism Central Deployment

The single VM deployment option for Prism Central offers a reduced footprint option for designs that do not require high availability for the control plane beyond that provided by the hypervisor. The single Prism Central VM option can be used with a small or large VM sizing. The size directly correlates to the size of environment it can manage.

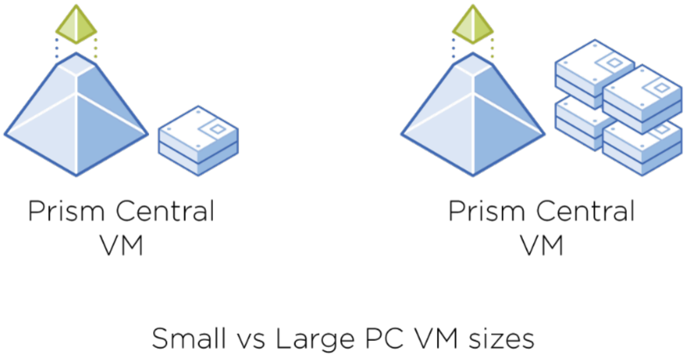


Figure 48: Small vs Large Prism Central VM Sizes

Scale-Out Prism Central Deployment

A scale out PC cluster is comprised of three VMs. All of the members of the cluster are running all of the Nutanix services and products in a distributed architecture.

This distributed architecture allows for the scaled-out Prism Central to manage a larger environment as well as handle more incoming requests from users and the API.

Additionally, it provides a higher level of availability. If one of the VMs is down or performing a maintenance operation, the other members remain available. You have the option of deploying the PC VMs as small- or large-sized VMs.

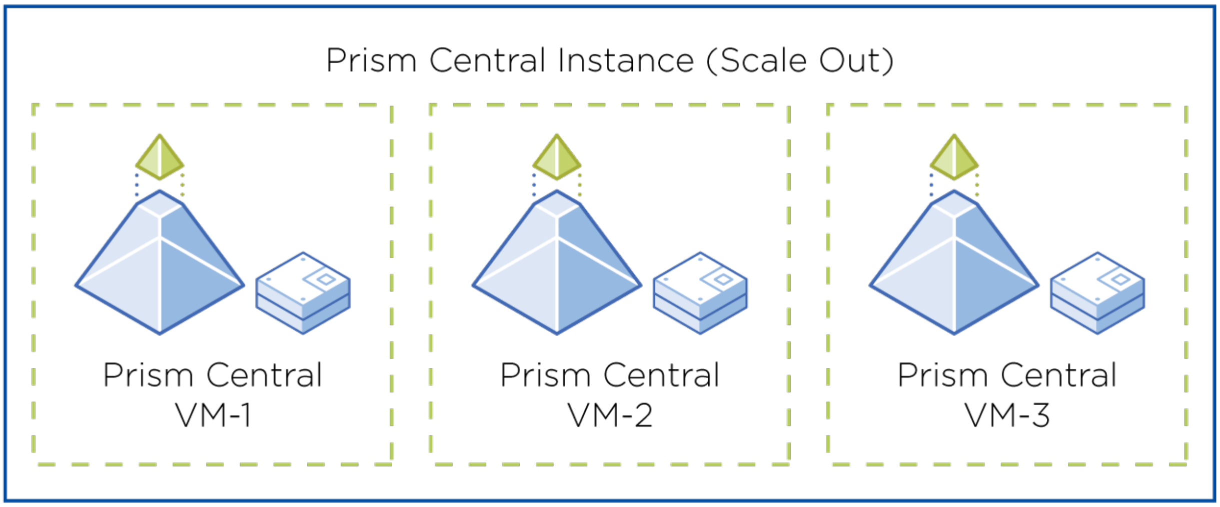


Figure 49: Scale Out Prism Central Deployment

The scale-out Prism Central architecture deploys 3 VMs on the same cluster and is superior in terms of availability and capacity. If scale-out is not the best option from the beginning, it is fairly simple to go from a single PC VM deployment to a scale-out architecture at a later time.

Image Templates

Within any environment operating at scale there is a need to keep approved template images available and in sync between clusters. For AHV clusters, image placement policies should be utilized for this. Image policies are configured to determine the clusters each image can or cannot be utilized in. This makes the initial roll out of new and updated image versions easy.

For ROBO deployments utilizing AHV, image placement policies should be configured to avoid unnecessary data being replicated between locations with limited bandwidth.

Prism Central Recommendations

Table 87: Prism Central Recommendations

Recommendation Area	Description
Prism Central instances	Deploy Prism Central in the following way: • Scale out mode (cluster of 3 VMs). The benefits: • Simplifies capacity planning. • Simplifies platform life cycle management. • Simplifies management for virtual networking. • Reduce management overhead.
Workload Domains	Deploy management workloads such as Prism Central on management clusters.
Access Control	• Integrate with Active Directory. • Leverage AD groups to optimize access management.

VMware vCenter Server

For designs utilizing ESXi as the hypervisor, vCenter is a critical part of the infrastructure for hypervisor cluster management and VM operations. It may be accessed often in environments that are not highly automated. In automated environments, vCenter is needed to perform operations that are sent to it from different orchestration layers.

The vCenter appliance can be deployed in many different sizes. This ultimately determines how large an environment you can support in terms of VMs, hosts, and clusters. Refer to the official VMware documentation for the version you are deploying to determine the proper sizing for your design.

Beyond the size of the environment, the vCenter appliance can be deployed as a single VM or as a vCenter High Availability (vCenter HA) for additional availability. The option to deploy vCenter HA does not increase the size of the environment it can manage as a single virtual appliance is active at any time. The size of the environment that can be managed is based on the size of the VMs deployed.

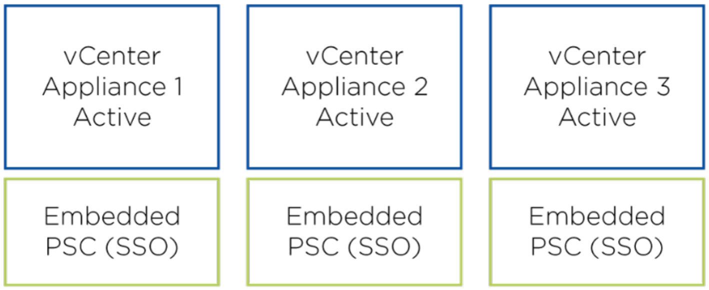

vCenter Appliance 1 Active
vCenter Appliance 2 Active
vCenter Appliance 3 Active
Embedded PSC (SSO)
Embedded PSC (SSO)
Embedded PSC (SSO)

Figure 50: vCenter Appliances

To ensure that the control plane is as highly available as possible, the clustered deployment option is preferred as long as its compatible with other layers of the solution.

vCenter Server Recommendations

Table 88: vCenter Server Recommendations

RECOMMENDATION AREA	DESCRIPTION
vCenter instances	Deploy one vCenter in an HA configuration for all workloads.Configure with Embedded PSC.
Workload domains	Deploy management workloads on management clusters.
Access Control	• Integrate with Active Directory. • Leverage AD groups to optimize access management.

Dependent Infrastructure

There are a variety of other infrastructure services that are necessary for a successful Nutanix deployment, such as NTP, DNS, and AD. You may already have these infrastructure services deployed and available for use when you deploy your Nutanix environment. If they do not exist, then you will probably need to deploy them as part of the new environment

NTP

Network Time Protocol (NTP) is used to synchronize computer clock times including network, storage, compute, and software. If clock times drift too far apart, some products may have trouble communicating across layers of the solution. Keeping time in-sync is beneficial when examining logs from different layers of the solution to determine the root cause of an event.

A minimum of three NTP servers should be configured and accessible to all solution layers with the NTP standard recommendation being five to detect a rogue time source. These include AOS, AHV, and Prism Central, plus vCenter and ESXi if using VMware vSphere as the virtualization layer. Use the same NTP servers for all infrastructure components.

Do not use an Active Directory Domain Controller as an NTP source.

If you are in a Dark site with no internet connectivity, consider using a switch or GPS time source.

Table 89: NTP Servers for Infrastructure Components

DEP-001	A MINIMUM OF THREE NTP SERVERS FOR ALL INFRASTRUCTURE COMPONENTS SHOULD BE PROVIDED.
Justification	Unreliable time services can lead to authentication failures, mis-aligned log entries, and other critical errors.
Implication	Provides reliable time services for all system components.

DNS

Domain Name System (DNS) is a directory service that translates domain names of the form domainname.ext to IP addresses. DNS is important to ensure that all layers can resolve names and communicate. A highly available DNS environment should be utilized to support this design. At least two DNS servers should be configured and accessible at all layers to ensure components can reliably resolve addresses at all times.

These layers include the following:

- AOS
- AHV
- Prism Central
- VMware ESXi
- VMware vCenter
- Network switches

Table 90: DNS Servers for Infrastructure Layers

DEP-002	A MINIMUM OF TWO DNS SERVERS SHOULD BE CONFIGURED ACCESSIBLE TO ALL INFRASTRUCTURE LAYERS
Justification	DNS is critical for a number of services and infrastructure failures will occur if DNS is unavailable.
Implication	The organization needs to deploy at least two DNS servers and point all services to them.

Active Directory

Active Directory (AD) is a directory service developed by Microsoft for Windows domain networks. AD often serves as the authoritative directory for all applications and infrastructure within an organization. For this design, all of the consoles and element managers will utilize RBAC and use AD as the directory service for user and group accounts. Where possible, AD groups should be utilized to assign privileges for easier operations. User access can then be controlled by adding or removing a user from the appropriate group.

The AD design should be highly available to ensure that directory servers are always available when authentication requests occur.

Logging Infrastructure

Capturing logs at all layers of the infrastructure is very important. For example, if there is a security incident, logs can be critical for forensics. An example of a robust log collector is Splunk, but there are other options. Nutanix recommends that you deploy a robust logging platform that meets your security requirements. All infrastructure logs should be forwarded to the centralized log repository.

Recommendation

It is commonly recommended to store the logs in a different cluster, or location, from where they are being collected. This protects the logs in case of a catastrophic cluster failure, ensuring they can later be used for forensics.

Security Layer Design

Nutanix Enterprise Cloud can be used to build private or multi-tenant solutions, and depending on the use case the security responsibility will vary. The security approach includes multiple components:

- Physical
  - Datacenter access
  - Equipment access (firewalls, load balancers, nodes, network switches, racks, routers)
- Virtual Infrastructure
  - Clusters
  - Nodes
  - Management components
  - Network switches
  - Firewalls
  - Load balancers

- Threat vectors
    - Active
    - Automated
    - Internal
    - External
- Workloads
    - Applications
    - Containers
    - Virtual Machines

Nutanix Enterprise Cloud infrastructure is designed to deliver a high level of security with less effort. Nutanix publishes custom security baseline documents for compliance, based on United States Department of Defense (DoD) RHEL 7 Security Technical Implementation Guides (STIGs) that cover the entire infrastructure stack and prescribe steps to secure deployment in the field. The STIGs uses machine-readable code to automate compliance against rigorous common standards.

Nutanix has implemented Security Configuration Management Automation (SCMA) which is installed, configured, and enabled during the Nutanix deployment. There are configuration options, however one key differentiator is that Nutanix provides this functionality by default. SCMA checks multiple security entities for both Nutanix storage and AHV. Nutanix automatically reports inconsistencies and reverts them to the baseline.

With SCMA, you can schedule the STIG to run hourly, daily, weekly, or monthly. STIG has the lowest system priority within the virtual storage controller, ensuring that security checks do not interfere with platform performance.

In addition, Nutanix releases Security Advisories which describe potential security threats and their associated risks plus any identified mitigation(s) or patches.

For more information about security, see Building Secure Platforms And Services With Nutanix Enterprise Cloud.

Authentication

Maintain as few user and group management systems as possible. A centrally managed authentication point is preferred to many separately managed systems. Based on that general recommendation you should at a minimum take advantage of the external LDAP support provided by Nutanix components.

Prism Central also provides both LDAP and Security Assertion Markup Language (SAML) and makes it possible for users to authenticate through a qualified Identify Provider (IdP). If none of the options are available Nutanix also provides local user management capabilities.

Table 91: Active Directory Authentication

SEC-001	USE ACTIVE DIRECTORY AUTHENTICATION. THIS APPLIES FOR USER AND SERVICE ACCOUNTS.
Justification	User activity logged for auditing purposes and account security is configured and maintained from a single centralized solution.
Implication	Requires a highly available active directory infrastructure and additional initial configuration.

Table 92: Using SSL/TLS Connection to Active Directory

SEC-002	USE SSL/TLS CONNECTION TO ACTIVE DIRECTORY.
Justification	Increases security by eliminating cleartext exchanges on the network.
Implication	Might require additional configuration.

Certificates

All components facing consumers should be protected with certificate authority (CA) signed certificates. Internal or external signed certificates depend on consumer classifi- cation and what kind of service the specific component provides.

Table 93: Using a Certificate Authority

SEC-003	USE CERTIFICATE AUTHORITY (CA) THAT ARE CONSIDERED "TRUSTED CA" BY YOUR ORGANIZATION FOR THE COMPONENTS WHERE CERTIFICATES CAN BE REPLACED. THIS CAN BE EITHER INTERNAL OR EXTERNAL SIGNED CERTIFICATES.
Justification	Certificates are required as part of Transport Layer Security (TLS) protocol and are negotiated during session establishment to establish secure session. They provide an extra layer of security and prevent man-in-the-middle attacks.
Implication	Implementation effort and potentially an extra cost if certificates need to be purchased.

Cluster Lockdown

Nutanix cluster lockdown lets you enforce SSH access to the CVMs and hosts using key pairs instead of passwords.

Table 94: Nutanix Cluster Lockdown

SEC-004	DO NOT USE NUTANIX CLUSTER LOCKDOWN.
Justification	Nutanix recommends that access, including SSH, directly to CVM and hypervisor should be restricted to as few entities as possible:Operators. Keep service account passwords secure.Remote workstations. Use e.g. jump hosts.Networks. Use network segmentation via firewall rules or IPtables rules.
Implication	Users can access the CVM with username and password.

If only password-less communication is required, then enable Nutanix cluster lockdown.

VMware Cluster Lockdown

Cluster lockdown can be enabled at the VMware layer as well. Enabling either normal or strict cluster lockdown mode can fail the Nutanix cluster functionality. Make the necessary ESXi configurations to guarantee Nutanix functionality if vSphere cluster lockdown mode is required.

Table 95: vSphere Cluster Lockdown

SEC-005	DO NOT USE VSPHERE CLUSTER LOCKDOWN.
Justification	No requirements exist in this design that justify vSphere cluster lockdown mode.
Implication	Users can access vSphere cluster with username and password.

Hardening

There are certain hardening configurations you can apply for the AHV host and the CVM if required:

- AHV and CVM
  - AIDE. Advanced Intrusion Detection Environment.
  - Core. Enable stack traces for cluster issues.
  - High Strength Passwords. Enforce complex password policy (Min length 15, different by at least 8 characters).
  - Enable Banner. Get specific login message via SSH.
- CVM only
  - Enforce SNMPv3 only.

Table 96: CVM and Hypervisor AIDE

SEC-006	ENABLE CVM AND HYPERVISOR AIDE
Justification	With AIDE enabled, the environment will perform checksum verification of all static binaries and libraries for improved security.
Implication	Additional, very limited, CVM and hypervisor resources will be required.

Table 97: Configuring SCMA

SEC-007	CONFIGURE SCMA TO RUN HOURLY
Justification	To more quickly capture unacceptable configuration drift. Default is to take actions daily.

Implication	SCMA is lightweight and adds very limited additional load to CVM and hypervisor. Security benefits outweigh the added resources.

Table 98: Unused ESXi Services and Unused Firewall Ports

SEC-008	STOP UNUSED ESXI SERVICES AND CLOSE UNUSED FIREWALL PORTS. Important: Make sure not to stop a service or close a firewall port required by Nutanix, such as SSH and NFS
Justification	To limit the attack surface.
Implication	Additional configuration required.

We recommend that you follow the Hardening Controller VM guide. If you are using AHV, follow the Hardening AHV guide.

For updated Hardening Guides, search portal.nutanix.com for the most recent guide that relates to your version of AOS and AHV.

Internet-facing Services

Security for Internet-facing services is out of the scope of this design, but highly important and mentioned here for awareness.

Internet-facing services are at constant risk of being attacked. Two common types of attacks are denial of service (DoS) and distributed denial of service (DDoS). To help mitigate the potential impact of a DoS or a DDoS attack a design can take advantage of:

- Multiple internet connections (active/backup).
- ISP provided DoS and DDoS filtering.

There are additional ways to implement protection against these types of attacks.

Logging

Logging is critical from an auditing and traceability perspective so making sure the virtual infrastructure, AOS, Prism Central, AHV, ESXi, any additional software in the environment send their log files to a highly available log infrastructure is critical.

Recommendation

A minimum one of the individual targets making up the highly available logging infrastructure should run outside of the virtual infrastructure itself so that if the virtual infrastructure is compromised, forensic investigations will need to access logs that might not be available on the cluster itself.

A single centralized activity logging solution for auditing purposes and account security should be configured and maintained.

Table 99: Logging

SEC-009	SEND LOG FILES TO A HIGHLY AVAILABLE SYSLOG INFRASTRUCTURE.
Justification	Logs are helpful in report creation, during troubleshooting, and for investigating potential security breaches.
Implication	Requires initial configuration and a highly available logging infrastructure.

Table 100: Including All Nutanix Modules in Logging

SEC-010	INCLUDE ALL NUTANIX MODULES IN LOGGING.
Justification	Ensures data from all modules are included and search- able via a logging system. Refer to Nutanix Syslog documentation for additional information. Modules can be excluded if needed.
Implication	Data will be generated per module meaning the more modules included, the more log data is generated.

Table 101: Error Log Level for Nutanix Components

SEC-011	USE ERROR LOG LEVEL FOR THE NUTANIX COMPONENTS.

Justification	Ensures data from all modules are included and searchable via a logging system. Refer to Nutanix Syslog documentation for additional information. Modules can be excluded if needed.
Implication	Will not send all logging information to syslog infrastructure. The following levels will be included:ErrorCriticalAlertEmergency

Table 102: ESXi Logging Level, Rotation, and File Size

SEC-012	USE DEFAULT ESXI LOGGING LEVEL, LOG ROTATION, AND LOG FILE SIZES.
Justification	No reason to change size and rotation unless a huge amount of logs are expected.
Implication	In rare situations, logs might rotate very fast.

Table 103: Protocols for Log Transport

SEC-013	IF EXTRA SECURITY AND RELIABILITY ARE REQUIRED, THEN USE TCP FOR LOG TRANSPORT. OTHERWISE, USE THE DEFAULT SYSLOG PROTOCOL, UDP.
Justification	Provides a reliable logging setup and ensures the logs are being received by the logging infrastructure.
Implication	Requires a sender and receiver to establish communication which generates minimal additional network traffic. May require configuration of logging infrastructure to accept TCP communication.

Table 104: Using Port 514 for Logging

SEC-014	USE PORT 514 FOR LOGGING.
Justification	Defined port in syslog RFC. No reason has been identified to change the port.

Implication	Traffic on well-known ports can be easy to locate.

Network Segmentation

To protect Nutanix CVM and hypervisor traffic, place them together in their own dedicated VLAN, separate from other VM traffic. This applies to all hosts in a cluster.

Nutanix recommends configuring the CVM and hypervisor host VLAN as a native, or untagged, VLAN on the connected switch ports. This native VLAN configuration allows for easy node addition and cluster expansion. By default, new Nutanix nodes send and receive untagged traffic. If you use a tagged VLAN for the CVM and hypervisor hosts instead, you must configure that VLAN while provisioning the new node, before adding that node to the Nutanix cluster.

Do not segment Nutanix storage and replication traffic, or iSCSI Volumes traffic, on separate interfaces (VLAN or physical) unless additional segmentation is required by mandatory security policy or the use of separate physical networks. The added complexity of configuring and maintaining separate networks with additional interfaces cannot be justified unless absolutely required.

Table 105: VLAN for Traffic Separation

SEC-015	USE VLAN FOR TRAFFIC SEPARATION OF MANAGEMENT AND USER WORKLOADS.
Justification	VLAN is a well-known standard for traffic separation when there are no requirements specifying physical separation.
Implication	This provides clear separation between: Management and end user traffic, Different management traffic types, and Different end user traffic types.

Table 106: Place CVM and Hypervisor on Same VLAN and Subnet

SEC-016	PLACE CVM AND HYPERVISOR ON THE SAME VLAN AND SUBNET.

Justification	Required for Nutanix functionality. Both hypervisor and CVM should be classified as core or management service traffic delivering the same service and can therefore be placed on the same VLAN.
Implication	Different components placed on the same network segment. Alternate configurations are not supported.

Table 107: Place Out-of-band Management on a Separate VLAN or Physical Network

SEC-017	PLACE OUT-OF-BAND MANAGEMENT ON A SEPARATE VLAN OR PHYSICAL NETWORK.
Justification	Out-of-band management is not required for core Nutanix functionality. To provide additional security, the management interface should not be on the same network segment as the CVM and hypervisor.
Implication	Network or VLAN management.

Role-based Access Control (RBAC)

Nutanix has built-in RBAC in both Prism Central and Prism Element, with the option to create custom roles in Prism Central. RBAC is used to limit access and control for various individuals and groups of administrative users. Use a least-privilege and separation-of- duties approach when assigning permissions to make sure each group or individual user has just enough permissions to perform their duties. Use pre-defined roles or create new roles as needed.

Configure RBAC at the Prism Central level since it provides the overlying management construct. Via Prism Central, consumers and administrators will be directed to underlying components, such as Prism Element, when needed. This ensures your least-privilege configuration stays in place, avoiding common mistakes that occur when RBAC is configured at multiple different levels.

The following table shows Prism Central default roles:

Table 108: Default Prism Central Roles

Role	Purpose

Super Admin	Highest-level admin with full infrastructure and tenant access. Manages a Nutanix deployment and can set up, configure, and make use of every feature in the platform.
Self-Service Admin	Cloud admin for a Nutanix tenant. Manages virtual infrastructure, oversees self-service, and can delegate end-user management.
Project Admin	Team lead to whom cloud administration is delegated in the context of a project. Manages end-users within a project and has full access to their entities.
Prism Viewer	View-only admin. Has access to all infrastructure and platform features but cannot make any changes.
Prism Admin	Day-to-day admin of a Nutanix deployment. Manages the infrastructure and platform but cannot entitle other users to be admins.
Operator	Lifecycle manager for team applications. Works on existing application deployments, exercises blueprint actions.
Developer	Application developer within a team. Authors blueprints, tests deployments, and publishes applications for other project members.
Consumer	Owner of team applications at runtime. Launches blueprints and controls their lifecycle and actions.

When creating custom roles in Prism Central there are multiple entities available, each with their own set of permission definitions:

- App
- VM
- Blueprint
- Marketplace Item
- Report
- Cluster
- Subnet
- Image

The following table describes Prism Element default roles:

Table 109: Prism Element Default Roles

Role	Purpose
User Admin	Able to view information, perform any administrative task, and create or modify user accounts.
Cluster Admin	Able to view information and perform any administrative task. Cannot create or modify user accounts.
Viewer	Able to view information only. No permission to perform any administrative tasks. Useful for auditing purposes.

In addition, there are pre-defined roles in VMware vCenter Server, with the option to create custom roles.

Table 110: Use Least-Privilege Approach when Determining Access

SEC-018	USE A LEAST-PRIVILEGE ACCESS APPROACH WHEN DECIDING WHO HAS ACCESS. ALIGN RBAC STRUCTURE AND USAGE OF DEFAULT PLUS CUSTOM ROLES ACCORDING TO COMPANY REQUIREMENTS.
Justification	This approach helps protect the environment from having users performing actions they are not authorized to perform.
Implication	Access requirements need to be defined, created, and implemented for each access group.

Table 111: Align RBAC to Company Structure

SEC-019	ALIGN RBAC STRUCTURE AND USAGE OF DEFAULT PLUS CUSTOM ROLES ACCORDING TO THE COMPANY REQUIREMENTS DEFINED VIA SEC-018.
Justification	There are numerous ways of applying RBAC. This design document can’t cover all possibilities.
Implication	If no company structure exists, fall back on description in SEC-018 and create a structure that meets your needs.

Data-at-Rest Encryption

Data-at-Rest Encryption (DaRE) is a security measure to prevent data from being stolen. DaRE offers protection against data exposure in the event of:

- Activities that require a full data copy that will be used outside the platform.
- Failed drives leaving the datacenter.
- Drive or node theft.

Keeping management traffic, including storage traffic, on a separate network is often an adequate data security practice. Nutanix supports four different options for Data at Rest Encryption (DaRE):

- Self-Encrypting Drives (SEDs) with an External Key Manager (EKM).
- Software-based encryption with an EKM.
- Software-based encryption and SED with an EKM (aka Dual Encryption).
- Software-based encryption with the Nutanix Native Key Manager (KMS).

With all software-based encryption options, encryption is performed in the software layer and data is encrypted over the wire, between CVMs. All or part of the user VM data can be encrypted.

- AHV encrypts the entire Nutanix cluster.
- ESXi gives you the option to define encryption on a per-Nutanix-container basis if required.

Self-Encrypting Drives (SEDs) provide FIPS 140-2 Level 2 compliance and can be used without any performance impact. Nutanix Software Based Encryption and Native Key Manager are FIPS 140-2 Level 1 Evaluated.

You can use both software-based encryption and SEDs. This requires an external key management server.

> Note: All methods of DaRE are FIPS 140-2 compliant however if levels 2, 3, or 4 are required a hardware component is necessary. Software-based encryption with the Nutanix Native Key Manager (KMS) can encrypt storage containers (ESXi or Hyper-V) or the entire Cluster (AHV).

Table 112: Do Not Use Storage Encryption

SEC-020	DO NOT USE STORAGE ENCRYPTION.
Justification	There are no technical requirements to justify implementing storage encryption in this design.
Implication	Data is never encrypted unless encrypted at the virtual machine level or via the ESXi VM encryption feature.

Key Management Server

A key management server is required when using Nutanix storage encryption.

When using an external key management service, you should have a minimum of two key management servers running. At least one server must not run on the infrastructure it is protecting.

Table 113: Do Not Use a Key Management Server

SEC-021	DO NOT USE A KEY MANAGEMENT SERVER.
Justification	No functionality will be used that requires a key management server.
Implication	N/A

Nutanix Flow

During the environment lifecycle, it’s important to ensure compliance is met and that your security configuration meets required standards. Nutanix Flow provides a centralized view of the security posture of your Nutanix environment and provides visibility into the security of your on-prem environment based on known compliance standards.

Flow can help you to comply with regulatory and business-specific compliance policies such as PCI-DSS, HIPAA, and NIST. (note: Xi Beam is not available for dark sites)

You gain deep insights into your on-premises Nutanix deployments based on over 300 audit checks and security best practices according to:

- Audit security checks for access, data, networking and virtual machines.

- Compliance checks against PCI-DSS v3.2.1 for AHV, AOS, Flow and Prism Central.
- For more information on Nutanix Flow, see the section Nutanix Flow for Microsegmentation.

Automation Layer Design

Nutanix supports intelligent IT operations and advanced automation that enable you to streamline operations, enable IT as a Service, and support the needs of developers and business teams. This section covers automation and orchestration of virtual infrastructure, focusing on provisioning and maintenance which are important aspects of the overall solution. Nutanix tools reduce the time required to perform initial setup plus software and firmware upgrades.

Upgrades

Upgrades can occur across a variety of components. The process to upgrade many of these components is primarily executed using LCM, which is able to understand dependencies without operator intervention.

It's important to understand the impact of upgrading certain software components, as their upgrade may affect the outside world.

- AOSWhen upgrading, each CVM is individually upgraded in a rolling fashion. While a CVM is rebooting, the host it is running on is redirected to a remote CVM to deliver storage IO. This is invisible to the VM. However, it does result in a loss of locality, which can potentially impact IO throughput and latency if the load on the system is high.
- HypervisorWhen upgrading the hypervisor on a node, each VM must be migrated off the host for the update to be performed so that a reboot of the host can occur. For vSphere, this requires vCenter integration to be configured to allow for a host to be put into maintenance mode.AHV live migration and vSphere vMotion are normally non-disruptive, however certain applications that utilize polling-style device drivers or that have near-real-time kernel operations cannot tolerate being migrated during the final

cutover step. Hypervisor updates will require downtime for these apps if they don't offer native failover functionality.

Firmware

Firmware can be updated for a variety of devices including:

- Board Management Controller (BMC)
- Motherboard (BIOS)
- Host Bus Adapter (HBA)
- Host Boot Device (SATA DOM)
- SSD or HDD
- NIC

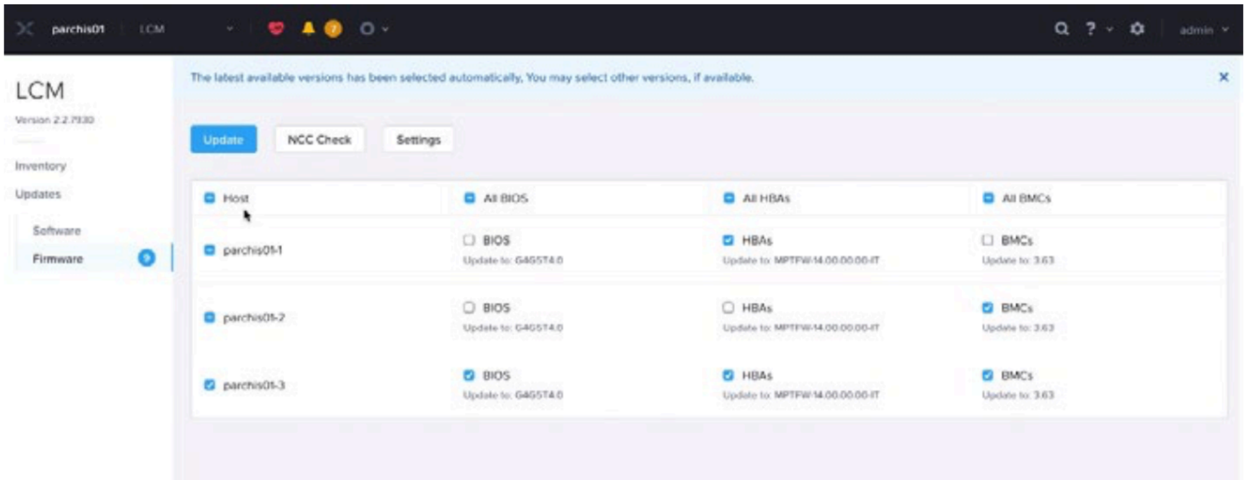
Software

Software updates can be performed for multiple components, including:

- Acropolis Operating System (Core AOS in CVM)
- Hypervisor (vSphere or AHV)
- Nutanix Cluster Check (NCC)
- Life Cycle Manager (LCM)
- Acropolis Operating System (AOS in Prism Central)
- MicroServices Platform (MSP in Prism Central)
- Services Functionality
  - Nutanix Files (Prism Element)
  - Calm/Epsilon (Prism Central)
  - Objects Manager (Prism Central)
  - Karbon (Prism Central)
  - Era (independent)
  - Move (independent)
  - X-Ray (independent)

Life Cycle Manager (LCM)

Nutanix Lifecycle Manager (LCM) improves the efficiency and reliability of IT infrastructure upgrades in modern datacenters. Nutanix LCM determines any software and firmware dependencies, intelligently prioritizes updates, and automates the entire upgrade process across all clustered hosts, without any impact to applications or data availability. It supports one-click upgrades across multiple qualified hardware manufacturers and configurations, so IT teams have the flexibility to deploy the best hardware for each use case – and still benefit from centralized upgrade capabilities. LCM supports AHV, ESXi and Hyper-V hypervisors, plus hardware support for Nutanix, Hewlett Packard Enterprise, Dell/EMC, Lenovo, Fujitsu, Inspur and Intel. LCM supports environments with internet connectivity as well as dark site deployments.



The latest available versions has been selected automatically, You may select other versions, if available.

▣ Host	▣ All BIOS	▣ All HBAs	▣ All BMCs
▣ parchis01-1	□ BIOS Update to: G4G5T4.0	☑ HBAs Update to: MPTFW-14.00.00.00-IT	□ BMCs Update to: 3.63
▣ parchis01-2	□ BIOS Update to: G4G5T4.0	□ HBAs Update to: MPTFW-14.00.00.00-IT	☑ BMCs Update to: 3.63
☑ parchis01-3	☑ BIOS Update to: G4G5T4.0	☑ HBAs Update to: MPTFW-14.00.00.00-IT	☑ BMCs Update to: 3.63

Figure 51: Life Cycle Manager

Foundation

Foundation manages the initial setup and configuration of a cluster. Nutanix nodes may come pre-installed with AHV and the Controller Virtual Machine (CVM) and you can:

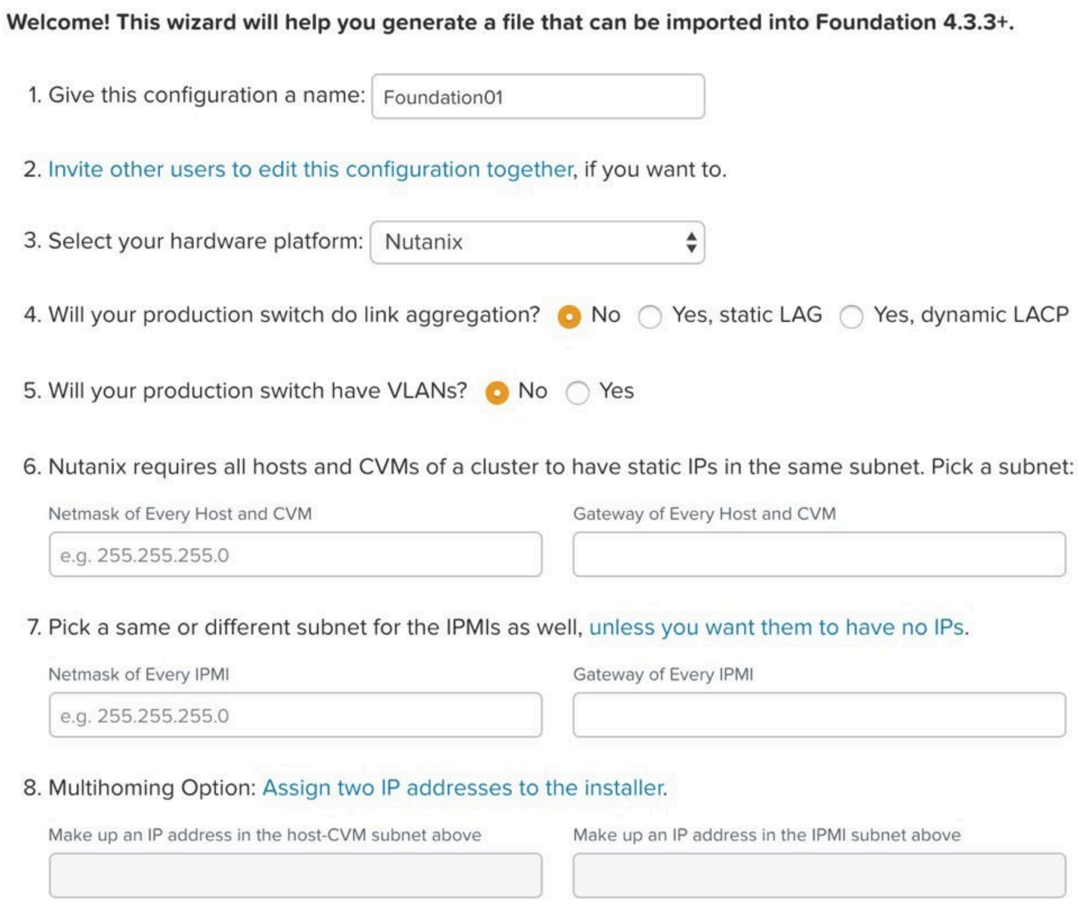
- Add the Nodes to an existing Nutanix cluster.
- Create a new Nutanix cluster.
- Re-image the nodes with different AHV and or AOS version or different hypervisor and create a Nutanix cluster.

There are five ways of invoking a Foundation process:

- Connect to a factory deployed Nutanix node's CVM via http://CVM_IP:8000/
- Use portable Foundation, Mac or Windows executable.
- Standalone Foundation VM.
- When adding a node to an existing Nutanix cluster.
- Call Foundation API via third party solution which provides a way to orchestrate the deployment.

> Note: The Foundation process may vary for different hardware vendors.

Nutanix provides a Foundation pre-configuration service which is accessible via install. nutanix.com. Via the service, you can define and download the Nutanix cluster configuration to be used during the Foundation process. The downloaded file contains all configuration required to perform the foundation operation and comes in json format. This makes it easy to keep track of configurations, document them, and share with your peers.



Welcome! This wizard will help you generate a file that can be imported into Foundation 4.3.3+.

1. Give this configuration a name: Foundation01

2. Invite other users to edit this configuration together, if you want to.

3. Select your hardware platform: Nutanix

4. Will your production switch do link aggregation? ● No ○ Yes, static LAG ○ Yes, dynamic LACP

5. Will your production switch have VLANs? ● No ○ Yes

6. Nutanix requires all hosts and CVMs of a cluster to have static IPs in the same subnet. Pick a subnet:

Netmask of Every Host and CVM: e.g. 255.255.255.0

Gateway of Every Host and CVM:

7. Pick a same or different subnet for the IPMIs as well, unless you want them to have no IPs.

Netmask of Every IPMI: e.g. 255.255.255.0

Gateway of Every IPMI:

8. Multihoming Option: Assign two IP addresses to the installer.

Make up an IP address in the host-CVM subnet above:

Make up an IP address in the IPMI subnet above:

Figure 52: Foundation Pre-Configuration Service

The following table provides an overview of the functions available for the different Foundation software options.

