v1.7 | October 2022 | BP-2083

BEST PRACTICES

ROBO Deployment and Operations

NUTANIX
YOUR ENTERPRISE CLOUD

Copyright

Contents

1. Executive Summary....4

2. Introduction....5
- Audience....5
- Purpose....5
- Document Version History....5

3. Remote Office and Branch Office Deployment....7
- Cluster Selection....7
- Hypervisor Selection....8
- Cluster Storage Capacity....9
- Witness Requirements....11
- Prism Central....11
- Seeding....14

4. Operating Nutanix in a Multisite Environment....16
- Prism Central Management....16
- Disaster Recovery and Backup....20
- Remote Site Setup....21
- Scheduling....26

5. Failure Scenarios....33
- Failed Hard Drive....33
- Failed Node....35

6. Appendix....36
- Best Practices Checklist....36

About Nutanix....41

List of Figures....42

1. Executive Summary

Nutanix provides a powerful converged compute and storage system that offers one-click simplicity and high availability for remote and branch offices. This document describes best practices for deploying and operating remote and branch offices on Nutanix, including guidance on picking the right Nutanix cluster for seeding customer data to overcome slow remote network links.

The Nutanix platform's self-healing design reduces operational and support costs, such as unnecessary site visits and overtime. With Nutanix, you can proactively schedule projects and site visits on a regular cadence, rather than working around emergencies. Prism Central, our end-to-end infrastructure management tool, streamlines remote cluster operations through one-click upgrades, while also providing simple orchestration for multiple cluster upgrades. Following the best practices in this document ensures that you can quickly restore your business services in the event of a disaster. Nutanix makes deploying and operating remote and branch offices as easy as deploying to the public cloud, but with control and security on your own terms.

2. Introduction

Audience

This best practice guide is part of the Nutanix Solutions Library. We wrote it for solution architects, administrators, and system engineers responsible for designing, managing, and supporting Nutanix infrastructures running remote office and branch locations. Readers should already be familiar with Nutanix.

Purpose

This document covers the following subject areas:

- Overview of the Nutanix solution for managing multiple remote offices.
- Nutanix cluster selection for remote sites.
- Network sizing for management and disaster recovery traffic.
- Best practices for remote management with Nutanix Cloud Manager (NCM) Pro.
- How to design for failure at remote sites.

Document Version History

Version Number	Published	Notes
1.0	June 2017	Original publication.
1.1	March 2018	Added information on one- and two-node clusters.
1.2	February 2019	Updated Nutanix overview.
1.3	April 2019	Updated witness requirements.

Version Number	Published	Notes
1.4	February 2020	Updated Best Practices Checklist section.
1.5	December 2020	Model updates.
1.6	January 2022	Updated Remote Office and Branch Office Deployment section.
1.7	October 2022	Updated product naming.

3. Remote Office and Branch Office Deployment

Cluster Selection

Picking the right solution always involves trade-offs. While a remote site isn't a datacenter, uptime is nonetheless a crucial concern. Financial constraints and physical layout also affect what counts as the best architecture for your environment. Nutanix offers a wide variety of clusters for remote locations. You can select single- and dual-socket cluster options, as well as options that can reduce licensing costs.

Three-Node Clusters

Although a three-node system may cost more money up front, it's the gold standard for remote and branch offices. Three-node clusters provide excellent data protection by always committing two copies of your data, which means that your data is safe even during failures. Three-node clusters also rebuild your data within 60 seconds of a node going down. Distributed storage rebuilds the data on the downed node and does so without any user intervention.

A self-healing Nutanix three-node cluster also prevents unnecessary trips to remote sites. We recommend designing these systems with enough capacity to handle an entire node going down, which allows the loss of multiple hard drives, one at a time. Because there is no reliance on RAID, the cluster can lose and heal drives, one after the next, until available space runs out. For sites with high availability requirements or sites that are difficult to visit, we recommend additional capacity above the n + 1 node count. Three-node clusters can scale up to eight nodes with 1 Gbps networking and up to any scale when using 10 Gbps and higher networking. With Nutanix reliability and availability, you can focus on expanding your business rather than wasting resources on emergency site visits.

Two-Node Clusters

Two-node clusters offer reliability for smaller sites that must be cost effective and run with tight margins. These clusters use a witness only in failure scenarios to coordinate rebuilding data and automatic upgrades. You can deploy the witness offsite up to 500 ms away for ROBO and 200 ms when you use Metro Availability. Multiple clusters can use the same witness for two-node and Metro clusters. Nutanix supports two-node clusters with ESXi and AHV only.

One-Node Clusters

One-node clusters are a perfect fit if you have low availability requirements and need strong overall management for multiple sites. One-node clusters provide resiliency against the loss of a hard drive while still offering great remote management. Nutanix supports one-node clusters with ESXi and AHV only.

Hypervisor Selection

Nutanix supports a wide range of hypervisors to meet your enterprise's needs. The three main considerations for choosing the right hypervisor for your environment are supportability, operations, and licensing costs.

Supportability can encompass support for your applications, the training your staff needs to support daily activities, and break-fix. When it comes to supportability, the path of least resistance often shapes hypervisor selection. Early on, where customers could afford to use virtualization, ESXi was a prime candidate. Because many environments run ESXi, the Nutanix 1000 series offers a mixed-hypervisor deployment consisting of two ESXi nodes and one AHV storage node. The mixed-hypervisor deployment option provides the same benefits as a full three-node cluster but removes the CPU licensing required by some hypervisor licensing models.

Operationally, Nutanix aspires to build infrastructure that's invisible to the people using it. Customers who want a fully integrated solution should select AHV as their hypervisor. With AHV, virtual machine (VM) placement and data placement happen automatically, without any required settings. Nutanix also hardens systems by default to meet security requirements and provides the automation necessary to maintain that security. Nutanix supplies STIGs

(Security Technical Information Guidelines) in machine-readable code for both AHV and the storage controller.

For environments that don't want to switch hypervisors in the main datacenter, Nutanix offers cross-hypervisor disaster recovery to replicate VMs from AHV to ESXi or ESXi to AHV. In the event of a disaster, administrators can restore their AHV VM to ESXi for quick recovery or replicate the VM back to the remote site with easy workflows.

Cluster Storage Capacity

For all two- and three-node clusters, Nutanix recommends n + 1 nodes to ensure sufficient space for rebuilding. You should also remove an additional five percent so that the system isn't completely full on rebuild. For single-node clusters, reserve 55 percent of usable space to recover from the loss of a disk.

NX-1175S-G8 - SPECIFICATION

MODEL	Nutanix : Per Node (per Block) NX-1175S-G8 (Configure to Order)
DEPLOYMENT MODEL	Factory Installed Software
USE CASE(S)	Remote Office/Branch Office, Test and Development, Analytics and Big Data
SERVER COMPUTE **	Single Intel Ice Lake: Silver 4309Y [8 cores / 2.80 GHz] Silver 4310 [12 cores / 2.10 GHz] Silver 4310T [10 cores / 2.30 GHz] Silver 4314 [16 cores / 2.40 GHz] Silver 4316 [20 cores / 2.30 GHz] Gold 5315Y [8 cores / 3.20 GHz] Gold 5317 [12 cores / 3.00 GHz] Gold 5318Y [24 cores / 2.10 GHz] Gold 5320T [20 cores / 2.30 GHz]... Show more
Boot	
Single M.2 NVMe	[1] x 512GB M.2 Boot Device
STORAGE CAPACITY	
All SSD SED	[2, 4] x SSD: [1.92 TB, 3.84 TB]
All SSD	[2, 4] x SSD: [1.92 TB, 3.84 TB, 7.68 TB]
SSD+HDD	[2] x SSD: [1.92 TB, 3.84 TB, 7.68 TB], [2] x HDD: [6.0 TB, 8.0 TB, 12.0 TB, 18.0 TB]
SSD+HDD SED	[2] x SSD: [1.92 TB, 3.84 TB], [2] x HDD: [6.0 TB, 8.0 TB, 12.0 TB]

Figure 1: Example NX-1175S Node Configuration

If you deploy the NX-1175S with 1.92 TB SSD and 6 TB HDD, the usable capacity for one node is 5.89 TB after accounting for all overhead. Therefore, the usable capacity while accounting for disk or node failure with replication factor 2 is as follows:

- One- and two-node clusters: 5.89 TB
- Three-node cluster: 11.19 TB

The NX-1000 series is built for ROBO environments. If you want to look at all available sizing options, refer to the dynamic sizing sheet.

