AHV

Nutanix Tech Note

Version 1.3 • May 2017 • TN-2038

Copyright

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

Contents

1. Executive Summary

Nutanix delivers the industry's most popular hyperconverged solution, natively converging compute and storage into a turnkey appliance that can be deployed in minutes to run any application out of the box. The Nutanix solution offers powerful virtualization capabilities, including core virtual machine operations, live migration, VM high availability, and virtual network management, as fully integrated features of the infrastructure stack rather than as standalone products that require separate deployment and management.

Until now, the core architecture of today's standalone virtualization solutions had not changed significantly in over 12 years, despite major technological advances and evolving user expectations. The native Nutanix hypervisor, AHV, represents a new approach to virtualization that offers substantial benefits to enterprise IT administrators by simplifying every step of the infrastructure life cycle, from buying and deploying to managing, scaling, and supporting.

Table 1: Document Version History

Version Number	Published	Notes
1.0	February 2016	Original publication.
1.1	July 2016	Updated platform information.
1.2	December 2016	Updated for AOS 5.0.
1.3	May 2017	Updated for AOS 5.1.

2. Nutanix Enterprise Cloud Platform

The Nutanix Enterprise Cloud Platform is a hyperconverged infrastructure system delivering enterprise-class storage, compute, and virtualization services for any application. As a fully integrated IT solution, it eliminates the cost and complexity of legacy datacenter products deployed individually and managed separately. Nutanix brings together web-scale engineering and consumer-grade design to make infrastructure invisible and to elevate IT teams so they can focus on what matters most—applications.

At the heart of the Nutanix platform are two product families: Nutanix Acropolis and Nutanix Prism. The Acropolis Distributed Storage Fabric (DSF) delivers enterprise storage services, and the highly available and scalable App Mobility Fabric (AMF) within Acropolis enables workloads to move freely between virtualization environments without penalty. Nutanix Prism, a comprehensive management solution for Acropolis, provides unprecedented one-click simplicity to the IT infrastructure life cycle.

Enterprise professionals have the choice of using VMware vSphere, Microsoft Hyper-V, or the natively integrated hypervisor, AHV, to run their applications on Nutanix. AHV is built on proven open source technology and has been hardened for security.

Together, AHV and AMF decouple applications from the underlying infrastructure, giving enterprise IT the flexibility to choose whichever runtime environment is best for their applications and services. Acropolis works with Nutanix Prism to give administrators consumer-grade simplicity in managing their entire virtual datacenter.

3. AHV

AHV is built on a proven open source CentOS KVM foundation and extends KVM's base functionality to include features such as high availability (HA) and live migration. AHV comes preinstalled on Nutanix appliances and can be configured within minutes to deploy applications.

3.1. Configuration Maximums and Scalability

The following configuration maximums and scalability limits apply:

- Maximum cluster size: N/A—same as Nutanix cluster size
- Maximum vCPUs per VM: Number of physical cores per host
- Maximum memory per VM: 2 TB
- Maximum VMs per host: N/A—Limited by memory
- Maximum VMs per cluster: N/A—Limited by memory

Within AHV there are three main components:

- KVM-kmod
  - KVM kernel module
- Libvirtd
  - An API, daemon, and management tool for managing KVM and QEMU. Communication between Acropolis and KVM and QEMU occurs through libvirtd.
- Qemu-KVM
  - A machine emulator and virtualizer that runs in userspace for every virtual machine (domain). In AHV it is used for hardware-assisted virtualization and VMs run as HVMs.

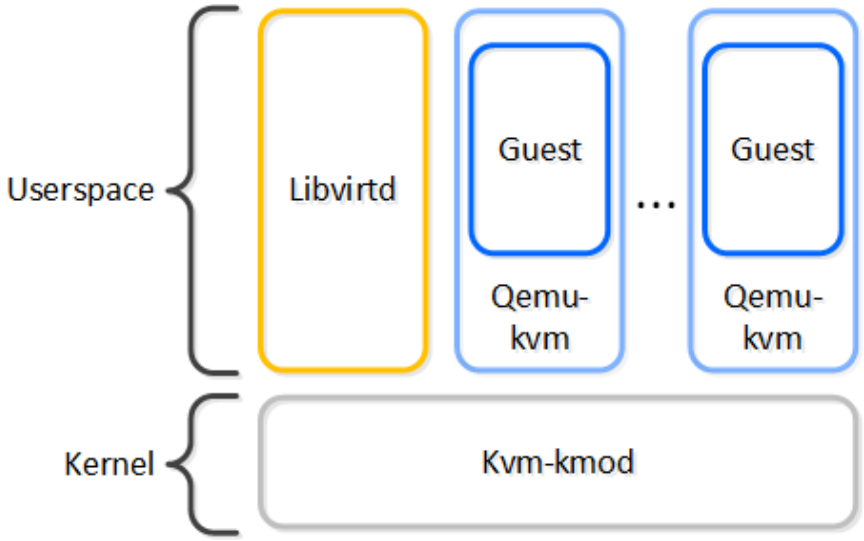

Figure 1: AHV Components

4. Acropolis App Mobility Fabric

The Acropolis App Mobility Fabric (AMF) is a collection of technologies built into the Nutanix solution that allows applications and data to move freely between runtime environments. AMF handles multiple migration scenarios, including from non-Nutanix infrastructure to Nutanix clusters, between Nutanix clusters running different hypervisors, and from Nutanix to a public cloud solution.