Table 114: Overview of Different Foundation Software Options

	CVM Foundation	Portable Foundation (Windows, Mac)	Standalone Foundation VM
Function	Factory-imaged nodes.	Factory-imaged nodes.	Maintains high availability in the event of the loss of one switch.

Bare-metal nodes	Factory-imaged nodes.		Requires at least two TOR switches, which will impact cost and rack space. Increases total bandwidth to each host to 20 Gbps.
Bare-metal nodes			
Hardware	Any	Nutanix (NX)	
Dell (XC)			
HPE (DX)	Any		
If IPV6 is disabled	Cannot image nodes.	IPMI IPv4	
required on the nodes.	IPMI IPv4 required on the nodes.		
VLAN Support	No	No	Yes
LACP Support	Yes	Yes	Yes
Multi-homing Support	N/A	Yes	Yes

Note: When available, Foundation Central will provide capabilities to perform Foundation operations via Prism Central.

Operations Design

Capacity and Resource Planning

After initial installation and migration of workloads to the platform, long term capacity planning should be enabled to avoid running out of resources. Many of the features discussed are integrated into Prism Pro.

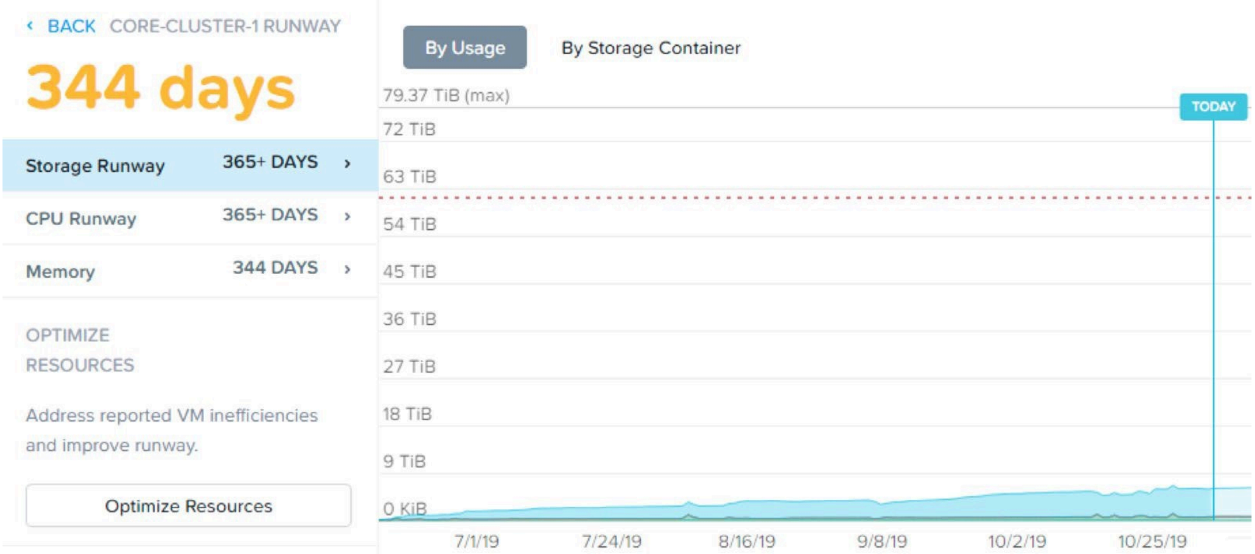
Table 115: Deploying Prism Pro for Enhanced Cluster Management

OPS-001	DEPLOY PRISM PRO FOR ENHANCED CLUSTER MANAGEMENT

Justification	Analytics, Capacity planning, Custom Dashboards, and Playbooks require the presence of Prism Pro.
Implication	Requires an additional license, which can increase cost.

Cluster capacity

Native runway calculations built into Prism Central will automatically calculate the remaining capacity of the system as soon as the cluster Prism Element is brought under management by Prism Central.



‹ BACK CORE-CLUSTER-1 RUNWAY

344 days

Storage Runway	365+ DAYS	›
CPU Runway	365+ DAYS	›
Memory	344 DAYS	›

OPTIMIZE RESOURCES

Address reported VM inefficiencies and improve runway.

Figure 53: Capacity Runway

These runway calculations should be configured to run as part of periodic reports and reviewed on a regular basis to ensure sufficient capacity exists. This is especially important for organizations that have a significant lag between the date a commitment to purchase additional gear occurs and when its online and available to use.

Expansion planning

When a new workload is identified to be onboarded, planning needs to occur if the new workload size or requirements are outside of established patterns. Prism Central offers a scenario simulation function that shows how the available capacity runway would change if this new workload was accommodated. Utilizing this planning functionality helps avoid unplanned capacity constraints.

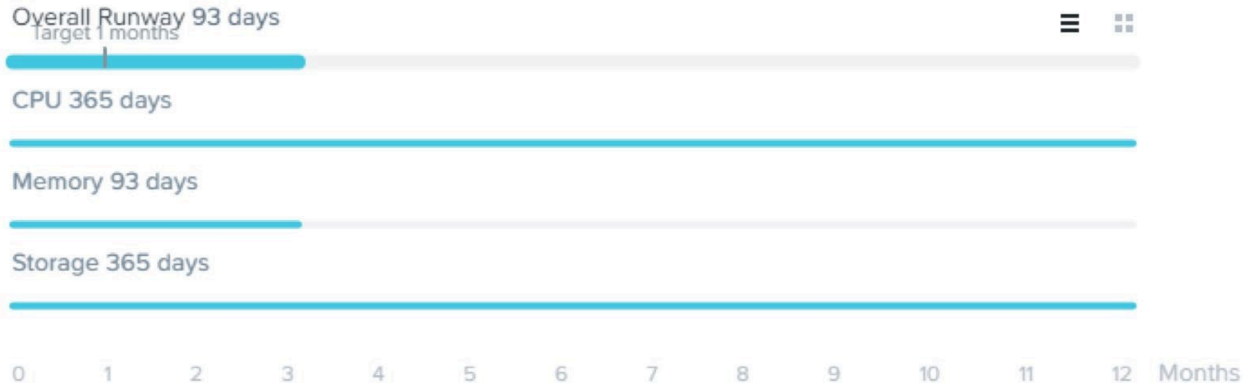


Figure 54: Expansion Planning

Table 116: Reviewing Monthly Capacity Planning

OPS-002	REVIEW MONTHLY CAPACITY PLANNING
Justification	Analytics, capacity planning, custom dashboards, and playbooks require Prism Pro.
Implication	Requires an additional license, which can impact cost.

Right-Sizing VMs

While general system capacity planning is useful, accurate and efficient system level planning requires accurate sizing for individual workloads. Machine learning in Prism Central provides anomaly detection for VMs when the workload crosses learned thresholds.

In addition, based on a number of thresholds, the system will categorize VMs based on their behavior. These categories include:

- Bully: A VM with potential to degrade the overall cluster performance by impacting the capability of the node they reside on.
- Constrained: A VM that is experiencing very high CPU or RAM utilization and potentially is unable to meet the needs of the application its running and is likely a candidate to increase the resources provided.

- Over-provisioned: A VM that is significantly underutilizing the resources provided to it and is likely a candidate to reduce in size.
- Inactive: A VM that has not powered on recently (dead) or is virtually idle (zombie) that is a candidate for reclamation.

In addition, custom alert policies can be created that match VMs by conditions.

Upgrade Methodology

This section describes the design decisions associated with upgrading the Nutanix infrastructure. Whether to perform upgrades during business hours or outside business hours comes down to a number of possible factors, including:

- Is the environment sized to tolerate failures during upgrades?
- Will performance be adequate during upgrades?
- Past experiences.
- Size of Nutanix cluster.

Table 117: Performing Updates

OPS-003	PERFORM UPDATES AFTER HOURS FOR PERFORMANCE OR MIGRATION SENSITIVE APPLICATIONS
Justification	Reduces load on the system to be handled during rolling CVM reboot from code upgrade.
Implication	Potential staffing impact.

When upgrading, AOS offers the choice of release trains to apply. Each of these is denoted with a major.minor.maintenance.patch numbering scheme, for example: 5.10.7.1 or 5.11.1.1.

A release train is based on the major and minor components. There are two types of release trains:

- Short Term Support (STS) releases which include new features and provide a regular and frequent upgrade path. These releases are maintained for shorter durations.

- Long Term Support (LTS) releases which provide bug fixes for features that have been generally available for a longer period of time. After features have been generally available in an STS for some time, they are included in and LTS, which are maintained for a longer duration.

KB 5505 on the Nutanix Support Portal covers the differences in greater detail.

Table 118: Utilizing the Current LTS Branch

OPS-004	UTILIZE THE CURRENT LTS BRANCH
Justification	Unless a new feature is required for a design, enterprise customers typically prefer a less frequent major version change of software components.
Implication	Lack of access to features present in or requiring new branch.

Updates to an existing train are released on a regular basis and should be applied on a standard cadence.

Table 119: Upgrading to Maintenance and Current Patch Versions

OPS-005	UPDATE TO THE NEXT MAINTENANCE VERSION 4 WEEKS AFTER RELEASE.
	UPDATE TO THE CURRENT PATCH VERSION 2 WEEKS AFTER RELEASE.

Justification	Unless a new feature is required for a design, enterprise customers typically prefer a less frequent major version change of software components.
Implication	Frequent smaller updates keep the amount of changes per update relatively small, minimizing the amount to troubleshoot in the event something goes wrong.

Testing

Environments which seek to achieve high overall uptime greater than 99.9% should build a pre-production environment to mimic production so that configuration changes can be tested before being pushed into production.

Table 120: Maintaining a Pre-Production Environment

OPS-006	MAINTAIN A PRE-PRODUCTION ENVIRONMENT FOR TESTING ANY CHANGES NEEDED (FIRMWARE, SOFTWARE, HARDWARE) PRIOR TO EXECUTING THE CHANGE IN PRODUCTION.
Justification	Environments typically have stringent uptime due to associated financial penalties or losses that occur from unavailable services. The cost of an outage with a realistic duration often outweighs the pre-production environment cost.
Implication	A pre-production environment will increase the overall cost of the solution.

Monitoring

Nutanix includes a variety of built-in, system-level monitoring functions. The relevant metrics for built-in monitoring are automatically gathered and stored without user intervention required.

Native cluster alerts can be sent from either the individual cluster or Prism Central.

Table 121: Configuring Alerts and Alert Policies

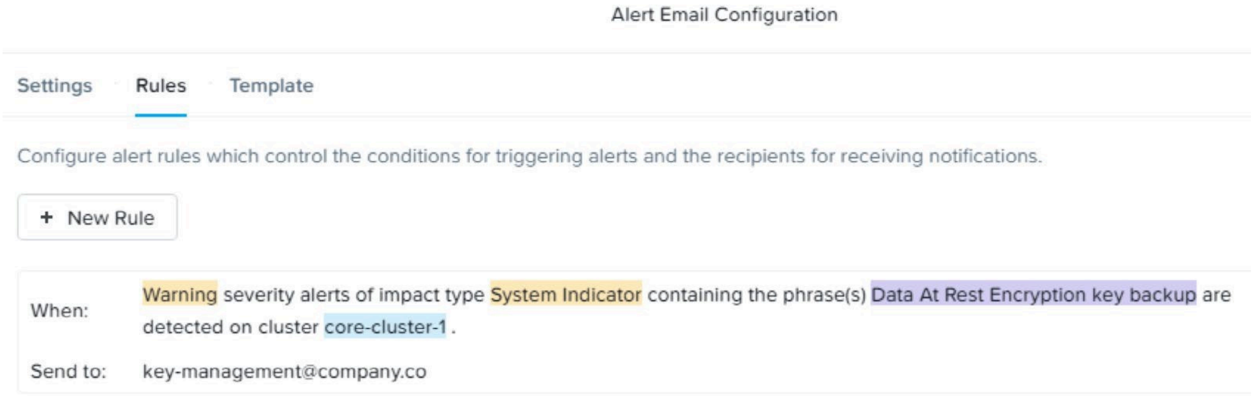
OPS-007	CONFIGURE ALERTS AND ALERT POLICIES IN PRISM CENTRAL, NOT PRISM ELEMENT.
Justification	Creates consistency across multiple clusters and reduces effort when making multiple changes. Also allows for anomaly detection.
Implication	Requires Prism Central to receive alerts.

When alerts are generated, in addition to raising the alert in Prism, the system can generate an outbound message. The two options available for sending alerts are SNMP and SMTP.

Table 122: Utilizing SMTP for Alert Transmission

OPS-008	UTILIZE SMTP FOR ALERT TRANSMISSION
Justification	Creates consistency across multiple clusters and reduces effort to make multiple changes. Also allows for anomaly detection. Offers more options to customize delivery.
Implication	Requires access to SMTP systems.

The following screen shows how to configure Prism to send alerts to a specified email account via SMTP.



Configure alert rules which control the conditions for triggering alerts and the recipients for receiving notifications.

When: Warning severity alerts of impact type System Indicator containing the phrase(s) Data At Rest Encryption key backup are detected on cluster core-cluster-1 .

Send to: key-management@company.co

Figure 55: Alert Email Configuration

5. Multi-Datacenter Design for Nutanix Core and BC/DR

Choosing the Right BC/DR Solution

General Guidance

Choosing the right solution for business continuity and disaster recovery is one of the key decisions every architect must make. A mistake can cause significant pain to the business and millions of dollars of losses.

The two key business requirements when designing BC/DR solutions are Recovery Point Objective (RPO) and Recovery Time Objective (RTO). In addition to RPO and RTO – Work Recovery Time (WRT) and Maximum Tolerable Downtime (MTD) should be taken into account.

- RPO: The maximum acceptable amount of data loss measured in time
- RTO: The maximum tolerable amount of time needed to bring a critical system back online
- WRT: The maximum tolerable amount of time that is needed to verify the system and/or data integrity
- MTD: The total amount of time that a business process can be disrupted without causing any unacceptable consequences

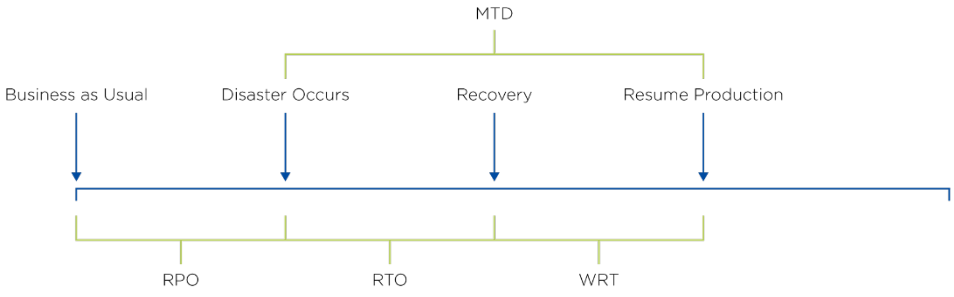


Figure 56: General Guidance for Business Continuity

When designing new BCDR solution make sure to understand your business requirements and technical constraints. Nutanix provides a range of protection mechanisms to help with data protection. The general advice when designing BC\DR is:

If an application has native data protection mechanisms (e.g. Exchange DAG, MS SQL AlwaysOn, Oracle RAC) leverage those to provide application data protection. If an application does not provide a data protection mechanism, use Nutanix built-in protection mechanisms.

The flowchart below can help guide your decision making.

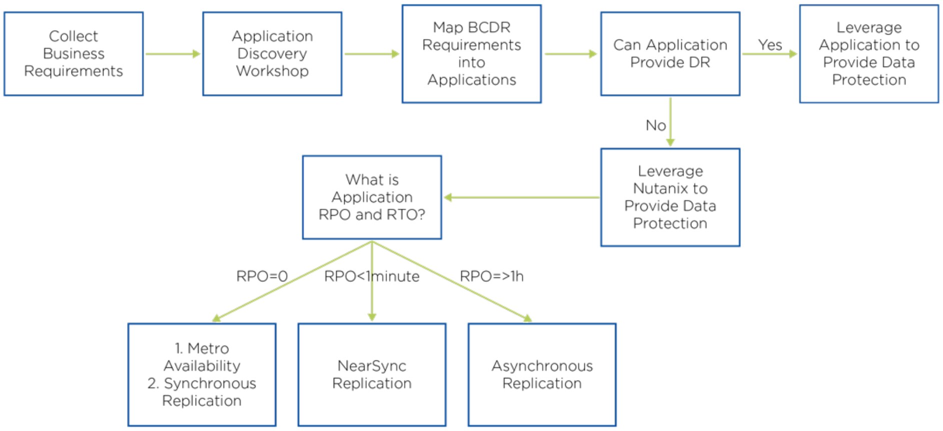

Collect Business Requirements
Application Discovery Workshop
Map BCDR Requirements into Applications
Can Application Provide DR
Yes
Leverage Application to Provide Data Protection
No
Leverage Nutanix to Provide Data Protection
What is Application RPO and RTO?
RPO=0
RPO<1minute
RPO=>1h
1. Metro Availability 2. Synchronous Replication
NearSync Replication
Asynchronous Replication

Figure 57: BCDR Workflow

DR Target Considerations

Disaster recovery requires a secondary location that is far enough away from the primary location to ensure that it won’t be affected by the disasters you need to protect against but close enough that latency and bandwidth don’t become a limiting factor. You can use multi-site DR, as described later to protect workloads using a combination of sites at different distances.

Another important consideration is whether DR will utilize on-premises locations or an option in the cloud:

- On-premises: DR using on-premises target locations gives you complete control over (and responsibility for) your DR operations. However, it may require significant CapEx and ongoing operating expense to maintain adequate resources to support failover if it becomes necessary.

- Nutanix Xi Leap: DR-as-a-Service (DRaaS) offloads much of the complexity and cost of DR. Xi Leap is a simple-to-configure DR service available to Nutanix customers. To learn more visit nutanix.com/leap
- Nutanix Clusters: Nutanix Clusters allow you to operate a full Nutanix environment in the public cloud. You can utilize Nutanix Clusters as a DR target and take advantage of cloud resources while maintaining full control and complete operational continuity with on-premises Nutanix AOS environments. (See the section Nutanix Clusters for Hybrid and Multicloud Operations for more information.)

Multi-Datacenter Design

Datacenter Resiliency Patterns

When operating applications across multiple physical locations for the purposes of improved resiliency, there are two common sets of operating patterns.

Active-Active

In an active/active pattern, two (or more) datacenters provide resources to actively support of a service. For this definition, the application is expected to be fully capable of continuing to service requests in the event one of the active AZs goes down. While it is up to the specific application architecture to determine exactly how the components are split up and their fault/partition tolerance, active sites are typically operated with a low latency constraint of <= 5ms RTT.

In the event of a localized disaster affecting only one availability zone, the application maximum throughput/capability is reduced proportional to the amount of resources lost in the down AZ.

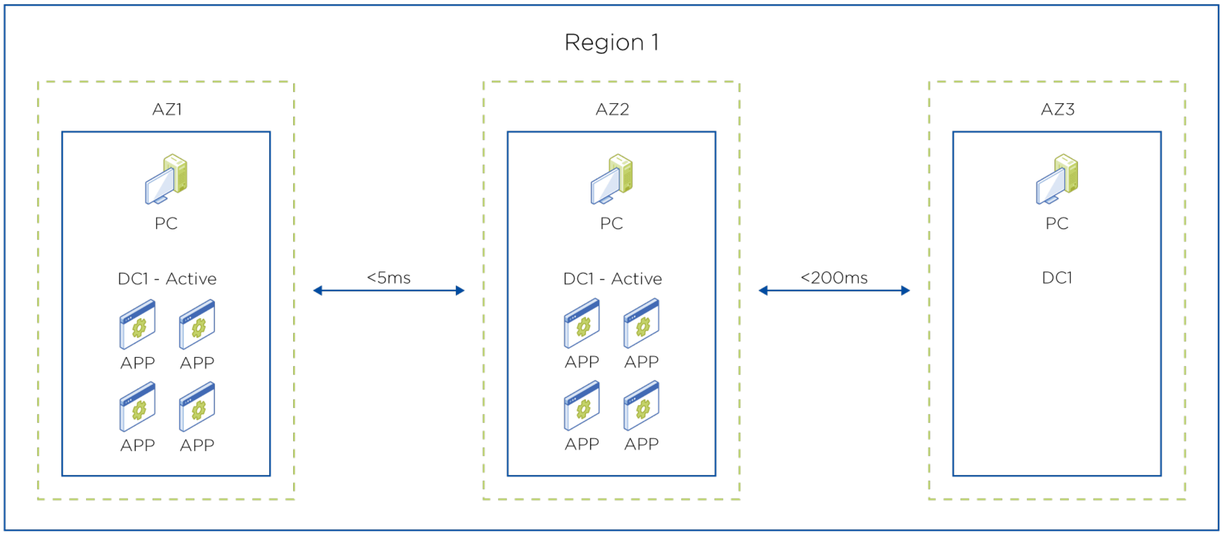


Figure 58: Active – Active

Active-Passive

In an active/passive scenario, only one datacenter has active resources supporting a service. An additional datacenter is available to provide services in the event that the primary datacenter goes down. For some application architectures (sometimes referred to as warm standby) the passive facility has resources configured and ready, however under normal conditions those resources are not servicing user requests. These warm standby applications are delineated from true passive systems, where the data is replicated, but the compute is not ready and waiting.

A hybrid model is also possible, where a portion of the environment in each datacenter is active and a portion is a standby for another site. For example, business unit 1 (BU1) may run actively out of AZ1 and passively out of AZ2 due to the application design, while BU2 is actively running out of AZ1 and AZ2.

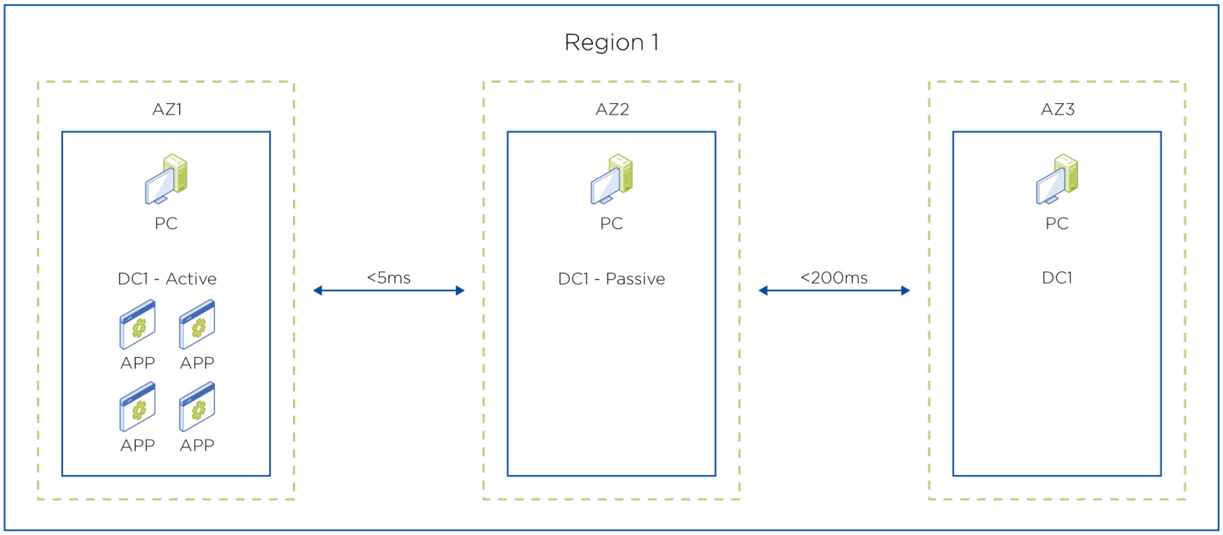


Figure 59: Active - Passive

Multi-Site DR

In either of the above patterns, application components may be further protected via an N+1 site providing DR capacity. Typically, the DR site is located either geographically far away from the primary system inside of the same region, or in another region entirely. This allows for better isolation from a widespread disaster (hurricane, large -scale flooding, tier 1 ISP issues), however the physical distances often create greater latency.

Replication Options

Snapshots & Clones

To provide full coverage data protection, Nutanix includes a number of native features, as well as integration with third-party backup vendors.

As a first level of protection (local recovery), an individual cluster can create a snapshot of a vDisk or an entire VM. Nutanix snapshots are redirect-on-write, where the pointers that indicate which physical disk extent maps to a virtual disk are updated only as data changes. This technique creates little to no performance overhead for a running workload when a snapshot exists. Cloned entities use the same principle and enjoy the same benefit. Additional data capacity is not used by the snapshot or clone unless changes are made to it.

The minimal overhead makes it low risk to protect as many workloads as necessary with snapshots. Protected VMs can have differing levels of retention and snapshot frequency, allowing for as much flexibility as needed to work around application scheduling requirements.

When choosing a retention schedule, choose the minimum number of snapshots that meet RPO requirements. Each snapshot requires additional metadata space to track the related extents on disk, and excessively deep snapshot chains can increase the background load on the CVM as part of its maintenance processing.

There are two options available for snapshots. Snapshots that invoke a method for quiescence inside the guest VM operating system are termed application-consistent, and those that do not inform the guest VM operating system are termed crash-consistent. AOS can trigger an application-consistent snapshot via Nutanix Guest Tools and/or VMware tools, depending on the hypervisor in use.

Guest limitations may prevent proper quiescence for very large machines (for example Windows VSS cannot handle more than 64 devices/drives concurrently).

Table 123: Utilizing Application-consistent Snapshots

BCN-001	UTILIZE APPLICATION CONSISTENT SNAPSHOTS WHEN NEEDED BY THE APPLICATION
Justification	VM/application quiescence allows for a “clean” snapshot from the application’s viewpoint.
Implication	Application quiescence can potentially take a significant amount of time, and the throughput of the application may be adversely affected. Consistency groups should have less than 20 members

For ESXi hypervisor triggered clones, AOS supports offloading the clone via VAAI. The VM must meet vSphere’s requirements to trigger offloading (VM must be powered off and the clone must reside on the same datastore as the primary VM).

Local snapshots and clones provide an easy restoration point on the local cluster in the event something goes wrong with the VM itself (OS bluescreen after patches, recover to a point before ransomware encryption, etc.). These snapshots can be instantiated in seconds, allowing for a very short RTO until VM boot. Via API integration, 3rd party backup vendors can utilize Prism REST API endpoints to minimize backup load and space required.

Nutanix also provides the ability to replicate data between clusters, where the data point referenced in the snapshot is stored on one or more remote clusters. As part of the snapshot schedule, these remote clusters can have a different retention period than the primary. An example of this would be keeping three daily copies locally on a performance-oriented all flash cluster, and 14 daily copies on a remote hybrid cluster.

There are two different forms of snapshots to support different modes of replication:

- Full snapshots for asynchronous replication (with RPO of 60 minutes or greater).
- Lightweight snapshots (LWS) for NearSync replication (when the RPO is between 15 minutes and 1 minute).

Full snapshots are efficient for keeping system resource usage low when you are creating many snapshots over an extended period of time. LWS reduces metadata management overhead and increases storage performance by decreasing the number of storage I/O operations that long snapshot chains can cause.

Nutanix snapshots are a vital first element of an overall data protection strategy, however they are not a substitute for a full backup methodology. Many organizations subscribe to what’s referred to as the 3-2-1 rule, which mandates 3 copies of data, on 2 backup mediums, with at least 1 copy offsite.

Snapshots can be replicated to multiple destination clusters to provide multiple redundant copies if needed. These arrangements are configured as part of the snapshot schedule by specifying each remote site that should receive a copy, and how many to retain at that site. One to many relationships are allowed (Site

A-> Site B, Site A -> Site B, Site A –> Site N), but cascading relationships (Site A -> Site B -> Site N) are not supported.

Direct Cluster to Cluster Backup and Recovery

When creating the configuration and schedules for snapshots, there are multiple options. Prism Element on each cluster offers a construct called a Protection Domain (PD), which is a grouping of VMs. During a failover event, all the entities in the Protection Domain are activated at the remote site as a group. Each VM must be placed manually in an applicable PD, either via Prism or by an API call.

The RPO (time between successive snapshots) chosen for these entities will determine whether the system uses traditional asynchronous snapshots, or lightweight snapshots.

Recommendations: Protection Domains

- No more than 200 VMs per PD.
- No more than 10 VMs per PD with NearSync.
- Group VMs with similar RPO requirements in the same PD.

Note that size and capacity of a node can limit the RPO achievable for VMs running on it. See AOS Snapshot Frequency for Nutanix Nodes on the Nutanix Support portal for details.

Consistency Groups

Administrators can create a Consistency Group for VMs and volume groups that are part of a Protection Domain where you want to snapshot all members of the group in a crash-consistent manner. All entities in the same Consistency Group will have their snapshot execute at exactly the same time. Entities that are not in a Consistency Group, but part of the same schedule will typically snapshot close together, but not at the exact same moment (potentially a few seconds apart).

Recommendations: Consistency Groups

- Keep Consistency Groups as small as possible. Limit Consistency Groups to fewer than 10 VMs.

- Consistency Groups should only be used for applications with a shared state, such as database replicas.
- A Consistency Groups using application-consistent snapshots can contain only 1 VM.

Table 124: Place Nearsync VMs in their Own Protection Domain

BCN-002	PLACE NEARSYNC VM'S IN THEIR OWN PROTECTION DOMAIN.
Justification	NearSync can only have one schedule so place NearSync VMs in their own PD.
Implication	Multiple PDs will need to be created for different NearSync VMs depending on their schedule requirements.

Snapshot Schedule and Retention Policy

Full Snapshots and Async Replication

Your RPO determines how much data you will lose in the event of a failure. You can create multiple schedules for a Protection Domain using full snapshots at various frequencies with different retention policies.

The snapshot interval should be shorter than the desired RPO to allow for failure and recovery, without manual intervention. This is ideally at least twice as fast as your desired RPO, including data transmission timing. For example:

- RPO objective: data should be no more than 4 hours old
- Snapshot schedule: every 4 hours, starting at 12 AM, with retention on a remote cluster
- Average replication of changes takes approximately 30 minutes.
- The source site experiences an outage at 4:20 AM

In the scenario above because replication of the 4AM snapshot has not completed, the most recent snapshot available is older than 4 hours, violating the RPO goal. A 2-hour (or more frequent) snapshot schedule would avoid this RPO violation.

Table 125: Configure Snapshot Schedules to be More Frequent

BCN-003	CONFIGURE SNAPSHOT SCHEDULES TO BE MORE FREQUENT THAN THE DESIRED RPO
Justification	Cross cluster replication prevents the instant transfer of snapshot contents, increasing the time before snapshot contents are available.
Implication	Desired RPO may be achievable given technology limitations, or the RPO may need to be adjusted to accommodate replication time.

Table 126: Configure Snapshot Schedules to Retain Lowest Number of Snapshots

BCN-004	CONFIGURE SNAPSHOT SCHEDULES TO RETAIN THE LOWEST NUMBER OF SNAPSHOTS WHILE STILL MEETING THE RETENTION POLICY.
Justification	Metadata space management on a cluster is more efficient with a lower number of snapshots.
Implication	Multiple schedules should be created for the same Protection Domain at different levels rather than a simple daily schedule.

LWS and NearSync Replication

Nutanix offers NearSync with a telescopic schedule (time-based retention). When the RPO is set to be #15 minutes and #one minute, you have the option to specify the maximum retention. The system will automatically roll up intermediate recovery points within the retention window specified, based on the table below. Multiple schedules cannot be created with NearSync.

The following table represents the schedule to save recovery points for 1 month:

Table 127: Schedule to Save Recovery Points

TYPE	FREQUENCY	RETENTION
Minute	Every minute	15 minutes
Hourly	Every hour	6 hours

Daily	Every 24 hours	7 days
Weekly	Every week	4 weeks
Monthly	Every month	1 month

See Nearsync Requirements and Limitations on the Nutanix Support Portal for more information.

Synchronous Data Transfer

For applications that need to meet an RPO goal of zero (simultaneous data updates at multiple sites), Nutanix supports a synchronous replication option. Configuration of synchronous replication also utilizes Protection Domains as described above.

Configuration of synchronous replication requires a container on the source and destination with the same name. The two clusters participating in synchronous replication are required to be within 5ms round trip time (RTT) of each other.

By default, systems protected synchronously will also automatically create a periodic snapshot schedule. These snapshot references are used to reduce the recovery time in the event that the data must be resynchronized, for example after a WAN outage.

Intermixing Protection Levels

Depending on the version of AOS in use, hypervisor, number of sites, and whether you are using Protection Domains or Protection Policies, the supported configuration may differ. For Protection Domains (Prism Element), refer to the following two tables for supported intermixing of protection levels.

Table 128: Protection Domain 2-Site Support Matrix

REPLICATION TYPE	2 SITES
Async	AHV, ESX
NEAR Sync	AHV, ESX
Synchronous	ESX

Table 129: Protection Domain Multi-site Support Matrix

REPLICATION COMBINATION	DIFFERENT ENTITIES	SAME ENTITIES
Multiple Async only combination (1:N)	AHV, ESXI	AHV, ESX
Multiple Near-Sync only combination (1:2)	AHV, ESX	AHV, ESX
Multiple Sync only combination (1:2)	ESX	Not Supported
Async + Near Sync (1:2) *	AHV, ESX	AHV, ESX
Async + Sync (1:2)	ESX	ESX
Sync + Near Sync (1:2)	ESX	ESX
Async + Near Sync + Sync (1:3) *	ESX	ESX

* STS only, 5.19 and later

If you are using Prism Central Protection Policies, the following two tables show the two site and multi-site intermixing support matrix.

Table 130: Protection Policy 2 Site Support

REPLICATION TYPE	2 SITES
Async	AHV, ESX
Near Sync	AHV, ESX
Synchronous	AHV*

* STS only, 5.17 and later

Table 131: Protection Policy Multi-site Support Matrix

REPLICATION COMBINATION	DIFFERENT/SAME ENTITIES
Multiple Async only combination (1:N)*	AHV, ESXI
Multiple Near-Sync only combination (1:2)	AHV, ESX
Multiple Sync only combination (1:2)	Not Supported

Async + Near Sync (1:2) *	AHV, ESX
Async + Sync (1:2)	Not Supported
Sync + Near Sync (1:2)	Not Supported
Async + Near Sync + Sync (1:3)	Not Supported

* STS only, 5.19 and later

For additional details, see Data Protection and Recovery with Prism Element on the Nutanix Support Portal.

In all cases, a failover action causes all items within the identified PD to be made available at the remote site as a collective unit. Only related entities should be grouped into a PD.

Table 132: Group Applications Together in Unique Protection Domains

BCN-005	GROUP APPLICATIONS TOGETHER IN UNIQUE PROTECTION DOMAINS. KEEP NUMBER OF VM'S PER PROTECTION DOMAIN AS SMALL AS REASONABLY POSSIBLE
Justification	Small groupings of related entities help minimize accidental collateral damage during a partial failover and reduce the need to track complex intermingled scheduling.
Implication	Entire site failover will require activating a larger number of Protection Domains, potentially increasing the time required to bring all entities online.

Leap

Protection Domains are a construct within Prism Element. One of their limitations however is that they don't provide additional controls around VMs associated with a large-scale failover, like power-on order control, or changing the VM IP address due to site change.

Prism Central introduced a feature called Leap, which allows for more advanced constructs relating to replication and failover.

Protection policies offer a way to configure replication grouping using categories. VMs that match a category are automatically protected, versus a Protection Domain where they must be manually added.

Leap also provides Recovery Plans. Recovery Plans allow for controlling the failover process. They allow you to specify:

- Power-on order
- Mapping networks when using DHCP
- Changing static IPs for sites that don't have access to stretch L2 networking
- Scripts to perform in guest actions as part of failover (if necessary)
- Perform test failovers and validation of the configuration without downtime
- With Leap, you can perform test failovers and validate your configuration without downtime.

Full coverage of leap is out of scope for this document. Additional details can be found in the Leap Administration Guide found on the Nutanix Support Portal.

Additional Options

Nutanix offers a number of additional options for ensuring business continuity that are beyond the scope of this design:

- Xi Leap: Nutanix (or partner) hosted datacenters provide a disaster recovery as a service option. This utilizes the Leap feature of Prism Central, where data is replicated to a managed cloud service location similar to how it would be replicated to an on-prem DR location.
- Nutanix Mine: As part of the conceptual 3-2-1 rule, Nutanix supports multiple backup vendors that can integrate via API to offload backup. Mine offers a bundle of backup vendor software and Nutanix platform to provide storage for the retention of backup data, independent of the primary cluster(s) in use. (See the section Nutanix Mine for Integrated Backup for more information on Mine.)

Disaster Recovery Decision Tree

A good DR solution should involve different levels of service for DR because applications will have different levels of criticality to the business and therefore different requirements for their disaster recovery needs. As a result, this document will include multiple architectures specific to meet a variety of scenarios that as a group can be configured to meet the requirements for all application needs. The following flow chart walks through the decision tree for choosing the best Nutanix DR solution to meet your requirements. Each DR solutions is discussed in the sections that follow.