Witness Requirements

There are several requirements when you set up a witness. The witness VM requires at least:

- 2 vCPU
- 6 GB of memory
- 25 GB of storage

The witness VM must reside in a separate failure domain, which means that you need independent power and network connections from each of the two-node clusters. We recommend locating the witness VM in a third physical site with dedicated network connections to site one and site two to avoid a single point of failure.

Communication with the witness happens over port TCP 9440; therefore, this port must be open for the Controller Virtual Machines (CVMs) on any two-node clusters using the witness.

Network latency between each two-node cluster and the witness VM must be less than 500 ms for ROBO.

The witness VM may reside on any supported hypervisor and run on either Nutanix or non-Nutanix hardware. You can register multiple (different) two-node and Metro Availability cluster pairs to a single witness VM, but no more than 50 combined witnessed Metro protection domains per cluster pair and two-node cluster.

Prism Central

Nutanix Prism provides central access for administrators to configure, monitor, and manage virtual environments. Powered by advanced data analytics, heuristics, and rich automation, Prism offers unprecedented simplicity by combining several aspects of datacenter management into a single, consumer-grade solution. Using innovative machine learning technology, Prism can mine large volumes of system data easily and quickly and generate actionable insights for optimizing all aspects of virtual infrastructure management. Prism

is a part of every Nutanix deployment and has two core components: Prism Element and Prism Central.

Prism Element

Prism Element is a service built into the platform for every Nutanix cluster deployed. It provides the ability to fully configure, manage, and monitor Nutanix clusters running any hypervisor.

Because Prism Element only manages the cluster it's part of, each Nutanix cluster in a deployment has a unique Prism Element instance for management. As organizations deploy multiple Nutanix clusters, they want to be able to manage all of them from a single Prism instance, so Nutanix introduced Prism Central.

Prism Central

Prism Central offers an organizational view into a distributed Nutanix deployment, with the ability to attach all remote and local Nutanix clusters to a single Prism Central deployment. This global management experience offers a single place to monitor performance, health, and inventory for all Nutanix clusters. Prism Central is available in a standard version included with every Nutanix deployment and as a separately licensed Pro version that enables several advanced features.

Nutanix Cloud Manager Standard offers all the great features of Prism Element under one umbrella, with single sign-on for your entire Nutanix environment. The standard version makes day-to-day management easier by placing all your applications at your fingertips with the entity explorer. The entity explorer offers customizable tagging for applications so that, even if they are dispersed among different sites, you can better analyze their aggregated data in one central location.

Nutanix Cloud Manager Pro has additional features that help manage large deployments and prevent emergencies and unnecessary site visits. Prism Central Pro includes:

- Customizable dashboards.
- Capacity runway to safeguard against exhausting resources.

- Capacity planning to safely reclaim resources from old projects and just-in-time forecasting for new projections.
- Advanced search to streamline access to features with minimal training.
- Simple multicluster upgrades.

We recommend the following best practices when you deploy Prism Central.

Network

- Prism Central uses TCP port 9440 to communicate with the CVMs in a Nutanix cluster. If your network and servers have a firewall enabled, open port 9440 between the CVMs and the Prism Central VM to allow access.
- Always deploy with DNS. Prism Central occasionally performs a request on itself; if it can't resolve the DNS name, some cluster statistics may not be present.
- If you use LDAP or LDAPS for authentication, open port 3268 (for LDAP) or 3269 (for LDAPS) on the firewall.

Initial Installation and Sizing

- Small environments: For fewer than 2,500 VMs, size Prism Central with 6 vCPU, 26 GB of memory, and 500 GiB of storage.
- Large environments: For up to 12,500 VMs, size Prism Central with 10 vCPU, 44 GB of memory, and 2,500 GiB of storage.
- If you install on Hyper-V, use the System Center Virtual Machine Manager (SCVMM) library on the same cluster to enable fast copy. Fast copy improves the deployment time.

Statistics

- Prism Central keeps 13 weeks of raw metrics and 53 weeks of hourly metrics.
- Nutanix Support can help you keep statistics over a longer period if needed. However, once you change the retention time, only stats written after the change have the new retention time.

Cluster Registration and Licensing

- Prism Central doesn't manage Cloud Connect (cloud-based) clusters.
- Each node registered to and managed by Prism Pro requires you to apply a Prism Pro license through the Prism Central web console. For example, if you have registered and are managing 10 Nutanix nodes (regardless of the individual node or cluster license level), you need to apply 10 Nutanix Cloud Manager Pro licenses through the Prism Central web console.

Seeding

When you deal with a remote site that has a limited network connection back to the main datacenter, you may need to seed data to overcome network speed deficits. Seeding involves using a separate device to ship the data to the remote location. Instead of replication taking weeks or months, depending on the amount of data you need to protect, you can copy the data locally to a separate Nutanix node and ship it to your remote site.

Nutanix checks the snapshot metadata before seeding the device to prevent unnecessary duplication. Nutanix can apply its native data protection to a seed cluster by placing VMs in a protection domain and replicating them to a seed cluster. A protection domain is a collection of VMs that have a similar recovery point objective (RPO). You must ensure, however, that the seeding snapshot doesn't expire before you can copy the data to the destination.

Seed Procedure

The following procedure lets you use seed cluster storage capacity to bypass the network replication step. During this procedure, the administrator stores a snapshot of the VMs on the seed cluster while it's installed in the ROBO site, then physically ships it to the main datacenter.

1. Install and configure application VMs on a ROBO cluster.
2. Create a protection domain called PD1 on the ROBO cluster for the VMs and volume groups.
3. Create an out-of-band snapshot (S1) for the protection domain on ROBO with no expiration.

4. Create an empty protection domain called PD1 (same name used in step 2) on the seed cluster.
5. Deactivate PD1 on the seed cluster.
6. Create remote sites on the ROBO cluster and the seed cluster.
7. Retrieve snapshot S1 from the ROBO cluster to the seed cluster (using Prism Element on the seed cluster).
8. Ship the seed cluster to the datacenter.
9. Re-IP the seed cluster.
10. Create remote sites on the ROBO cluster and on the datacenter main cluster (DC1).
11. Create PD1 (same name used in steps 2 and 4) on DC1.
12. Deactivate PD1 on DC1.
13. Retrieve S1 from the seed cluster to DC1 (using Prism on DC1). Prism generates an alert here, but though it appears to be a full data replication, the seed cluster transferred metadata information only.
14. Create remote sites on DC1 and the ROBO cluster.
15. Set up a replication schedule for PD1 on the ROBO cluster in Prism.
16. Once the first scheduled replication finishes, you can delete snapshot S1 to reclaim space.

Note the following requirements and best practices for seeding:

- Don't enable deduplication on the containers on any of the sites. You can enable it after seeding finishes.
- The seed cluster can use any hypervisor.