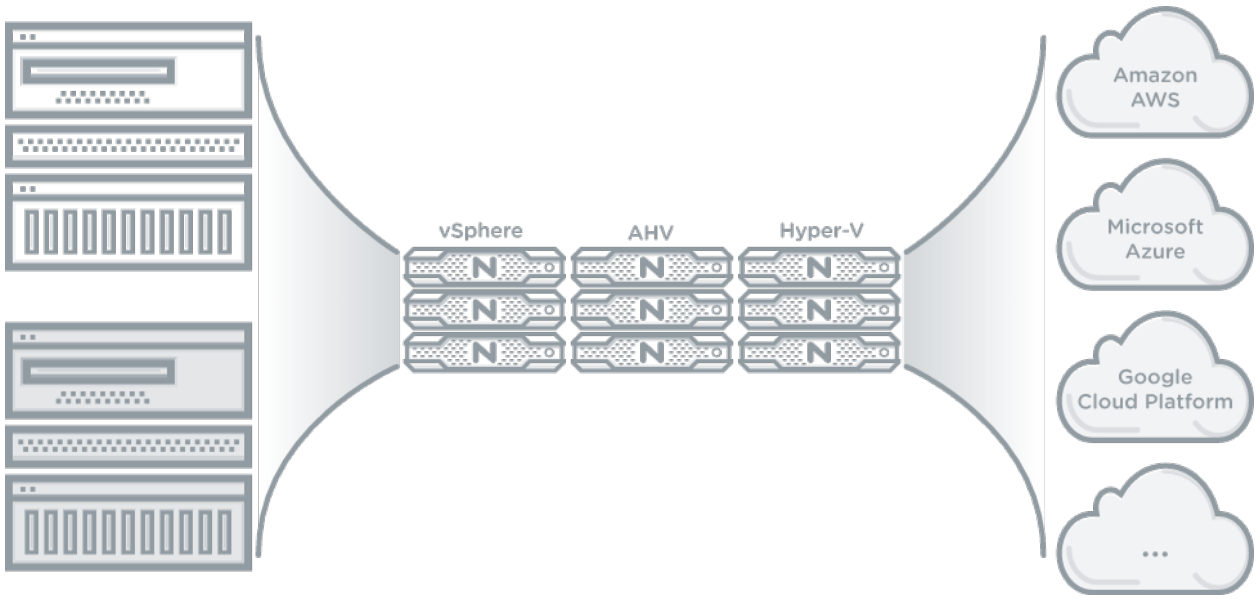

Figure 2: Acropolis App Mobility Fabric

AMF includes powerful features such as the Nutanix Sizer that allows administrators to select the right Nutanix system and deployment configuration to meet the needs of each workload; the Foundation tool that simplifies installation of any hypervisor on a Nutanix cluster; and Cloud Connect, the built-in hybrid cloud technology that seamlessly backs up data to public cloud services. We'll describe additional AMF features in the next sections.

5. Integrated Management Capabilities

Nutanix prioritizes making infrastructure management and operations uncompromisingly simple. The Nutanix platform natively converges compute, storage, and virtualization in a ready-to-use product that can be managed from a single pane of glass with Nutanix Prism. Prism provides integrated capabilities for cluster management and virtual machine management that are available via the Prism graphical user interface (GUI), command line interface (CLI), PowerShell, and the Acropolis REST API.

5.1. Cluster Management

Managing clusters on AHV focuses on creation, updates, deletion, and monitoring of cluster resources. These resources include hosts, storage, and networks.

Host Profiles

Prism provides a central location for administrators to update host settings such as virtual networking and high availability across all nodes in an AHV cluster. Controlling configuration at the cluster level eliminates the need for manual compliance checks and reduces the risk of having a cluster that is not uniformly configured.

Storage Configuration

Nutanix Acropolis relies on the hypervisor-agnostic Distributed Storage Fabric (DSF) to provide data services such as storage provisioning, snapshots, clones, and data protection for virtual machines directly, rather than using the hypervisor's storage stack. On each AHV host, an iSCSI redirector service provides a highly resilient storage path from each VM to storage across the Nutanix cluster.