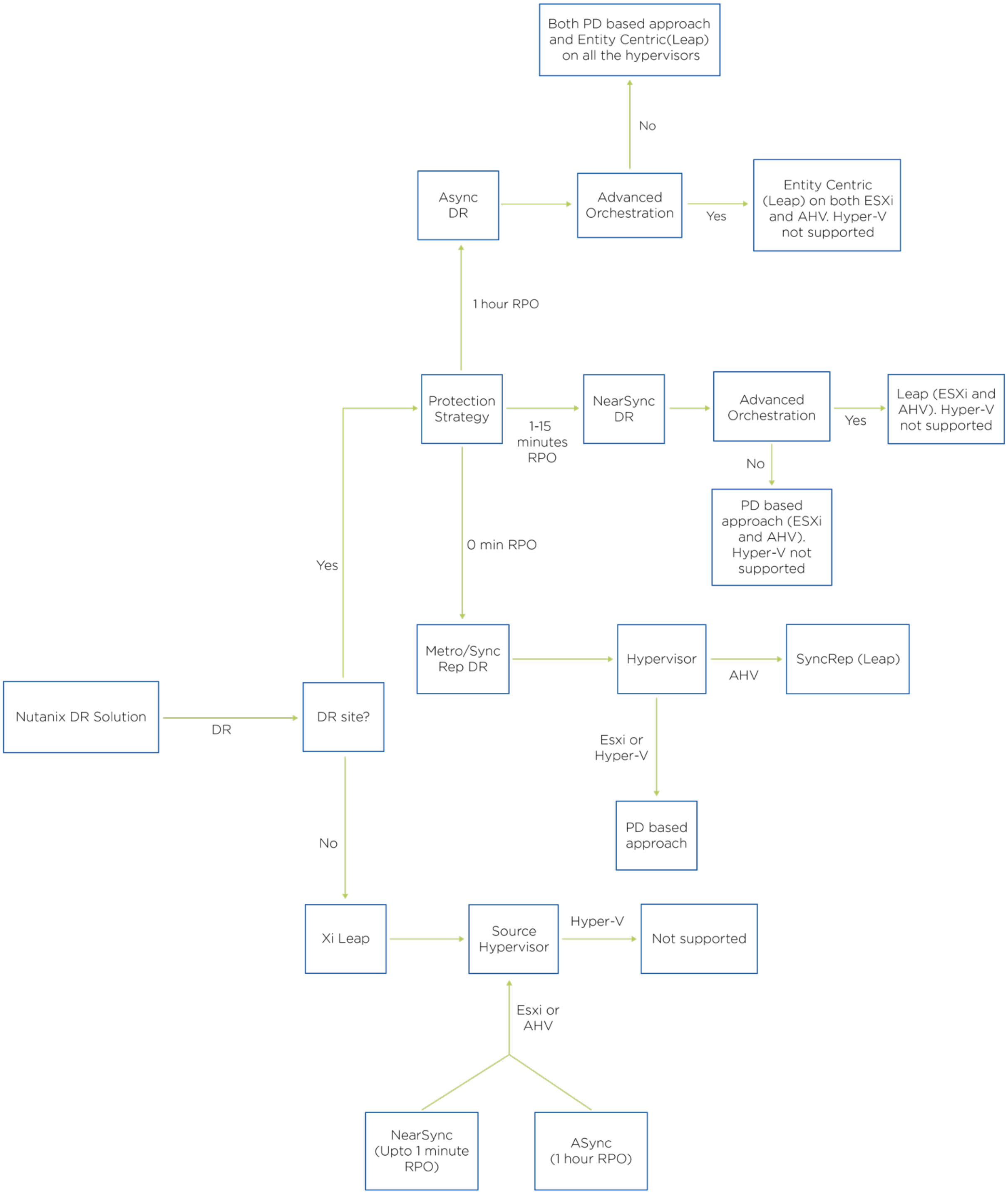


Figure 60: Disaster Recovery Decision Tree

Nutanix DR solutions

Nutanix Metro Availability

Metro Availability (MA) allows vSphere administrators to leverage hypervisor clustering technology across datacenters. This type of configuration presents two distinct Nutanix cluster to vSphere as a single stretched cluster, and it helps to minimize downtime during unplanned outages. In the event an entire site goes down, vSphere HA automatically starts the virtual machines on the secondary site if a witness has been properly configured.

This configuration is the most complicated choice, requiring stretched layer 2 networking of some form between sites. In addition, while this provides the most rapid failover, the configuration is highly dependent on the network connectivity being resilient and providing a high enough throughput to accommodate application change rates. Systems with frequent intermittent network issues may be better served via a NearSync style connection that is less impacted by network drops.

See the Metro Availability Guide on the Nutanix Support Portal for more information.

Table 133: Application Requires RPO

BCN-006	APPLICATION REQUIRES RPO=0 AND RTO NEAR ZERO
Justification	Business critical applications requires zero data loss and minimal to zero recovery time.
Implication	Nutanix MA requires high bandwidth and low latency (under 5ms). To comply with near 0 RTO, stretched layer 2 networks are required. Nutanix MA is supported only with VMware ESXi, which may bring additional management overhead and increase licensing costs.

Table 134: Application Requires Zero Downtime

BCN-007	APPLICATION REQUIRES ZERO DOWNTIME DR AVOIDANCE SOLUTION
Justification	Business critical application requires zero data loss and zero recovery time during DR avoidance event.
Implication	Nutanix MA requires high bandwidth and low latency (under 5ms). To comply with ZERO downtime during DR avoidance, stretched layer2 networks are required. Nutanix MA is supported only with VMware ESXi, which may bring additional management overhead and increase licensing costs. Nutanix AOS enterprise license is required. With Nutanix Metro Availability you can protect entire datastore. You cannot choose VM to protect.

Synchronous replication

Nutanix synchronous replication is available with Nutanix AHV (using Leap and AOS 5.17) or VMware vSphere. With synchronous replication workloads are protected with RPO=0. Synchronous replication is used when you cannot provide spanned L2 across datacenters (required by Nutanix Metro Availability) but still need to provide RPO=0 to applications.

Table 135: Application Requires RPO

BCN-008	APPLICATION REQUIRES RPO=0
Justification	Applications require 0 RPO and minimal RTO.
Implication	Nutanix Sync replication requires high bandwidth and low latency (under 5ms) datacenter connectivity. Nutanix Sync replication with AHV is supported only with Nutanix Leap. Nutanix synchronous replication requires Nutanix AOS enterprise license or Leap Advanced addon license.

NearSync replication

NearSync replication enables you to protect your data with an RPO as low as 1 minute. You configure a Protection Policy with NearSync replication by defining the VMs or the Categories of VMs. The policy creates a Recovery Point of the VMs in minutes (1–15 minutes) and replicates it to the recovery site.

Table 136: Provide RPO Between 1 and 15 Min

BCN-009	PROVIDE RPO BETWEEN 1 MIN AND 15 MIN TO AN APPLICATION
Justification	Provide low RPO to storage intensive workloads over limited connectivity.
Implication	A cluster is NearSync capable if the capacity of each SSD in the cluster is at least 1.2 TB. All-flash clusters do not have any specific SSD sizing requirements. Nutanix NearSync replication with AHV is supported only with Nutanix Leap. Nutanix NearSync replication requires Nutanix AOS enterprise license or Leap Advanced addon license.

Asynchronous replication

Nutanix asynchronous replication enables you to protect your data with an RPO of 1h or more. You can configure a Protection Policy or Protection Domain with asynchronous replication by defining the VMs or the Categories of VMs. The policy or protection domain scheduler creates a Recovery Point of the VMs in minutes (60 minutes) and replicates it to the recovery site.

Table 137: Application Requires RPO

BCN-010	APPLICATION REQUIRES RPO=>1H
Justification	Provide RPO=>1h to the application utilizing low bandwidth\high latency datacenter links.
Implication	Nutanix asynchronous replication requires Nutanix AOS enterprise license or Leap Advance addon license.

DR Patterns

1-to-1

For smaller environments with only a pair of locations, protect applications between two availability zones. Availability zones can be the same region or across regions. The same workload or application can be protected with different RPO thresholds. Choosing availability zones locations (same region or across regions) defines RPO value for protected application\workloads. If application or workload requires RPO=0, choose availability zones in the same region (latency requirement for RPO=0 is 5ms RTT or less).

1. Two AZ within the same region

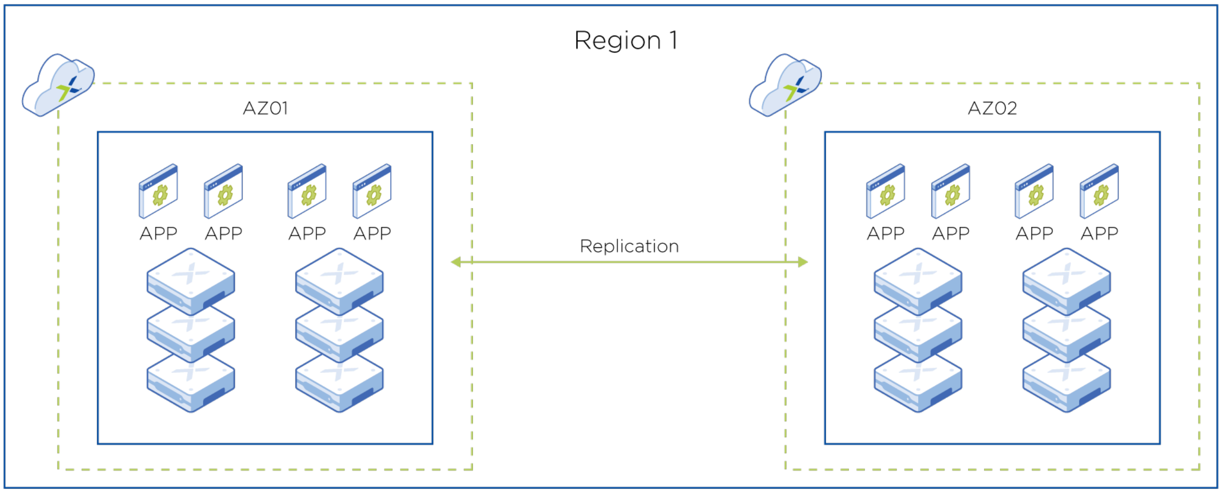


Figure 61: Two AZ within the Same Region

1. Two AZ across multiple regions

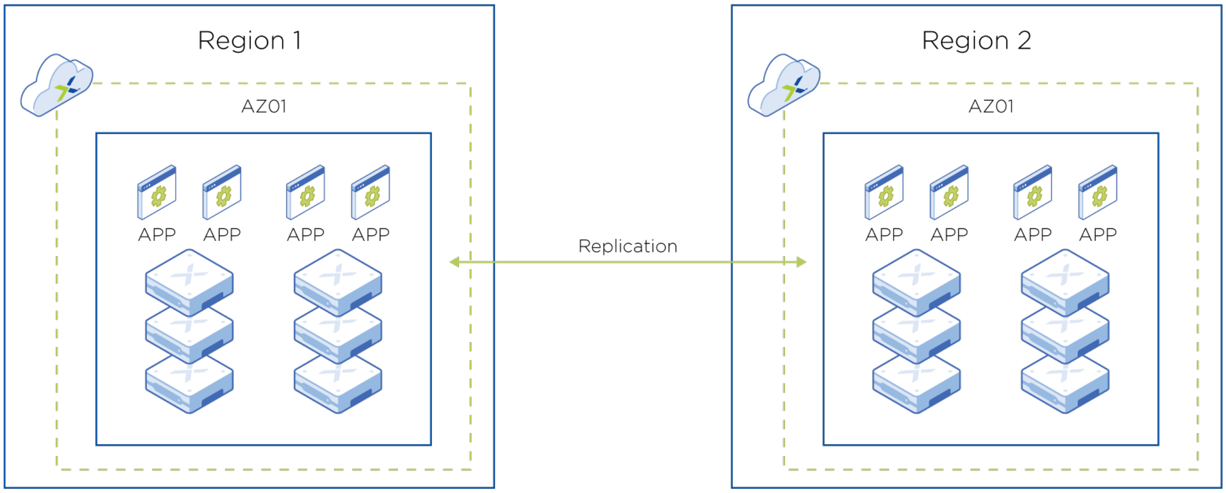


Figure 62: Two AZ across Multiple Regions

1-to-Many

The 1-to-Many DR pattern maps a single source Nutanix cluster to multiple DR target Nutanix clusters. This use case is applicable where an application requires multiple levels of data protection and at the same time providing high uptime and minimal to zero data loss. Availability zones can be located within the same region or across multiple regions. Choosing the availability zone location (same region or across regions) defines minimum RPO value for protected application\workloads.

If an application or workload requires RPO=0 and protection against natural disasters like hurricanes or earth quakes, choose an availability zone in the same region to provide RPO=0 (latency requirement for RPO=0 is 5ms RTT or less) and second availability zone in different region (to provide application resiliency against wide spread natural disasters like hurricanes, floods or earth quakes).

1. Multiple AZ within a single region

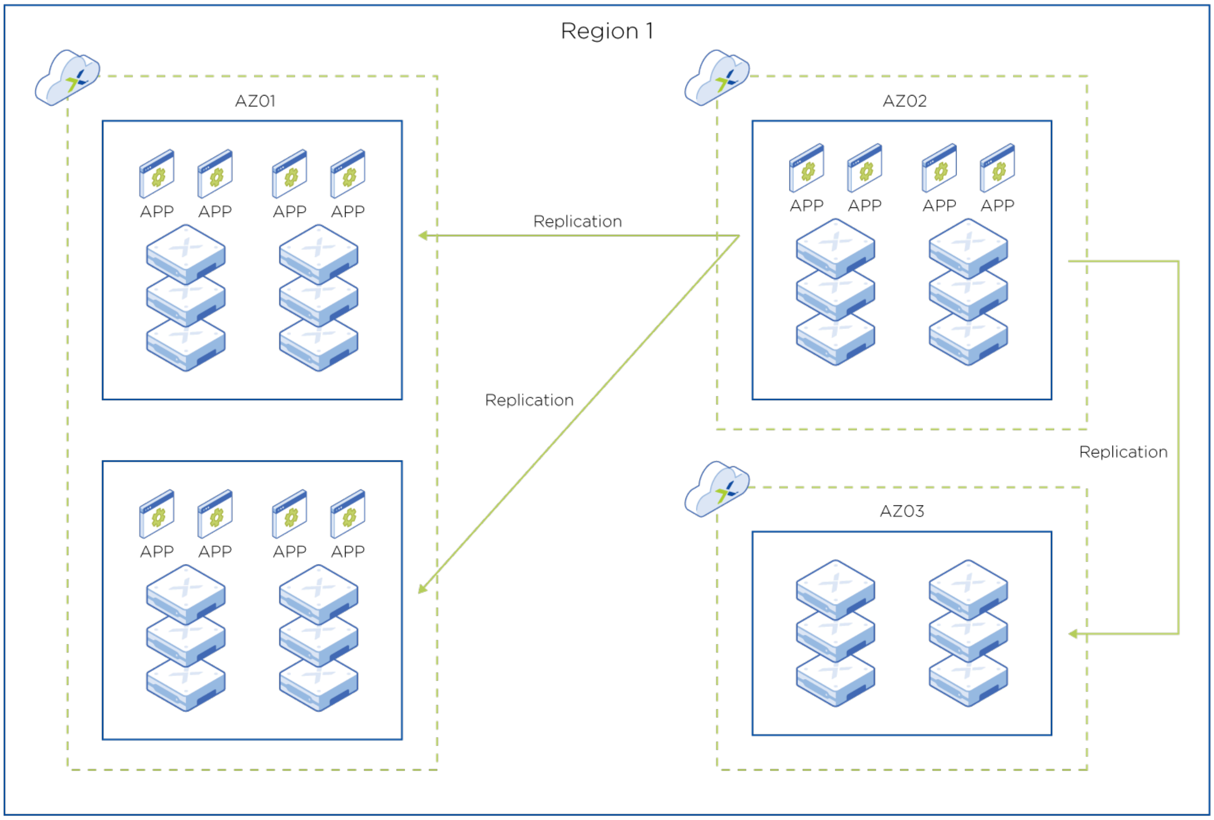


Figure 63: Multiple AZ within a Single Region

1. Multiple AZ across multiple regions

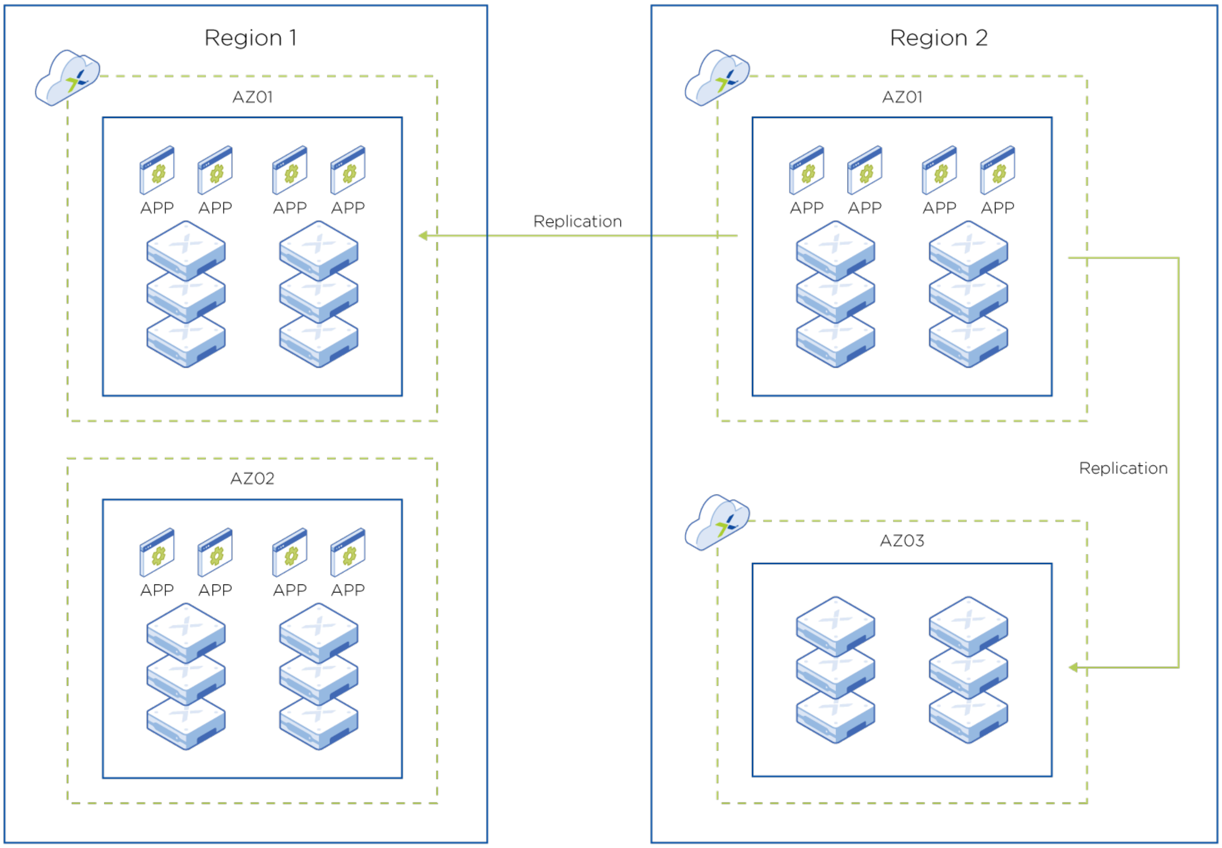


Figure 64: Multiple AZ across Multiple Regions

Many-to-1

The Many-to-1 DR pattern maps multiple source Nutanix clusters to single DR target Nutanix cluster. The best use case for the many-to-1 DR pattern is to protect multiple ROBO locations to a single Nutanix target in a DR hub, which is usually located in a regional or GEO datacenter. Each ROBO location is a separate availability zone which replicates data to the DR hub (which is located in separate availability zone) but the same region.

1. Multiple AZ within a single region

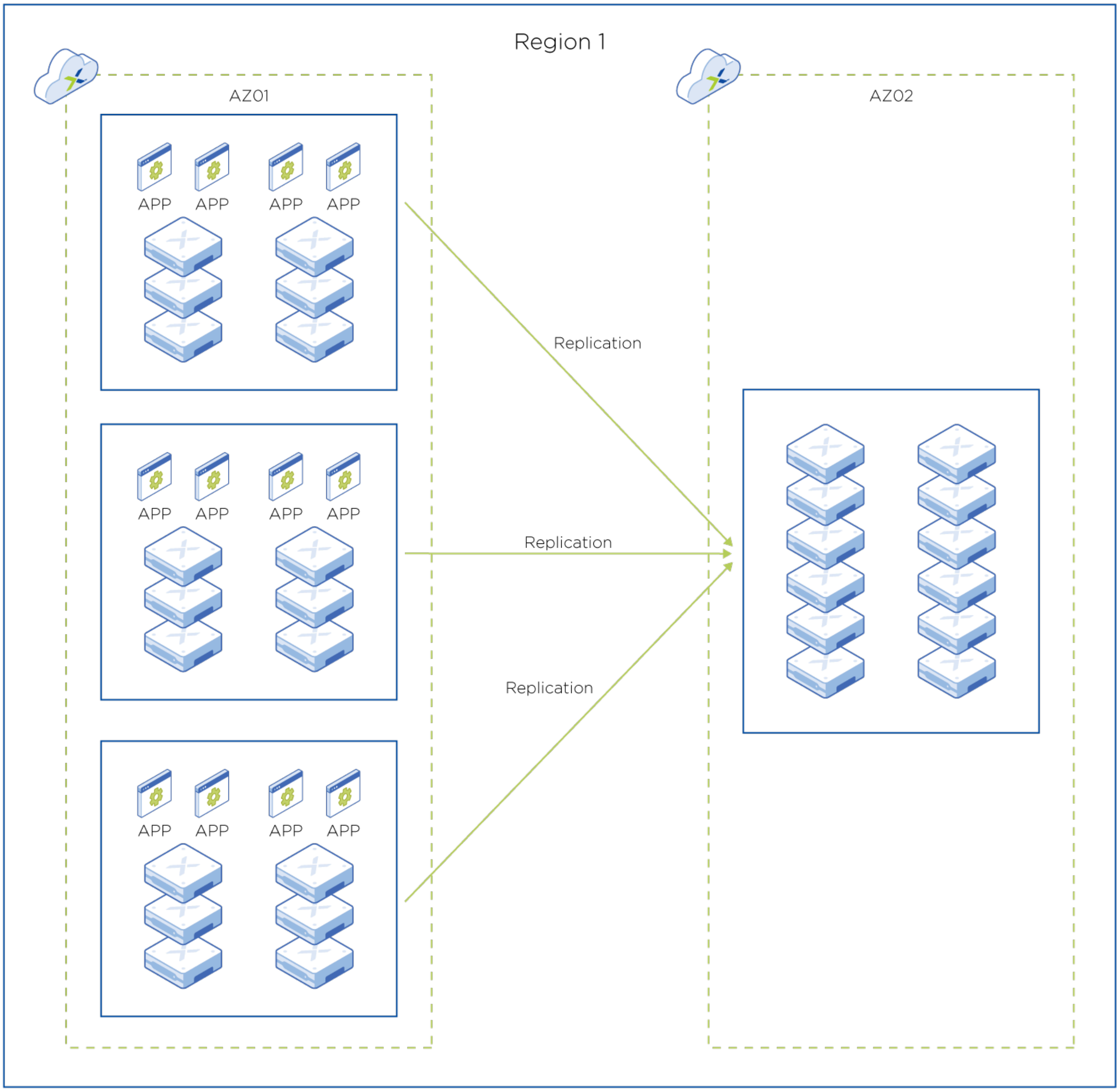


Figure 65: Multiple AZ within a Single Region

1. Multiple AZ across multiple regions

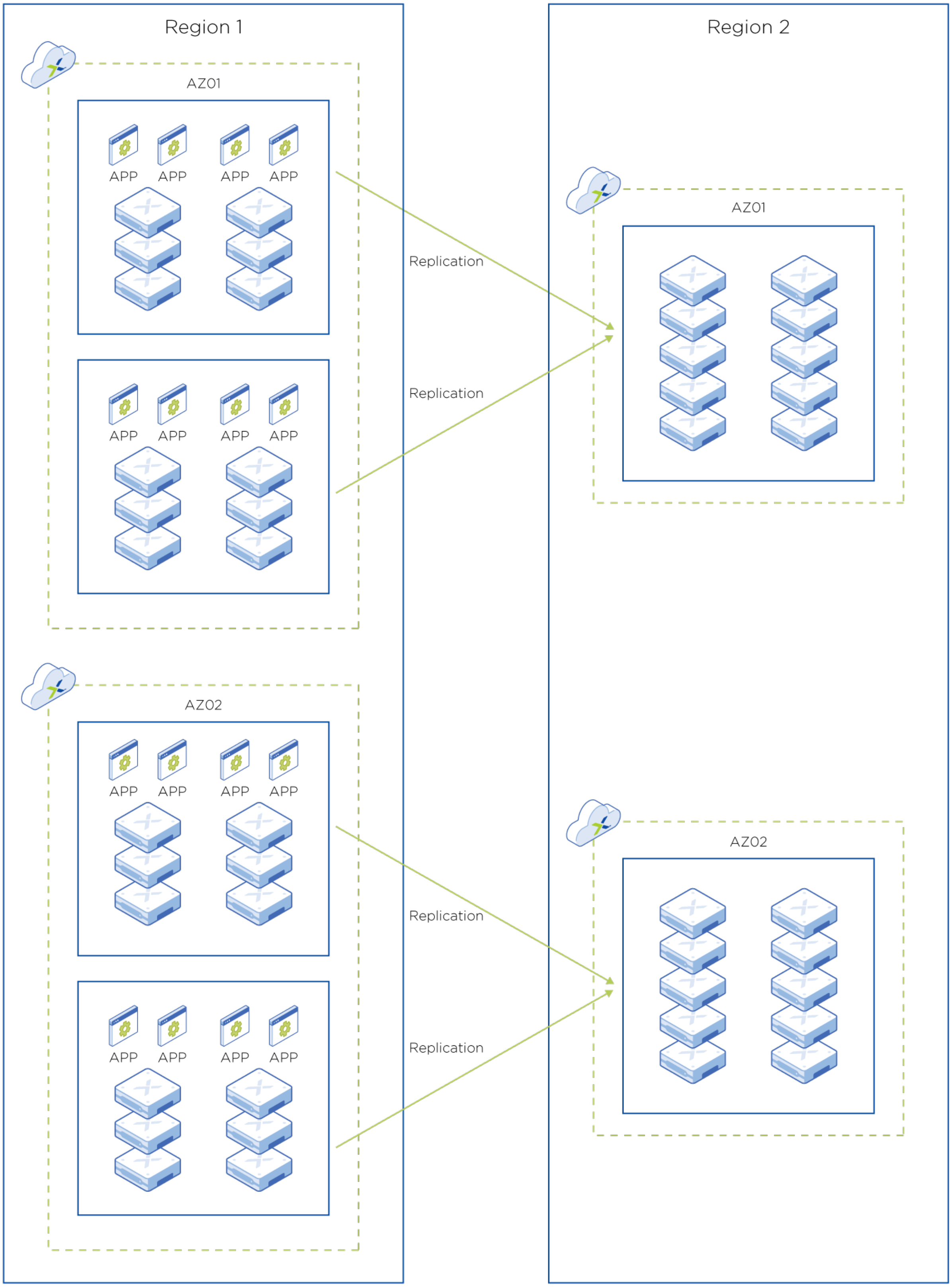


Figure 66: Multiple AZ across Multiple Regions

Many-to-Many

Description

The Many-to-Many DR pattern maps multiple source DR clusters with multiple target DR Nutanix clusters.

This pattern is suitable to applications which can run concurrently in more than one location. These applications which are require zero data loss and zero to minimal downtime. Leveraging multiple RPO schedules (synchronous, asynchronous, and nearsync) you can protect each application\workload with the most appropriate method to meet availability requirements.

1. Single region

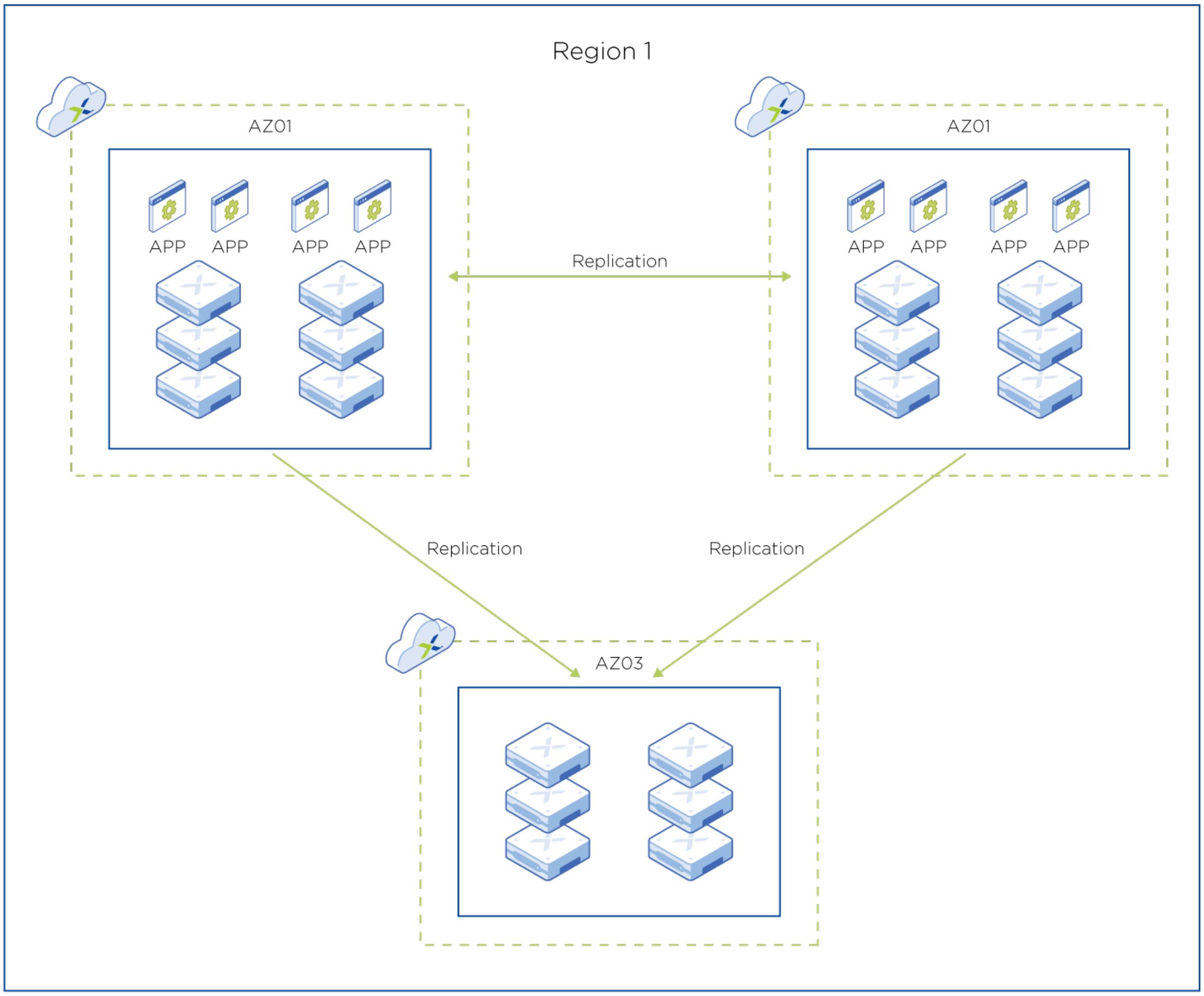


Figure 67: Single Region

1. Multiple regions

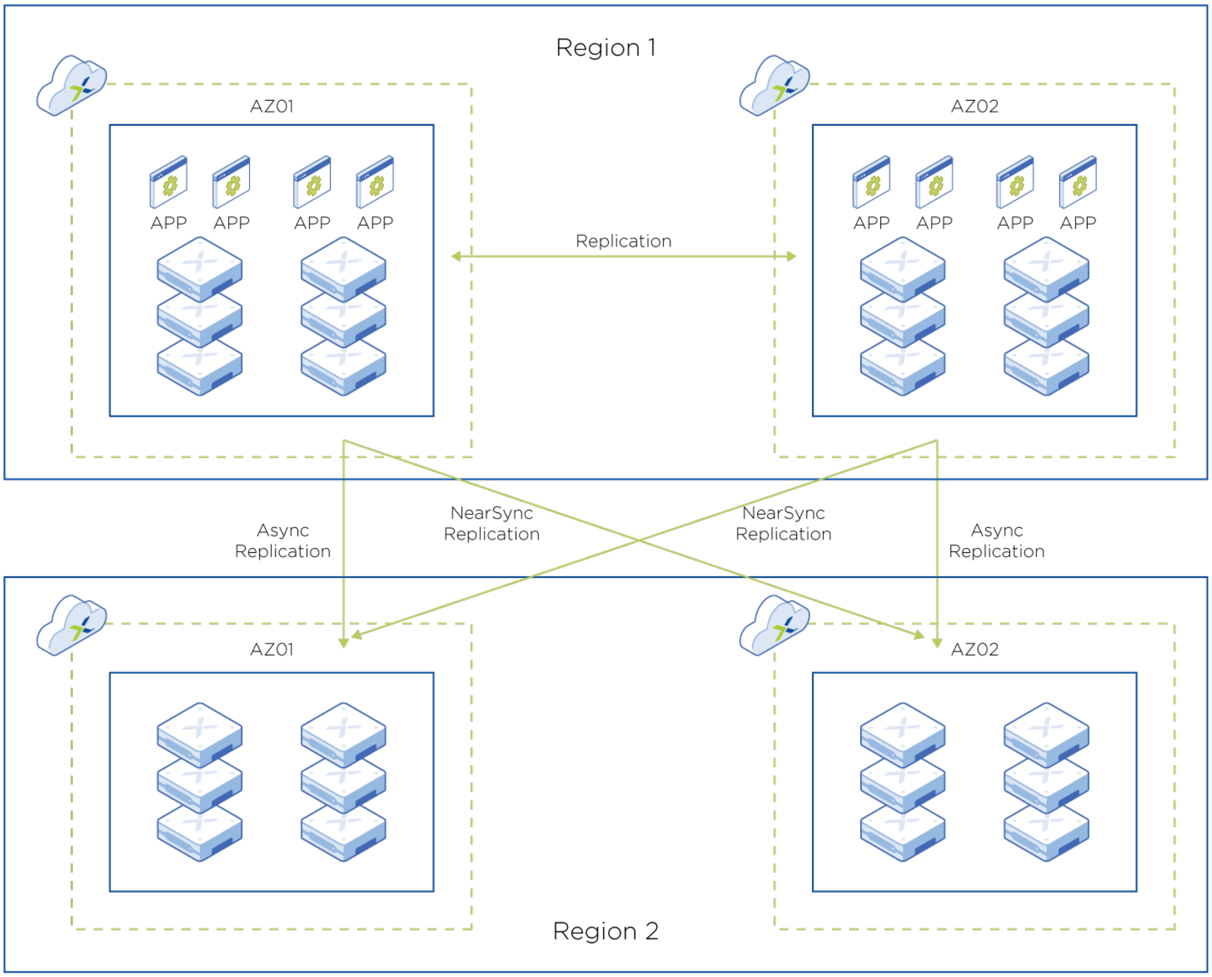


Figure 68: Multiple Regions

ROBO Patterns

For remote and branch offices, there are a few specific considerations. For sites utilizing 3 or more nodes in a cluster, the operation of that cluster is similar to a datacenter, while typically constrained by bandwidth/latency and facilities limitations.

A limited number of models are supported in either a single-node (no HA) or two-node (basic HA) configuration. These configurations are suitable for small footprint sites running a limited number of VMs (5-10, depending on load).

A two-node solution offers the same capacity sizing options as one-node, while allowing for failover in the event one of the nodes goes down. Automatic failover of VMs will require connecting the two-node cluster with a remotely hosted witness VM.

In addition, there are limitations on feature support related to one and two-node clusters. Details on these are maintained in KB 5943 on the Nutanix Support Portal.

Some of the most relevant limitations are:

- Lack of expandability (cannot grow to 3-nodes or larger without re-foundation)
- Limited VM sizing capacity due to limitations of the supported hardware models
- RPO support of 6 hours or greater

For ROBO sites, this creates a decision point. One and two-node solutions are appropriate for systems where a reduced feature set and performance still meet necessary requirements. Two-node systems can have their intra cluster traffic configured using directly connected links to avoid the need for high speed switches. This reduction in equipment is a valid tradeoff to sustain a lower price point compared to traditional 3+ node datacenter clusters.

An additional option is per VM ROBO licensing where the physical hardware is still composed of 3+ hosts. Each VM is explicitly licensed and limited to 32 GB of RAM.

Table 138: Decide on Cluster Licensing Model

DCD-001	DECIDE ON CLUSTER LICENSING MODEL (CBL OR PER VM)
Justification	The two licensing models cannot be mixed on a single cluster.
Implication	Per-VM ROBO licensing model allows for greater hardware capacity and feature set capabilities, while potentially increasing the footprint and hardware cost.

Table 139: Decide on Cluster Type Used

DCD-002	DECIDE ON CLUSTER TYPE USED (IF NOT ROBO PER VM)
Justification	One and two-node clusters are restricted to specific models, 3+ node models are not restricted.
Implication	Limitations imposed by a one/two-node solution cannot be corrected later (in place) by upgrading to a 3+ node.

Fan Architecture

In a fan-in/out style architecture, multiple ROBO sites are configured to replicate data back to a centralized location. The main location is normally larger, with enough capacity to meet the storage requirements for all the remote sites' expected replication and retention needs.

When building out a central location, at a minimum it needs enough storage capacity to accommodate the data from each ROBO site that requires protection. An important decision is whether to provide compute capacity to actively sustain the workload in the central site in a DR scenario, or only to provide capacity to store the data.

Table 140: Decide Whether to Provide Compute for ROBO Failover

DCD-003	DECIDE WHETHER TO PROVIDE COMPUTE FOR ROBO FAILOVER
Justification	A data replication only ROBO DR strategy will be sustainable by a smaller number of data dense storage nodes. An active failover capable ROBO DR strategy will require additional compute capacity to bring the failed services online in the event of a disaster.
Implication	Failover compute capacity must be reserved, potentially increasing solution cost and footprint.

Table 141: Decide on ROBO Failover Compute

DCD-004	DECIDE ON ROBO FAILOVER COMPUTE OVER COMMIT

Justification	The recovery site will likely only have to provide compute failover capacity for a limited number of concurrent site failures.
Implication	For a centralized location serving many ROBOs, the added cost of failover hardware may be relatively modest and worth the investment.