4. Operating Nutanix in a Multisite Environment

Once your cluster is running, you can see how the Nutanix commitment to one-click upgrades and maintenance frees IT to focus on the applications that make money for your enterprise. The following sections describe best practices for operating Nutanix in a multisite environment.

Prism Central Management

When dealing with multiple sites, a naming standard is useful. However, naming standards seldom last because of their rigidity, human error, and business changes. To add more management flexibility, Prism Central's entity explorer lets you tag VMs with one or more labels. When the entity explorer displays results, it represents your labels with a symbol; it also offers labels as an additional method for filtering results. You can use labels to tag VMs that belong to a single application, business owner, or customer.

In the following figure, we created the scada label to track supervisory control and data acquisition (SCADA) machines at remote sites. Support staff can search for the label to quickly look at the relevant information and drill down for more detail.

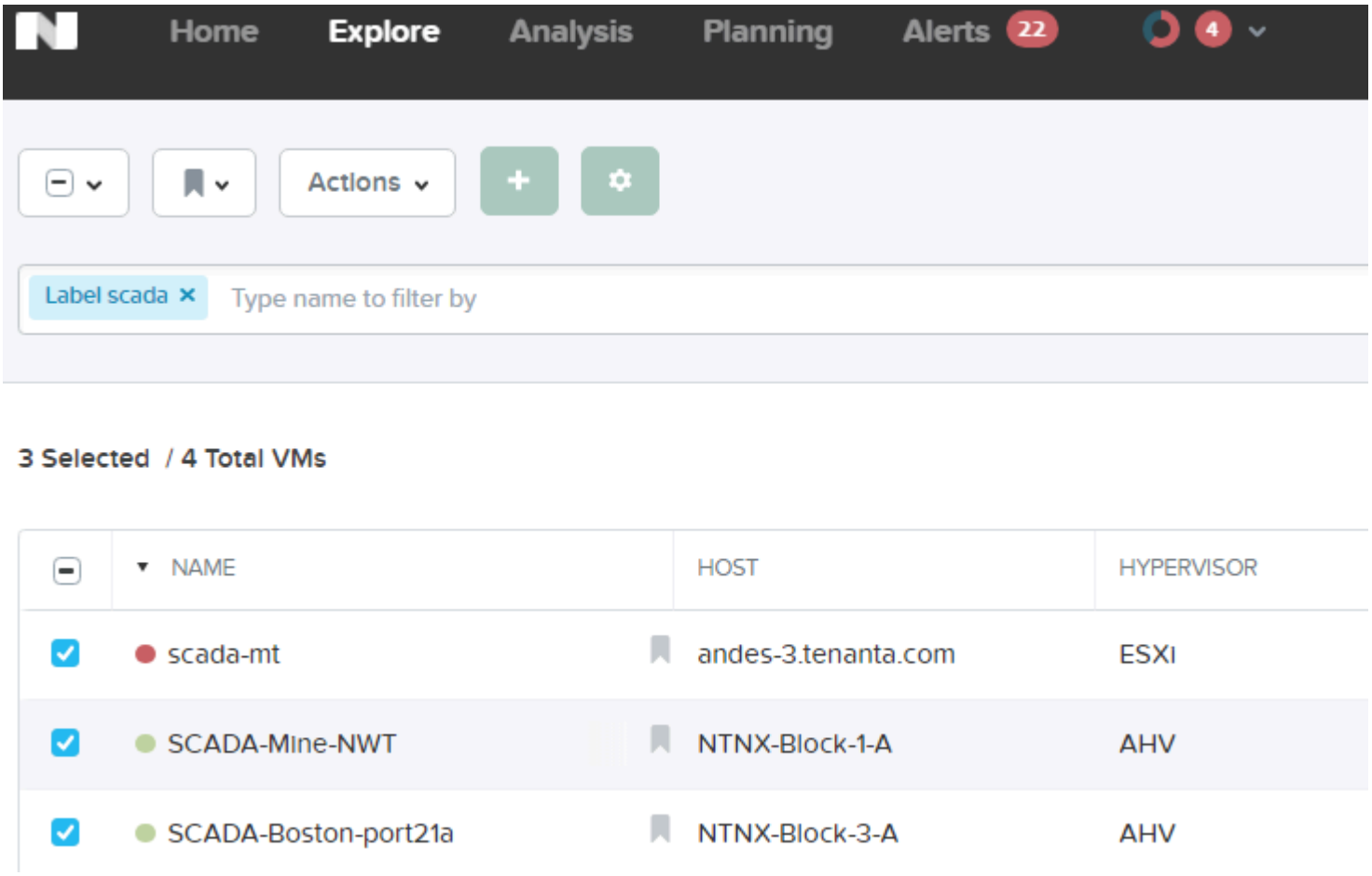

Figure 2: Using Labels for Management

Prism Central also lets you tag a cluster, which is very helpful for ROBO environments because Prism manages at site-level granularity.