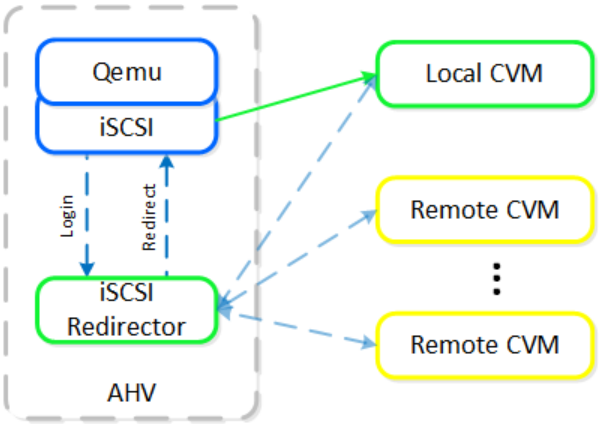

Figure 3: Storage Configuration

Virtual Networking

AHV leverages Open vSwitch (OVS) for all VM networking. When creating a new AHV cluster, the Controller VM (CVM) and management networking paths are configured automatically. Administrators can easily create new VLAN-backed layer-2 networks through Prism. Once you've created a network, you can assign it to existing and newly created VMs.

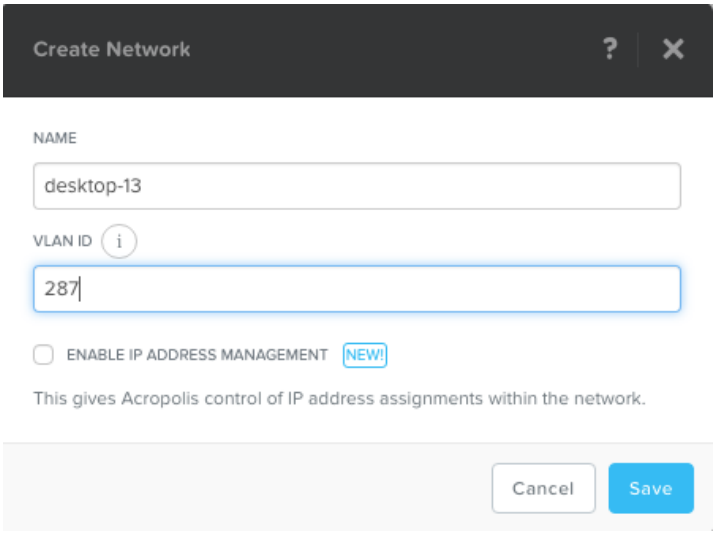

Figure 4: Creating a Network in Prism

Along with streamlining virtual machine network creation, Acropolis can provide DHCP address management for each network that is created. This functionality allows administrators to configure pools of addresses for each network that they can automatically assign to VMs.

Rolling Upgrades

Nutanix provides an incredibly simple and reliable one-click upgrade process for all software within the Nutanix platform. This feature includes updates for the Acropolis base software, AHV, firmware, and Nutanix Cluster Check (NCC). Upgrades are nondisruptive and allow the cluster to run continuously while nodes upgrade on a rolling basis in the background, thus ensuring always-on cluster operation during software maintenance. Nutanix qualifies firmware updates from the manufacturers of the hard or solid-state disk drives in the cluster and makes them available via the same upgrade process.

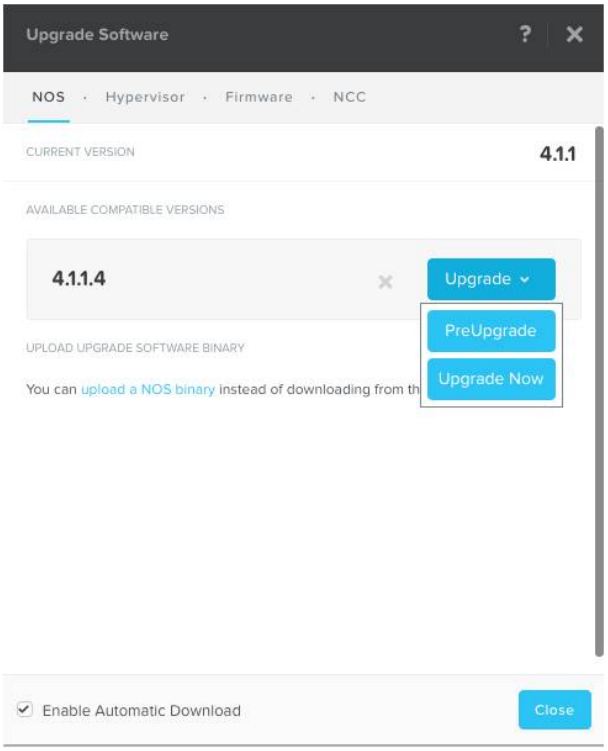

Figure 5: Nutanix Upgrade Process

Host Maintenance Mode

Administrators can place AHV hosts in maintenance mode during upgrades and maintenance-related operations. Maintenance mode live migrates all VMs running on the node to other nodes within the AHV cluster, and the Controller VM (CVM) can safely shut down if required. Once the maintenance process has completed all of the steps for the node, it returns the CVM to service and synchronizes with other CVMs in the cluster. Maintenance mode enables graceful suspension of hosts for routine cluster maintenance.