DR Pairing

For configurations lacking a centralized replication target, an alternative is to create replication partner sites. These sites are typically within a region/ availability zone and may have improved network speeds between sites (for example via metro ethernet) versus to a main datacenter. Each of the paired sites should be far enough apart from each other to support recoverability requirements in the event of a disaster.

This configuration can also be used for ROBO in a different country/legal entity than the main datacenter, where data privacy laws make replication back to a larger datacenter impossible.

Control Planes

Control plane design in a multi-site environment is a delicate balance between having a centralized view of everything versus a completely disaggregated view. This design suggests striking a balance, with separate control planes when necessary, to meet availability and product (e.g. VMware SRM or Prism Central) requirements.

vCenter

The guidance for vCenter placement in a multi-datacenter environment is to generally keep the vCenter instance and the clusters its managing within the same availability zone. An exception to this is in ROBO designs where having a vCenter at each remote site is not always desirable since it lacks a single management point for a larger number of sites.

If protecting workloads across availability zones—whether in the same region or across regions—you need to have at least a single vCenter to manage workloads at the source and target sites. This allows for Nutanix DR playbooks or VMware Site Recovery Manager (SRM) to automate the failover of these workloads.

Prism Central

The guidance for Prism Central (PC) placement is to have at least one PC deployment within each availability zone depending on the number of VMs. This allows the PC instance to manage local clusters and resources while utilizing DR runbook automation to connect to PC instances in other regions for protecting workloads. Having separate PC instances at the source and target locations is a requirement for DR runbook automation available within PC. The only time you would deploy more than one PC per availability zone is when you scale past the limits of a single PC.

There are only two situations where you might consider having a PC instance manage clusters in other availability zones:

- ROBO architecture where each site typically has a small cluster and it's desirable to have a single central PC instance to manage them.
- If you have multiple AZs and are not replicating between them, they can be managed by a single PC instance.

Multi-DC Networking

Networking is critical for disaster recovery, since the network forms the backbone carrying all replicated data. Each component in the path from the physical layer of the source cluster, to the physical layer of the destination cluster must be considered. Especially important are the site-to-site links that connect protected sites.

Each Nutanix replication technology has specific requirements to ensure that the connection between sites has sufficient bandwidth, low enough latency, and meets redundancy requirements. Please note that the formulas below do not take into account link latency, and as such, replication over links with high latency can see significantly reduced performance.

The following table summarizes the required networking characteristics for each replication type:

Table 142: Required Networking Characteristics

REPLICATION TYPE	LATENCY REQUIREMENTS	BANDWIDTH CALCULATION	NOTES
Async DR	NA	Bandwidth needed = (RPO change rate * (1 - compression on wire savings %)) / RPO Max of 800mbps per node	
Near Sync DR	NA	Bandwidth needed = (RPO change rate * (1 - compression on wire savings %)) / RPO Max - Unbounded	
ESXi Metro Availability	Sites: 5msec RTT Witness: 200msec	Metro Container data change rate Max - unbounded	Redundant site links recommended Witness: Separate links
AHV Synchronous Replication	Sites: 5msec RTT	Data change rate Max - unbounded	Redundant site links recommended

In addition to the characteristics of the network transport between protected sites, careful attention must be paid to the VM networking at each site. When a workload moves from one site to another there are many possible options for maintaining workload connectivity to the required customers or endpoints.

Technologies like stretched networking, load balancing, NAT, and more should be considered to provide seamless replication and application connectivity.

Bandwidth Requirements

Replicating data between sites requires that enough bandwidth is available to replicate all of the data changed during the RPO window within the RPO interval. Refer to the Data Protection and Disaster Recovery Guide for a complete explanation of bandwidth calculation.

Calculate this data change rate for both asynchronous and NearSync replication. Here is an example of the bandwidth needed for an RPO of 1 hour with 15GB of data changed during that hour. We assume 30% savings for compression.

Table 143: Bandwidth Needed for an RPO

BANDWIDTH NEEDED = (RPO CHANGE RATE * (1 - COMPRESSION ON WIRE SAVINGS %)) / RPO
Example: RPO: 1hr (3,600s) RPO Change Rate: 15GB Compression Savings: 30% (15GB * (1 - 0.3)) / 3,600s (15GB * 0.7) / 3,600s 10.5GB / 3,600s (10.5GB * 1,000 MB/GB * 8 bits/byte) / 3,600s - converting to Mb/s 84,000 Mb / 3,600s = 23.33 Mb/s Bandwidth needed = 23.33 Mb/s

The bandwidth needed is easier to calculate when the RPO is zero, since it is equal to the data change rate. Data will be compressed before being sent to another site, so assume compression in this calculation as well.

Table 144: Calculate Required Storage Replication Bandwidth

MULTI-NET-001	CALCULATE REQUIRED STORAGE REPLICATION BANDWIDTH BASED ON RPO.
Justification	Bandwidth must be available to keep up with data change rate and meet the desired RPO.
Implication	Achieving the desired RPO can require significant bandwidth and requires knowledge of the workload’s current usage.

In addition to storage replication between sites, bandwidth must be allocated for workload traffic and application level replication traffic. First determine how much workload traffic and application replication traffic will exist. This is largely dependent on the applications in use. Next determine if there are any priority or latency requirements that are unique to the applications. Applications with latency sensitive traffic may require WAN quality of service, or even dedicated WAN links.

Table 145: Calculate Application Specific Bandwidth

MULTI-NET-002	CALCULATE APPLICATION SPECIFIC BANDWIDTH.
Justification	WAN links must also carry application level replication traffic and application traffic.
Implication	Critical application traffic may require more WAN links or WAN quality of service policies.

Latency Requirements

Latency between protected sites will have an impact on the speed of data replication. Latency is a function of physical distance, network cabling plus equipment, and the speed of light.

Asynchronous DR has no minimum network latency requirements as long as the bandwidth requirements can be met. This makes asynchronous replication a great choice for protecting sites a long distance apart such as between regions.

Near-sync DR also has no fixed requirement for minimum latency between sites, but the bandwidth requirement must be met.

Metro Availability has two separate requirements for network latency, one between the protected sites, and another between the protected sites and the optional witness. Between protected sites the latency must be less than 5 milliseconds round trip, or 2.5 milliseconds one way. The sites must therefore be located relatively close to each other, but in separate availability zones. Each protected site must also have a connection to the witness that is within 200 milliseconds or less to allow for reliable message exchange during site failures.

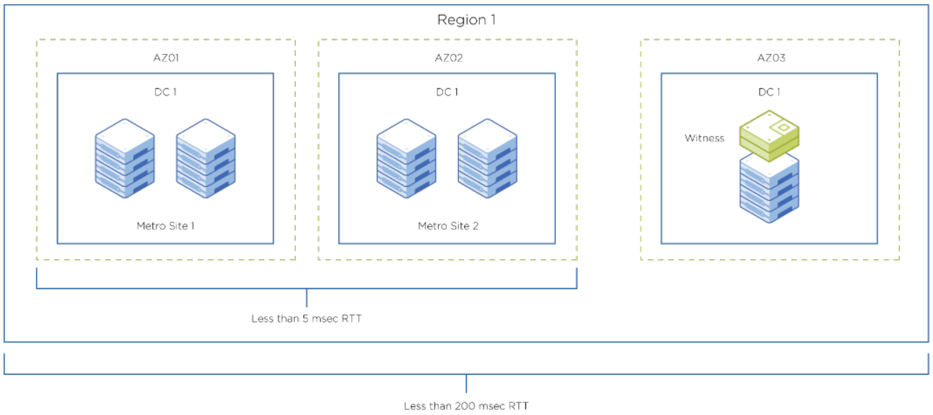


Figure 69: Latency Requirements

Every bit of latency on a link between Metro Availability sites also affects storage write latency for workloads. Keep this latency as low as possible by choosing MA sites that are as close to each other as possible while still providing the necessary physical resilience. In this guide we assume that MA sites exist within the same availability zone, but the witness can be in another AZ, or even in another region.

Table 146: Place Metro and Synchronous Replication Sites within the Same Region

MULTI-NET-003	PLACE METRO AND SYNCHRONOUS REPLICATION SITES WITHIN THE SAME REGION WITHIN 100KM OR LESS THAN 5MSEC RTT
Justification	Latency on metro links is directly added to storage write latency.
Implication	Metro Availability site design must consider physical location carefully.

Table 147: Place Metro Witness within 200msec

MULTI-NET-004	PLACE METRO WITNESS WITHIN 200MSEC
Justification	Sites must receive responses from the witness to determine failure handling.
Implication	Witness can be relatively far from protected sites.

Separating DR Traffic

Nutanix does not recommend isolating DR traffic on separate WAN links because it adds complexity, but if security or business requirements dictate a separated WAN, keep the following in mind.

Consider the scenario where a company has several separate links between protected sites where some are for low latency applications that take high priority. This customer may decide to also purchase an entirely separate physical link between sites that is for higher bandwidth replication traffic that is not latency sensitive and at a lower cost. In essence this is a form of QoS for the application. The customer can ensure that critical application traffic is never stopped by lower priority replication traffic.

It's also possible that the reverse is true for a case like Metro Availability, where certain application and storage replication traffic may take higher priority than other application traffic, and we dedicate an entire WAN link to this storage traffic.

In these cases, the desire is to segment traffic for one purpose over one WAN link between sites, and traffic for another purpose over another link between sites. There are several ways to achieve this goal.

Network Segmentation in the Nutanix CVM

Network Segmentation in the CVM creates another virtual NIC on the CVM and the hypervisor, allowing the separation of user, replication, and DR traffic to different VLANs or physical interfaces. Every CVM and every host needs additional IP addresses in this method, so the configuration for each host becomes slightly more complex.

Network Routing: Manual or Policy Based

Using network routing, the network devices use criteria like source and destination address, and even TCP ports to determine which link is used for routing traffic. This method requires more advanced network configuration but simplifies the CVM and host configuration.

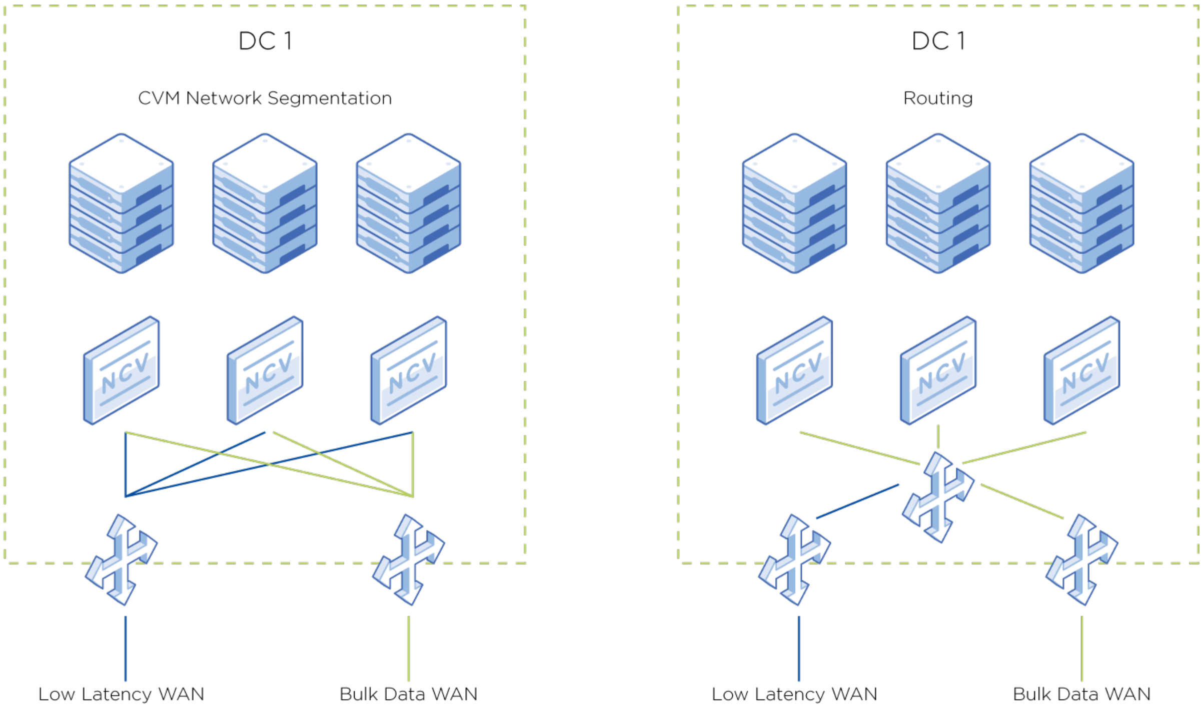


Figure 70: Network Routing

Table 148: Do Not Use Network Segmentation for DR

MULTI-NET-005	DO NOT USE NETWORK SEGMENTATION FOR DR UNLESS REQUIRED. ROUTING SEGMENTATION IS PREFERRED OVER CVM SEGMENTATION
Justification	Network segmentation in the CVM and host adds complexity that must be managed during the system lifecycle.
Implication	Unless specifically required to do otherwise, traffic from applications and DR will share the same network links.

Network Redundancy

Ensure there are redundant paths between protected sites and minimize or eliminate single points of failure. This prevents the failure of a single networking component from stopping replication of storage and application traffic. At each network hop between two protected sites, evaluate whether failure of that component would be gracefully handled, or if would lead to an outage.

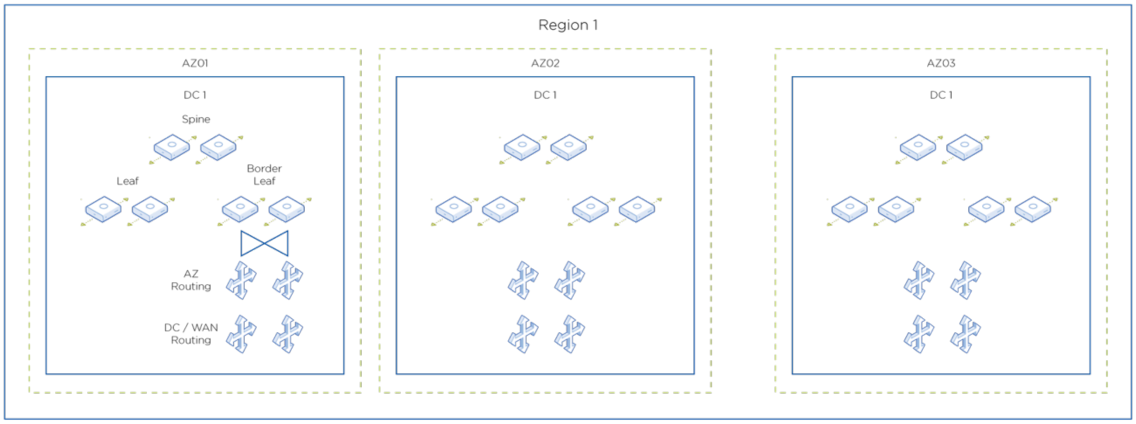


Figure 71: Network Redundancy

For example, in the path between two sites connected over a WAN we have the following components:

- Site 1: Top-of-rack switch (leaf)
- Site 1: Spine switches
- Site 1: Border leaf
- Site 1: DC router
- Site 1: WAN router
- Site 2: WAN router
- Site 2: DC router
- Site 2: Border leaf
- Site 2: Spine switches
- Site 2: Top-of-rack switch (leaf)

This means we need to ensure redundancy in the following components:

Top-of-rack switch

Ensure that each Nutanix host in every cluster is connected to at least two top-of-rack switches.

Spine switch

Ensure that there are at least two spine switches, and that all leaf switches are connected to all spine switches.

Border leaf

The border leaf is where external services such as routing connect. Ensure that there are at least two border leaf switches, and that they connect to all spines. The border leaf switches should connect to at least two routers, to allow redundant external routing.

DC router

Depending on the size of the deployment, there will be one or more levels of routing between the rack networks and the WAN edge. Each routing tier should provide high availability to allow external routing from the racks to the rest of the DC.

WAN router

The WAN router provides connectivity external to the datacenter. These should be deployed in a redundant fashion, allowing for the failure of any single routing component.

The WAN also presents a special case for WAN Link redundancy. In addition to providing multiple WAN routers, designs should use multiple WAN links, potentially from different network providers.

The WAN provider may be an external entity, or in the case of connection between sites within a campus, they may be entirely within the control of one entity. In these cases where the links are under your control, it is still desirable to use multiple separate physical connections, such as separate fiber links between sites. It's desirable to route these links to the WAN through separate physical paths to protect against ground or aerial utility faults (such as backhoes). It may not always be possible to use different physical paths based on datacenter location, design, or infrastructure provider constraints. Where a single point of failure exists, it should be well documented and understood. One example could be a single piece of conduit carrying multiple WAN links inside it between datacenters.

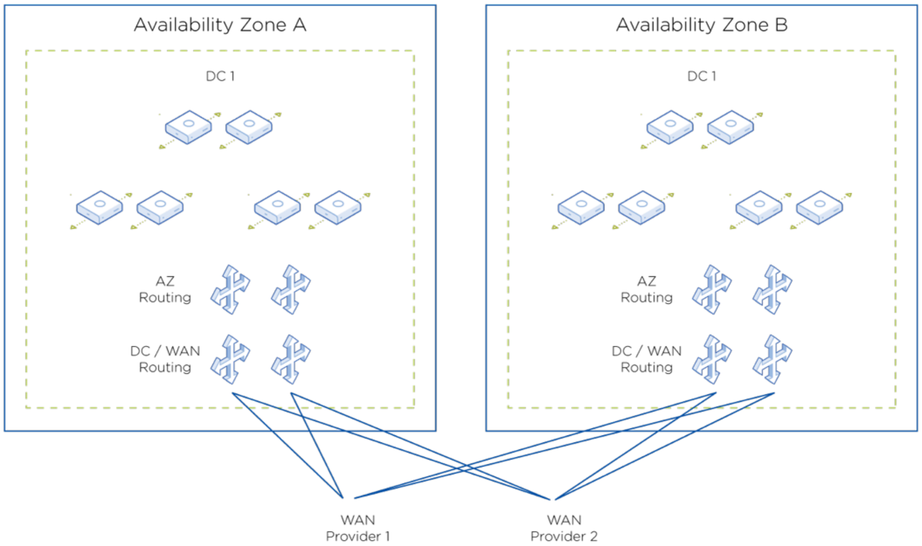


Figure 72: Single Piece of Conduit with Multiple WAN Links

Testing is an important part of the redundancy in any system and networking is no exception. Create and maintain a regular schedule to test the redundancy of each individual component such as routers and switches. Maintenance such as vendor software upgrades and patches as well as regular power maintenance provide great opportunities to verify the resilience during planned single component outages.

Table 149: Ensure Redundancy of Each Network Device

MULTI-NET-006	ENSURE REDUNDANCY OF EACH NETWORK DEVICE. TRACK THE COMPLETE NETWORK PATH BETWEEN PROTECTED SITES
Justification	Failure of a single component must not lead to connectivity failure between protected sites.
Implication	Redundant components must be purchased and configured for high availability and failover.

Table 150: Create and Follow a Redundancy Test Plan

MULTI-NET-007	CREATE AND FOLLOW A REDUNDANCY TEST PLAN. CAPTURE NETWORK COMPONENTS IN PLAN
Justification	High availability configurations are often complex and must be tested to ensure they work as expected.
Implication	Regularly scheduled maintenance must include additional protocols to test network components during this time.

Witness Path Requirements

Using Metro Availability with a witness creates resiliency requirements between the protected sites and the witness site.

First, the witness should be at a separate site from either of the protected sites. This allows the witness to monitor the availability of both sites independently and means that the witness should survive the failure of one of the protected sites. The witness provides no extra benefit if it goes down during a site failure.

In addition, the network path between the protected sites should be a separate path from the path to reach the witness. That means each protected site should have at least two routing paths, one to reach the witness, and another to reach the other protected site.

Finally, the connection between the protected site and the witness site should have no more than 200msec of latency. This helps ensure timely event handling during a failure scenario.

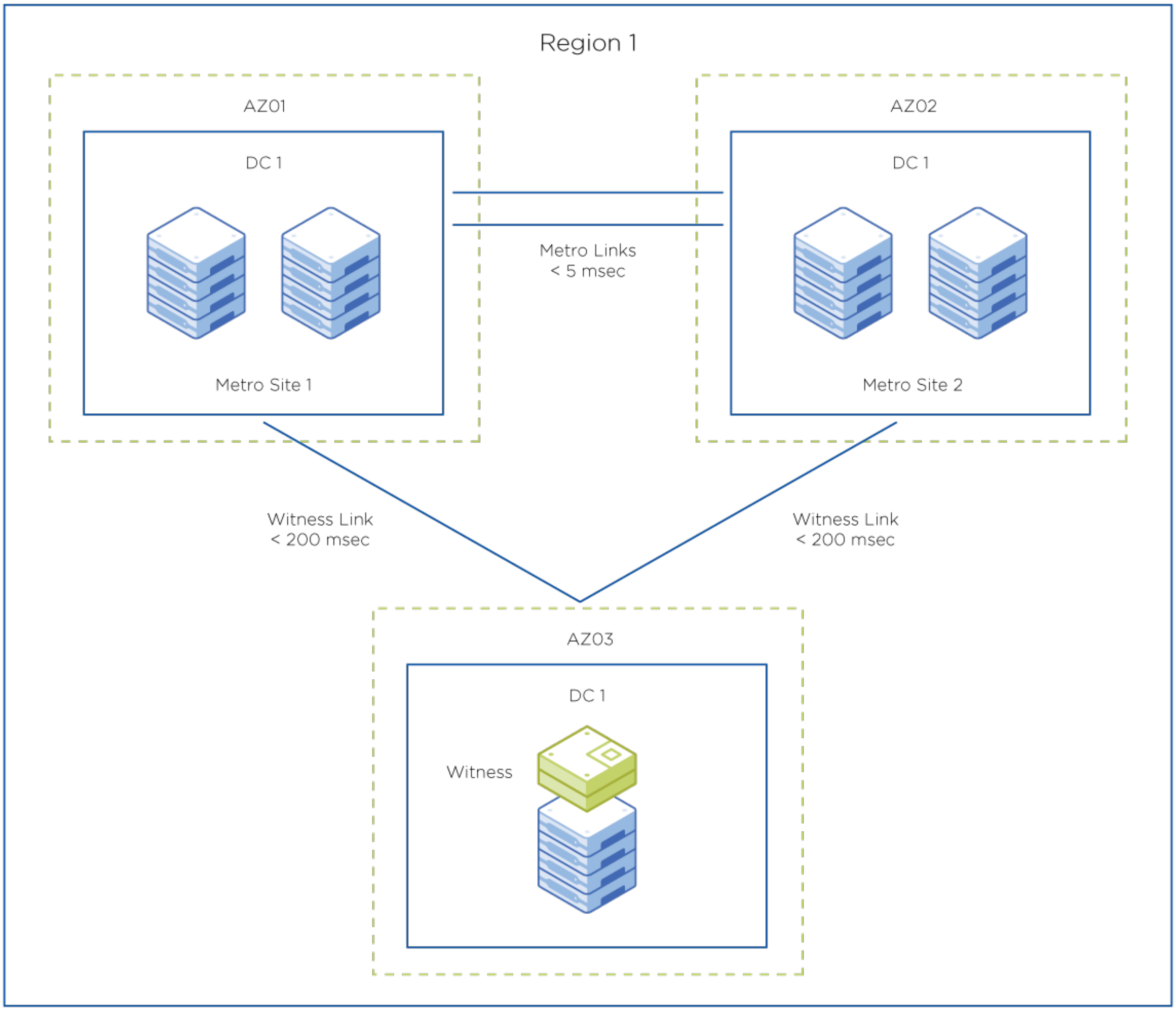


Figure 73: Metro Witness Link Requirements

Table 151: Ensure that the Link Between Protected Sites and a Witness does not Use the Link Between Sites

MULTI-NET-008	ENSURE THAT THE LINK BETWEEN PROTECTED SITES AND A WITNESS DOES NOT USE THE LINK BETWEEN SITES.

Justification	The path to the witness should be a separate, protected link. The witness should be reachable during failures of the link between the protected sites.
Implication	Additional WAN connectivity may be required to maintain a separate external connection to the witness.

Workload Networking

In addition to transporting critical storage traffic, the network also provides access to running applications. During site failures it's important to consider how the application or workload will provide constant access for clients, even when the workload may be running in a secondary location.

IP Address Assignment

There are two approaches for managing workload connectivity during site failures.In the first approach, workloads keep their existing IP addresses and the network provides connectivity in the backup site even when using previous IP addresses. In the second approach, workloads fail over to the second site and receive new IP addresses belonging to the backup site.

The network tools used are outside the scope of this document but consider the following when moving workloads between sites using these tools. The workload and the network capabilities will determine the chosen method.

Case 1: Workloads keep the same IP address

Preserving the IP addresses of workloads during failover is helpful for several reasons. No change is required inside the workloads, which may sometimes not even support IP address changes. This also means any scripts or infrastructure that rely on hardcoded IP addresses will continue to function. The network infrastructure seamlessly handles the transition by migrating the network to the new site, or by maintaining a stretched network at all times. The responsibility of failover reachability is shifted away from the application and to the network.

The following approaches are frequently used to preserve IP addresses:

- Stretched layer 2 networks
- NAT

- Layer 3 route advertisement updates

Case 2: Workloads receive a new IP address

Provisioning workloads with new IP addresses allows more network flexibility, for instances where moving a network between sites is not feasible. When the workload easily supports IP address changes, or if the network does not support network mobility, changing workload addresses may be the best option. Using a tool to automate these address changes ensures a smooth recovery.

The following tools are often used to provision new addresses at the DR site.

- DHCP / IPAM
- Site Recovery Manager
- Nutanix Runbook Automation
- Manual or scripted steps
- Vendor tools such as Zerto

Table 152: For Each Workload Decide Whether to Maintain or Change

MULTI-NET-009	FOR EACH WORKLOAD, DECIDE WHETHER TO MAINTAIN OR CHANGE IP ADDRESSES DURING FAILOVER.
Justification	
Implication	

Service Advertisement

Service availability is the advertisement of the workload and its availability to external clients at the DR site. There are a number of network and application technologies that can provide service advertisement, and this will be specific to your network and workload. For each application, consider how clients will reach this application at the DR site. This may involve application multi-site awareness, DNS updates, routing updates, load balancers, or some combination of these.

6. Incorporating Optional Nutanix Products and Services

Nutanix offers a number of optional products and services that you may choose to incorporate when implementing private, hybrid, or multicloud solutions on the Nutanix platform. These options can help you quickly satisfy unique organizational requirements.

The subsections that follow introduce the following solutions and provide operational guidelines consistent with earlier sections of this guide:

- Prism Operations: Prism Pro and Prism Ultimate add to the base capabilities of Nutanix Prism to provide monitoring, analytics, and automation capabilities from a single management interface.
- Nutanix Clusters: Clusters extends the Nutanix stack to the public cloud, creating a single management domain that spans private and public clouds and removes the friction from multicloud operations.
- Flow: Microsegmentation and other capabilities offered by Flow further enhance the security of your Nutanix operations.
- Files: Files enables a Nutanix cluster to incorporate full-featured file services capabilities, eliminating the need for standalone NAS appliances or file servers.
- Objects: Objects enables a Nutanix cluster to provide S3-compatible, software-defined, scale-out object stores for a variety of use cases, eliminating the need for standalone object storage while supporting applications that rely on object stores with an on-premises alternative.
- Mine: Nutanix Mine is an integrated backup solution, combining the benefits of the Nutanix HCI architecture with the capabilities of proven backup vendors.
- Calm: Calm provides application-level orchestration and lifecycle management, simplifying application management and enabling self-service.

- Karbon: Karbon integrates certified Kubernetes with the Nutanix operating environment, enabling Nutanix clusters to support both VM- and container-based applications while simplifying Kubernetes deployment and management.
- Era: Era automates and simplifies database administration, provisioning, and lifecycle management, facilitating database as a service (DBaaS).

Each subsection includes: Key Design Objectives, Architecture Overview, and Detailed Technical Design Considerations.

Prism Operations

Key Design Objectives

Table 153: Prism Ops Design Objectives

DESIGN OBJECTIVES	DESCRIPTION	NEW / UPDATE / REASONING
Enable proactive operations management	Prism Pro and Prism Ultimate licenses provide: - Operational automation in response to common triggers - Monitoring for non-Nutanix VMware environments - Application discovery and application-level monitoring	New

Architecture Overview

The operations tier in Nutanix Prism (provided by Prism Pro & Ultimate) focus on the entire IT Ops triangle—monitoring, analytics, and automation. Prism offers a single pane of glass that combines monitoring and automation to isolate pain points and ease the deployment of operational automation. This allows you to harness the full power of HCI. The solution aims to remove silos between different monitoring and automation tools and avoid the noise created by dynamic virtual infrastructure environments.

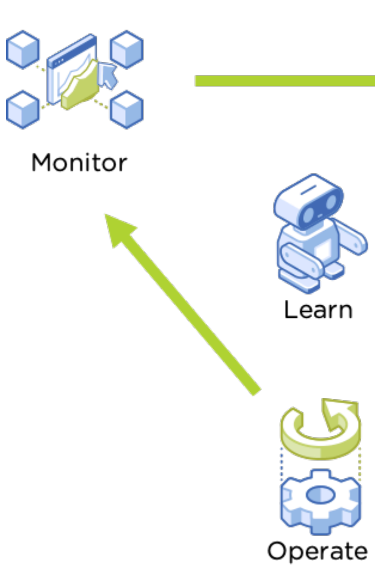


Figure 74: IT Ops Triangle

The goal is to provide broad observability and actionable signals powered by machine learning and automation for a seamless operations experience.

Prism Operations (Prism Ops) is powered by X-FIT, a distributed time series analysis and forecasting system that uses an ensemble of models. X-FIT is an enabler for building a management fabric; it can autonomously optimize datacenter performance and intelligently manage application resource demands using operational data.

Virtualization

- Prism Pro and Ultimate are hypervisor agnostic, working with both AHV and ESXi.
- With the release of Prism Central 2020.9, it is possible to monitor non-AOS vCenter VMs with Prism Pro. Similar capabilities available for an AOS environment, such as runway forecasting and playbook execution, are extended to VMware when it is monitored with Prism.

Management

All Prism Operations features are accessed through Prism Central, which removes the need for separate management tools, providing one tool for monitoring, operational analysis, and automation. With Prism Ultimate, this visibility and functionality has been extended to the application layer.

Automation

- X-Play is a simple-to-use automation engine geared toward the average IT admin managing Nutanix and non-Nutanix infrastructure (as of PC 2020.9). It is based on various triggers, including actionable alerts generated by machine learning insights. As X-Play runs through a Playbook, it initiates a sequence of actions that can be tracked through a corresponding Play.
- Six triggers are currently available: alert-based, manual, time-based, alerts matching multiple criteria, event-based, and webhook triggers. There are 27 out-of-box actions provided that can be customized according to your needs and integrated with third-party tools. Playbooks can also be imported and exported across different Prism Central instances.

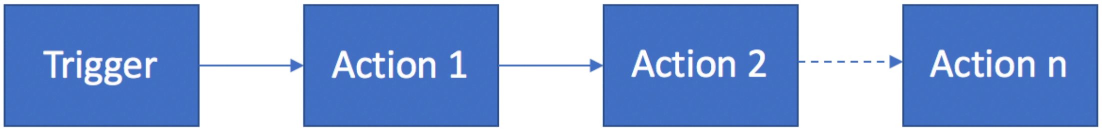


Figure 75: With Prism X-Play, a single trigger can precipitate multiple actions

Prism Licenses

Prism has three licensing tiers:

- Starter: Included free with AOS and deployed automatically with Prism Central. Features include:
  - Monitoring and troubleshooting
  - RBAC
  - LCM upgrades
  - Comprehensive search

- Pro: Separate add-on license. Enabled via Prism Central. Features include:
  - Customizable dashboards and reporting
  - Capacity forecast and runway planning
  - VM inefficiency detection and right-sizing
  - Anomaly detection and advanced troubleshooting
  - Monitoring non-AOS infrastructure
  - Smart operations automation
- Ultimate. Separate add-on license. Enabled via Prism Central. Features include:
  - Application discovery
  - Application monitoring – SQL Server content pack
  - Cost-metering, budgeting, and chargeback for resources

Prism Ultimate

Administrators may have zero visibility into the applications running inside VMs deployed on Nutanix infrastructure. Troubleshooting application issues becomes onerous without this application data and often requires separate software for application monitoring.

Prism Ultimate collects application metrics using third-party/Nutanix collectors (agentless), providing a single pane of glass to view both application and infrastructure data. With collected performance metrics, Prism Central can now provide a holistic view of IT operations that allows admins to easily monitor and diagnose problems caused by app performance issues without finger pointing.

High-Level Design Considerations

Currently, Prism's operations features are tied to a single instance of Prism Central. This means all nodes registered to a Prism Central instance must be licensed for Prism Pro & Ultimate in order for those nodes to be license compliant.

- To enable Prism Ultimate features, you must increase Prism Central resources to 14 vCPUs and 30 GiB memory on a small Prism Central instance or 18 vCPUs and 48 GiB memory on a large Prism Central instance.
- Prism Ultimate is available beginning with release PC 2020.8. Non-AOS VM monitoring is available beginning with PC 2020.9.

Operational Automation

Table 154: Leverage X-Play Playbooks

OPS-001	LEVERAGE X-PLAY FOR POWERFUL LOW-CODE/NO-CODE OPERATIONAL AUTOMATION
Justification	Playbooks require the presence of Prism Pro.
Implication	Requires an additional license that can increase cost.

The webhook trigger gives a third-party tool the ability to execute actions within Prism Central. Conditional execution based on a string that is read and parsed allows for if-this-then-that (IFTTT) workflows, giving complete control to the admin for building powerful automations. For example, you can create a service ticket in your preferred tool (ServiceNow, etc.) based on alerts or events and trigger further actions in Prism Central based on approval or denial.

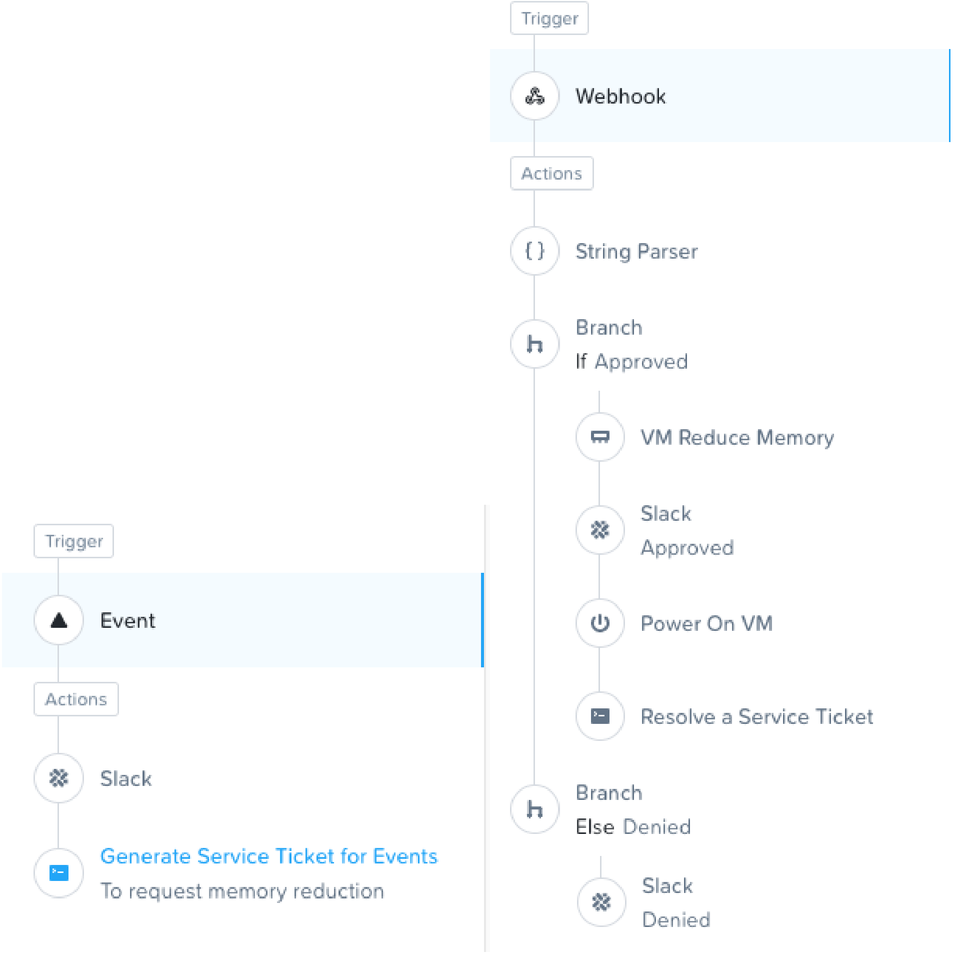


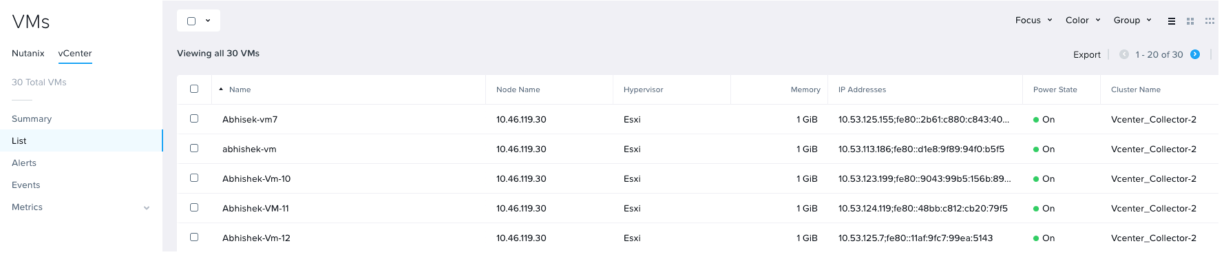
Figure 76: Example Workflows Based on Various Triggers

Non-AOS VM Monitoring

Table 155: Monitor Non-AOS VMs

OPS-002	MONITOR NON-AOS VMs
Justification	VMware virtualized environments not running under AOS can be monitored with Prism Pro.
Implication	Requires an additional license that may increase cost.