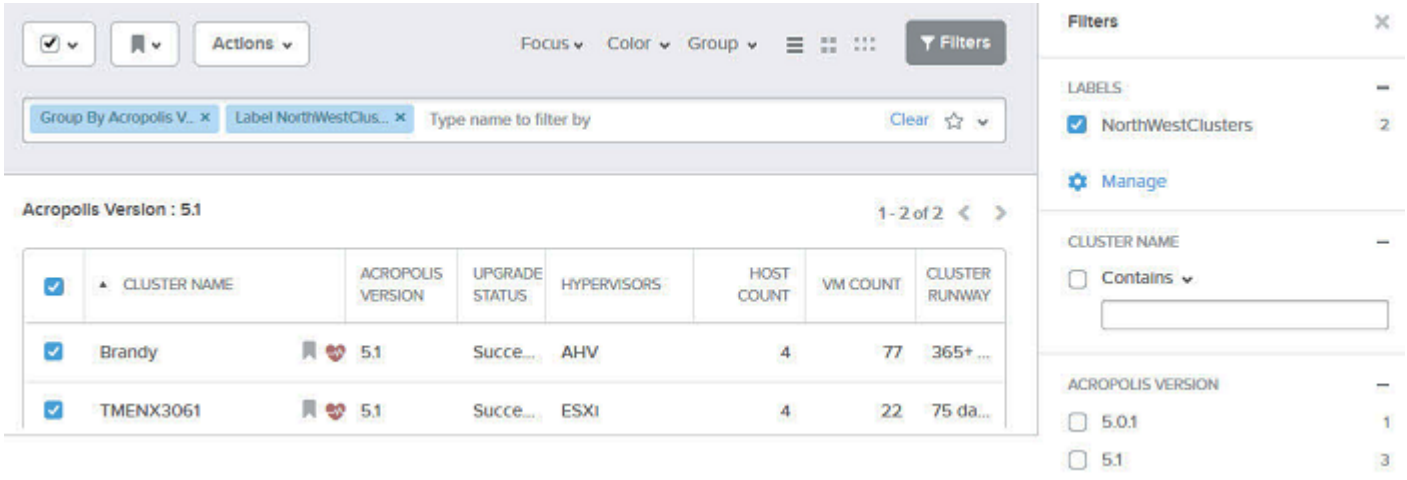

Figure 3: Grouping Clusters Using Labels

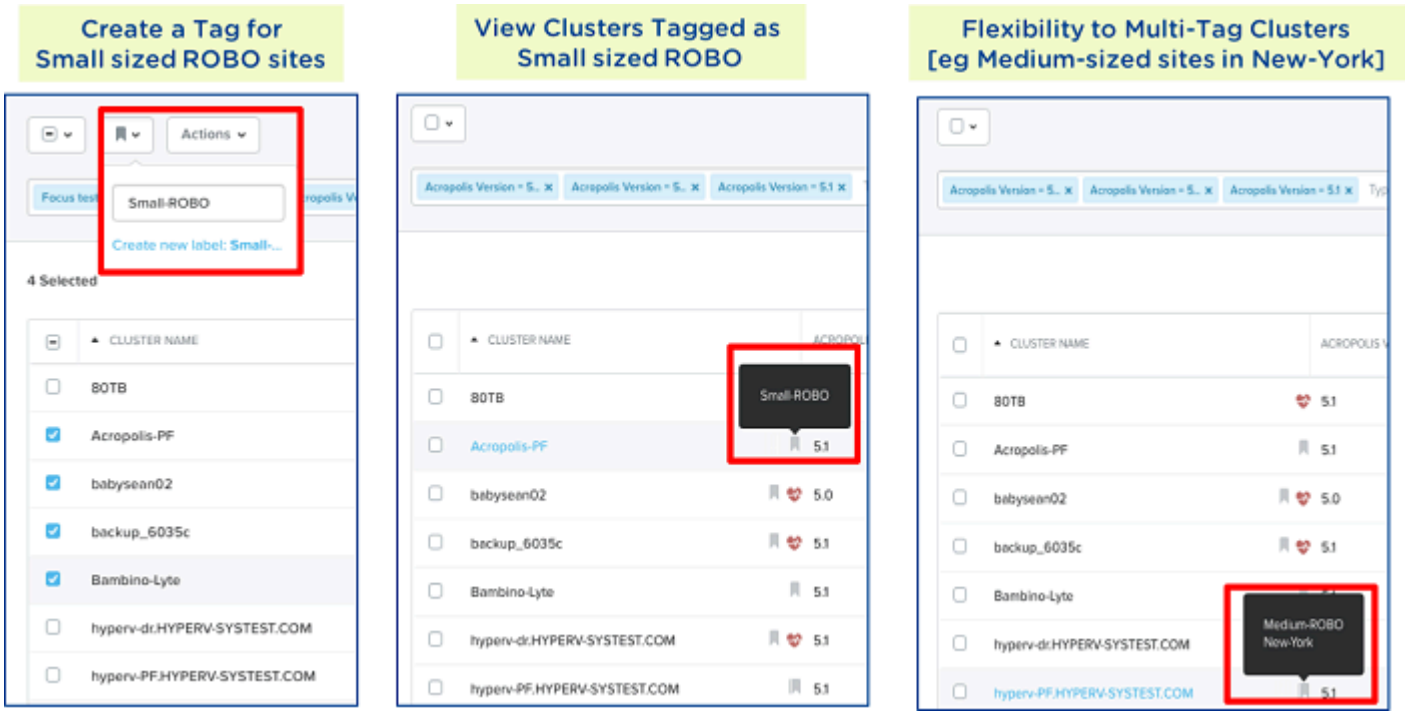

Figure 4: Tagging Clusters

You can use entity explorer in Prism to perform operations or actions on multiple entities at the same time. For example, as the figure above illustrates, you can specify tags such as medium ROBO sites in New York and then run the upgrade task with a single click.

Upgrade Modes

For environment-specific service-level agreements (SLAs), there are two upgrade modes: simultaneous mode or staggered mode.

Simultaneous Mode

Simultaneous mode is important when you're short on time—for example, when performing a critical update or a security update that you must push to all ROBO sites and clusters quickly. Simultaneous mode upgrades all the clusters immediately, in parallel.

Staggered Mode

Staggered mode allows rolling, sequential upgrades of ROBO sites as a batch job, without any manual intervention—one site upgrades only after the previous site has upgraded successfully. This feature is advantageous because it limits exposure to an issue (if one emerges) to one site rather than multiple sites. This safeguard is especially valuable for centralized administrators and others managing multiple ROBO sites. Staggered mode also lets administrators choose a custom order for the upgrade sequence.

Upgrade Best Practices

While administrators often forget remote sites after they're deployed, Prism makes it easy to keep them up to date with one-click upgrades. The following best practices help ensure that you meet your maintenance window:

- If WAN links are congested, predownload your upgrade packages near the end of your business day.
- Perform preupgrade checks before attempting the upgrade.

When running preupgrade checks, you have two options:

- Run the checks from the Cluster Health area in the UI.

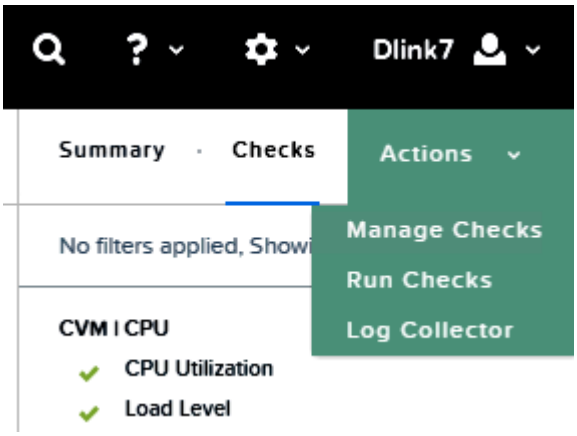

Figure 5: Cluster Health

- Log on to a CVM and run Nutanix Cluster Check (NCC) from the Nutanix command-line interface (nCLI):

```
nutanix@cvm$ ncc health_checks run_all
```

If the check reports a status other than PASS, resolve the reported issues before you proceed. If you can't resolve the issues, contact Nutanix Support for assistance.

If you need to upgrade multiple clusters, use Prism Central.

> Note: Create cluster labels for clusters in similar business units. If you need to meet an upgrade window, you can upgrade all the selected clusters to run in parallel. We recommend running one upgrade first before continuing to all the clusters.

Disaster Recovery and Backup

Nutanix allows administrators to set up remote sites and select whether they use those remote sites for simple backup or for both backup and disaster recovery.

Remote sites are a logical construct. Admins must first configure any AHV cluster—either physical or cloud-based—that functions as the snapshot destination and as a remote site from the perspective of the source cluster. Similarly, on this secondary cluster, configure the primary cluster as a remote site before snapshots from the secondary cluster start replicating to it.

Configuring backup on Nutanix lets an organization use its remote site as a replication target. You can back up data to this site and retrieve snapshots from it to restore locally, but you can't enable failover protection (running failover VMs directly from the remote site). Backup also supports using multiple hypervisors. For example, an enterprise might have ESXi in the main datacenter but use Hyper-V at a remote location. With the backup option configured, the Hyper-V cluster could use storage on the ESXi cluster for backup. Using this method, Nutanix can also back up to AWS from Hyper-V or ESXi.

Configuring the disaster recovery option allows you to use the remote site as both a backup target and a source for dynamic recovery. In this arrangement, failover VMs can run directly from the remote site. Nutanix provides cross-hypervisor disaster recovery between ESXi and AHV clusters. Hyper-V clusters can only provide disaster recovery to other Hyper-V-based clusters.

For data replication to succeed, configure forward (DNS A) and reverse (DNS PTR) DNS entries for each ESXi management host on the DNS servers that the Nutanix cluster uses.

Remote Site Setup

There are several options available when you set up a remote site. Protection domains inherit all remote site properties during replication.