Scaling

The Nutanix solution's scale-out architecture enables incremental, predictable scaling of capacity and performance in a Nutanix cluster running any hypervisor, including AHV. Administrators can start with as few as three nodes and scale out without limits. The system automatically discovers new nodes and makes them available for use. Expanding clusters is as simple as selecting the discovered nodes you want to add and providing network configuration details. Through Prism, administrators can image or update new nodes to match the AHV version of their preexisting nodes to allow seamless node integration, no matter what version was originally installed.

5.2. Virtual Machine Management

Virtual machine management on AHV focuses on creation, updates, deletion, data protection, and monitoring of VMs and their resources. These cluster services and features are all available via the Prism interface, a distributed management layer that is available on the CVM on every AHV host.

VM Operations

Prism provides a list of all VMs in an AHV cluster along with a wealth of configuration, resource usage, and performance details on a per-VM basis. Administrators can create VMs and perform a number of operations on selected VMs, including power on/off, power cycle, reset, shutdown, reboot, snapshots and clones, migration, pause, update, delete, and launch a remote console.

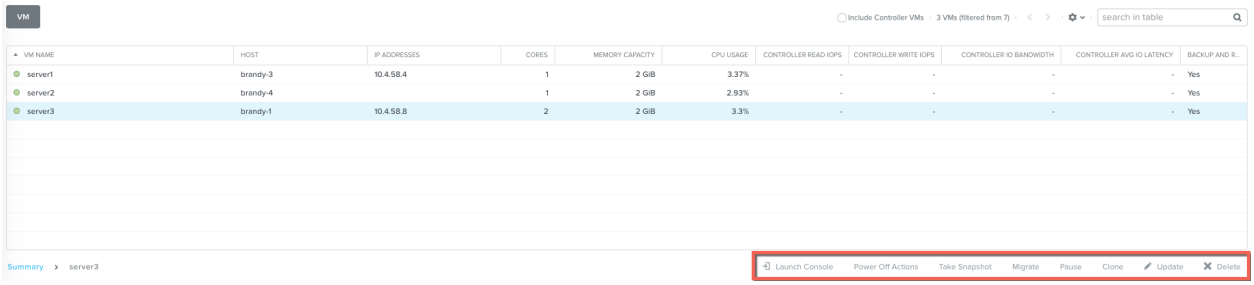

Figure 6: VM Operations in Prism

Image Management

The image management service within AHV is a centralized repository that provides access to virtual media and disk images, as well as the ability to import from external sources. It allows you to store virtual machines as templates or master images, which you can then use to create new VMs quickly from a known good base image. The image management service can store the virtual disk files that are used to create fully functioning VMs or operating system installation media as an .iso file that you can mount to provide a fresh operating system install experience. Incorporated into Prism, the image service can import and convert existing virtual disk formats,

including .raw, .vhd, .vmdk, .vdi and .qcow2. The previous virtual hardware settings do not constrain an imported virtual disk, allowing administrators the flexibility to fully configure CPU, memory, virtual disks, and network settings at the time of VM provisioning.