A containerized Nutanix collector is deployed on the Prism Central VM and gathers data for each configured VMware cluster, using the X-Stream interface to store the data. X-FIT generates the same intelligence as for a Nutanix cluster, and X-Play can be used to connect to automation.



Viewing all 30 VMs

Name	Node Name	Hypervisor	Memory	IP Addresses	Power State	Cluster Name
Abhisek-vm7	10.46.119.30	Esxi	1 GiB	10.53.125.155;fe80::2b61:c880:c843:40...	● On	Vcenter_Collector-2
abhishek-vm	10.46.119.30	Esxi	1 GiB	10.53.113.186;fe80::d1e8:9f89:94f0:b5f5	● On	Vcenter_Collector-2
Abhishek-Vm-10	10.46.119.30	Esxi	1 GiB	10.53.123.199;fe80::9043:99b5:156b:89...	● On	Vcenter_Collector-2
Abhishek-VM-11	10.46.119.30	Esxi	1 GiB	10.53.124.119;fe80::48bb:c812:cb20:79f5	● On	Vcenter_Collector-2
Abhishek-Vm-12	10.46.119.30	Esxi	1 GiB	10.53.125.7;fe80::11af:9fc7:99ea:5143	● On	Vcenter_Collector-2

Figure 77: Monitoring non-AOS VMs with Prism Ultimate

Application Discovery

Table 156: Discover Applications with Prism Ultimate

OPS-003	DISCOVER APPLICATIONS RUNNING ON YOUR INFRASTRUCTURE.
Justification	You can discover and export applications running on your Nutanix cluster with Prism Ultimate.
Implication	Requires an additional license that may increase cost.

Prism Ultimate’s agentless discovery allows you to identify application-to-VM dependencies using an API for CMDB integration. (A cloud connection is required.) Internet Protocol Flow Information Export (IPFIX) data is collected from VMs and sent to the Nutanix cloud which analyzes and syncs the findings back to Prism.

App Monitoring

Table 157: Monitor Microsoft SQL Server

OPS-004	MONITOR SQL SERVER

Justification	Prism Ultimate can provide greater insights into SQL Server operation, enabling you to understand how infrastructure impacts applications and to isolate problems between the two layers.
Implication	Requires an additional license that may increase cost.

Prism Ultimate provides visibility into databases, queries, and SQL-specific metrics plus behavior learning and anomaly detection. The containerized Nutanix collector is deployed on the Prism Central VM and gathers data for each configured SQL instance, using the X-Stream interface. X-FIT then analyzes the data to generate insights and alerts for troubleshooting and automation.

SQL Server is the first app monitoring content pack to be released, providing deeper insights into application performance metrics. It currently supports 100 SQL instances or 94,000 metrics.

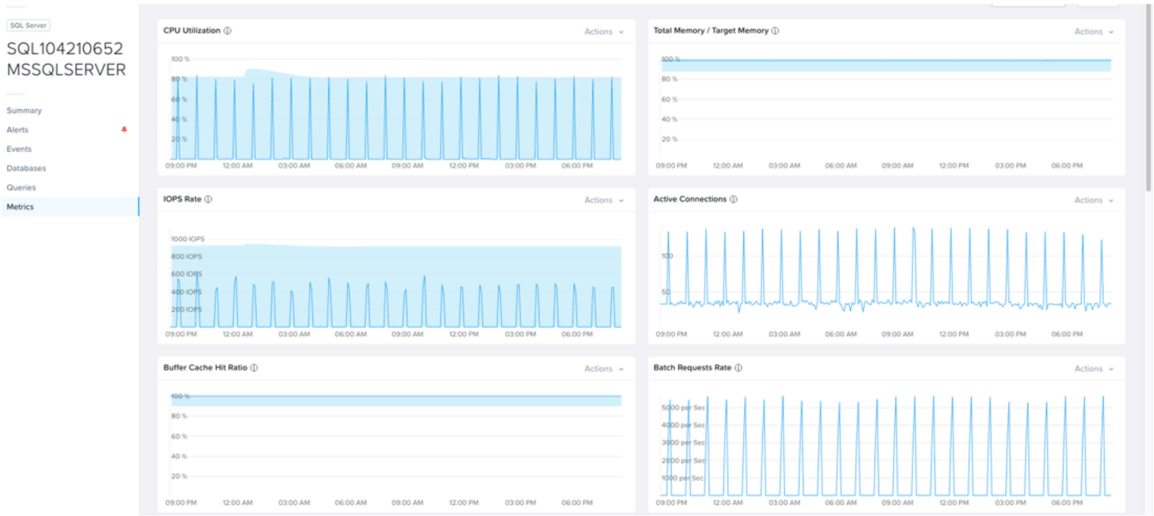
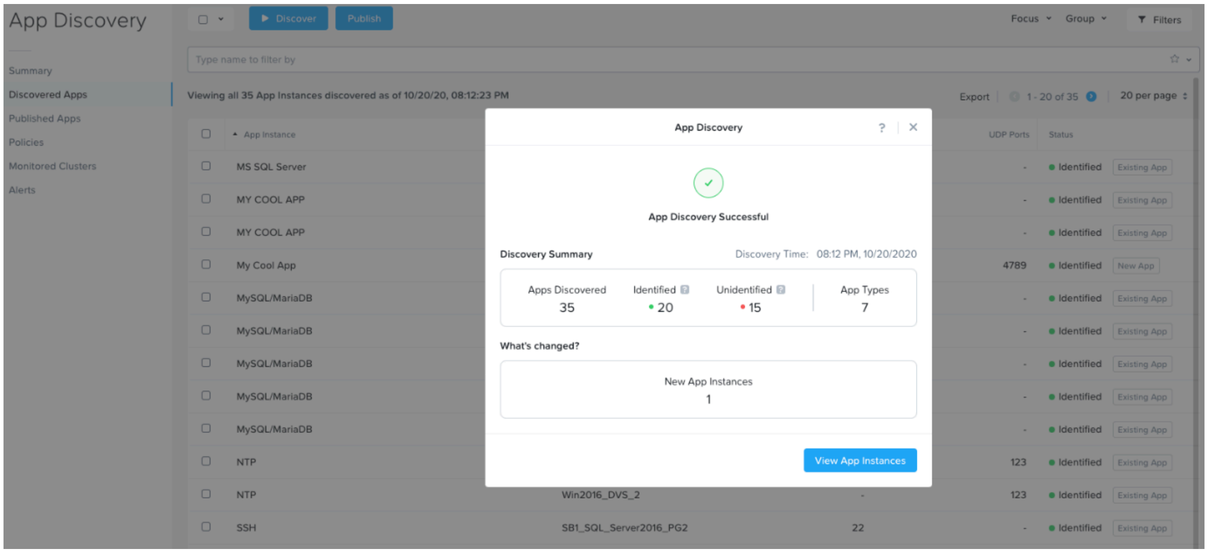


Figure 78: Monitoring SQL Server with Prism Ultimate



App Discovery

Summary
Discovered Apps
Published Apps
Policies
Monitored Clusters
Alerts

Viewing all 35 App Instances discovered as of 10/20/20, 08:12:23 PM

Export | 1 - 20 of 35 | 20 per page

☐	App Instance			UDP Ports	Status	
☐	MS SQL Server			-	● Identified	Existing App
☐	MY COOL APP			-	● Identified	Existing App
☐	MY COOL APP			-	● Identified	Existing App
☐	My Cool App			4789	● Identified	New App
☐	MySQL/MariaDB			-	● Identified	Existing App
☐	MySQL/MariaDB			-	● Identified	Existing App
☐	MySQL/MariaDB			-	● Identified	Existing App
☐	MySQL/MariaDB			-	● Identified	Existing App
☐	MySQL/MariaDB			-	● Identified	Existing App
☐	NTP			123	● Identified	Existing App
☐	NTP	Win2016_DVS_2	-	123	● Identified	Existing App
☐	SSH	SB1_SQL_Server2016_PG2	22	-	● Identified	Existing App

App Discovery

App Discovery Successful

Discovery Summary — Discovery Time: 08:12 PM, 10/20/2020

Apps Discovered	Identified	Unidentified	App Types
35	● 20	● 15	7

What's changed?

New App Instances
1

Figure 79: App Discovery

Nutanix Clusters for Hybrid and Multicloud Operations

Key Design Objectives

Table 158: Nutanix Clusters Design Objectives

DESIGN OBJECTIVES	DESCRIPTION	NEW / UPDATE / REASONING
Extend Nutanix operations to encompass public cloud	Utilizing Nutanix Clusters extends the Nutanix operational domain to encompass AWS in addition to on-premises locations	New

Architecture Overview

Nutanix enables customers running Nutanix software on-premises to deploy the same software stack on cloud computing providers such as Amazon Web Services (AWS), resulting in a more uniform hybrid cloud environment. Because Nutanix Clusters on AWS (NCA) runs Nutanix AOS and AHV —with the same

CLI, UI, and APIs—existing IT processes and third-party integrations that work on-premises operate in the cloud without modification.

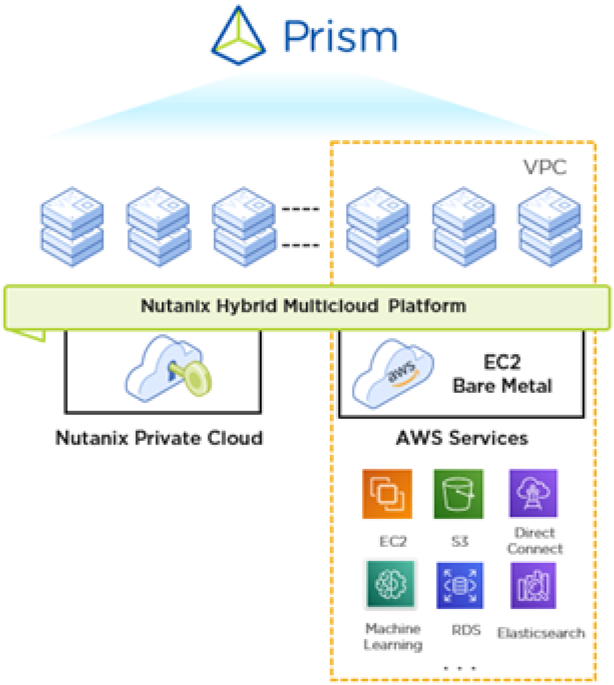


Figure 80: Nutanix Clusters on AWS

NCA implements the complete Nutanix hyperconverged infrastructure (HCI) stack using Amazon Elastic Compute Cloud (EC2) bare-metal instances. Each bare-metal instance runs the same Controller VM (CVM) and the Nutanix AHV hypervisor as used on-premises, using the AWS elastic network interface (ENI) to connect to networking. AHV user VMs do not require any additional configuration to access AWS services or other EC2 instances.

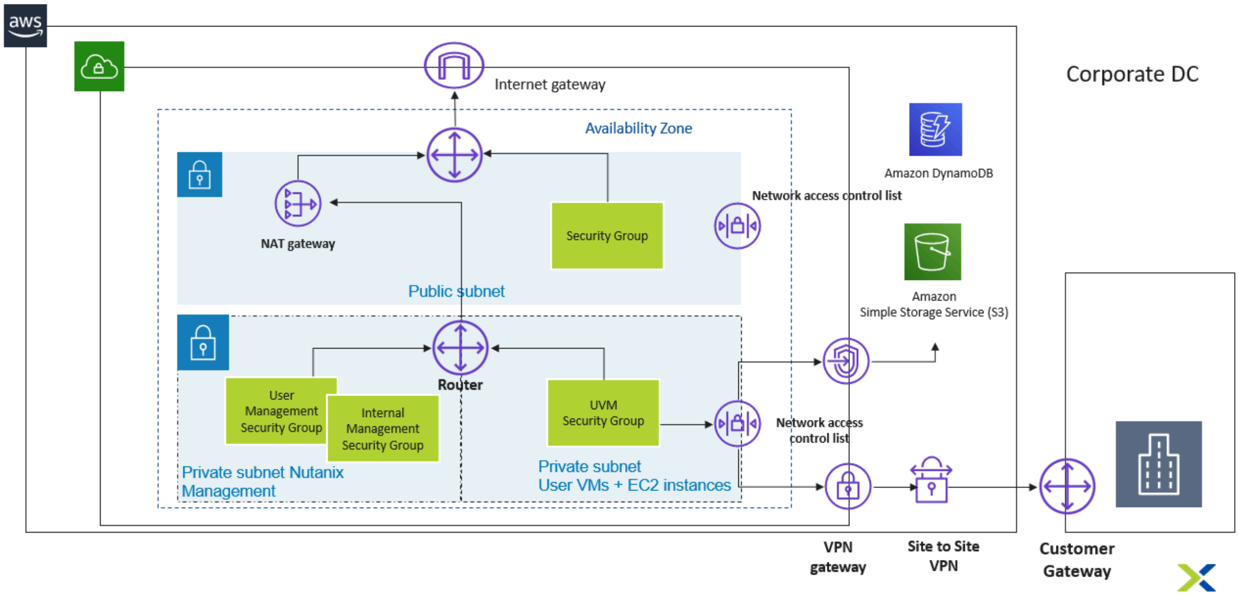

aws
Internet gateway
Availability Zone
NAT gateway
Security Group
Network access control list
Public subnet
Router
User Management Security Group
Internal Management Security Group
UVM Security Group
Network access control list
Private subnet Nutanix Management
Private subnet User VMs + EC2 instances
Corporate DC
Amazon DynamoDB
Amazon Simple Storage Service (S3)
VPN gateway
Site to Site VPN
Customer Gateway

Figure 81: Nutanix Clusters on AWS Architecture

In most ways, Nutanix Clusters on AWS is similar to running Nutanix AOS on any hardware OEM. Almost everything you know about AOS is the same. This reduces the need for different systems administrators, tooling, and scripting between private and public cloud. The following description focuses on the differences related to deployment in AWS.

Networking

AHV runs an embedded distributed network controller that efficiently integrates user VM networking with AWS networking. AHV assigns all user VM IPs to the bare-metal host where VMs are running. Instead of creating an overlay network, the AHV embedded network controller simply provides the networking information for VMs running on NCA, even as VMs move between AHV hosts. Because NCA IP address management is integrated with AWS Virtual Private Cloud (VPC), AWS allocates all user VM IPs from the AWS subnets in your existing VPCs.

NCA utilizes AWS Elastic Network Interfaces (ENI) to provide IP addresses for virtual machines that get created on EC2 bare-metal nodes using AHV's built-in IPAM. One EC2 bare-metal host can have a maximum of 15 ENIs; NCA uses the first ENI for host management (CVMs and AHV). The remaining 14 ENIs can be used as VM IPs. An ENI on the EC2 instance will only be consumed if there is a VM present on the subnet.

Networking limits and restrictions

The following limits and restrictions apply to NCA deployments:

- One subnet per ENI.
- An ENI can have up to 50 IP addresses, one of which must be the primary IP.
- The primary IP of an ENI cannot be changed or migrated to another ENI, leaving 49 IP addresses per ENI.
- The private management subnet cannot be used for deploying VMs.
- The user VM subnets cannot be shared between clusters.
- The user VMs should use a separate subnet from other AWS EC2 instances.

Currently, the maximum number of VMs supported is:

14 ENI * 49 IPs = 686 VMs per node

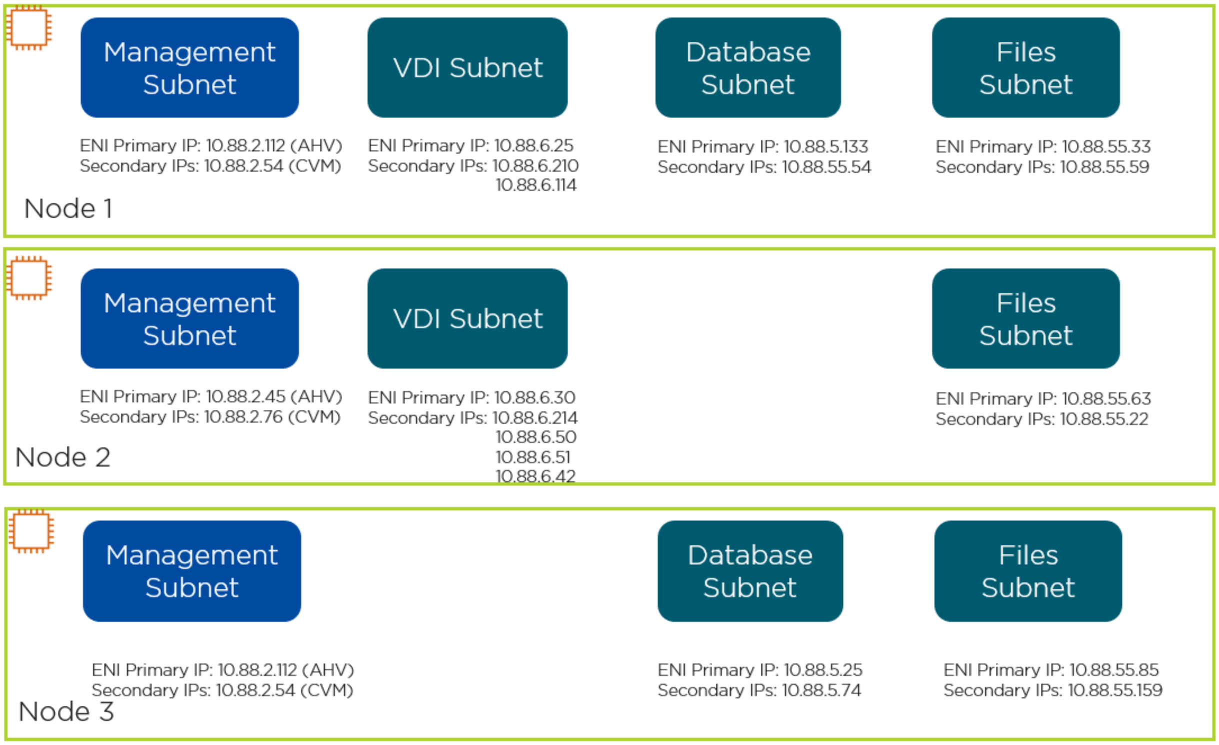

Management Subnet
VDI Subnet
Database Subnet
Files Subnet
ENI Primary IP: 10.88.2.112 (AHV)
Secondary IPs: 10.88.2.54 (CVM)
ENI Primary IP: 10.88.6.25
Secondary IPs: 10.88.6.210
10.88.6.114
ENI Primary IP: 10.88.5.133
Secondary IPs: 10.88.55.54
ENI Primary IP: 10.88.55.33
Secondary IPs: 10.88.55.59
Node 1
Management Subnet
VDI Subnet
Files Subnet
ENI Primary IP: 10.88.2.45 (AHV)
Secondary IPs: 10.88.2.76 (CVM)
ENI Primary IP: 10.88.6.30
Secondary IPs: 10.88.6.214
10.88.6.50
10.88.6.51
10.88.6.42
ENI Primary IP: 10.88.55.63
Secondary IPs: 10.88.55.22
Node 2
Management Subnet
Database Subnet
Files Subnet
ENI Primary IP: 10.88.2.112 (AHV)
Secondary IPs: 10.88.2.54 (CVM)
ENI Primary IP: 10.88.5.25
Secondary IPs: 10.88.5.74
ENI Primary IP: 10.88.55.85
Secondary IPs: 10.88.55.159
Node 3

Figure 82: Example IP Address Assignment for an NCA Deployment

The figure above illustrates that nothing is running on the third node of the cluster on the VDI subnet. We can tell this because no ENI/IPs are being used on that node. Likewise, we only have two database VMs deployed on the cluster on nodes 1 and 3.

IP Mobility

Nutanix does not currently provide a stretched layer-2 network connection to AWS. The Nutanix Leap Disaster Recovery (DR) service preserves IP and MAC addresses and can be paired with third-party software to stretch the network. Leap can automate assigning new IPs to your VMs, and it uses recovery plans that provide network mappings for remote subnets.

Compute

NCA only supports homogeneous clusters today, so you must use the same EC2 bare-metal instance type for all nodes within a cluster. However, you can have multiple clusters with different instance types that are optimally suited to the workload(s) in each cluster.

Hardware Choice

The instance types available will depend on the AWS region you select. NCA supports four possible EC2 bare-metal instances:

Table 159: Supported EC2 Bare-Metal Instances

INSTANCE	vCPU	MEMORY (GiB)	NETWORK PERFORMANCE	SSD STORAGE (GB)
z1d.metal	48	384	25 Gigabit	2 * 900 NVMe
m5d.metal	96	384	25 Gigabit	4 * 900 NVMe
i3.metal	72	512	25 Gigabit	8 x 1,900 NVMe
i3en.metal	96	768	100 Gigabit	8 x 7,500 NVMe

The i3.metal instance is a great choice for virtual desktop deployments, offering the best price/performance. The i3en.metal instance is good for database workloads and can provide lower TCO when high storage capacity is needed. The z1d.metal and m5d.metal instances are well positioned for general compute and dev/test workloads that are temporary in nature.

Minimum cluster size is 3 nodes, up to a max of 16 nodes. Each instance type consumes a certain amount of vCPU that goes against your configured limits in AWS. You need to ensure that you have sufficient headroom so that you are not prevented from deploying additional nodes in a cluster. If you are deploying a cluster for N+1 availability, your vCPU limits should account for both the additional node and any future growth.

Business Continuity Layer

Protection against site failures

- A Nutanix AWS cluster with three or more nodes is deployed with a Replication Factor (RF) of 2 by default, with nodes explicitly distributed across racks to protect against failure of a single device, node, or rack. You can also configure RF=3 to provide higher levels of redundancy.

- Similar to Nutanix on-premises operation, a single Nutanix cluster on AWS cannot protect against infrastructure failures that affect more than one rack or an entire AWS availability zone. To learn more about Rack Awareness see KB 9723 .

To protect data against large-scale infrastructure failures, Nutanix recommends you regularly back up and/or replicate data to another Nutanix cluster running on premises or in a different AWS Availability Zone (AZ). Consult Data Protection and Recovery with Prism Element for more information. You can also leverage Nutanix Xi Cloud services such as Xi Leap for DRaaS. See Nutanix Clusters on AWS for more documentation.

Protection against disruptive actions

- A simple reboot (intentional or not) of an EC2 bare metal instance only temporarily affects a Nutanix cluster running on AWS. The node automatically rejoins the cluster.
- Stopping or terminating an EC2 bare metal instance that is part of a Nutanix cluster on AWS effectively removes that node from the cluster. With the default RF=2 configuration, removing a single node does not affect availability. However, stopping or terminating more than one EC2 instance can result in data loss.
- Shutting down or powering off an AHV host running on AWS from the AHV shell results in the permanent removal of that node from the cluster. The cluster will add a node back when it self-heals. The default RF=2 configuration can tolerate unavailability of one node. Shutting down or powering off more than one AHV host from the AHV shell can result in data loss.

Security

Securing traffic to and from your AWS cluster can be accomplished using a combination of third-party firewalls, Nutanix built-in security groups, and Nutanix Flow (see the section Nutanix Flow for Microsegmentation for details).

The Nutanix Clusters console creates the following security groups at the time a cluster launches in AWS:

- Internal Management

- User Management
- UVM

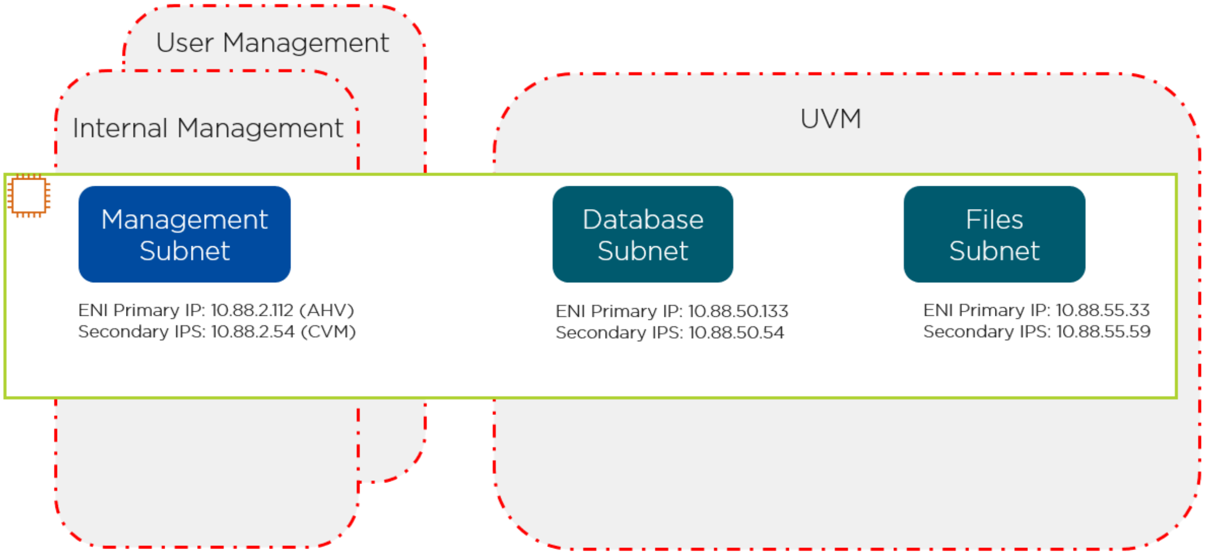

User Management
Internal Management
UVM
Management Subnet
ENI Primary IP: 10.88.2.112 (AHV)
Secondary IPS: 10.88.2.54 (CVM)
Database Subnet
ENI Primary IP: 10.88.50.133
Secondary IPS: 10.88.50.54
Files Subnet
ENI Primary IP: 10.88.55.33
Secondary IPS: 10.88.55.59

Figure 83: Security groups used by NCA

These security groups have the recommended default rules. Nutanix recommends that you do not modify internal management.

All management network ENIs created have the following security groups attached, even after the cluster is running:

- internal_management: Allows all communication between AHV and CVM within a cluster
- user_management: Allows specific ports for UVM to CVM communication

ENIs created for user VM subnets have the following security group attached:

- UVM: Allows communication between user VMs

Detailed information on allowed port groups can be found in the Nutanix Clusters on AWS Deployment and User Guide. Nutanix Flow works both in the cloud and on-prem and can be used to segment operations and to block east/west traffic.

High-Level Design Considerations

Networking

Internet Access

Portal Access: Determine how you will enable outbound access from your NCA cluster to the Cluster Portal. You can route all Internet traffic through your corporate firewall or set up an Internet Gateway in AWS. If your company is blocking outbound traffic, you also need a firewall that can block domain traffic versus subnet ranges since some of the Cluster Portal services are dynamic.

On-Premises Connectivity: There are three native ways to connect to on-premises operations from AWS:

- VPC VPN: Works well for connecting to a single site.
- Transit Gateway: Preferred when you plan to connect multiple sites to your NCA cluster.
- Direct Connect: Typically, more expensive but provides a very stable connection at your desired speed.

Table 160: Separate Subnets

AWS-01	CREATE SEPARATE SUBNETS FOR USER VMS
Justification	AWS clusters cannot deploy user VMs on the same subnet as the CVMs and AHV hosts. New subnets are also needed if you are using Leap to failover VMs into AWS. You should also use different subnets for native AWS EC2 instances.
Implication	Unless using third-party software to create a stretched layer 2 network, you will have to change your VMs' IP addresses when they move to AWS.

Business Continuity

Determine your availability needs. If you are running dev/test, you probably don't need to replicate data to another region or to an on-premises datacenter.

If you are running production workloads, you need to replicate a copy of your data on-prem, to another AWS region, or use an AHV backup product.

Table 161: Create Availability Zones

AWS-02	USE AWS DIRECT CONNECT OR CREATE A 2ND AWS CLUSTER IN A DIFFERENT AVAILABILITY ZONE.
Justification	To achieve an RPO of zero, you will need to ensure you meet the latency requirements for synchronous replication. If your latency is over 5ms round trip from your primary datacenter, you will have to create a secondary AWS cluster in a different AZ.
Implication	A disaster that affects an entire geography or a system wide event might affect more than one AZ in a given region.

Table 162: Use Replication

AWS-03	PROTECT YOUR AWS CLUSTER BY REPLICATING BACK TO ON-PREM.
Justification	You need to protect yourself from an outage in AWS. You can back up your AWS cluster using AHV-based backup software, but the region where you place the backups has to be able to be restored in a timely fashion. Also, the region you restore service to has to be close enough to meet any on-prem SLAs you might have.
Justification	You need to protect yourself from an outage in AWS. You can back up your AWS cluster using AHV-based backup software, but the region where you place the backups has to be able to be restored in a timely fashion. Also, the region you restore service to has to be close enough to meet any on-prem SLAs you might have.

Table 163: Don't Overlap Subnets

AWS-04	PLAN YOUR AWS VIRTUAL PRIVATE CLOUD NOT TO OVERLAP WITH YOUR ON-PREM SUBNETS.
Justification	Nutanix Clusters does not provide stretched layer 2 networking natively. If you want to failover or migrate your on-prem applications and also keep a subnet active on-prem, you will need a subnet that does not overlap. If your goal is only to achieve a complete failover into AWS, you can create the same subnet in AWS, since your on-premises environment won't be active after a failover occurs.
Implication	Failing over an entire subnet without re-IPing requires you to shut down those subnets on-prem if you intend to keep the connection alive.

Security

AWS Accounts: Determine if your level of risk is acceptable with the privileges delegated to the Cluster Portal. While the Cluster Portal is built with the utmost attention to security, some customers may still want a greater level of separation. To achieve this, customers can use a separate AWS account with VPC pairing to their current AWS account.

Table 164: Allow Outbound Internet Access

AWS-05	ALLOW OUTBOUND INTERNET ACCESS THROUGH AN EXISTING CONNECTION TO THE INTERNET TO ALLOW THE CLUSTER PORTAL TO MANAGE YOUR CLUSTER.
Justification	You should limit the number of attack vectors into your environment. In a hybrid cloud environment, you most likely already have an existing outbound connection. If you only have a public cloud environment, you will have to setup NAT and an Internet gateway.

Implication	You will have to pay for egress networking costs to your on-prem environment(s).

VPC Limits

Table 165: Plan VPC Limits for Growth and Availability

AWS-06	You will have to pay for egress networking costs to your on-prem environment(s).
Justification	Your vCPU limits are set at the regional level and cover many instance types. You should ensure that you have enough headroom for expansion and to accommodate a node failure. If your cluster is using RF=3, then you must ensure that you have enough spare vCPU capacity to accommodate the vCPU needs for two failed nodes.
Implication	Increasing limits affects a variety of EC2 services. You should monitor for any suspicious activity since the limit will be higher.

Nutanix Flow for Microsegmentation

Key Design Objectives

Table 166: Flow Design Objectives

DESIGN OBJECTIVES	DESCRIPTION	NEW / UPDATE / REASONING
Implement microsegmentation for greater security	Provide stateful firewalls to segment applications running in VMs on Nutanix clusters	New

Architecture Overview

Flow provides microsegmentation functionality using a distributed stateful firewall that enables granular network monitoring and enforcement between VMs running on the AHV platform as well as external entities they communicate with. Flow provides security rules, which enforce what kind of traffic is allowed between VMs. Flow configuration is done via Prism Central (PC) by defining security rules based on categories. Doing the configuration from PC allows centralized management and distributed enforcement, where the enforcement is done on multiple Nutanix clusters and multiple Nutanix AHV nodes. Flow configures OpenFlow rules in each AHV host in order to implement the security rules.

Nutanix Flow uses categories and security rules that are configurable based on defined categories.

Categories

Categories are used to define groups of entities. The category schema consists of keys and values:

- Example Key: Department
- Example Value: Engineering

A {Key, Value} pair together forms a unique tuple in the PC management domain and this tuple is used to tag VMs.

For example, a VM providing production database services may have the following categories assigned to it:

- Key: Value
- AppTier: Database
- AppType: MySQL
- Environment: Production

These categories can then be utilized by Flow to determine what security rules to apply.

Security Rules

Security rules specify the Flow policies which determine how VMs can talk with each other. These rules govern firewalls for applications and many other use cases, including isolating environments or quarantining VMs that are suspected to be infected. Based on the most common use cases, Flow provides four types of policies as shown in Table 164.

Table 167: Flow Policy Types

POLICY TYPE	USE CASE	EXAMPLE DEPLOYMENT
Quarantine	Quarantine Apps/VMs	Fault isolation and control
Isolation	Environment separation	Geographical isolations
App-Type	Application segmentation	App policy or inter-tier
AD-Type	AD/VDI segmentation	ID firewalls

Quarantine Policy

A Quarantine policy denies all traffic to/from specified VM(s)/categories.

Example: VMs A, B, and C are infected with a virus; isolate them to stop the virus from further infecting the network

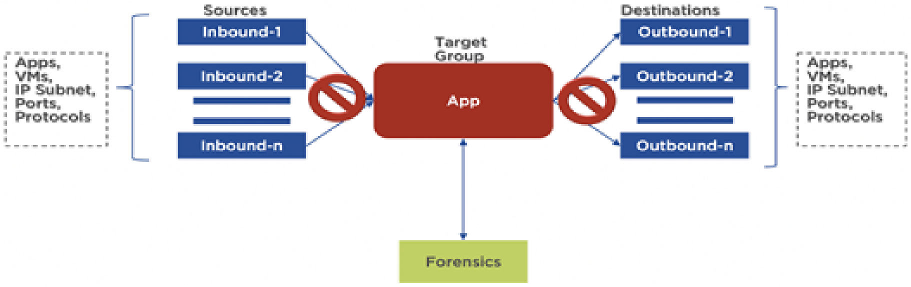


Figure 84: Effects of a Quarantine Policy

Isolation Policy

An Isolation policy denies traffic between two categories, while allowing traffic within a category.

Example: Separate tenant A from tenant B. Each environment is isolated and runs in parallel without affecting normal network communication.

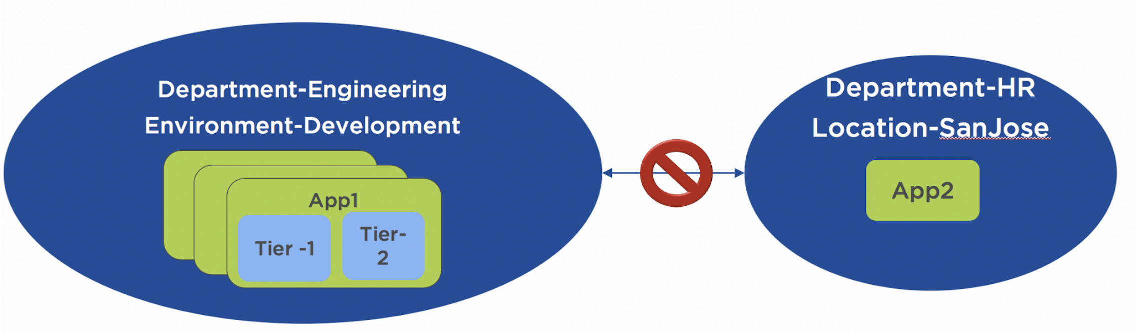


Figure 85: Effects of an Isolation Policy

App-Type Policy

An App-Type policy is used for configuring firewalls for applications running inside VMs. This rule allows you to define what transport (TCP/UDP), Port, and source/destination is allowed/denied.

- [Allow|Deny] Transport: Port(s) [To|From]

Example: Allow transport on TCP port 8080 from Category:Tier:Web to Category:Tier:App

The policy schema consists of Inbounds, Target Group, and Outbounds. Inbounds identify who can talk with the Target Group. Outbounds identify who the Target Group can talk with. Target Group identifies the VMs or applications that are being protected.

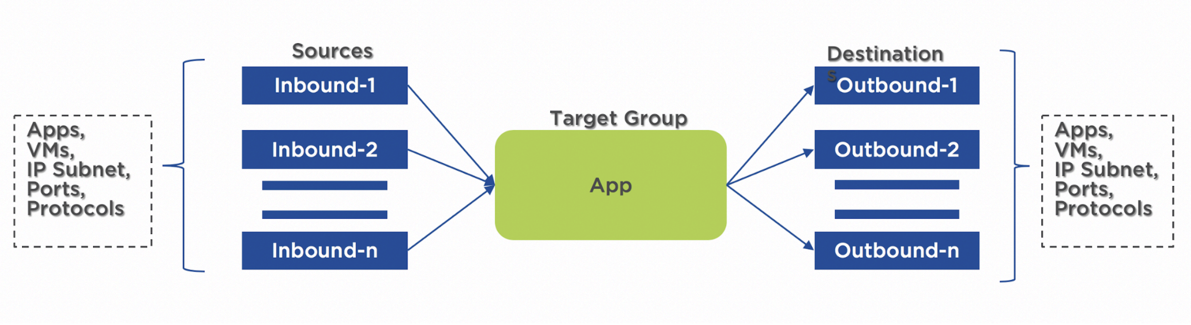


Figure 86: Illustration of an App-Type Policy

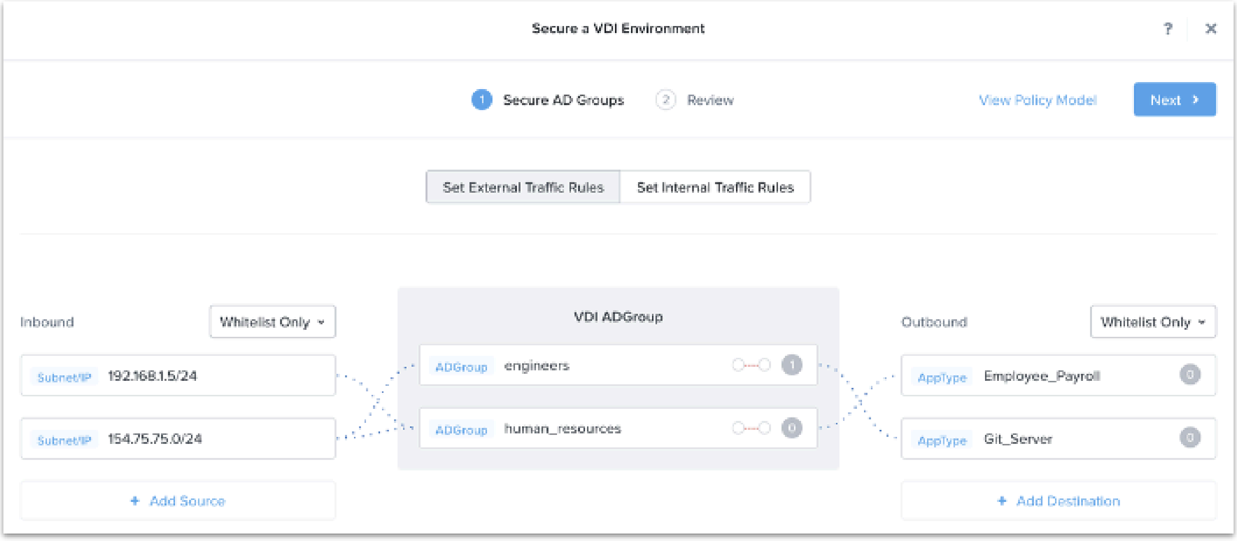
AD-Type Policy

An AD-Type policy is used for configuring firewalls for AD/LDAP users. This rule allows you to define what transport (TCP/UDP), Port, and source/destination is allowed/denied.