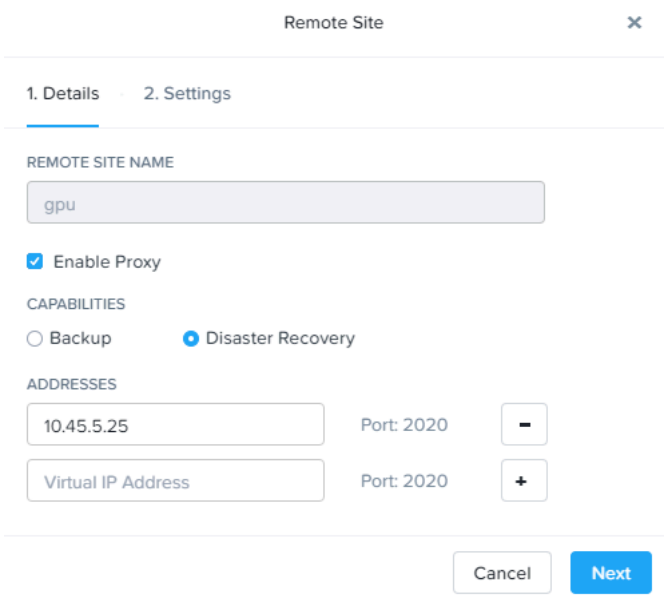

Figure 6: Setup Options for a Remote Site

Remote Site Address

Use the cluster virtual IP address as the address for the remote site. The cluster virtual IP is highly available, as it creates a virtual IP address for all the virtual storage controllers. You can configure the external cluster IP in the Prism UI under Cluster Details.

We also recommend that you try to keep both sites at the same AHV version. If both sites require compression, both must have the compression feature licensed and enabled.

Enable Proxy

The enable proxy option redirects all egress remote replication traffic through one node. This remote site proxy is different from the Prism proxy. When you select Enable Proxy, replication traffic goes to the remote site proxy, which forwards it to other nodes in the cluster. This arrangement significantly reduces the number of firewall rules you need to set up and maintain.

It is best practice to use the remote site proxy with the external address.

Capabilities

The disaster recovery option requires that both sites either support cross-hypervisor disaster recovery or have the same hypervisor. Today, Nutanix supports only ESXi and AHV for cross-hypervisor disaster recovery. When you use the backup option, the sites can use different hypervisors, but you can't restore VMs on the remote side. The backup option also works when you back up to AWS and Azure.

Maximum Bandwidth

Maximum bandwidth throttles traffic between sites when no network device can limit replication traffic. The maximum bandwidth option allows different settings throughout the day, so you can assign a maximum bandwidth policy when your sites are busy with production data and disable the policy when they're less busy. Maximum bandwidth doesn't imply maximum observed throughput. If you plan to replicate data during the day, create separate policies for business hours to avoid flooding the outbound network connection.

> Note: When talking with your networking teams, be sure to note that this setting is in megabytes per second (MBps), not megabits per second (Mbps).

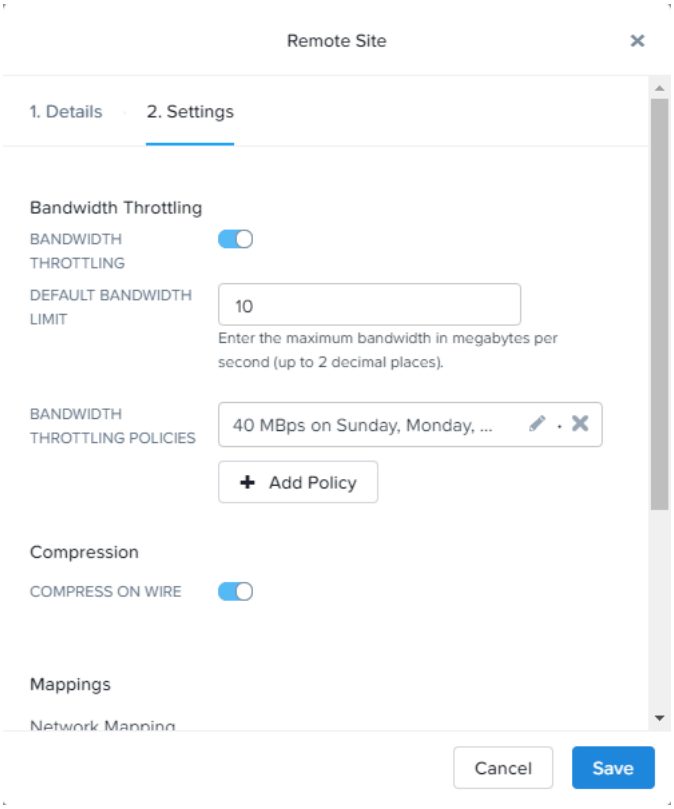

Figure 7: Maximum Bandwidth

Remote Container

vStore name mapping identifies the container on the remote cluster used as the replication target. When you establish the vStore mapping, we recommend creating a new, separate remote container with no VMs running on it on the remote side. This configuration allows the hypervisor administrator to quickly distinguish failed-over VMs and apply policies on the remote side in case of a failover.

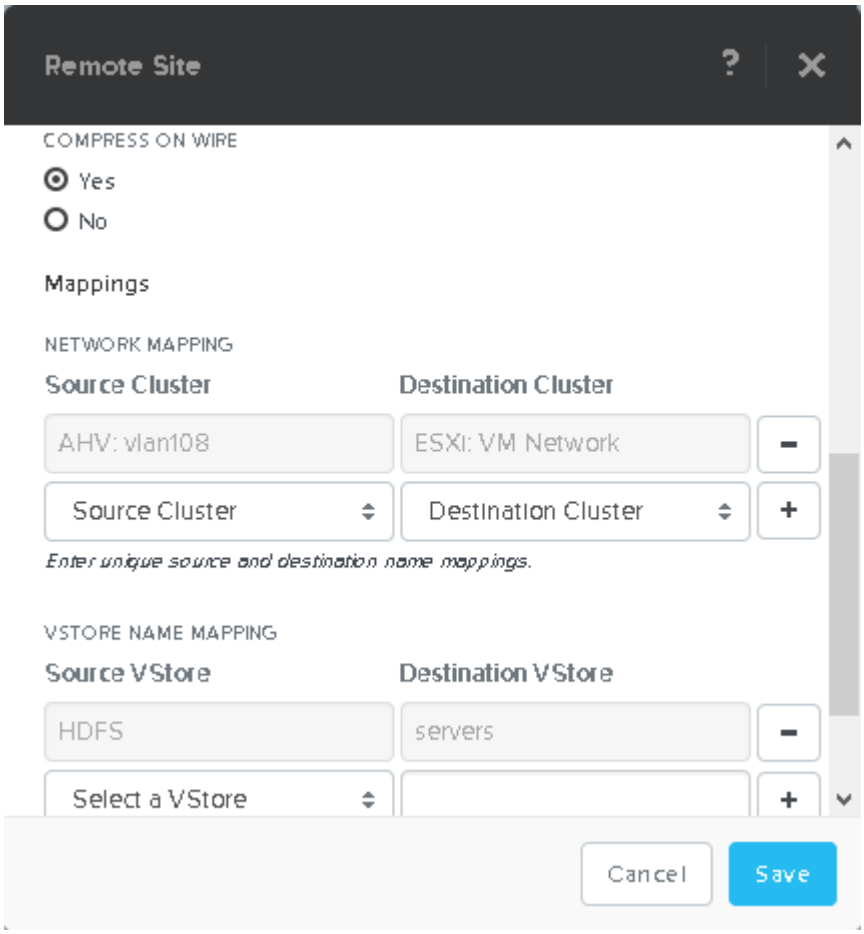

Figure 8: vStore and Container Mappings for Replication

The following are best practices for using remote containers:

- Create a new remote container as the target for the vStore mapping.
- If multiple clusters are backing up to one destination cluster, use only one destination container if the source containers have similar advanced settings.
- Enable compression if licensing permits.
- If the aggregate incoming bandwidth required to maintain the current change rate is less than 125 Mbps per node, we recommend skipping the performance tier. This setting saves flash for other workloads while also saving on SSD write endurance. For example, with a 20-node cluster, the

HDD tier could serve 2,500 Mbps throughput. To skip the performance tier, use the following command from the nCLI:

```
ncli ctr edit sequential-io-priority-order=DAS-SATA,SSD-SATA,SSD-PCIe
 name=<container-name>
```

You can reverse this command at any time.