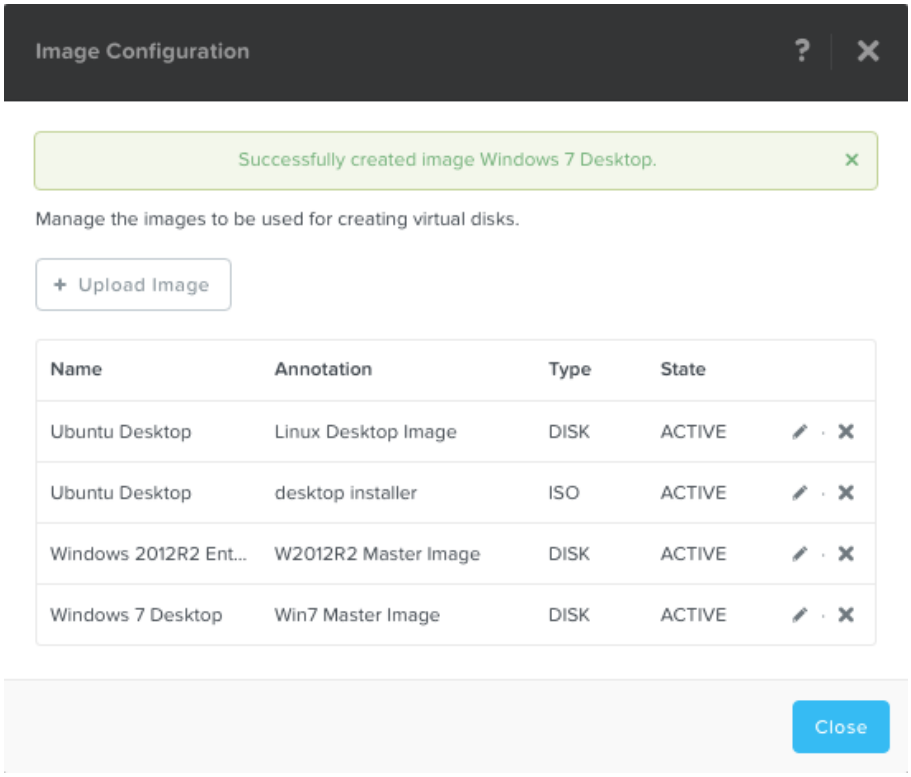

Figure 7: Image Configuration in Prism

Acropolis Dynamic Scheduling

Acropolis Dynamic Scheduling (ADS) is an automatic function enabled on every AHV cluster to avoid hot spots within cluster nodes. ADS continually monitors CPU, memory, and storage data points to make migration and initial placement decisions for VMs and ABS volumes. Starting with existing statistical data for the cluster, ADS watches for anomalies, honors affinity controls, and only makes move decisions to avoid hot spots. Using machine learning, ADS can adjust move thresholds over time from their initial fixed values to achieve the greatest efficiency without sacrificing performance.

ADS tracks each individual node's CPU and memory utilization. When a node's CPU allocation breaches its threshold (currently 85 percent of CVM CPU), Nutanix migrates VMs or ABS volumes as needed off that host to rebalance the workload.

Note: Migration only occurs when there is contention. If there is skew in utilization between nodes (for example, three nodes at 10 percent and one at 50 percent),

migration does not occur, as it offers no benefit unless there is contention for resources.

Intelligent VM Placement

When you create, restore, or recover VMs, Acropolis assigns them to an AHV host within the cluster based on recommendation from ADS. This VM placement process also takes into account the AHV cluster's high availability (HA) configuration, so it doesn't violate any failover host or segment reservations. We explain these HA constructs further in the high availability section below.

Affinity and Antiaffinity

Affinity controls provide the ability to govern where VMs run. AHV has two types of affinity controls:

1. VM-host affinity strictly ties a VM to a host or group of hosts, so the VM only runs on that host or group. Affinity is particularly applicable for use cases that involve software licensing or VM appliances. In such cases, you often need to limit the number of hosts an application can run on or bind a VM appliance to a single host.
2. Antiaffinity lets you declare that a given list of VMs should not run on the same hosts. Antiaffinity provides a mechanism for allowing VMs running a distributed application or clustered VMs to run on different hosts, thereby increasing the application's availability and resiliency. To prefer VM availability over VM separation, the system overrides this type of rule when a cluster becomes constrained.

Live Migration

Live migration allows the system to move VMs from one Acropolis host to another while the VM is powered on, whether the movement is initiated manually or through an automatic process. Live migration can also occur when a host is placed in maintenance mode, which evacuates all VMs.

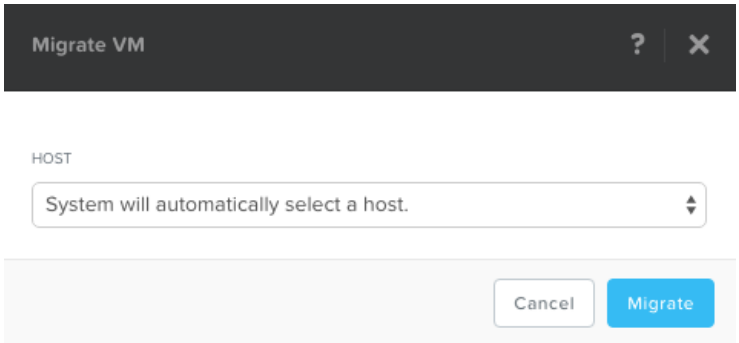

Figure 8: Migrating VMs

Cross-Hypervisor Migration

The Acropolis App Mobility Fabric (AMF) simplifies the process of migrating existing VMs between an ESXi cluster and an AHV cluster using built-in data protection capabilities. You can create one or more protection domains on the source cluster and set the AHV cluster as the target remote cluster. Then, snapshot VMs on the source ESXi cluster and replicate them to the AHV cluster, where you can restore them and bring them online as AHV VMs.

Automated High Availability

Acropolis offers virtual machine high availability (VMHA) to ensure VM availability in the event of a host or block outage. If a host fails, the VMs previously running on that host restart on healthy nodes throughout the cluster. There are multiple HA configuration options available to account for different cluster scenarios.