- [Allow|Deny] Transport: Port(s) [To|From]

Example: Allow subnets (example: clients) to talk with AD groups and allow AD groups to access payroll data

The policy schema consists of Inbounds, AD Groups, and Outbounds. Inbounds identify who can talk with the AD Groups. Outbounds identify who the AD Groups can talk with. AD Groups identify the users (based on VMs that the users have logged into) that are being protected.



Inbound: Whitelist Only

Subnet/IP 192.168.1.5/24

Subnet/IP 154.75.75.0/24

VDI ADGroup

ADGroup engineers 1

ADGroup human_resources 0

Outbound: Whitelist Only

AppType Employee_Payroll 0

AppType Git_Server 0

Figure 87: Using an AD-Type Policy to Secure a VDI Environment

The following figure shows an example of using Flow microsegmentation to control traffic for an application:

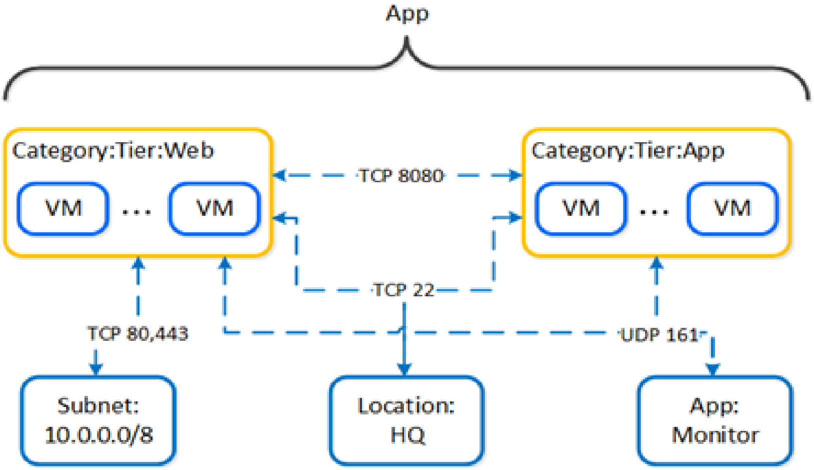


Figure 88: Using an AD-Type Policy to Secure a VDI Environment

Security Rule Actions

A security rule has an action field which determines what action is taken when a rule is matched. With AHV and Flow microsegmentation, there are two types of actions:

- Apply: Enforce the policy by permitting defined flows and blocking all others.
- Monitor: Allow all flows but highlight any packets that violate the policy on the policy visualization page.

Other Flow Features

While microsegmentation—including policy management, policy distribution, policy enforcement, and monitoring— is a core feature of Flow, the product offers many other capabilities such as:

- ID firewalls for VDI security
- Flow policy export/import
- Policy hit logs
- IPV6 enable/disable
- Service groups
- Address groups

- Audits
- Logging
- NCC alerts
- Service chains for advanced firewalls and integration with third-party devices
- Flow security central for advanced auditing, policy automation, governance
- L2 isolation
- External IPAM support
- Live migration connection tracking
- Nutanix Clusters integration

High-Level Design Considerations

Nutanix recommends the use of microsegmentation to isolate applications that don’t interact or talk to each other directly.

Table 168: Use Microsegmentation to Isolate Applications

FLOW-001	USE MICROSEGMENTATION WITH APP-TYPE POLICY TO SEGMENT APPLICATIONS
Justification	Microsegmentation between applications can decrease the attack surface if a datacenter is penetrated and also prevent ransomware and other malware from spreading.
Implication	Nutanix AHV is required to utilize Flow microsegmentation.

If you are running VDI in your environment, using flow to isolate traffic based on AD groups is also extremely useful.

Table 169: Use AD-Type Policy to Isolate VDI

FLOW-002	USE MICROSEGMENTATION WITH AD-TYPE POLICY FOR VDI DEPLOYMENTS

Justification	Microsegmentation based on AD groups can isolate VDI environments, prevent users from accessing restricted resources, and help prevent ransomware and other malware from spreading.
Implication	Nutanix AHV is required to utilize microsegmentation.

Nutanix Files for Integrated File Services

Key Design Objectives

Table 170: Nutanix Files Design Objectives

DESIGN OBJECTIVES	DESCRIPTION	NEW / UPDATE / REASONING
Enable file services from Nutanix clusters	Nutanix Files provides flexible file services implementations including: - Mixed or dedicated clusters - Flexible storage options - Multiple, separate file servers per cluster - Simple data protection and DR	New

Architecture Overview

Nutanix Files is a scale-out, software-defined file server that provides Server Message Block (SMB) and Network File System (NFS) file services to clients. Nutanix Files leverages Nutanix HCI for its core storage, networking, virtualization (AHV and ESXi) and compute requirements. Management for Nutanix Files is performed from Prism Element.

Nutanix Files instances are composed of a set of VMs (called FSVMs). For a given file server instance, Files requires at least three FSVMs running on three physical nodes to satisfy a quorum for high availability. You can scale out a given file server at any time, by adding FSVMs, up to the number of physical nodes in the HCI cluster or a maximum of 16 FSVMs. You can also scale the CPU and memory of the set of FSVMs at any time.

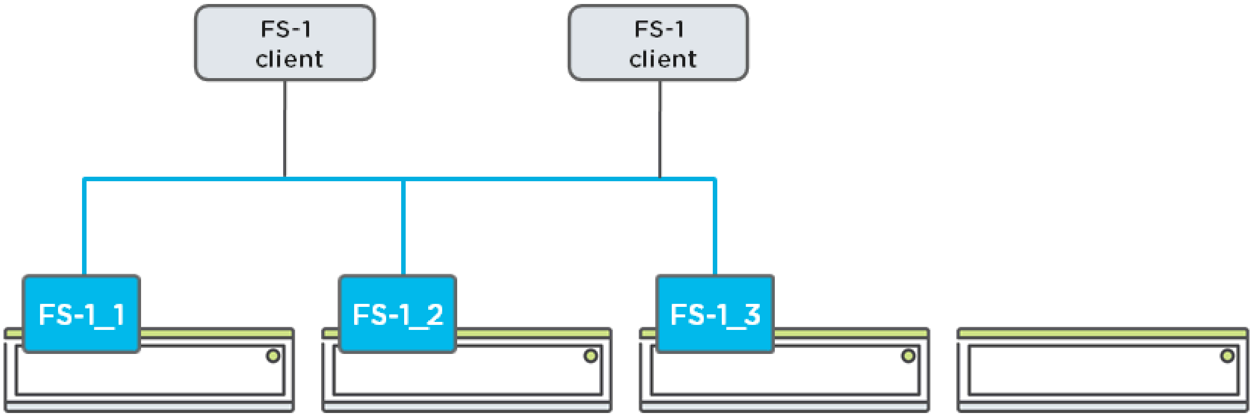


Figure 89: Example of a Nutanix Files Instance

Multiple Files instances can run on the same HCI cluster. Each instance represents an individual namespace. Each instance can scale from three FSVMs up to 16 FSVMs, assuming there are enough resources in the HCI cluster.

You can also deploy single FSVM file servers. Note that single FSVM file servers cannot scale out.

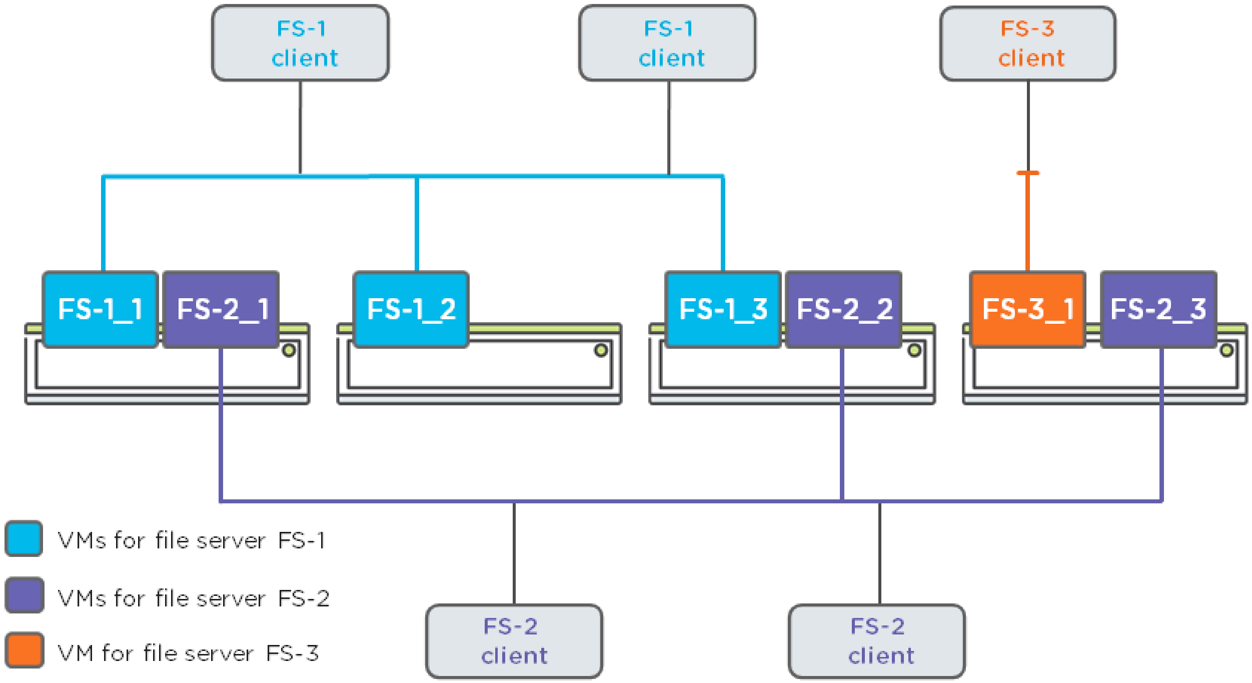


Figure 90: Three Separate Nutanix Files Instances Deployed on a Single Cluster

For file servers with three or more FSVMs, high availability during node failure and Files upgrades is achieved at the file-server level. Resources controlled by a failed or upgrading FSVM are moved temporarily to another FSVM. For single FSVM instances, hypervisor HA provides protection against node failure events.

Each file server has two networks: a client network and a storage network. The client network is used for SMB and NFS traffic and other required services such as active directory, DNS, and NTP. The storage network provides shared storage services for file shares and exports. Nutanix Volumes is used on the backend to provide shared storage, leveraging the iSCSI protocol over the storage network.

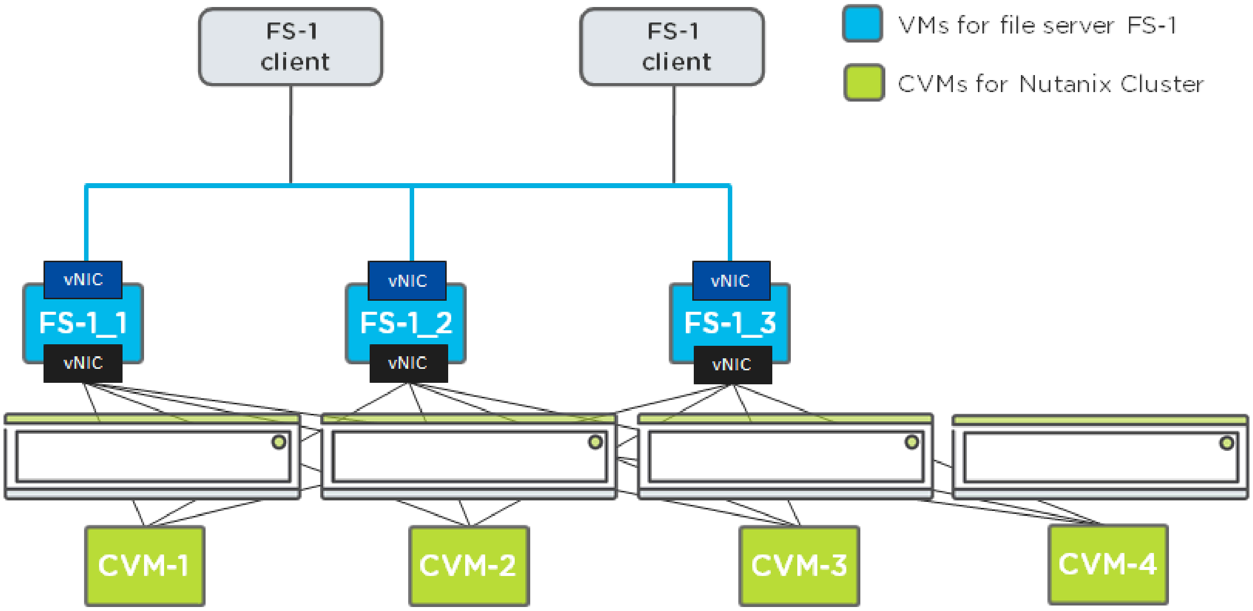


Figure 91: Relationship Between File Server VMs, CVMs, and Cluster Nodes

File shares are load balanced across FSVMs based on share type. Nutanix Files has two types of shares, distributed shares and standard shares.

Distributed shares have distributed ownership across all FSVMs in a Files cluster. Distribution occurs from the top-level directory (TLD) (\\ClusterName\DistributedShare\TLD). Each TLD can belong to a different FSVM and reside in its own volume group. Several volume groups support a distributed share; the total number of volume groups depends on the size of the Files cluster. Each FSVM has five volume groups to support distributed shares. For example, a three-node Files cluster has 15 volume groups to support shares.

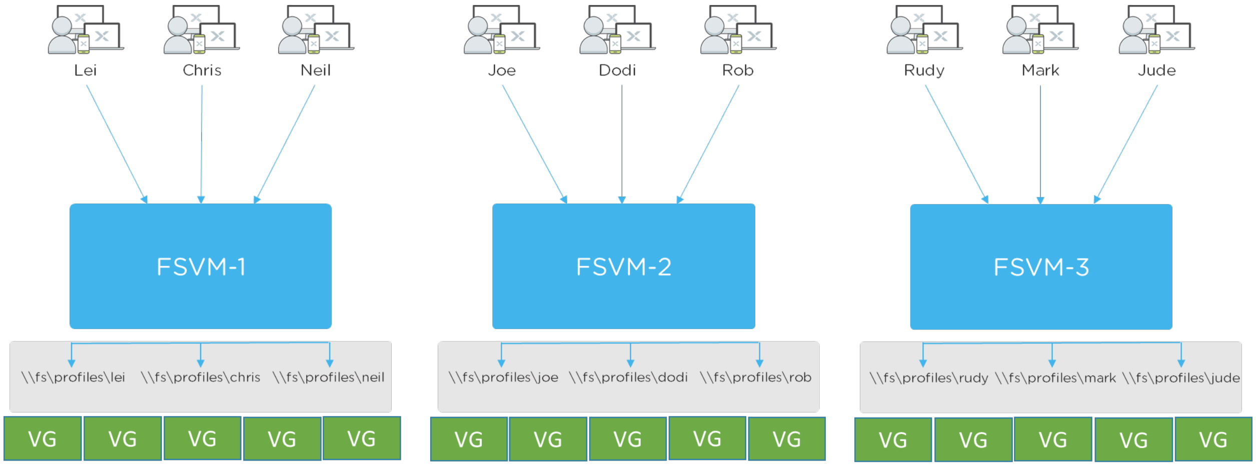


Figure 92: Distributed Shares are Distributed Across Multiple Volume Groups and FSVMs

Standard shares are owned by one file server VM (FSVM) at a time and have a single volume group to store data.

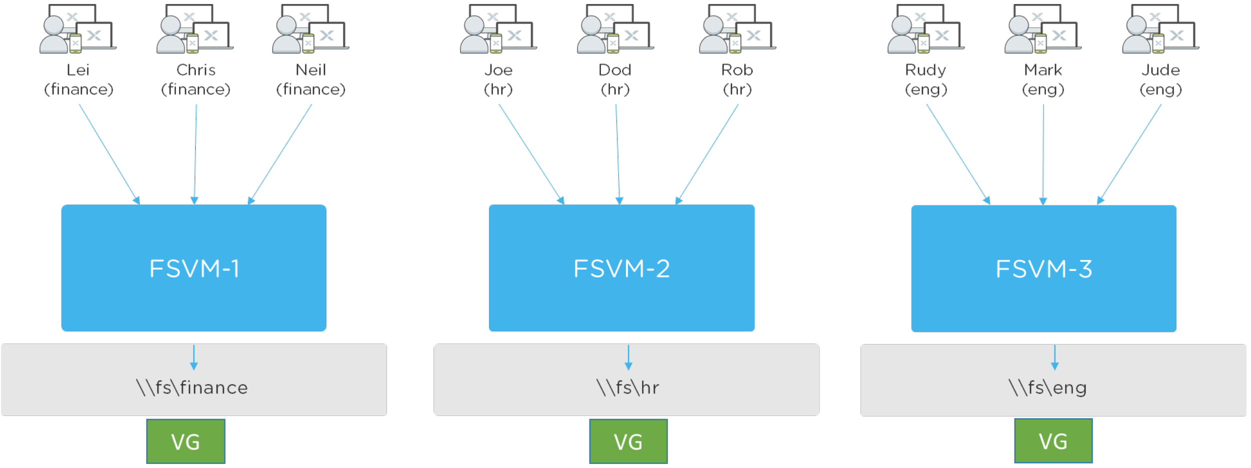


Figure 93: Standard Shares are Assigned to a Single FSVM and a Single Volume Group

For business continuity, Nutanix Files leverages the native data protection of Nutanix HCI. A protection domain is configured for each file server and native Nutanix HCI async and nearsync replication is used to protect configuration and user data.

Nutanix Files can run in a mixed environment along with other workloads, or on a dedicated cluster just for Files. Mixed environments require a Nutanix Files for AOS license, which is an add-on to your AOS license. Dedicated environments require a Nutanix Files Dedicated license. In both cases Nutanix Files is licensed based on usable TiB from the perspective of the SMB or NFS protocol, plus any space consumed by snapshots.

High-Level Design Considerations

Mixed or Dedicated Clusters

There is no hard and fast rule regarding whether to deploy Nutanix Files in a dedicated cluster or a mixed cluster with other workloads. Like any other workload, the compute and storage density of your applications, performance requirements, failure domains, management domains, and other considerations need to be weighed.

Table 171: Mixed or Dedicated Environments

FILES-001	USE NUTANIX CLUSTERS DEDICATED TO FILE SERVICES WORKLOADS
Justification	File services tend to need storage dense configurations, with a higher ratio of HDD to SSD. Unstructured data workloads typically have limited data overwrites and large amounts of cold data, which benefit from erasure coding (EC-X).
Implication	Dedicating clusters to specific workloads can potentially lead to inefficient compute and storage utilization.

Hybrid Storage or All Flash clusters?

Every environment is different, but unstructured data has a tendency to have a small amount of active data and a large amount of inactive or cold data and also tends to be storage dense. High storage density with a majority of cold data leads many Nutanix Files deployments to utilize clusters with hybrid storage (a combination of SSD and HDD in each node). Considerations for sizing hybrid clusters is discussed in the Storage Capacity Utilization and Sizing section below. The Nutanix Files Sizing Guide (Nutanix Portal account required) can also help.

Table 172: Choose Hybrid Clusters

FILES-002	USE HYBRID STORAGE NUTANIX CLUSTERS
Justification	Most unstructured data use cases work well with hybrid storage solutions where hot data resides on SSD and cold data resides on HDD. The Nutanix Distributed Storage Fabric will tier data as required between local or remote nodes and SSD or HDD based on the workload.
Implication	End user performance will be negatively impacted if an insufficient SSD tier is sized for the use case.

Number of FSVMs per File Server Instance

Each file server instance consists of some number of FSVMs. It's recommended to deploy at least 3 FSVMs for high availability at the file server level. The amount of compute you assign to each FSVM, and the number of FSVMs beyond the minimum of three, depend on your performance requirements. You need to ensure there is enough compute assigned to FSVMs to support the expected number of concurrent user connections. The release notes (System Limits section) for each version of Files contains a table that provides guidance. For example, the number of supported SMB connections per FSVM in the Nutanix Files 3.8 release is:

- 4 vCPU/12 GB memory = (supports up to) 500
- 4 vCPU/16 GB memory = 1000
- 6 vCPU/24 GB memory = 1500
- 6 vCPU/32 GB memory = 2000
- 8 vCPU/40 GB memory = 2750
- 12 vCPU/48 GB memory = 3250
- 2 vCPU/64 - 96 GB memory = 4000

You also need to ensure that enough storage IOPs and throughput are available for your specific workload. The number of FSVMs also plays a role in performance. Recommendations for sustained read and write performance per FSVM can be found in the Nutanix Files Sizing Guide (Nutanix Portal account required) and the Nutanix Files Performance Guide (Nutanix Portal account required).

Once the minimum number of FSVMs and compute resources are determined, it is recommended that you take a balanced approach to the number of FSVMs, allowing you to scale-up or scale-out if needed. Having a 16-node AOS cluster does not mean you need a 16 FSVM file server.

Table 173: Minimum FSVMs

FILES-003	USE A MINIMUM OF 3 FSVMs
Justification	Three FSVMs or more provides improved high availability with less-disruptive maintenance operations versus single FSVM configurations.
Implication	Using a minimum of three FSVMs may lead to excessive resource utilization in environments with multiple file servers. Also, ROBO configurations require a minimum of three physical nodes which could lead to the over allocation of resources.

Table 174: N+1 Configuration

FILES-004	USE AN N+1 CONFIGURATION OF FSVMs TO PHYSICAL NODES
Justification	Having one physical node more than the number of FSVMs reduces performance impact during HA events.
Implication	Maintaining an additional node leads to unused compute resources during normal operations.

Table 175: CPU and Memory Allocation

FILES-005	ALLOCATE MORE THAN THE MINIMUM FSVM CPU AND MEMORY RECOMMENDED FOR YOUR PLANNED WORKLOAD
Justification	The recommended FSVM configuration for SMB environments is the minimum to support the specified number of users. Allocating more than the minimum will help ensure better performance and avoid under sizing.
Implication	Allocating more than the minimum memory and CPU for FSVMs may lead to over-allocated resources. You can always scale-up FSVM resources to improve performance or support more users online when needed.

Number of File Server Instances

Many environments will need to deploy multiple file server instances on the same hardware. Some of the considerations for deploying multiple file servers include:

- Replacing an existing file server environment with multiple namespaces. Each file server instance represents a namespace.
- Different remote replication (RPO) requirements.
- Multiple untrusted Active Directory domains. Nutanix Files supports one registered Active Directory per file server instance. Domain trusts are required for accessing a common file server.
- File server administration between different teams or tenants may require multiple file servers with separate management domains.
- Nutanix Files supports one external VLAN per file server instance. Non-routed VLANs may require their own file server instances.

Once the number of file server instances is determined, you can then determine the total compute requirements per physical node.

Table 176: Minimum Number of File Servers

FILES-006	USE THE MINIMUM NUMBER OF FILE SERVER INSTANCES FOR YOUR NEEDS.
Justification	Using the minimum number of file server instances helps simplify management and reduce wasted resources.
Implication	Single, large file server instances can result in longer maintenance operations, such as scaling-up or upgrades.

File Server Networking

It's recommended that the file server client network be in close relation to the clients accessing the shares for best performance. It's also recommended that you place the file server storage network on the same VLAN as eth0 of the CVM. A flat non-routed storage network for iSCSI traffic will ensure the best performance.

Table 177: File Server Storage Network

FILES-007	PLACE THE FILE SERVER STORAGE NETWORK ON THE SAME VLAN AS THE eth0 INTERFACE OF THE CONTROLLER VMS.
Justification	Using the same network VLAN for the FSVM storage network and the CVM iSCSI network ensures the best performance.
Implication	Files does not currently support the iSCSI network segmentation feature of AOS. For security purposes, some environments may require complete separation of the iSCSI network from any management network.

Table 178: Shortest Route Between Client and Server

FILES-008	PLACE THE FSVM CLIENT NETWORK SO IT USES THE SHORTEST NETWORK ROUTE FOR THE MAJORITY OF FILE SERVER CLIENTS.

Justification	Ensuring the shortest route between client and server will give the best performance for your end users and applications.
Implication	Creating file servers with networks local to your end users can improve performance but may lead to more complex environments, increasing administrative burden.

Standard Shares or Distributed Shares?

Distributed shares are recommended for the majority of use cases. Distributed shares will help to ensure data is hosted by all FSVMs. Data must be spread across multiple top-level directories within a distributed share to ensure load balancing. Standard shares use an individual volume group and can only scale to 280TB (with the Files 3.7 release). Standard shares are also only hosted by one FSVM at a time. Multiple standard shares would be needed to load balance an environment across FSVMs.

Table 179: Utilize Distributed Shares

FILES-009	USE DISTRIBUTED SHARES
Justification	Distributed shares provide automatic load balancing of top-level directories across FSVMs. Distributed shares help to limit the number of shares needed to load balance an environment, simplifying administration.
Implication	Distributed shares do not support files in the root of SMB shares today (Files 3.7 has support for NFS). Using the Nutanix Files MMC to manage top level directories can add some complexity to administration.

Storage Capacity Utilization and Sizing

Nutanix Files leverages the core data services of the Nutanix platform, including fault tolerance and replication factor options. For data reduction, inline LZ4-based compression per file share (performed by Nutanix Files) and erasure coding at the storage container level are recommended.

Deduplication is not recommended. Any storage sizing exercise for Files should take into account the potential for savings based on both compression and erasure coding. File types should also be taken into account. For example, file types like video and image data will not benefit from storage compression. Cold data benefits from erasure coding.

Storage capacity consumed by self-service snapshots and/or data protection snapshots should also be factored into storage sizing. Nutanix Sizer helps to determine potential capacity when considering data reduction, snapshot overheads and change rates for Nutanix Files workloads.

Table 180: Compression

FILES-010	USE FILE SERVER SHARE-LEVEL INLINE COMPRESSION
Justification	Compression provides considerable space savings with minimal performance impact in Files environments.
Implication	Small potential for performance impact.

Table 181: Erasure Coding (EC-X)

FILES-011	USE AOS CONTAINER-LEVEL ERASURE CODING
Justification	Considerable storage space savings will be obtained by ensuring EC-X is enabled in Files environments.
Implication	Files environments with heavy overwrite activity could see a performance impact with EC-X enabled.

Table 182: Deduplication

FILES-012	DO NOT ENABLE AOS CONTAINER-LEVEL DEDUPLICATION
Justification	Testing has shown minimal space savings benefits when using storage container deduplication with Nutanix Files.
Implication	A small amount of additional space utilization may occur.

Table 183: Scheduling Snapshots

FILES-013	USE TELESCOPING SNAPSHOTS SCHEDULES
Justification	Telescoping schedules (x hourly, y daily, z weekly) ensures efficient space utilization while minimizing performance impact.
Implication	Keeping fewer long-term snapshots limits fine-grained recoverability over time.

Disaster Recovery

Nutanix Files supports both async and nearsync replication. Files deployments commonly use storage-dense nodes which historically have had snapshot frequency limitations when using AOS based data protection.

Nutanix Files 3.8 introduces Smart DR. With Files Smart DR the replication engine is no longer managed by the AOS clusters using native Nutanix protection domains. Nutanix Files now directly manages the replication. Because Files manages replication, node density limits specific to AOS snapshots and replication no longer apply. You can now use our most storage-dense nodes, which support up to 350TB of hybrid storage, with the benefits of native remote replication.

If you choose to continue to use AOS based snapshots and replication, follow the node density, SSD, and CVM compute requirements for your expected RPO. Read more in the Data Protection Guide.

Table 184: Disaster Recovery

FILES-014	USE THE MINIMUM NUMBER OF DISASTER RECOVERY SNAPSHOTS NEEDED FOR YOUR BUSINESS RPO REQUIREMENTS.
Justification	Keeping a smaller number of short duration snapshots ensures the most efficient storage utilization.
Implication	Retaining fewer long-term snapshots limits fine-grained recoverability over time.

Table 185: Disaster Recovery

FILES-015	USE FILE ANALYTICS
Justification	File Analytics provides native auditing, anomaly detection, data age categorization, reporting and ransomware protection. These features help you better understand your unstructured data and can protect you from unwanted activity.
Implication	File Analytics consumes additional compute and storage resources on your cluster. The File Analytics release notes provides guidance on sizing your File Analytics VM.

Nutanix Objects for Integrated Object Storage

Key Design Objectives

Table 186: Design Objectives

DESIGN OBJECTIVES	DESCRIPTION	NEW / UPDATE / REASONING

Enable S3-compatible object store within Nutanix clusters	Nutanix Objects lets you create flexible object stores suitable for a variety of use cases. Features include: - Mixed or dedicated clusters - Flexible storage options - Advanced storage efficiency - Multiple object stores per cluster - Object stores spanning multiple clusters - Versioning - WORM - Object lock - Data protection and DR	New

Architecture Overview

Nutanix Objects is an S3-compatible, software-defined, scale-out object storage solution. Nutanix Objects leverages Nutanix AOS for its core storage, network, virtualization (AHV and ESXi) and compute requirements. Management of Nutanix Objects is performed from Prism Central.

Nutanix Objects is deployed as a set of Kubernetes Pods running on top of the Nutanix Microservices Platform (MSP). MSP is a fully managed Kubernetes deployment, where the entire lifecycle is managed by the Nutanix platform. There are two types of services which are deployed on MSP, Objects Worker Nodes (Workers) and Objects Load Balancers.

Nutanix Objects deployments can start as small as three nodes and scale-out across multiple AOS clusters. This multi-cluster scalability ensures that you can build large-scale Objects deployments with petabytes of data and billions of objects. Objects deployments can live on dedicated AOS clusters (Objects Dedicated) or be deployed alongside existing workloads.

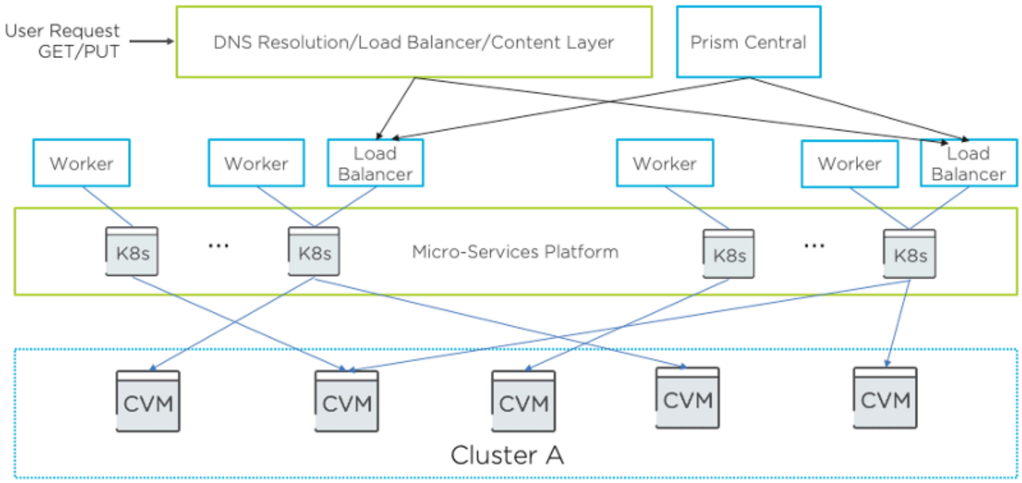


Figure 94: Nutanix Objects Deployments

Multiple Objects instances can run on the same HCI cluster. These instances can represent different namespaces (ex. customer-a.domain.tld, customer-b.domain.tld) or different use-cases (ex. backups.domain.tld, splunk.domain.tld). Each instance can scale from three Workers to 31 Workers in a single cluster.

For Object Stores with three or more Workers, high availability during node failure and Objects upgrades are performed at the object server level. All API requests for objects are redirected to other Workers by the Load Balancers during an HA or Upgrade event. For single-node Objects instances (Dev/Test use cases), hypervisor HA is utilized during HA or Upgrade events.

Each Worker has two networks: An Internal network and an External network. Each Load Balancer (has two networks as well: An Internal network and an External network, used for client access. The Worker utilizes the Internal network interface to communicate to the Load Balancer, while the Load Balancer utilizes its External network interface to communicate with clients utilizing the S3 API protocol.

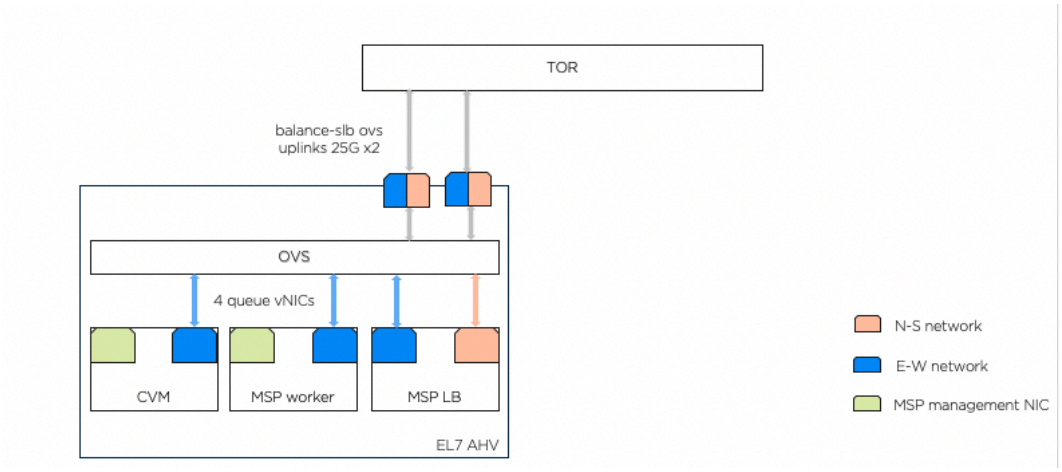


Figure 95: Internal and External Networks

Hints about the physical location of the metadata are stored at the Objects layer, enabling faster reads. This is because the hint can be passed for the lookup at the CVM layer, and we can skip the vBlock lookup. This keeps the latency for time-to-first-byte low.

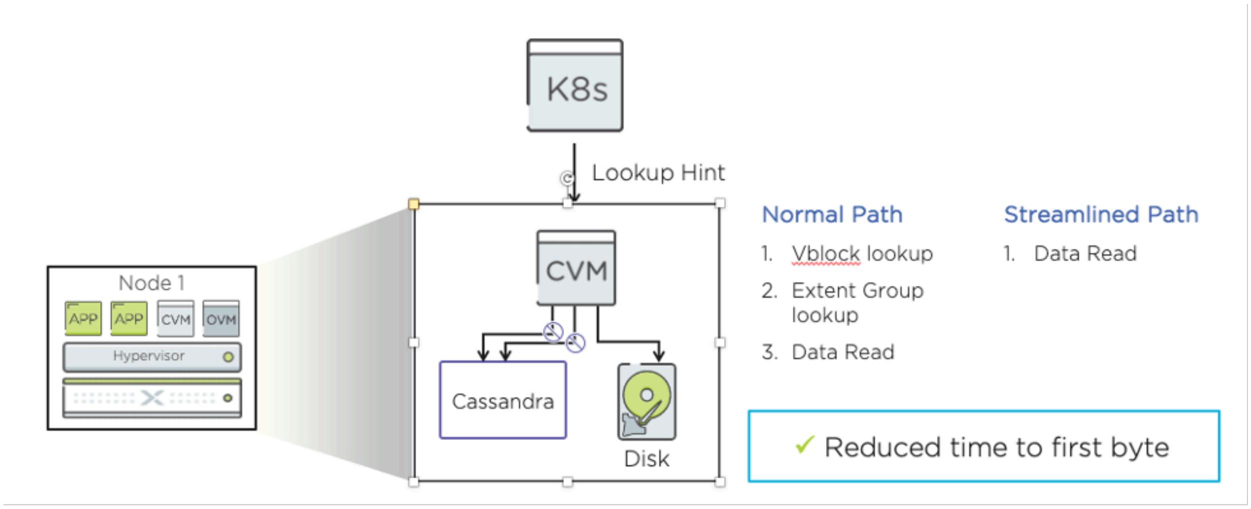


Figure 96: Lookup Hints

For business continuity, Nutanix Objects has replication built in. Streaming replication allows objects to be replicated at the bucket level from one cluster to another. Given that objects are inherently immutable—meaning a particular object cannot be edited—having a replicated copy of the object that may or may not be versioned gives an organization a powerful business continuity option. Coupling Versioning and WORM (Write Once, Read Many) allows you to keep and store copies of objects for a set period. WORM ensures that even an administrator or Nutanix support engineer cannot delete your objects, allowing you to meet compliance needs or protect against ransomware.