Network Mapping

AHV supports network mapping for disaster recovery migrations moving to and from AHV. When you delete or change the network attached to a VM specified in the network map, modify the network map accordingly.

Scheduling

Make the snapshot schedule the same as your desired RPO. In practical terms, the RPO determines how much data you can afford to lose in the event of a failure caused by a hardware, human, or environmental issue. Taking a snapshot every 60 minutes for a server that changes infrequently or when you don't need a low RPO takes up resources that could benefit more critical services.

Set the RPO from the local site. If you schedule a snapshot every hour, bandwidth and available space at the remote site determine if you can achieve the RPO. In constrained environments, limited bandwidth may cause the replication to take longer than the one-hour RPO, increasing the RPO. We list guidelines for sizing bandwidth and capacity to avoid this scenario later in this document.

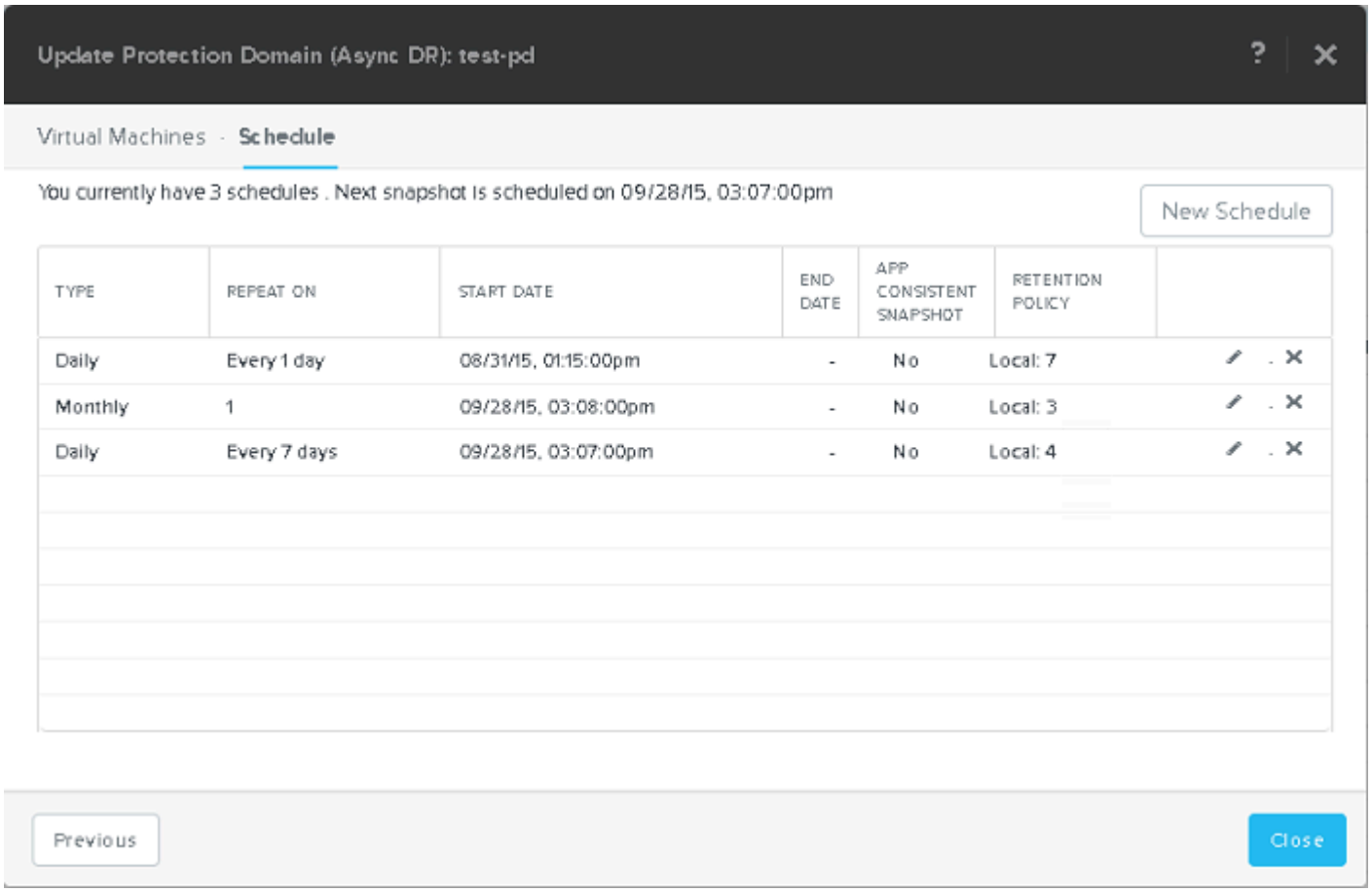

Figure 9: Multiple Schedules for a Protection Domain

You can create multiple schedules for a protection domain, and you can have multiple protection domains. The figure above shows seven daily snapshots, four weekly snapshots, and three monthly snapshots to cover a three-month retention policy. This policy is more efficient for managing metadata on the cluster than a daily snapshot with a 180-day retention policy.

The following are the best practices for scheduling:

- Stagger replication schedules across protection domains to spread out replication impact on performance and bandwidth. If you have a protection domain that starts every hour, stagger the protection domains by half of the most commonly used RPO.

- Configure snapshot schedules to retain the lowest number of snapshots while still meeting the retention policy, as shown in the previous figure.

Remote snapshots expire based on how many snapshots there are and how frequently you take them. For example, if you take daily snapshots and keep a maximum of five, on the sixth day the first snapshot expires. At that point, you can't recover from the first snapshot because the system deletes it automatically.

In case of a prolonged network outage, Nutanix always retains the last snapshot to ensure that you never lose all the snapshots. You can modify the retention schedule from the nCLI by changing the `min-snap-retention-count`. This value ensures that you retain at least the specified number of snapshots, even if all the snapshots have reached the expiry time. This setting works at the protection domain level.

Sizing Storage Space

Local Snapshots

To size storage for local snapshots at the remote site, you need to account for the rate of change in your environment and how long you plan to keep your snapshots on the cluster. Reduced snapshot frequency may increase the rate of change due to the greater chance of common blocks changing before the next snapshot.

To find the space needed to meet your RPO, you can use the following formula. As you decrease the RPO for asynchronous replication, you may need to account for an increased rate of transformed garbage. Transformed garbage is space the system allocated for I/O optimization or assigned but that no longer has associated metadata. If you're replicating only once each day, you can remove `(change rate per frequency * # of snapshots in a full Curator scan * 0.1)` from the following formula. A full Curator scan runs every six hours.

```
snapshot reserve = (frequency of snapshots * full change rate per frequency) +
 (change rate per frequency * # of snapshots in a Curator scan * 0.1)
```

You can look at your backups and compare their incremental differences to find the change rate.