1. By default, all Acropolis clusters provide best effort HA, even when the cluster is not configured for HA. Best effort HA works without reserving any resources. Admission control is not enforced, so there may not be sufficient capacity available to start all the VMs from the failed host.
2. You can also configure an Acropolis cluster for HA with resource reservation to guarantee that the resources required to restart VMs are always available. Acropolis offers two modes of resource reservation: host reservations and segment reservations. Clusters with uniform host configurations (for example, RAM on each node) use host reservation, while clusters with heterogeneous configurations use segment reservation.

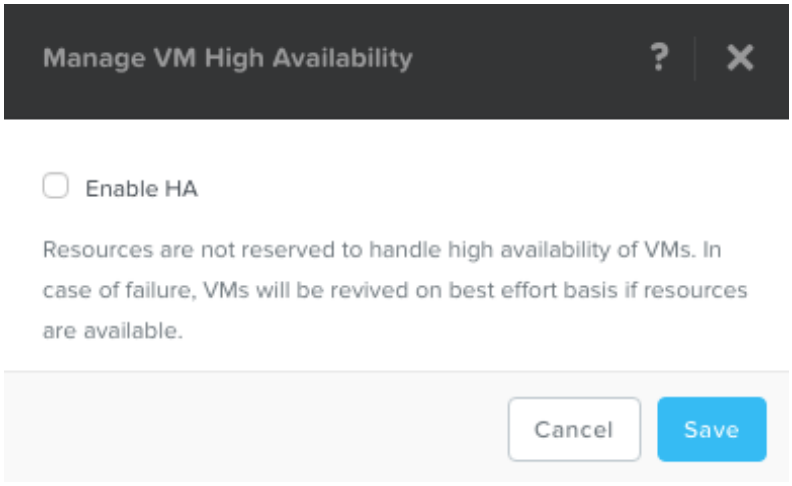

Figure 9: High Availability

- Host reservations.

  This method reserves an entire host for failover protection. Acropolis selects the least used host in the cluster as a reserve node, and all VMs on that node are migrated to other nodes

in the cluster so that the reserve node's full capacity is available for VM failover. Prism determines the number of failover hosts needed to match the number of failures the cluster will tolerate for the configured replication factor (RF).

- Segment reservations.

  This method divides the cluster into fixed-size segments of CPU and memory. Each segment corresponds to the largest VM that is guaranteed to be restarted after a host failure. The scheduler, also taking into account the number of host failures that can be tolerated, implements admission control to ensure that there are enough resources reserved to restart VMs upon failure of any host in the cluster.

The Acropolis Master CVM restarts the VMs on the healthy hosts. The Acropolis Master tracks host health by monitoring connections to the libvirt on all cluster hosts. If the Acropolis Master becomes partitioned or isolated, or if it fails, the healthy portion of the cluster elects a new Acropolis Master.