Nutanix Objects may run in a mixed environment along with other workloads, or in a dedicated environment just for Objects. You can license for mixed environments with the Nutanix Objects for AOS add-on to your AOS license. Licensing for dedicated environments requires the Nutanix Objects Dedicated license. In both cases Nutanix Objects is licensed based on Used TiB from the perspective of the S3 protocol, not the amount of storage used on disk after Compression or Erasure Coding.

High-Level Design Considerations

Objects Networking

Table 187: An Object Store Requires Internal and External Networks

OBJECTS-001	PROVIDE INTERNAL AND EXTERNAL NETWORKS
Justification	The internal network is used for communication between the components of an object store. The external network is used for client communication.
Implication	N/A

The Internal network needs to have IPAM enabled for that network in AHV. See the Network Configuration section of the documentation for more info.

Mixed or Dedicated Environments?

Table 188: Mixed or Dedicated Objects Clusters

OBJECTS-002	MIXED OR DEDICATED OBJECTS CLUSTERS?

Justification	Mixed Objects clusters, combining Objects and production workloads, are most often used for small deployments or alongside applications that are actively utilizing object-based storage. Dedicated Objects clusters are for large object-based storage deployments, such as backup/archive, big data, or object stores where a high level of performance is desired.
Implication	Mixed clusters will be utilizing resources alongside production workloads, so proper sizing of existing environments must be considered. Dedicated clusters will require their own hardware, which includes the need to size for rack space, power, cooling, networking, and many other physical factors.

Whether to deploy Nutanix Objects in a mixed or dedicated environment is mostly determined by your workload needs. Like any other workload, the compute and storage density of your applications, performance requirements, failure domain, management domain and other considerations need to be weighed.

Hybrid Storage or All Flash Clusters

Table 189: Hybrid or All Flash Clusters

OBJECTS-003	HYBRID OR ALL FLASH AOS CLUSTERS

Justification	Hybrid clusters that combine HDD and flash storage are used for very large deployments where the cost per TB of raw capacity is the most important concern. Use-cases for Hybrid clusters often include backup, archive, and big data. All Flash clusters are used for smaller deployments or when high levels of performance and the lowest application latency are required.
Implication	Hardware costs are often much higher for All Flash clusters vs Hybrid.

Given that objects are write-cold, the decision between Hybrid vs All Flash infrastructure primarily comes down to workload access patterns and cost. If access patterns are highly random and need very low latency, then consider All Flash. If workloads are highly sequential (e.g., Backup/Archival) hybrid may be best. Utilizing Nutanix's standard sizing methodologies is usually the best way to size infrastructure for Nutanix Objects.

Storage Efficiency

Table 190: Enable Compression

OBJECTS-004	Enable Compression
Justification	Enable compression for all workloads. Erasure Coding is enabled for all Nutanix Objects deployments by default.
Implication	Compression can significantly reduce storage consumption for many data types.

Nutanix Objects utilizes AOS and all its storage efficiency capabilities to ensure that your object store consumes storage as efficiently as possible. After seven days, all objects are erasure-coded because Objects utilizes a "never overwrite" architecture by design.

Number of Worker Nodes Per Object Server Instance

Table 191: Number of Worker Nodes

OBJECTS-005	NUMBER OF WORKER NODES
Justification	The number of worker nodes should first be based on resiliency requirements. For anything other than Dev/Test, multiple workers are recommended. Light, medium, or heavy performance configurations can be chosen based on your application's required requests per second.
Implication	More memory and CPU capacity are needed as the number of worker nodes increase.

Nutanix Objects is a scale-out platform rather than scale-up. As performance needs increase, you simply add more workers.

Recommended Configurations:

- Light Performance (1,000 requests per second) – (3) Workers, 10vCPU/32GB Memory each, (2) Load Balancers, 2vCPU/4GB Memory each
- Medium Performance (10,000 requests per second) – (4) Workers, 10vCPU/32GB Memory each, (3) Load Balancers, 2vCPU/4GB Memory each
- Heavy Performance (20,000 requests per second) – (5) Workers, 10vCPU/32GB Memory each, (4) Load Balancers, 2vCPU/4GB Memory each
- Custom – Scale out worker nodes as performance needs increase. Maximum number of worker nodes is equal to the number of nodes in an AOS cluster. A maximum of (4) Load Balancers can be deployed
- Single Node – (1) node deployments for Dev/Test can be created by choosing a custom deployment and setting the resources to 10vCPU/32GB Memory

Multiple Object Store instances

Table 192: Multiple Object Stores

OBJECTS-006	MULTIPLE OBJECT STORES
Justification	Multiple object stores can be deployed on a single cluster to support: - Multi-tenancy: Requires authentication across multiple domains. - Different VLANs: In a highly segmented network with routing concerns between VLANs, a customer may choose to deploy multiple object store instances across each network.
Implication	More resources (CPU, Memory) will be required for the additional Worker and Load Balancer nodes. Prism Central and all AOS nodes will need to have trunked access to all VLANs needed for deployment.

Nutanix Objects supports deploying multiple Object Store servers. This allows you to deploy multiple instances based on your requirements. This could be used in a multi-tenant situation, different networks, etc.

Multi-Cluster Object Stores

Table 193: Multi-Cluster Object Stores

OBJECTS-007	MULTI-CLUSTER OBJECT STORES

Justification	Very large, multi-petabyte object stores can be spread across more than one AOS cluster. This can also allow you to utilize land-locked storage in other clusters to scale out the storage available to Nutanix Objects.
Implication	All clusters utilized for multi-cluster object stores must be within the same datacenter and registered to the same Prism Central instance.

Since Nutanix Objects utilizes Nutanix AOS for storage, it is important that clusters are sized to meet AOS recommendations. For very large multi-petabyte object stores, you may split the object store across multiple AOS clusters.

Objects enables you to utilize storage from other AOS clusters within your datacenter. For large, multi-rack object stores you can utilize Nutanix Objects to deliver a very large namespace spread across multiple clusters and even multiple hypervisors.

Multi-cluster object stores are also useful when you need to utilize “land-locked” storage from other Nutanix clusters, ensuring you utilize your existing capacity efficiently.

Versioning

Table 194: Enabling Versions

OBJECTS-008	ENABLE VERSIONING?
Justification	Versioning is used when an application or organization needs to retain the old versions of changed objects.
Implication	Each version is another copy of the data that lives on disk. Additional capacity is utilized to store the old versions.

Object versioning ensures that previous versions of an object are kept when the object is overwritten with a PUT or POST. Previous versions can be set to be deleted after they reach a certain age.

WORM (Write Once, Read Many)

Table 195: WORM

OBJECTS-009	WORM (WRITE-ONCE, READ-MANY)
Justification	Versioned WORM can be used when an application requires that a DELETE request for an object must occur, but the organization needs to retain the previous version of that object for a specified period. Versioned WORM can be used when an application requires that a DELETE request for an object must occur, but the organization needs to retain the previous version of that object for a specified period. Non-versioned WORM is used when an organization wants to block any DELETE operations on the objects in a particular bucket.
Implication	Versioning utilizes additional storage.Objects effected by a WORM policy cannot be deleted, utilizing cluster storage for the period specified in the policy. Objects effected by a WORM policy cannot be deleted, utilizing cluster storage for the period specified in the policy.

Nutanix Objects supports versioned and non-versioned WORM. WORM ensures that no one can permanently delete any object within a specified window of time (e.g., 3 years). WORM is set at the bucket level.

WORM is very useful for protecting against ransomware, as the object store will deny any object changes or deletions.

Versioned WORM allows DELETE operations to occur, causing the current version of an object to be labeled as a previous version, while retaining no current version. This is useful when the application needs to be able to do

DELETE operations, but the organization requires that the object be retained for a certain period.

Non-versioned WORM will not allow any DELETE operations to occur on any object in a bucket with WORM enabled.

Object Lock

Table 196: Object Lock

OBJECTS-010	OBJECT LOCK
Justification	Used for applications that require the ability to set WORM policies on individual objects via API (e.g., Veeam).
Implication	Objects clusters must be sized for additional storage requirements for objects marked with WORM policy.

Object Lock is object-level WORM that can be set per object by applications via APIs.

Lifecycle Policies

Table 197: Lifecycle Policies

OBJECTS-011	LIFECYCLE POLICY
Justification	A Lifecycle policy can be set when an application or organization requires the ability to delete objects after they reach a certain age. This is often used in conjunction with a WORM policy to ensure that objects are kept for the amount of time necessary to meet regulations, but no longer than that.
Implication	Objects are automatically deleted once they reach the specified age. Ensure this is the desired behavior or retain replica copies on another cluster using streaming replication.

Lifecycle settings on a bucket allow you define the time when objects are permanently deleted. WORM settings always take precedence over Lifecycle settings, ensuring objects are kept for the required amount of time.

Streaming Replication

Table 198: Streaming Replication

OBJECTS-012	STREAMING REPLICATION
Justification	Backups of objects need to be stored off-site or be made highly available across multiple sites.
Implication	Duplicate hardware required for secondary object store. Prism Central at both sites must be set up as an Availability Zone in order for Streaming Replication to see both sites.

Nutanix Objects has built in, object-based replication. Streaming Replication enables you to replicate objects from one cluster to another with an RPO that can be measured in seconds.

Nutanix Mine for Integrated Backup

Key Design Objectives

Table 199: Nutanix Mine Design Objectives

DESIGN OBJECTIVES	DESCRIPTION	NEW / UPDATE / REASONING
Enable integrated backup	Nutanix Mine integrates backup into the Nutanix ecosystem using Nutanix-controlled infrastructure as a backup target and integrating with proven third-party data protection solutions.	New

Architecture Overview

Nutanix Mine is an integrated backup solution, combining the performance, scalability, and ease of use of Nutanix HCI with the robust feature set of industry-proven backup vendors. Nutanix Mine automates the provisioning of your secondary storage solution, optimizing performance and ensuring adherence to backup vendor best practices.

Mine can be purchased and configured in a number of predefined sizes and can be scaled-out as needed:

Table 200: Nutanix Mine Configurations

	XSmall	Small	Medium	Scale-out
Rack Size	2U	2U	4U	2U
# of Nodes	4	4	4	2
Raw Capacity	48 TB	96 TB	192 TB	96 TB
Usable Capacity (Extent store in N+1)	15TB	30TB	60TB	+40TB
Effective Capacity (assuming 2:1 comp/ dedup	30-50 TB	60-100 TB	120-200TB	60-100TB
Scale-out unit	XS	Single NX8235-G7	Single NX8235-G7	-

At the time of this writing, Nutanix Mine supports two backup vendors: Veeam and HYCU. For future Mine supported backup vendors, reference the Mine section in the Nutanix Portal Documentation

Veeam

Veeam Backup & Replication (VBR) provides a full suite of backup capabilities for all Nutanix-supported hypervisors as well as physical workloads. In addition to VM workloads, Veeam supports the ability to back up Nutanix Files.

- VBR leverages a distributed architecture in which backup components such as Backup Proxies, Backup Repositories, and management consoles can be deployed in multiple sites as needed, to maximize performance and minimize performance bottlenecks.
- Veeam leverages Microsoft Windows as its underlying operating system for the VBR server as well as Backup Proxies for vSphere and Hyper-V hypervisors. Configuring backup jobs for vSphere, Hyper-V, and physical workloads is performed within the VBR management console.
- For Nutanix AHV, a separate Linux-based Veeam AHV Backup Proxy is deployed on the AHV cluster(s) you wish to back up, and backup job configuration and management is performed via a separate web interface.
- Backup jobs are defined on a per-VM-group basis, with backup schedules, archival implementation, and RTO and RPO targets defined by the user for each group.
- Veeam supports the ability to tier to S3-compatible object storage, which, with its WORM capabilities, can help defend against ransomware attacks.

HYCU

HYCU provides native Nutanix backup support with an emphasis on simplicity and ease of use via web-based administration and configuration. HYCU supports:

- Backup for Nutanix AHV and VMware vSphere workloads with native Nutanix snapshot awareness.
- Use of Nutanix Change File Tracking (CFT) APIs for optimized backup of Nutanix Files.
- Ability to back up physical servers as well as integrated application awareness for Oracle Database, SAP HANA, Microsoft SQL Server, and others.
- The ability to backup directly to an S3-compatible object store as well as the ability to schedule an archival job to replicate backups from the Mine appliance to object storage.

HYCU closely integrates with Nutanix Disaster Recovery Replication technology to allow for a distributed architecture and to accommodate ROBO workloads in which native Nutanix snapshots replicated via Nutanix Protection Domains can be leveraged within the HYCU backup solution.

High-Level Design Considerations

Choosing a Backup Vendor

Table 201: Choose a Backup Vendor

BAC-001	CHOOSE A BACKUP VENDOR TO USE WITH NUTANIX MINE
Justification	Choose a backup vendor from the list of qualified vendors for use with Mine.
Implication	Your current backup vendor may not be qualified.

When deciding which Mine Backup solution to use in your deployment, the choice comes down to which solution best meets your business and technical requirements and your budget. Key areas to consider when choosing a backup vendor are:

- Primary workload support (including legacy and current OSes, applications, and File Storage)
- Backup and retention requirements
- Staff skill set and training (architecture, administration, daily operations)
- Scalability of the solution
- Licensing costs and model (including existing licensing agreements)
- Integration with the existing environment
- Migration of existing backup repositories and data
- Management plane technical features and performance
- Simplicity or complexity of the solution
- Features or products that support only one hypervisor or the other

- Satisfaction with existing backup solution(s)
- ROI/TCO of the full stack

Freedom of choice is a key tenet of Nutanix. After considering these factors, it is likely that one backup platform stands out as the best option for your deployment. No matter which backup software you choose, the solution is backed by world-class Nutanix support as well as the chosen backup vendor's support.

Deploy Mine in Conjunction with Nutanix Objects

Table 202: Utilize Objects for Archival

BAC-002	UTILIZE NUTANIX OBJECTS AS AN ARCHIVAL TIER FOR BACKUPS
Justification	Both Veeam and HYCU support S3-compatible object stores, so you can deploy Nutanix Objects as an archival tier with either option for ransomware protection and operational and cost efficiency. Optionally HYCU can be deployed to backup directly to an S3-compatible object store
Implication	May require additional licensing for Nutanix Objects.

Nutanix Mine can be deployed in a number of configurations depending on backup storage capacity and performance. Because both Veeam and HYCU support offloading backups to an S3-compatible Object Store, Nutanix Objects can be leveraged as an archival tier for backups. Using Nutanix Objects in conjunction with a Nutanix Mine deployment offers the following benefits:

- Ransomware protection: By using WORM-enabled Buckets and Objects' S3-compatible Object Lock API, backups can be protected against ransomware and deliberate or accidental deletion.
- Greater flexibility: With the use of Nutanix Objects, backup administrators can choose to keep backup data on either or both the Mine Cluster and

Object Store and leverage the benefits of Objects Replication to distribute backup copies across datacenters or regions.

- Cost-efficiency: For long-term retention requirements, administrators can choose to tier off backups to Objects and free up Mine Secondary storage for more recent or higher priority backups

The following diagram highlights the different architectural patterns in which Nutanix Mine can be deployed across single or multiple datacenters and/or regions.

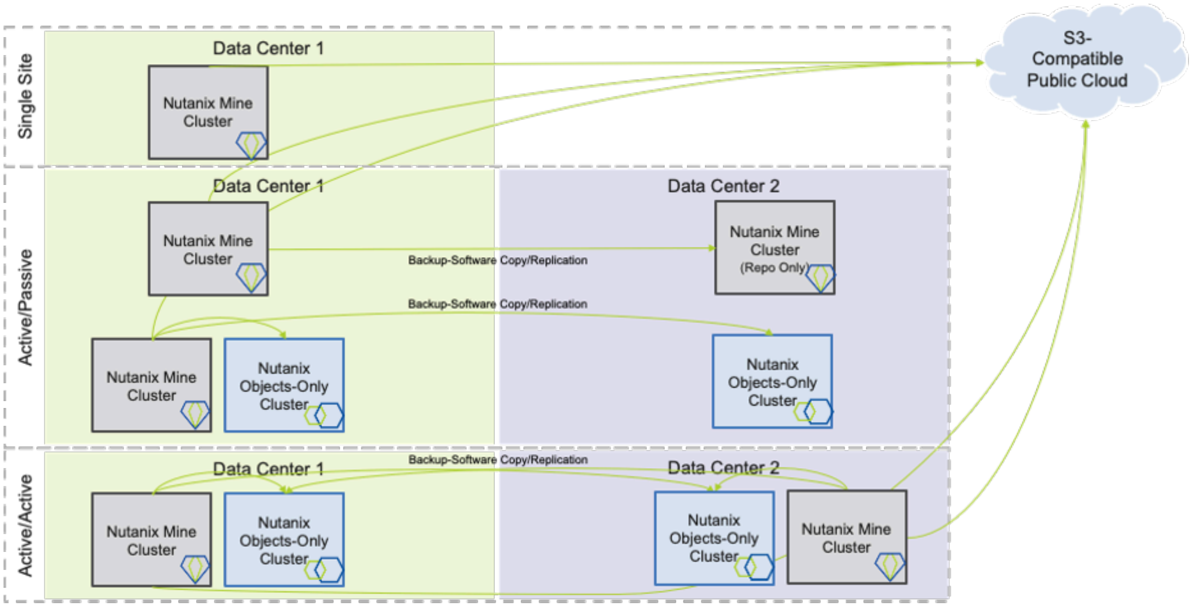

Single Site
Active/Passive
Active/Active
Data Center 1
Data Center 2
Nutanix Mine Cluster
Nutanix Mine Cluster (Repo Only)
Nutanix Objects-Only Cluster
Backup-Software Copy/Replication
S3-Compatible Public Cloud

Figure 97: Mine Deployments Across Datacenters or Regions

Calm Application Orchestration

Key Design Objectives

Table 203: Nutanix Calm Design Objectives

DESIGN OBJECTIVES	DESCRIPTION	NEW / UPDATE / REASONING

Private/Public cloud automation and lifecycle management	Virtual Machines, Bare Metal endpoints, Containers	Update

Architecture Overview

Nutanix Cloud Application Lifecycle Management (Calm) is an enterprise-grade, multi-cloud application and infrastructure lifecycle management framework. Nutanix Calm provides application lifecycle management, monitoring, and remediation to manage heterogeneous infrastructure, including VMs, bare-metal servers, and containers. Nutanix Calm supports multiple platforms and endpoints so that you can have a single self-service and automation framework that can manage your applications and infrastructure. The core automation constructs available in Calm are:

- Blueprints. Blueprints provide automation instructions for single or multiple instance(s) (VMs, containers, or native cloud services) governing application creation and lifecycle management of the instance(s). Blueprints can support instance creation, deletion, and any number of standard or custom day 2 management actions.
- Runbooks. Available with Calm version 3.x and later, runbooks provide the ability to create a set of automation tasks that are endpoint agnostic. Any day-to-day management operation that is performed manually, can be encapsulated in a Calm Runbook and executed against any service in your IT estate (VMs, bare-metal instances, containers, or native cloud services.)

Nutanix Calm includes several additional capabilities for efficient management of Hybrid Cloud automation:

- Self-service via the Calm Marketplace, the ServiceNow plugin for Nutanix Calm, or any self-service platform of your choice via Calm API integration.
- Governance via built in RBAC and project resource management
- Resource quota management (for on-premises endpoints only) via Policy Engine
- An intuitive and user-friendly GUI

- A powerful code-based approach with a Python-based Domain Specific Language (DSL) that Nutanix has released and open-sourced.
- The API-centric design of Calm enables extensibility for integration with many industry automation, configuration management, and self-service solutions such as Ansible, Terraform, Puppet, Chef, SCCM, ServiceNow, Cherwell, BMC Remedy, etc. This enables organizations to leverage existing intellectual capital from any of these solutions, allowing for rapid deployment and a short learning/adoption curve.

Hybrid Cloud management enables you to leverage all of the Calm capabilities on non-Nutanix providers (VMWare on Nutanix and non-Nutanix infrastructure) and the major public cloud solutions (AWS, GCP, and Azure). Nutanix Calm is a native service in Prism Central and is automatically enabled in Prism Central for both single Prism Central and scale-out Prism Central deployments.

Design Considerations

Sizing Considerations and Lifecycle Management

When enabling Calm, you must choose between a small or large deployment.

Table 204: Choose a Small or Large Calm Deployment

CALM-001	SMALL OR LARGE CALM DEPLOYMENT
Justification	You can start with a small deployment and easily scale to a large deployment at a later date.
Implication	To make an informed decision, you must first familiarize yourself with the sizing parameters, requirements, and Certified Configurations in the Nutanix Calm Administration and Operations Guide. Your chosen sizing allows you to hot-add necessary resources to Prism Central instance(s) with no disruption to Prism Central services.

It is recommended to upgrade Calm to the latest version as quickly as possible when new releases become available. This allows you to leverage new features and capabilities quickly and ensures you receive the latest available fixes/ patches, etc. Calm is deployed through Prism Central and tied to the same user interface. However, the two can be upgraded independently of one another. As with Prism Central, all Calm releases are Long Term Support (LTS); Calm does not provide Short Term Support (STS) releases. See the earlier section Operations Design for information on LTS and STS.)

Table 205: Upgrade to Most Recent Calm Version

CALM-002	UPGRADE TO THE MOST RECENT VERSION OF NUTANIX CALM
Justification	The Calm product team delivers fixes, enhancements, and new features on a regular cadence. You will want to leverage these enhancements as quickly as possible. Upgrading Calm is a non-disruptive LCM function that allows Calm to be upgraded separately from Prism Central or other application services running on Prism Central.
Implication	Calm blueprints from a newer version of Calm are usable with older versions of Calm.

Infrastructure-as-Code solutions require the creation of software code and a developer mindset. Robust processes for code creation and release management are imperative. It is recommended to deploy Calm in a minimum of two separate Prism Central physical instances, one for development and one for production. Larger enterprise deployments may consider up to four separate environments if their existing datacenter architecture includes separate environments and processes for Development, Test, QA, and Prod.

With this implementation methodology, you can follow the recommendation Calm-002 and always work with the latest Calm release in the development environment, and then roll the Calm upgrades through your other environments as part of a lifecycle management. This enables development and testing of existing and new blueprints with the latest version, while at the same time supporting existing blueprints and applications via the Production Calm instance. This also allows for maintaining a previous Calm version for backwards compatibility for existing blueprints.

Table 206: Deploy Calm for Development and Production

CALM-003	SEPARATE CALM INSTANCES FOR DEVELOPMENT AND PRODUCTION
Justification	Maintaining a development instance of Prism Central and Calm allows for an independent upgrade of Calm and integration testing against all blueprints and runbooks prior to upgrading in production. This provides an additional advantage in cases where new features become available, and you wish to update blueprints or even new automation blueprints or runbooks leveraging the new features prior to upgrading production.
Implication	This requires a separate instance of Prism Central (this could be a single instance even on a single node cluster) on a separate Nutanix cluster. This also adds an additional instance of Calm to manage from the point of view of RBAC access for developers, etc. This is minimal overhead especially since only automation engineers and developers will be leveraging the Dev instance. If you currently leverage a release management process, it may require updates to include Calm.

Business Continuity and Disaster Recovery

It is recommended that you create a process (automated or manual) to execute Nutanix-provided backup scripts via SSH from Prism Central to protect Calm. These scripts are documented in the Calm support documentation. If a cold Prism Central instance is not immediately available for a restore, then a new Prism central/cluster instance can be provisioned and used as a target for a Calm restore. The only risk with this approach is the impact to RTO due to having to rebuild the Prism Central instance as a new target.

If you have enabled the policy engine for your Calm instance, then there is a separate script to create backups of your policy engine database. The policy engine database backup can be used to restore the policy engine to either an earlier state or to a new policy engine VM.

Table 207: Calm for DR and Backup

CALM-004	DEFINE A PROCESS FOR CALM DR AND BACKUP/ RESTORE
Justification	In the event of data corruption or loss of Prism Central, Nutanix Calm blueprints, runbooks, and configurations need to be backed up. It is recommended that the backup scripts are scheduled to run at a set time each day.
Implication	A separate “cold” Prism Central instance(s) with Calm enabled is required as a restore target for the backup/ restore script. An external scheduler is required to run the scripts via SSH. Customers with Prism Pro can use an X-Play Playbook to call a Calm Runbook to execute the backup script. Keep in mind that Prism Central has security features that limit its ability to remotely execute some SSH commands, and the Calm Runbook will need to SSH to an independent endpoint (VM or container) to run the backup command.

Management and Operations

Calm requires integration with an AD/LDAP source, and Calm can leverage Prism Central's existing AD/LDAP configuration. Local Prism Central users cannot be used for user and role assignments in Calm. The AD/LDAP integration for Calm is tied to the Project construct. It is a best practice to align your AD/LDAP security groups to projects and roles within Calm in order to achieve the desired self-service consumption model for each project. You can create security groups for Calm projects and roles and then add AD users to these groups to simplify access management via standard AD user authentication. (For example, for "ProjectA" all developers are assigned to a group called "ProjectADev" that is associated with the project. Adding new developers to this security group grants them access to the project resources based on their AD/LDAP credentials).

Control of project configuration (creation, group membership, etc.) can also be performed via Calm blueprints and Runbooks via API if you need a more dynamic RBAC project/resource model. This way, self-service features in Calm can be used to automate RBAC policies.

The Prism Central "admin" account can be used to gain/restore access to Calm if something happens to the Active Directory LDAP connection(s).

Table 208: Calm and RBAC Configurations

CALM-005	ALIGN PROJECT STRUCTURE AND RBAC CONFIGURATION
Justification	Design Calm Projects and Security Groups to align with AD/LDAP users and roles that are specific to entitlements for administering, developing, managing, and consuming automation constructs (blueprints, runbooks. etc.)
Implication	AD/LDAP configurations in Prism Central are required in order to leverage Calm RBAC features.

Calm has a native feature to help manage and share code across environments. End-users can select blueprints to provision into any environment in any geography at runtime. This allows blueprints to be created once and delivered to any Calm physical instance without having to manage multiple copies of the same blueprint. Through RBAC, administrators can assign users/groups to projects that control access to who can provision which blueprints in which environments. It is a good design principle to align your physical Calm instances to both your environments (e.g., Dev, Test, QA, Prod) and your deployment geographies (e.g., Americas, Europe, Asia). Here is an example of what this might look like for a global enterprise deployment of Calm.

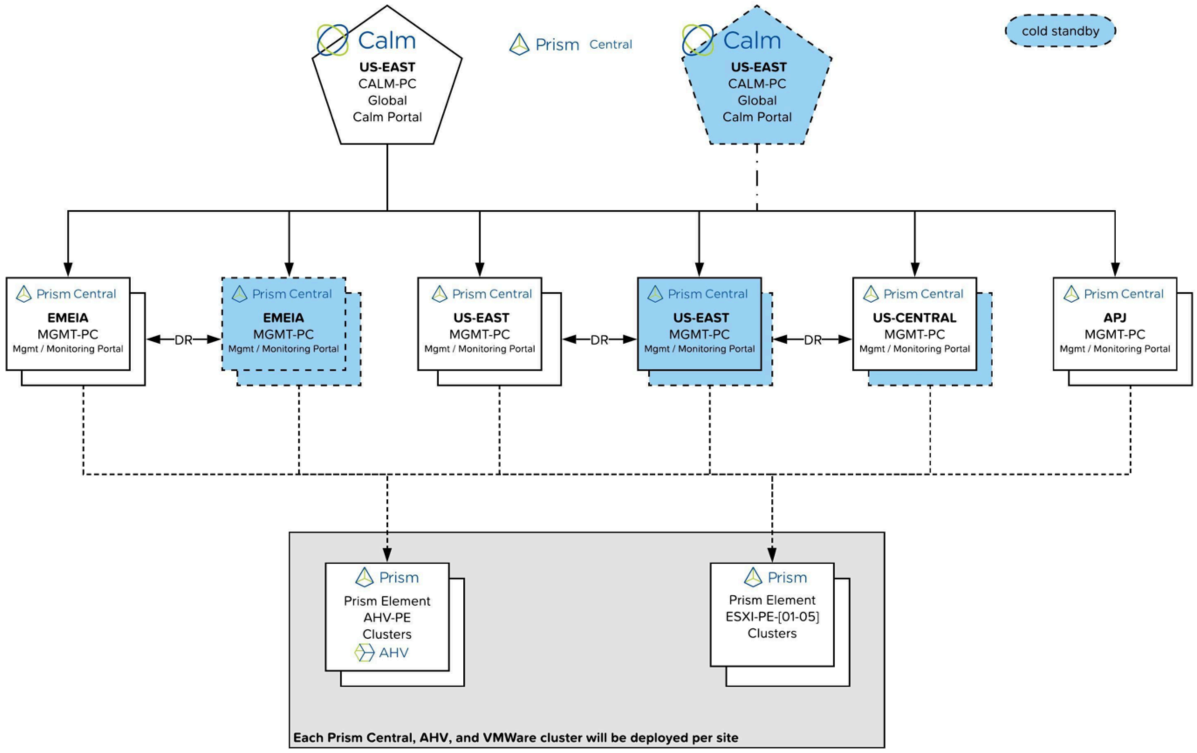


Figure 98: Global Enterprise Calm Deployment

Table 209: Calm Releases and Geographic Regions

CALM-006	DESIGN CALM ENVIRONMENT TO ALIGN WITH RELEASE METHODOLOGY AND GEOGRAPHIC REGIONS

Justification	This feature maximizes efficiency by allowing you to have one automation blueprint that can be chosen as a self-service request from end-users in the same cost center, but in different sites or geographies.
Implication	A Calm decision should be made while planning your infrastructure to establish physical clusters that align with your needs.

Nutanix Calm has the ability to set resource quotas for projects via its Policy Engine. Maximum quotas for compute, memory, and storage can be set for projects so that the total amount provisioned by all users in the project does not exceed the threshold. When you enable the Policy Engine for a Calm instance, a new VM is created and deployed. For quota policy enforcement, VMs must be provisioned through Calm. In order to enforce the desired policy for VMs that are not provisioned through Calm, migrate the VMs to Calm once you have enabled the Policy Engine.

Table 210: Calm Policy Engine

CALM-007	ENABLE CALM POLICY ENGINE
Justification	Prevents teams from over-provisioning or hogging resources and allows you to better manage consumption among business units and projects.
Implication	You need an available IP address that belongs to the same network as that of your Prism Central VM for the Policy Engine VM.

Integration with Nutanix Beam enables showback for Calm without any additional licensing costs. Nutanix Beam is a SaaS product that helps customers optimize their multi-cloud spend. Cost calculations in Nutanix Beam are configured for you, and Beam will automatically calculate your compute, memory, and storage costs. Since Beam runs in the Public Cloud, the automatic configuration of your cost data requires an active internet connection to Prism Central. In the event you do not have an active internet connection, you can still leverage Beam showback features by manually entering your cost data. Showback is currently available for Nutanix and VMware environments.

Table 211: Showback with Nutanix Beam

CALM-008	ENABLE SHOWBACK VIA NUTANIX BEAM
Justification	Showback helps you control IT spend by tracking utilization of IT resources to each business unit. Reporting can identify how much each business unit is consuming and helps in budgeting and cost optimization. You may also be able to implement Chargeback processes whereby business units pay for their consumption on a monthly or yearly basis.
Justification	Showback helps you control IT spend by tracking utilization of IT resources to each business unit. Reporting can identify how much each business unit is consuming and helps in budgeting and cost optimization. You may also be able to implement Chargeback processes whereby business units pay for their consumption on a monthly or yearly basis.

Self-service enables end-users to create new workloads and manage existing workloads provisioned by Calm without accessing the administrative consoles or needing help from IT Admins with privileged access. Self-service can be achieved via a variety of mechanisms: the Calm Marketplace, the ServiceNow plugin for Nutanix Calm, or the self-service platform of your choice via Calm API integration. The Calm Marketplace is native to Nutanix Prism and included with Calm.

The ServiceNow plug-in for Calm may be ideal for companies who already use the ServiceNow ITSM platform; you can expose Calm services to all end-users that have access to ServiceNow. In addition, ServiceNow approvals, incidents, and CMDB updates are integrated with the Calm Plug-in. The ServiceNow Plug-in for Calm is available from the ServiceNow store and adds no additional Nutanix licensing costs. For customers who prefer to use an in -house self-service interface or some other third-party interface, Calm blueprints and runbooks can be invoked via REST API, enabling easy integration.

Table 212: Self-Service with Calm

CALM-009	CHOOSE A METHOD FOR SELF-SERVICE
Justification	Self-service allows end-users to perform provisioning and Day 2 operations without the need for privileged access or assistance from IT Admins.
Implication	Potential third-party licenses needed for self-service integration are not provided by the Calm Marketplace.

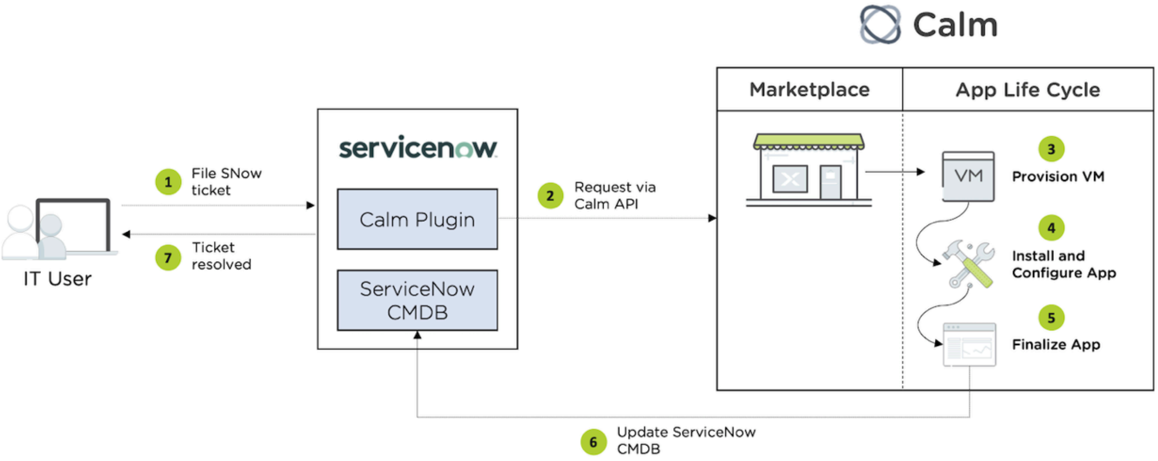


Figure 99: Self-Service with Calm

Karbon for Kubernetes and Container Support on Nutanix

Key Design Objectives

Table 213: Nutanix Karbon Design Objectives

DESIGN OBJECTIVES	DESCRIPTION	NEW / UPDATE / REASONING

Enable support for containers and Kubernetes	Nutanix Karbon enables teams to incorporate containers and Kubernetes as part of the Nutanix ecosystem with simplified provisioning and management.	New

Architecture Overview

Nutanix Karbon is an enterprise-grade Certified Kubernetes distribution that simplifies the provisioning, operations and lifecycle management of Kubernetes, all from within Nutanix Prism. Karbon is an opt-in service running on Prism Central; upon enablement, Karbon Core and Karbon UI docker containers are instantiated on Prism Central VM(s), allowing you to be up and running in 5 to 10 minutes.

Once enabled, Nutanix Karbon can deploy and manage Kubernetes clusters to any AHV cluster that's registered to Prism Central. Any number of Kubernetes clusters can be deployed with any configuration, assuming adequate resources on the physical clusters. Day 2 management operations, such as host OS upgrades, Kubernetes version upgrades, storage class management, and worker node additions or removals are feasible via the UI or APIs.

The UI provides a mechanism to download a hardened host OS image provided by Nutanix (which is similar to the CVM) in a single click. Nutanix quickly reacts to any published security vulnerability and will push out a new host OS image with the appropriate vulnerabilities patched. Karbon admins can quickly download these new host OS images, and initiate host OS upgrades to bring a given Kubernetes cluster to the updated OS level in a couple of clicks. Alternatively, if the vulnerability is at the Kubernetes level, admins can also easily perform a Kubernetes version upgrade.

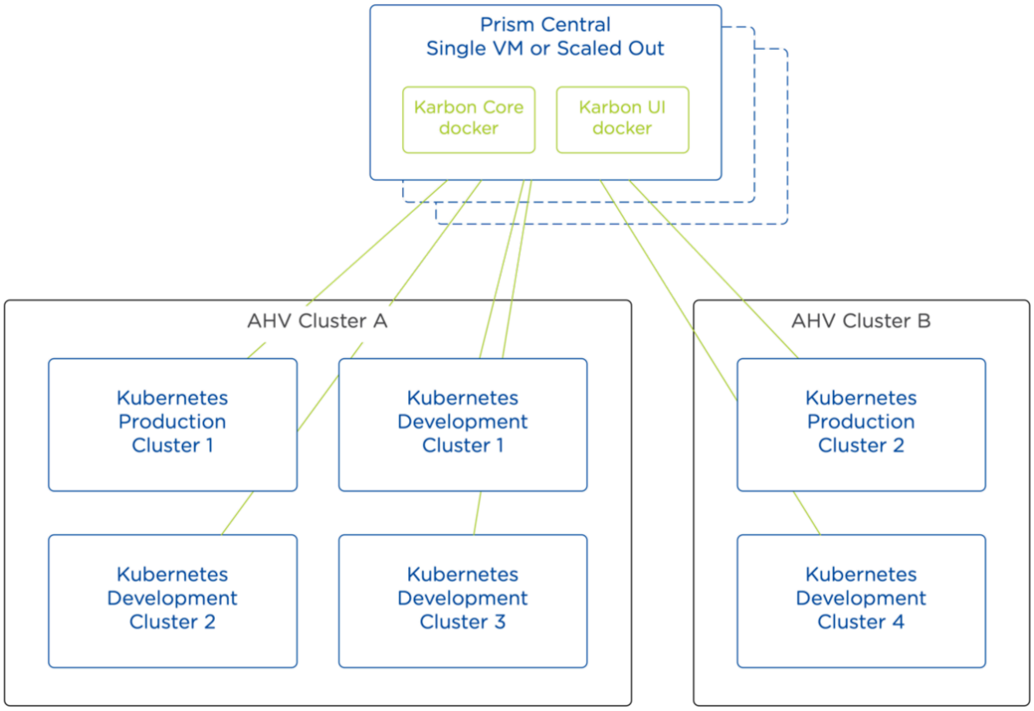

Prism Central
Single VM or Scaled Out
Karbon Core docker
Karbon UI docker
AHV Cluster A
Kubernetes Production Cluster 1
Kubernetes Development Cluster 1
Kubernetes Development Cluster 2
Kubernetes Development Cluster 3
AHV Cluster B
Kubernetes Production Cluster 2
Kubernetes Development Cluster 4

Figure 100: Karbon Provides a Single Control Plane for Multiple Kubernetes Clusters and Locations

Karbon can deploy both development and production Kubernetes clusters, depending upon use case. Development clusters are restricted to a single master node (which runs the Kubernetes control plane), a single ETCD node (Kubernetes' key-value store), and any number of worker nodes (which run end-user applications). The number of worker nodes can be specified during cluster creation and can be increased or decreased at any point in the cluster's life. Due to the single master and ETCD nodes, these clusters cannot tolerate failures in the control plane, and applications will experience downtime during upgrades. However, they're light on resources, making them perfect for development.