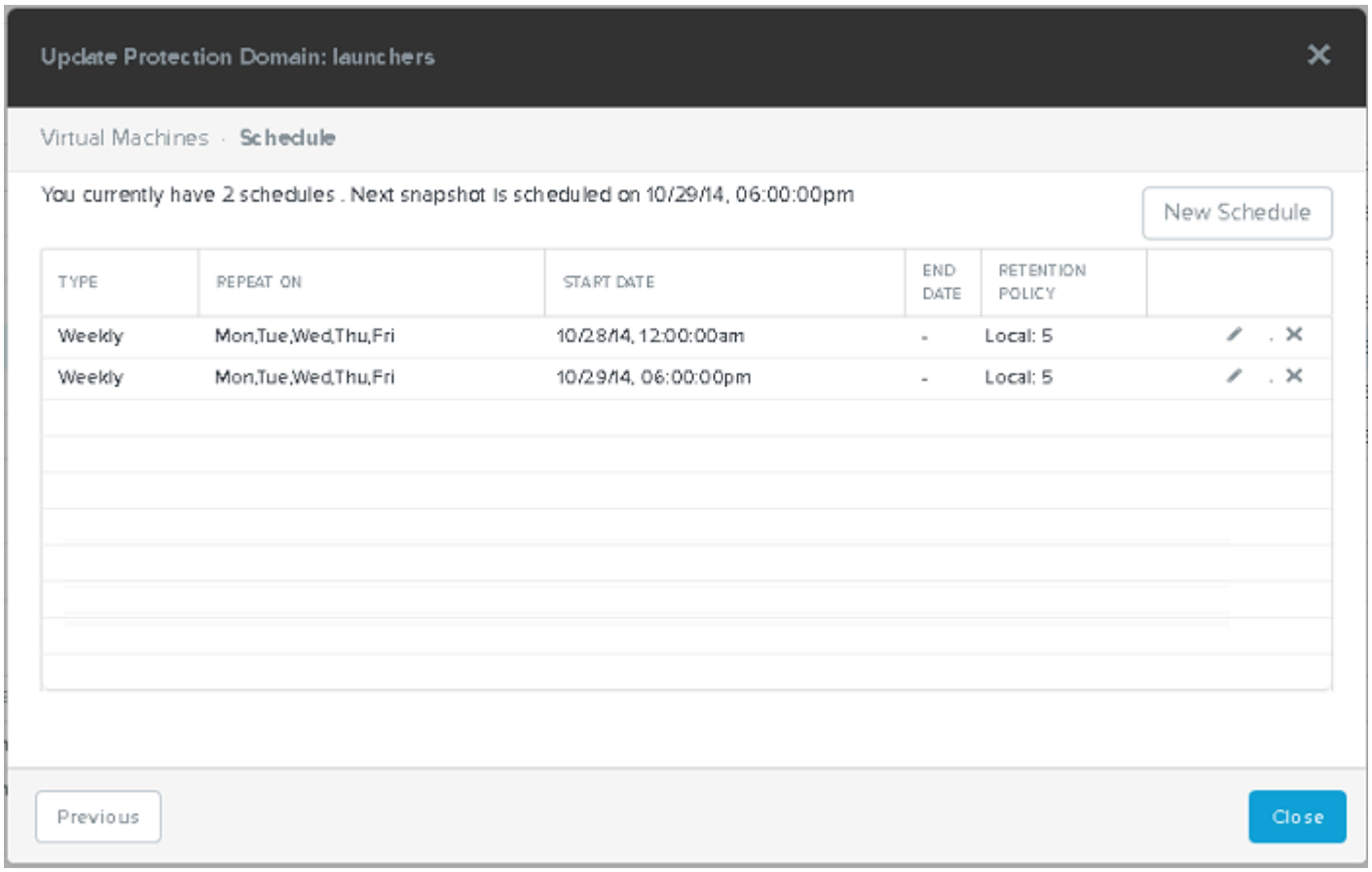

Figure 10: Example Snapshot Schedule

Using the local snapshot reserve formula presented above and assuming for demonstration purposes that the change rate is 35 GB of data every six hours and that we keep 10 snapshots, we get a 363 GB snapshot reserve:

```
snapshot reserve = (frequency of snapshots * change rate per frequency) +
(change rate per frequency * # of snapshots in a full curator scan * 0.1)
= (10 * 35,980 MB) + (35,980 MB * 1 * 0.1)
= 359,800 + (35,980 * 1 * 0.1)
= 359,800 + 3,598
= 363,398 MB
= 363 GB
```

Remote Snapshots

Remote snapshots use the same process, but you must include the first full copy of the protection domain plus delta changes based on the set schedule.


```
snapshot reserve = (frequency of snapshots * change rate per frequency) + (change
 rate per frequency * # of snapshots in a full Curator scan * 0.2) + total size
 of the source protection domain
```

To minimize the storage space you need at the remote site, use 130 percent of the protection domain as an average.

Bandwidth

You must have enough available bandwidth to keep up with the replication schedule. If you are still replicating when the next snapshot is scheduled, the current replication job finishes first. The newest outstanding snapshot then starts to get the newest data to the remote side first. To help replication run faster when you have limited bandwidth, you can seed data on a secondary cluster at the primary site before you ship that cluster to the remote site.

To figure out the needed throughput, you must know your RPO. If you set the RPO to one hour, you must be able to replicate the changes within that time.

Assuming you know your change rate based on incremental backups or local snapshots, you can calculate the bandwidth you need. The next example uses a 15 GB change rate and a one-hour RPO. We didn't use deduplication in the calculation, partly so the dedupe savings could serve as a buffer in the overall calculation and partly because the one-time cost for deduped data going over the wire has less impact once the data is present at the remote site. We assumed an average of 30 percent bandwidth savings for compression on the wire.

```
Bandwidth needed = (RPO change rate * (1 – compression on wire savings %)) / RPO
Example:
(15 GB * (1 – 0.3)) / 3,600 s
(15 GB * 0.7) / 3,600 s
10.5 GB / 3,600 s
(10.5 * 1,000 MB) / 3,600 s (changing the unit to MBps)
(10,500 MB) / 3,600 s
10,500 MB / 3,600 = 2.92 MBps
Bandwidth needed = 23.33 Mbps
```

You can perform the calculation online using the WolframAlpha computational knowledge engine.

If you have problems meeting your replication schedule, either increase your bandwidth or increase your RPO. To allow more RPO flexibility, you can run different schedules on the same protection domain. For example, set one daily

replication schedule and create a separate schedule to take local snapshots every few hours.

Single-Node Backup Target

Nutanix offers the ability to use an NX-1175S or NX-8155 appliance as a single-node backup target for an existing Nutanix cluster. Because this target has different resources than the original cluster, you primarily use it to provide backup for a small set of VMs. This utility gives small and medium-size business (SMB) and ROBO customers a fully integrated backup option.

The following are best practices for using a single-node backup target:

- All protection domains combined should be under 30 VMs.
- To speed up restores, limit the number of VMs in each protection domain.
- Limit backup retention to a three-month policy.
  - We recommend seven daily, four weekly, and three monthly backups.
- Map a single-node backup target to only one physical cluster.
- Set the snapshot schedule to six hours or more.
- Turn off deduplication.

One- and Two-Node Clusters

Nutanix one- and two-node clusters follow the same best practices as the single-node backup target because of limited resources on the nodes. The only difference for one- and two-node clusters is that all protection domains should have only five VMs per node.

Cloud Connect

Because the CVM running in AWS and Azure has limited SSD space, we recommend the following best practices when sizing:

- Limit each protection domain to one VM to speed up restores. This approach also saves money, as it limits the amount of data going across the WAN.
- The RPO shouldn't be less than four hours.

- Turn off deduplication.
- Use Cloud Connect to protect workloads that have an average change rate of less than 0.5 percent.

5. Failure Scenarios

Companies need to resolve problems at remote branches as quickly as possible, but some branches are harder to access. Accordingly, you need a system that can self-heal and that has break-fix procedures that less technically skilled staff can manage. In the following section, we cover the most basic remote branch failure scenarios: losing a hard drive and losing a node in a three-node cluster.