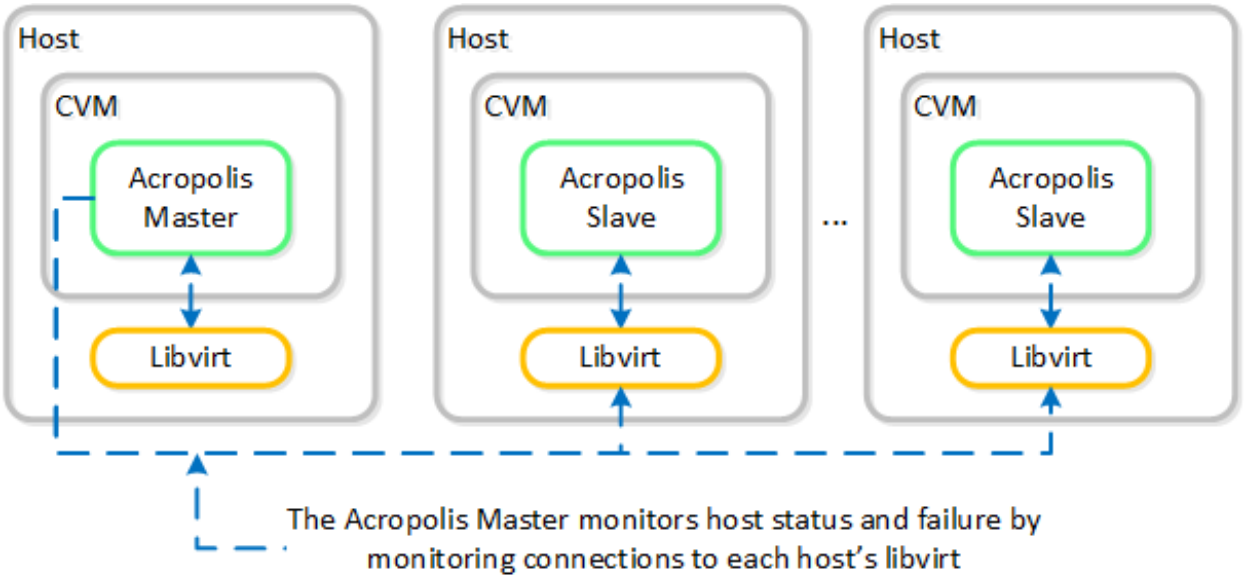

Figure 10: Acropolis Master Monitoring

Converged Backup and Disaster Recovery

The Acropolis App Mobility Fabric's converged backup and disaster recovery (DR) services protect your clusters. Nutanix clusters running any hypervisor have access to these features, which safeguard VMs both locally and remotely for use cases ranging from basic file protection to recovery from a complete site outage. To learn more about the built-in backup and DR capabilities in the Nutanix platform, read the Data Protection and Disaster Recovery technical note.

Backup APIs

To complement the integrated backup that the Enterprise Cloud Platform provides, AHV also offers a rich set of APIs to support external backup vendors. The AHV backup APIs utilize changed region tracking to allow backup vendors to back up only the data that has changed since the last backup job for each individual VM. Changed region tracking also allows backup jobs to skip reading zeros, further reducing backup times and bandwidth consumed.

Nutanix backup APIs allow backup vendors that build integration to provide full, incremental, and differential backups. Changed region tracking is always on in AHV clusters and does not require you to enable it on each VM. Backups can be either crash-consistent or application-consistent.

Analytics

Nutanix Prism provides in-depth analytics for every element in the infrastructure stack, including hardware, storage, and VMs. Administrators can use Element views to monitor these infrastructure stack components, and they can use the Analysis view to get an integrated assessment of cluster resources or to drill down to specific metrics on a given element.

Prism makes detailed VM data available, grouping it into the following categories:

- VM Performance: Multiple charts with CPU and storage-based reports around resource usage and performance.
- Virtual Disks: In-depth data points that focus on I/O types, I/O metrics, read source, cache hits, working set size, and latency on a per-virtual disk level.
- VM NICs: vNIC configuration summary for a virtual machine.
- VM Snapshots: A list of all snapshots for a virtual machine with the ability to clone or restore from the snapshot or to delete the snapshot.
- VM Tasks: A time-based list of all operational actions performed against the selected virtual machine. Details include task summary, percent complete, start time, duration, and status.
- Console: Administrators can open a pop-up console session or an inline console session for a virtual machine.

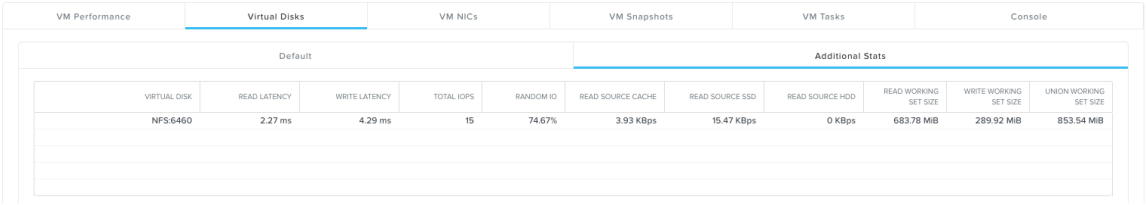

Figure 11: Prism Analytics

The Storage tab provides a direct view into the Distributed Storage Fabric (DSF) running on an AHV cluster. Administrators can look at detailed storage configurations, capacity usage over time, space efficiency, and performance, as well as a list of alerts and events related to storage.

The Hardware tab provides a direct view into the Nutanix blocks and nodes that make up an Acropolis cluster. These reports are available in both a diagram and a table format for easy consumption.