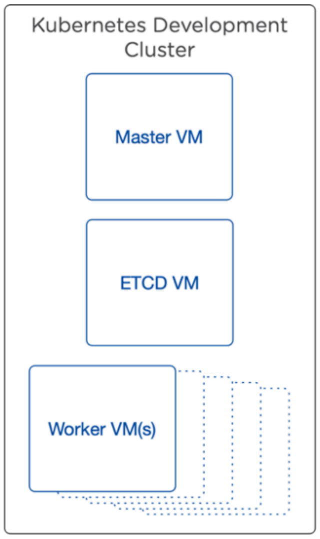


Figure 101: A Karbon Development Cluster is not Highly Available

Production Kubernetes clusters deployed by Nutanix Karbon have two master nodes in active-passive mode by default, utilizing VRRP for high availability. If the active master node goes down (either due to an upgrade or loss), the passive node will assume the virtual IP and the cluster will continue to function.

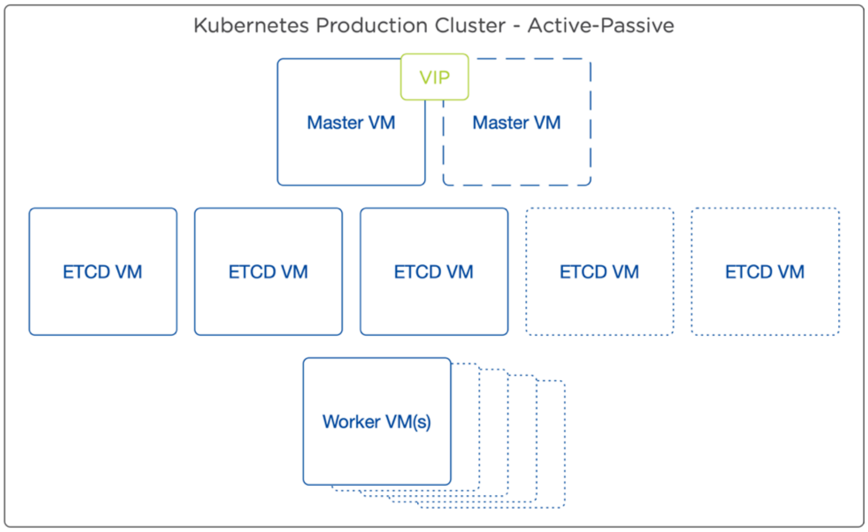


Figure 102: A Karbon Cluster for Production is Highly Available

Alternatively, if you have an external load balancer, anywhere from two to five master nodes can be deployed in an active-active configuration. This increases performance, and as long as a single master node is online, the cluster continues to function.

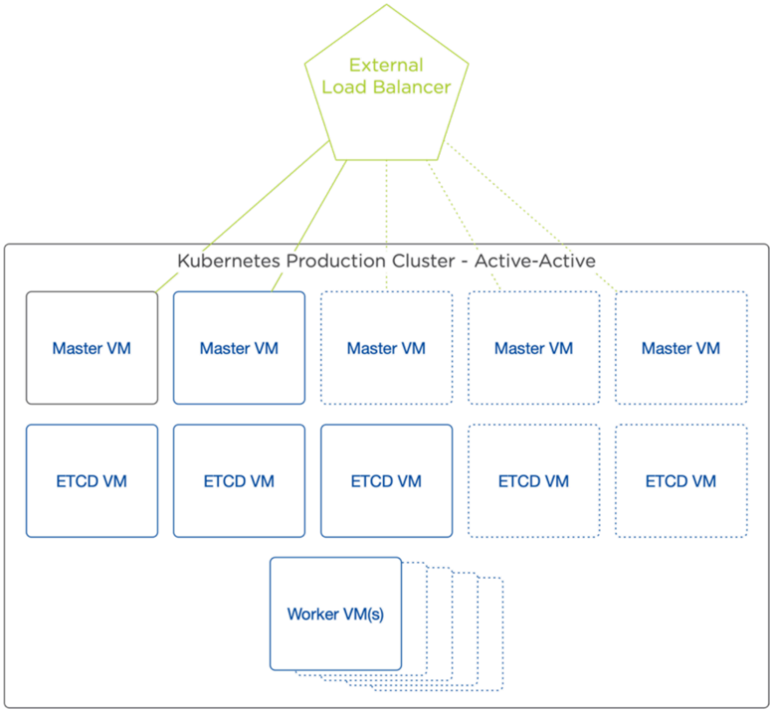


Figure 103: An External Load Balancer Enables Karbon to Have More than Two Master Nodes

Either three or five ETCD nodes can be deployed for production clusters; three enables the cluster to sustain a single ETCD node failure, and five enables two ETCD nodes to fail without causing a disruption. Stated differently, a quorum of two is required when three ETCD nodes are used, and a quorum of three is required when five nodes are used. As with development clusters, any number of worker nodes can be deployed at cluster creation, and later modified to scale with workload changes.

During deployment, admins are able to choose the physical network and VLAN which the nodes that make up the Kubernetes cluster belong to. All nodes must live in the same network. For development clusters, these networks can be backed by a DHCP server, or can utilize AHV managed networks (IPAM). For production clusters, you are required to select an AHV managed network. For Kubernetes cluster networking, either the Flannel or Calico CNIs can be utilized.

Nutanix Karbon Kubernetes clusters have the Nutanix CSI driver installed during cluster creation for persistent storage. A default storage class based on Nutanix Volumes is also defined and created during cluster creation, which provides Read Write Once storage (RWO). After cluster creation, any number of additional storage classes based on either Nutanix Volumes or Nutanix Files, which provides Read Write Many (RWX), can be created.

Karbon admins are also able to download a secure, 24-hour valid kubeconfig file from the Karbon UI, which allows for kubectl usage, which is the standard Kubernetes command line interface. This effectively provides root access to the cluster. Utilizing this, Karbon admins can create Kubernetes authentication objects, such as a Role, ClusterRole, RoleBinding, and ClusterRoleBinding, for end users. End users are added (either directly or as part of group membership) via Active Directory as Nutanix Read-Only users within the role mapping configuration section of Prism Central. This allows end users to log into the Karbon UI and download kubeconfig files.

Once these authentication objects are appropriately created, Karbon end users can easily create and manage applications utilizing kubectl, assuming they have been granted privilege to do so. This includes, but is not limited to, creating Kubernetes deployments, services, persistent volume claims, persistent volumes, and stateful sets. Ultimately the types of objects end users can create or manage is dependent upon the roles configured for them by the Karbon admin.

Since Kubernetes is a cloud native technology, it is recommended to implement Business Continuity at the application level. This means running multiple instances of your applications in different regions so that, in the event of a disaster, the secondary application continues to operate. For environments or applications where this is not possible, open source tools like Velero can provide Business Continuity at the Kubernetes cluster level.

Nutanix Karbon is included in every licensed version of Nutanix software. The only requirements are at least one AHV cluster and Prism Central.

High-Level Design Considerations

When configuring Karbon, first determine the availability needs of your applications and Kubernetes clusters. Development and test workloads are likely okay to run on development clusters with single points of failure. Any production workloads should run on production clusters designed to ensure that a single VM loss will not bring down an application or an entire cluster. Business critical workloads should be distributed across multiple production clusters in multiple availability zones to ensure availability in the event of a disaster.

Next, determine your preferred CNI, either Calico or Flannel. Generally, Flannel offers simpler configuration, while Calico has more features. It is possible to have some clusters with Flannel and some with Calico to address specific use cases, however this may complicate management.

It is recommended to use the latest version of Kubernetes supported by Karbon unless there's a specific reason for using an older version. When possible, upgrade existing Kubernetes clusters to newer versions of Kubernetes.

Utilize a consistent naming convention for Storage Classes, based on volume type (Volumes or Files), reclaim policy, file system, and performance.

Table 214: Minimize Admin Kubeconfig Use

KARBON-001	MINIMIZE USE OF THE ADMIN KUBECONFIG
Justification	The admin kubeconfig provides root access to the cluster, so regular use of it is a security vulnerability. It should only be used for initial cluster configuration. Use AD-based Prism Read-Only credentials (with the corresponding Role and RoleBinding) instead, or Kubernetes Service Accounts for automation use cases.
Implication	Configuration of Prism role mappings, and Kubernetes objects (such as Roles, ClusterRoles, RoleBindings, ClusterRoleBindings, and ServiceAccounts) must be performed before running end-user applications on a cluster.

Table 215: Use Multiple Availability Zones for Production

KARBON-002	DISTRIBUTE PRODUCTION APPLICATIONS ACROSS AVAILABILITY ZONES
Justification	Production applications should run in an active-active or active-passive mode on multiple Nutanix clusters in different availability zones. In the event of a disaster, the secondary site can take over transparently.
Implication	A minimum of 2 AHV clusters and AZs are required.

Table 216: Use Namespaces to Segment Users and Apps

KARBON-003	UTILIZE KUBERNETES NAMESPACES TO SEGMENT USERS AND APPLICATIONS RATHER THAN CLUSTERS
Justification	Where possible, use of namespaces to segment work is preferable to reduce Kubernetes cluster sprawl and resource utilization.
Implication	This requires an approach to cluster configuration similar to that prescribed above in Karbon-001.

Table 217: Leave Room for a Worker Node to Fail

KARBON-004	LEAVE ENOUGH SPACE IN A CLUSTER TO TOLERATE A WORKER NODE FAILURE
Justification	Whether there is an unplanned outage, or a planned upgrade, sizing the cluster to tolerate a failure of a worker node allows applications to continue to run in the event of a failure.
Implication	At minimum, the cluster must be sized for n-+1, and workers must be added if the workload grows.

Table 218: Use Consistent CIDRs

KARBON-005	UTILIZE CONSISTENT POD AND SERVICE CIDRS
Justification	Whether utilizing Calico or Flannel as the CNI provider, during cluster deployment the admin must provide a large (/16 by default) subnet for use by Kubernetes Pods and Services. While these are internal subnets which do not go over the wire, they must not overlap with physical host networking. It is recommended to utilize the same subnets for all clusters for consistency and to minimize the chance of IP overlap.
Implication	Two large subnets are needed for Karbon Kubernetes clusters which do not overlap with physical host networking.

Table 219: Keep the Host OS Updated

KARBON-006	UPGRADE THE HOST OS AS SOON AS POSSIBLE
Justification	Due to the nature of Common Vulnerabilities and Exposures (CVEs), it is critical to upgrade Host operating systems quickly whenever a new version is released. This greatly reduces the chances of a cluster being compromised.
Implication	For development clusters, an outage is required. For production clusters, there will be a single point of failure during upgrades.

Era for Database as a Service

Key Design Objectives

Table 220: Nutanix Era Design Objectives

DESIGN OBJECTIVES	DESCRIPTION	UPDATE

Enable Database as a Service (DaaS	• Deliver Database as a Service (DBaaS) consumable via a user interface and REST API. • Perform a standardized and repeatable deployment of Postgres database instances. • Deploy and manage Postgres database instances in both: › HA mode › Single mode • Deploy Postgres HA instances across Nutanix clusters. • Protect (back up) database data to a location remote from the database instance. • Perform lifecycle and day2 operations for Postgres databases. • Provide copy data management functionality such as the ability to clone database instances. • Manage 100+ DB instances and DB VMs • Network communication is required between the Era management plane and the database instances managed by Era.	New

Architecture Overview

Era automates and simplifies database administration and brings one-click simplicity and invisible operations to database provisioning and lifecycle management, facilitating database as a service (DBaaS). Era enables database administrators to perform operations such as database registration, provisioning, cloning, patching, restore, and much more. It allows administrators to define standards for database provisioning with end-state driven functionality that includes HA database deployments.

Era enables multi-cluster database management. Database administration for different databases across multiple Nutanix clusters can be performed with a single Era instance. With the extension of support for Nutanix Clusters, you can utilize all the capabilities of Era both in the cloud and on-prem.

The Era management plane operates on one or more VMs running in a Nutanix cluster with at least three Nutanix nodes (physical servers). The default Era management plane deployment in a single Nutanix cluster consists of one VM running all services:

- Front-end services (API, agent, web service).
- Back-end service (database or repository).

The following figure presents a logical implementation of Era DBaaS, including the Era management VM, database server VMs (DB), Nutanix cluster, and physical hardware

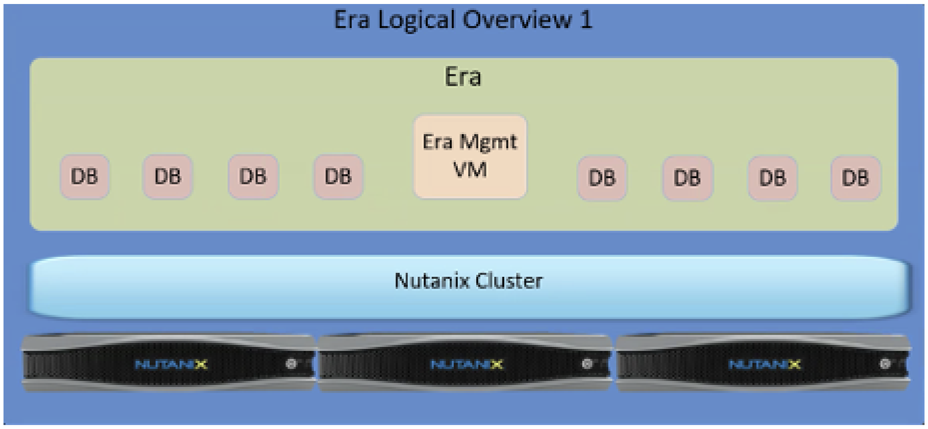


Figure 104: Nutanix Era DBaaS

In a multicluster Era environment, where Era manages database server VMs in multiple Nutanix clusters, the Era management plane requires one Era management agent VM running only the agent service in each additional Nutanix cluster it manages.

The following figure expands on the previous one to include the two Nutanix clusters and two Era management agent VMs.

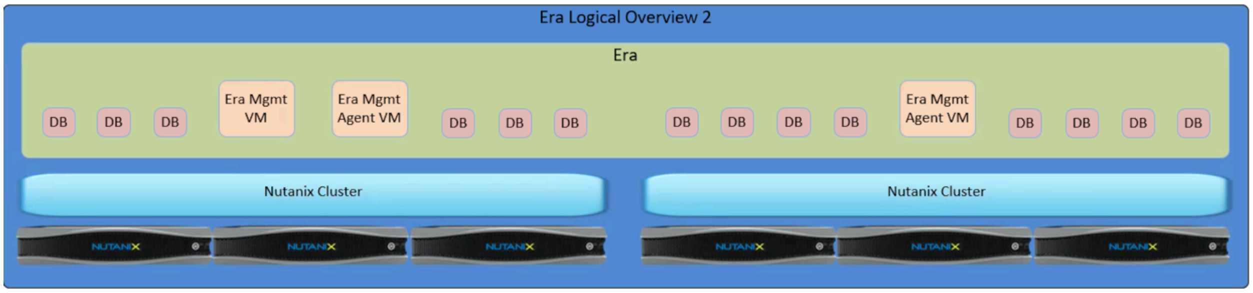


Figure 105: Nutanix Era VMs and DBs

It is possible to run the Era management plane in a highly available mode to avoid service disruption from a node failure. This is done by provisioning an HA PostgreSQL instance where Era deploys a multi-node PostgreSQL instance. One of the nodes in the cluster acts as the leader, and other nodes serve as replicas. The leader node is used for both read and write operations while the replicas are only used for read-only requests. When a primary node becomes unavailable, one of the replicas is promoted to the role of master, ensuring availability of the database for write operations.

When Era is deployed in a high availability (HA) configuration, it includes six VMs:

As described earlier, Era supports managing database instances across multiple Nutanix clusters. The figure below illustrates the Era management architecture in a multi-cluster scenario where C1, C2, and C3 are different Nutanix clusters.

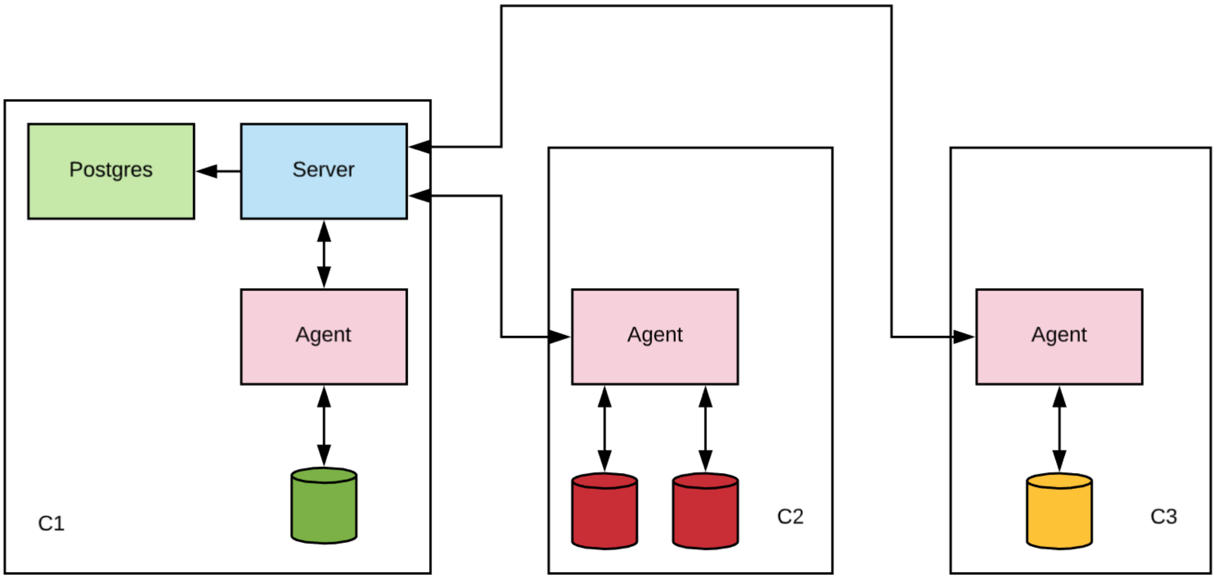


Figure 106: Era Management Layer in a Multi-Cluster Deployment

Based on the rich API provided by Era, you can automate any action available via the Era UI so that Era can easily be consumed from an overlying automation, self-service, or orchestration solution such as:

- Nutanix Calm
- ServiceNow
- VMware vRealize Automation

The Era management plane and the database instances managed by Era communicate during initial provisioning and registration, as well as during ongoing operation. Communication is mostly initiated from the database instances to the Era management plane, accept during provisioning and registration.

No specific hardware is required to run Era; it runs as a VM on any Nutanix AOS-supported hardware running Nutanix AHV or VMware ESXi.

Era employs role-based access control to govern access to an Era system. RBAC allows system administrators to restrict user access and limit the operations they can perform. For example, a database operator can be granted access to only perform database-related operations.

With RBAC, you can add users either in Era or Active Directory and then later add those users to a group. Groups allow you to assign the same roles to multiple users at the same time. After adding a user, you can assign one or more roles to the user based on the operations they can perform.

There are multiple license options available for Era:

- License the entire Nutanix cluster
- License per physical CPU cores
- License per DB VM vCPU

High-Level Design Considerations

The design decisions presented in this section cover the Era management plane but not the database instances managed by Era.

Table 221: Era to Enable DBaaS

DBAAS-001	ENABLE DATABASE AS A SERVICE
Justification	Era can easily be integrated with overlying automation, self-service, and orchestration solutions.Era deploys DB instances with a predefined operating system, database engine binary and DB instance configuration in a repeatable fashion. Era can deploy and manage (clone, take a snapshot, perform log catchup, refresh) Postgres in single and highly availability mode in one or multiple Nutanix clusters. Snapshots of the managed database instances can be stored in the same Nutanix cluster where they are running or in remote Nutanix clusters.
Implication	N/A

Table 222: Era Hardware Requirements

DBAAS-002	MINIMUM HARDWARE REQS FOR NUTANIX CLUSTERS RUNNING ERA
Justification	The minimum configuration requirements shown below provide a solution with predictable performance. Since no specific performance requirements are established for this section, these requirements can serve as a baseline for running DB workloads.
Implication	Since workload characteristics are undefined the configuration below could be either too performant (increasing CapEx unnecessarily) or not performant enough and therefore need to be upgraded. Careful sizing should be performed for your intended DB workloads.

A Nutanix cluster supporting Era and database instances should meet the following minimum hardware requirements:

- Minimum 2.7GHz, 16-core CPUs
- An optimal memory configuration with minimum 768 GB memory per node. See Nutanix Physical Memory Configuration document for Nutanix NX recommendations.
- Minimum All flash storage
- Minimum 2x10 Gbps networks.

Table 223: Dedicated DB Clusters

DBAAS-003	USE DEDICATED NUTANIX CLUSTERS TO SUPPORT DATABASE WORKLOADS

Justification	The requirement to host more than 100 VMs justifies a minimum of two separate Nutanix clusters for database workloads to ensure that database workloads do not have to compete with non-database workloads for resources. The vCPU:pCPU ratio is typically lower for database workloads compared to other server virtualization workloads.
Implication	Hardware may be underutilized if not enough workloads are provisioned in the Nutanix clusters.

Table 224: HA Management Plane

DBAAS-004	DEPLOY ERA MANAGEMENT PLANE FOR HIGH AVAILABILITY
Justification	Provides the highest possible uptime for the Era management plane, ensuring that no changes are needed to address future changes in SLA uptime requirements. The solution protects against single-node failures.
Implication	When an Era VM providing front-end services or the hypervisor host (where the Era VM runs) becomes unavailable, there will be a short period of downtime for any user or service accessing Era.

Table 225: Management Network

DBAAS-005	PLACE ERA ON A SEPARATE MANAGEMENT NETWORK
Justification	The Era service is treated like any other management service and placed on a separate management network accordingly.

Implication	Firewall openings between the network where an Era server is hosted and the network(s) where the DB workloads are hosted are required if the networks are not the same.

Table 226: Era Platform License

DBAAS-006	USE THE ERA PLATFORM LICENSE
Justification	The Platform license lets you run as many DB workloads in a cluster as permitted by the underlying hardware and your acceptable vCPU:pCPU ratio. In addition to the Era license, the Platform license also includes the Nutanix AOS Ultimate license, so both Era and AOS licenses are managed as one entity.
Implication	Some AOS Ultimate features may go unused.

Table 227: Active Directory Integration

DBAAS-007	ACTIVE DIRECTORY IS USED FOR USER, GROUP, AND SERVICE ACCOUNT ACCESS
Justification	Using AD provides a single source of truth for user, group, and service management for traceability and auditing purposes. The local Era admin account can be used to gain access to the environment for super admins if the Active Directory connection is lost.
Implication	Era requires a connection to Active Directory to provide end user functionality.

Table 228: Active Directory Access

DBAAS-008	ACTIVE DIRECTORY GROUPS ARE USED TO PROVIDE ACCESS TO ERA
Justification	One place—Active Directory—is used to configure Era permissions (add and remove users from Active Directory groups) during normal operations.
Implication	Era requires a connection to Active Directory to provide end user functionality.

Table 229: Era Test Environment

DBAAS-009	AN ERA TEST ENVIRONMENT MUST EXIST
Justification	Upgrades and configuration changes must be applied and tested in a test environment before being applied in the production environment to avoid problems with new versions or upgrade processes.
Implication	An Era test system must be available.

Table 230: Service Windows

DBAAS-010	ERA UPGRADES ARE PERFORMED DURING SERVICE WINDOWS
Justification	There will be a very short service interruption during an upgrade. If a critical patch must be applied for stability or security reasons an upgrade can be performed during business hours after the service window is approved by the change control function
Implication	Could create some delay before a patch can be applied since a service window has to be requested.

Table 231: Era Roles

DBAAS-0011	USE ERA SUPER ADMIN AND DATABASE ADMIN ROLES
Justification	Provides a clear division between those managing the Era solution and those consuming DBaaS
Implication	N/A

7. Conclusion

This document is intended to demonstrate valuable methods and practices that organizations, both large and small, can use to implement Nutanix Solutions to solve their IT and business problems. There is no one-size-fits-all solution for Nutanix Hyperconverged Infrastructure, and the contents of this document are informational only and intended to provide customers with suggested best practices from which they can design and evolve their private, hybrid, and multi-cloud solutions.

For more information on any details of this document, please visit our website at www.nutanix.com or reach out to our sales team. You may also contact one of our many global support phone numbers listed on our website.

Appendix

Table of Design Decisions

The following table summarizes all the design decisions described in this document.

Table 232: Design Decisions

REFERENCE	DECISION NAME	DECISION
Region-001	Number of regions to be used	
AZ-001	Number of availability zones to be used	
DC-001	Number of datacenters to be used	
VRT-001	Choose Nutanix AHV or VMware ESXi as the hypervisor for your deployment.	
PFM-001	Management Cluster Architecture: Deploy a separate Management Cluster or share a cluster with other workloads. When choosing a separate management cluster, consider a redundant configuration.	
PFM-002	Mixed or dedicated workload per cluster.	
PFM-003	Select Nutanix software licensing level	
PFM-004	Select physical node vendor	
PFM-005	Select node model(s) per use case	
PFM-006	Number, type, and size of clusters	
PFM-007	Decide which workload domains will span a single or multiple racks	
NET-001	Use a large buffer datacenter switch at 10Gbps or faster.	

NET-002	Use a leaf-spine network topology for new environments.	
NET-003	Populate each rack with two 10GbE or faster ToR switches.	
NET-004	Avoid switch stacking to ensure network availability during individual device failure.	
NET-005	Ensure that there are no more than three switches between any two Nutanix nodes in the same cluster.	
NET-006	Reduce network oversubscription to achieve as close to a 1:1 ratio as possible.	
NET-007	Configure the CVM and hypervisor VLAN as native, or untagged on server facing switch ports.	
NET-008	Use tagged VLANs on the switch ports for all guest workloads.	
NET-009	Use a Layer 2 network design.	
NET-010	Connect at least one 10 GbE or faster NIC to each top-of-rack switch.	
NET-011	Use a single br0 bridge with at least two of the fastest uplinks of the same speed.	
NET-012	Use NICs of the same vendor within a bond	
NET-013	Use VLANs to separate logical networks.	
NET-014	Use active-backup uplink load balancing.	
NET-015	Use standard 1,500-byte MTU and do not use jumbo frames.	
NET-016	Use virtual distributed switch (vDS).	
NET-017	Connect at least one 10 GbE NIC to each top-of-rack switch.	
NET-018	Use a single vSwitch0 with at least two of the fastest uplinks of the same speed.	

NET-019	Use Route Based on Physical NIC Load uplink load balancing.	
NET-020	Use standard 1,500-byte MTU and do not use jumbo frames.	
CMP-001	If running a non-NUMA aware application on a VM, configure the VM's memory and vCPU to be within a NUMA node on AHV host.	
STR-01	When creating vDisks in ESXi, always use thin- provisioned vDisks.	
STR-02	When sizing a hybrid cluster, make sure to have enough usable SSD capacity to meet active data set of application.	
STR-03	Do not mix node types from different vendors in the same cluster.	
STR-04	Do not mix nodes that contain NVMe SSDs in same cluster with hybrid SSD/ HDD nodes.	
STR-05	Minimum 2:1 HDD to SSD ratio required for Hybrid clusters.	
STR-06	Size for N+1 node redundancy for storage and compute when sizing. For mission critical workloads that need higher SLAs, use N+2 node redundancy.	
STR-07	Use FT=2 and RF=3 for workloads and clusters that need higher SLAs or for cluster sizes >32.	
STR-08	Enable Inline Compression.	
STR-09	Enable/Disable Deduplication.	
VRT-001	Deploy scale-out Prism Central for enhanced cluster management.	
VRT-002	Deploy prism central in each region or az using runbook dr automation.	
VRT-003	Use VM-HA Guarantee.	

VRT-004	Deploy HA vCenter instance with embedded PSC to manage all ESXi based Nutanix clusters.
VRT-005	Deploy a vCenter in each region or AZ using runbook DR automation.
VRT-006	Enable EVC mode and set to the highest compatibility level the processors in the cluster will support.
VRT-007	Enable HA.
VRT-008	Enable DRS with default automation level.
VRT-009	Disable VM PDL and APD component protection.
VRT-010	Configure das.ignoreInsufficientHbDatastore if one Nutanix container is presented to the ESXi hosts.
VRT-011	Disable Automation level for all CVMs.
VRT-012	Set host failures clusters tolerate to 1 for RF2 and to 2 for RF3
VRT-013	Set host isolation response to “Power off and restart VMs”.
VRT-014	Set host isolation response to “Leave Powered on” for CVMs.
VRT-015	Disable HA/DRS for each Nutanix CVM.
VRT-016	Disable SIOC.
DEP-001	A minimum of three NTP servers for all infrastructure components should be provided.
DEP-002	A minimum of two DNS Servers should be configured accessible to all infrastructure layers.
SEC-001	Use active directory authentication. This applies for user and service accounts.
SEC-002	Use SSL/TLS connection to Active Directory.

SEC-003	Use signed Certificate Authority (CA) certificates for the components where certificates can be replaced. This can be either internal or external signed certificates.
SEC-004	Do not use Nutanix Cluster lockdown.
SEC-005	Do not use vSphere Cluster lockdown.
SEC-006	Enable CVM and hypervisor AIDE.
SEC-007	Configure SCMA to run hourly.
SEC-008	Stop unused ESXi services and close unused firewall ports.
SEC-009	Send log files to a highly available syslog infrastructure.
SEC-010	Include all Nutanix modules in the logging.
SEC-011	Use Error log level for the Nutanix components.
SEC-012	Use default esxi logging level, log rotation, and log file sizes.
SEC-013	If extra security and reliability are required, then use tcp for log transport. Otherwise, use the default syslog protocol, udp.
SEC-014	Use port 514 for logging.
SEC-015	Use VLAN for traffic separation of management and user workloads.
SEC-016	Place CVM and hypervisor on the same VLAN and subnet.
SEC-017	Place out of band management on separate VLAN or physical network.
SEC-018	Use the least privileged access approach when providing access. and Align RBAC structure and usage of default plus customer roles according to the company requirement.

SEC-019	Align RBAC structure and usage of default plus custom roles according to the company requirements defined via SEC-019.	
SEC-020	Do not use storage encryption.	
SEC-021	Do not use a key management server.	
OPS-001	Deploy Prism Central Pro for enhanced cluster management.	
OPS-002	Review monthly capacity planning.	
OPS-003	Perform updates after hours for performance or migration sensitive applications.	
OPS-004	Utilize the current LTS branch.	
OPS-005	Update to the next maintenance version 4 weeks after release. Update to the current patch version 2 weeks after release.	
OPS-006	Maintain a pre-production environment for testing any changes needed (firmware, software, hardware) prior to executing the change actual production.	
OPS-007	Configure alerts and alert policies in prism central, not prism element.	
OPS-008	Utilize SMTP for alert transmission.	
BCN-001	Utilize application- consistent snapshots when needed by the application	
BCN-002	Place nearsync vm's in their own protection domain.	
BCN-003	Configure snapshot schedules to be more frequent than the desired rpo	
BCN-004	Configure snapshot schedules to retain the lowest number of snapshots while still meeting the retention policy.	

BCN-005	Group applications together in unique protection domains. Keep number of vm’s per protection domain as small as reasonably possible	
BCN-006	Application requires RPO=0 and RTO near zero	
BCN-007	Application requires zZero downtime DR avoidance solution	
BCN-008	Application requires RPO=0	
BCN-009	Provide RPO between 1 min and 15 min for the application	
BCN-010	Application requires RPO=>1h	
DCD-001	Decide on cluster licensing model for each ROBO (CBL or per-VM)	
DCD-002	Decide on cluster type used (if not ROBO per- VM)	
DCD-003	Decide whether to provide compute for ROBO failover	
DCD-004	Decide on ROBO failover compute over commit	
MULTI-NET-001	Calculate required storage replication bandwidth based on RPO.	
MULTI-NET-002	Calculate application- specific bandwidth.	
MULTI-NET-003	Place metro Availability and synchronous replication sites within the same region (Within 100km or less than 5msec RTT)	
MULTI-NET-004	Place metro witness within 200msec	
MULTI-NET-005	Do not use network segmentation for DR unless required. Routing segmentation is preferred over CVM segmentation	

MULTI-NET-006	Ensure redundancy of each network device. Track the complete network path between protected sites.	
MULTI-NET-007	Create and follow a redundancy test plan. Capture network components in plan.	
MULTI-NET-008	Ensure that the link between protected sites and a witness does not use the link between sites.	
MULTI-NET-009	For each workload, decide whether to keep the same IP addresses or change them during failover.	
OPS-001	Use X-PLAY for low-code/no-code operational automation.	
OPS-002	Monitor non-AOS VMs	
OPS-003	Discover the applications running on your infrastructure	
OPS-004	Monitor SQL Server from Prism	
AWS-001	Create separate subnets for user VMS	
AWS-002	Use AWS Direct Connect or create a second AWS cluster in a different AZ	
AWS-003	Protect your AWS cluster by replication back to on-prem.	
AWS-004	Plan your AWS VPC not to overlap with on-prem subnets	
AWS-005	Allow outbound internet access through an existing connection to the internet to allow the cluster portal to manage your AWS cluster	
AWS-006	Set AWS VPC limits to accommodate growth and auto-remediate features	
FLOW-001	Use Microsegmentation with App-Type Policy to segment applications	
FLOW-002	Use Microsegmentation with AD-Type Policy For VDI Deployments	

FILES-001	Dedicate Nutanix clusters to file services workloads
FILES-002	Use Nutanix clusters with hybrid storage for Files
FILES-003	Use a minimum of 3 FSVMs
FILES-004	Use an N+1 configuration of FSVMs to physical nodes
FILES-005	Allocate more than the minimum FSVM CPU and memory recommended for your planned workload
FILES-006	Use the minimum number of file server instances for your needs
FILES-007	Place the file server storage network on the same VLAN as the eth0 interface of the controller VMs.
FILES-008	Place the FSVM client network so it uses the shortest network route for the majority of file services clients
FILES-009	Use Files distributed shares
FILES-010	Use file server share-level inline compression
FILES-011	Use AOS container-level erasure coding (EC-X)
FILES-012	Do not enable AOS container-level deduplication
FILES-013	Use telescoping snapshot schedules
FILES-014	Use the minimum number of DR snapshots needed for your RPO requirements
OBJECTS-001	Provide internal and external networks
OBJECTS-002	Choosing between mixed and dedicated Objects clusters
OBJECTS-003	Hybrid or all-flash Nutanix clusters for Objects?
OBJECTS-004	Enable compression

OBJECTS-005	Number of worker nodes	
OBJECTS-006	Multiple object stores	
OBJECTS-007	Multi-cluster object stores	
OBJECTS-008	Enable versioning?	
OBJECTS-009	WORM (write-once, read-many)	
OBJECTS-010	Object lock	
OBJECTS-011	Lifecycle policy	
OBJECTS-012	Streaming replication	
BAC-001	Choose a backup vendor to use with Nutanix Mine	
BAC-002	Use Nutanix Objects as an archival tier for backups	
CALM-001	Small or Large Calm deployment	
CALM-002	Upgrade to the most recent version of Calm	
CALM-003	Create separate Calm instances for development and production	
CALM-004	Define processes for Calm DR and backup/ restore	
CALM-005	Align Calm project structure and RBAC configuration	
CALM-006	Design Calm environment to align with release methodology and geographic regions	
CALM-007	Enable the Calm Policy Engine	
CALM-008	Enable Showback via Nutanix Beam	
CALM-009	Choose a method for self-service	
KARBON-001	Minimize use of the admin kubeconfig	
KARBON-002	Distribute production applications across AZs	

KARBON-003	Use Kubernetes namespaces to segment users and applications rather than clusters	
KARBON-004	Leave enough space in a cluster to tolerate a worker node failure	
KARBON-005	Utilize consistent pod and service CIDRs	
KARBON-006	Upgrade the host OS as soon as possible	
DBAAS-001	Enable Database as a Service (DBaaS)	
DBAAS-002	Minimum hardware reqs for Nutanix clusters running Era	
DBAAS-003	Use dedicated Nutanix Clusters for Database	
DBAAS-004	Deploy Era management plane for HA	
DBAAS-005	Place Era on a separate management network	
DBAAS-006	Use the Era Platform license	
DBAAS-007	Use Active Directory for user, group, and service account access	
DBAAS-008	Use Active Directory groups to provide access to Era	
DBAAS-009	Provide an Era test environment	
DBAAS-010	Perform Era upgrades during service windows	
DBAAS-011	Use Era super admin and database admin roles	

List of Figures

List of Tables