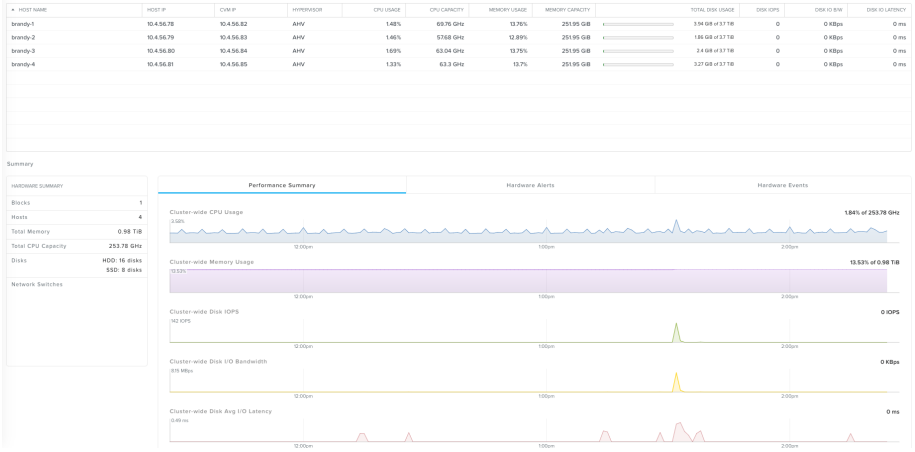

Figure 12: Performance Summary in Prism

The Prism Analysis tab gives administrators the tools they need to explore and understand what is going on in their clusters quickly and to identify steps for remediation as required. You can create custom interactive charts using hundreds of metrics available for elements such as hosts, disks, storage pools, containers, virtual machines, protection domains, remote sites, replication links, clusters, and virtual disks, then correlate trends in the charts with alerts and events in the system. You can also choose specific metrics and elements and set a desired time frame when building reports, so you can focus precisely on the data that you're looking for.

Figure 13: Prism Analysis

6. GPU Support

A graphics processing unit (GPU) is the hardware or software that displays graphical content to end users. In laptops and desktops, GPUs are either a physical card or built directly into the CPU hardware, while GPU functions in the virtualized world have historically been software-driven and consumed additional CPU cycles. With modern operating systems and applications as well as 3D tools, more and more organizations find themselves needing hardware GPUs in the virtualized world.

6.1. GPU Passthrough

The GPU cards deployed in server nodes for virtualized use cases typically combine multiple GPUs in a single PCI card. With GPU passthrough, AHV can pass a GPU through to a VM, allowing the VM to own that physical device in a 1:1 relationship. Configuring nodes with one or more GPU cards that attach multiple GPUs to a larger number of VMs allows you to consolidate applications and users on each node. AHV currently supports NVIDIA Grid cards for GPU passthrough and is certified for the M10 and M60 cards.

With passthrough, you can also use GPUs for offloading computational workloads—a more specialized situation than the typical graphical use cases. GPU compute scenarios assign one or more GPUs for a VM to use for processing. AHV allows you to assign up to 16 GPUs to a single VM, whereas competing hypervisors permit you to assign only one GPU per VM.

7. Security

Nutanix has adopted a holistic approach to infrastructure security. The fully integrated infrastructure stack eliminates security risks associated with legacy solutions that involve many vendors with a narrow, fragmented view of security. For example, Nutanix designed AHV to be an integral component of the converged infrastructure stack rather than a general-purpose hypervisor. Consequently, many of the services that the Acropolis solution makes unnecessary are turned off in order to reduce the security surface area.

7.1. Security Development Life Cycle

To maintain agile and comprehensive security, Nutanix has developed its own Security Development Life Cycle (SecDL), which addresses security at every step of the development process instead of applying it at the end as an afterthought. SecDL integrates security features into the software development process, including automated security testing during development and threat modeling to assess and mitigate customer risk from code changes. SecDL makes security a first-class citizen that drives best practices within Nutanix and for our customers, providing both defense in depth and a “hardened by default” posture for releases.

7.2. Security Baseline and Self-Healing

Nutanix has developed custom Security Technical Implementation Guides (STIGs), security tools based on well-established National Institute of Standards and Technology (NIST) standards, that administrators can apply to multiple baseline requirements for DoD and PCI-DSS. Unlike general-purpose STIGs that provide blanket security recommendations, Nutanix STIGs are specific to the Acropolis platform and therefore more effective. Encoded in a machine-readable format, Nutanix STIGs allow for automated validation, ongoing monitoring, and self-remediation, reducing the time required to verify security compliance from weeks or months to days.

You can read more about the Nutanix approach to information security in the Information Security technical note.

8. Conclusion

The Nutanix Enterprise Cloud Platform embodies a radically new approach to enterprise infrastructure—one that simplifies every step of the infrastructure life cycle from buying and deploying to managing, scaling, and supporting. The Nutanix solution's web-scale technologies and architecture let you run any workload at any scale. With Nutanix Acropolis and Nutanix Prism, administrators get powerful virtualization capabilities that are fully integrated into the converged infrastructure stack and can be managed from a single pane of glass.

Appendix

References

Data Protection and Disaster Recovery, Nutanix Tech Note (2016): http://go.nutanix.com/data-protection-disaster-recovery.html

Information Security, Nutanix Tech Note (2016): http://go.nutanix.com/information-security-with-nutanix.html

About Nutanix

Nutanix makes infrastructure invisible, elevating IT to focus on the applications and services that power their business. The Nutanix Enterprise Cloud Platform leverages web-scale engineering and consumer-grade design to natively converge compute, virtualization, and storage into a resilient, software-defined solution with rich machine intelligence. The result is predictable performance, cloud-like infrastructure consumption, robust security, and seamless application mobility for a broad range of enterprise applications. Learn more at www.nutanix.com or follow up on Twitter @nutanix.

List of Figures

List of Tables