Failed Hard Drive

The Hades service, which runs in the CVM, enables AOS storage to detect accumulating disk errors (for example, I/O errors or bad sectors). It simplifies the break-fix procedures for disks and automates several tasks that previously required manual user actions. Hades helps fix failing devices before they become unrecoverable.

Nutanix also has a unified component called Stargate that manages receiving and processing data. The system sends all read and write requests to the Stargate process running on the node. Stargate marks a disk offline if there are three consecutive I/O errors when writing an extent, taking the disks offline well before a catastrophic disk failure, as increasing I/O errors happen first as a disk drive ages. Hades then automatically removes the disk from the data path and runs smartctl checks against it. If the checks pass, Hades marks the disk online and returns it to service. If the smartctl checks fail or if Stargate marks a disk offline three times in one hour (regardless of the smartctl check results), Hades removes the disk from the cluster, and the following sequence occurs:

1. Hades marks the disk for removal in the cluster's Zeus configuration.
2. The system unmounts the disk.
3. The disk's red LED turns on to provide a visual indication of the failure.
4. The cluster automatically begins to create new replicas of the data stored on the disk.

The system marks the disk as tombstoned to prevent the cluster from using it again without manual intervention.

Marking a disk offline triggers an alert, and the system immediately removes the offline disk from the storage pool. Curator then identifies all extents stored on the failed disk and the distributed storage fabric and makes additional copies of the associated replicas to restore the desired replication factor. By the time administrators learn of the disk failure from Prism Element, SNMP trap, or email notification, distributed storage is already healing the cluster.

The distributed storage data rebuild architecture provides faster rebuild times than traditional RAID data protection schemes, with no performance impact to workloads. RAID groups or sets usually have a small number of disks. When a RAID set performs a rebuild operation, it typically selects one disk as the rebuild target. The other disks in the RAID set must divert enough resources to quickly rebuild the data on the failed disk. This process can lead to performance penalties for workloads served by the degraded RAID set. Distributed storage allocates remote copies found on any individual disk to the remaining disks in the Nutanix cluster. As a result, distributed storage replication operations are background processes with no impact to cluster operations or performance. Moreover, AOS distributed storage accesses all disks in the cluster at any given time as a single, unified pool of storage resources.

Additionally, every node in the cluster participates in replication, which means that as the cluster size grows, disk failure recovery time decreases. Because distributed storage allocates the data needed to rebuild a disk throughout the cluster, more disks contribute to the rebuild process and accelerate the additional replication of affected extents.

Nutanix maintains consistent performance during the rebuild operations. For hybrid systems, Nutanix rebuilds cold data to cold data so that large hard drives do not flood the SSD caches. For all-flash systems, Nutanix protects user I/O by implementing quality of service for back-end I/O.

In addition to a many-to-many rebuild approach to data availability, the distributed storage data rebuild architecture ensures that all healthy disks are always available. Unlike most traditional storage arrays, Nutanix clusters don't need hot spare or standby drives. Because data can rebuild to any of

the remaining healthy disks, you don't need to reserve physical resources for failures. Once healed, you can lose the next drive or node.

Failed Node

A Nutanix cluster must have at least three nodes. Minimum configuration clusters provide the same protections as larger clusters, and a three-node cluster can continue normally after a node failure. The cluster starts to rebuild all the user data right away because of the separation of cluster data and metadata. However, one condition applies to three-node clusters: when a node failure occurs in a three-node cluster, you can't dynamically remove the failed node from the cluster. The cluster continues running without interruption on two healthy nodes and one failed node, but you can't remove the failed node until there are three or more healthy nodes. Therefore, the cluster isn't fully protected until you either fix the problem with the existing node or add a new node to the cluster and remove the failed one. This condition doesn't apply to clusters with four or more nodes, where you can dynamically remove the failed node to bring the cluster back to full health. The newly configured cluster still has at least three nodes, so the cluster is fully protected. You can then replace the failed hardware for that node as needed and add the node back into the cluster as a new node.

6. Appendix

Best Practices Checklist

Cluster Storage Capacity

- Size for n + 1 nodes for usable space, including an additional five percent for system overhead.

Prism Central

Network

- Prism Central uses TCP port 9440 to communicate with the CVMs in a Nutanix cluster. If your network or servers have a firewall enabled, open port 9440 between the CVMs and the Prism Central VM to allow access.
- Always deploy with DNS. Prism Central occasionally does a request on itself; if it can't resolve the DNS name, some cluster statistics may not be present.
- If you use LDAP or LDAPS for authentication, open port 3268 (for LDAP) or 3269 (for LDAPS) on the firewall.

Initial Installation and Sizing

- Small environments: For fewer than 2,500 VMs, size Prism Central with 6 vCPU, 26 GB of memory, and 500 GiB of storage.
- Large environments: For up to 12,500 VMs, size Prism Central with 19 vCPU, 44 GB of memory, and 2,500 GiB of storage.
- If you install on Hyper-V, use the SCVMM library on the same cluster to enable fast copy. Fast copy improves deployment time.

Statistics

- Prism Central keeps 13 weeks of raw metrics and 53 weeks of hourly metrics.

- Nutanix Support can help you keep statistics over a longer period if needed. However, once you change the retention time, only stats written after the change have the new retention time.

Cluster Registration and Licensing

- Prism Central doesn't manage Cloud Connect (cloud-based) clusters.
- Each node registered to and managed by Prism Pro requires you to apply a Prism Pro license through the Prism Central web console. For example, if you register and manage 10 Nutanix nodes (regardless of the individual node or cluster license level), you need to apply 10 Prism Pro licenses through the Prism Central web console.

Disaster Recovery: General

Protection Domains

- Protection domain names must be unique across sites.
- Group VMs with similar RPO requirements.
- Each protection domain can have a maximum of 200 VMs.
- VMware Site Recovery Manager and Metro Availability protection domains can only have 50 VMs.
- Remove unused protection domains to reclaim space.
- If you must activate a protection domain rather than migrate it, deactivate the old primary protection domain when the site comes back up.

Consistency Groups

- Keep consistency groups as small as possible. Keep dependent applications in one consistency group to ensure that the system recovers them in a consistent state and timely manner (for example, App and DB).
- Each consistency group using application-consistent snapshots can contain only one VM.

Disaster Recovery and Backup

- Configure forward (DNS A) and reverse (DNS PTR) DNS entries for each ESXi management host on the DNS servers used by the Nutanix cluster.

Disaster Recovery: ROBO

Remote Sites

- Use the external cluster IP as the address for the remote site.
- Use the remote site proxy to limit firewall rules.
- Use maximum bandwidth to limit replication traffic.
- When you activate protection domains, use intelligent placement for Hyper-V and DRS for ESXi clusters on the remote site. Intelligent placement evenly spreads out the VMs upon start during a failover. AHV starts VMs uniformly at start time.

Remote Containers

- Create a new remote container as the target for the vStore mapping.
- When you back up many clusters to one destination cluster, use only one destination container if the source containers have similar advanced settings.
- Enable compression if licensing permits.
- If you can satisfy the aggregate incoming bandwidth required to maintain the current change rate from the hard drive tier, skip the performance tier to save flash capacity and increase device longevity.

Network Mapping

- Whenever you delete or change the network attached to a VM specified in the network map, modify the network map accordingly.

Scheduling

- To spread out the impact replication has on performance and bandwidth, stagger replication schedules across protection domains. If you have a protection domain starting each hour, stagger the protection domains by half of the most commonly used RPO.

- Configure snapshot schedules to retain the fewest snapshots while still meeting the retention policy.
- Configure the CVM external IP address.
- Obtain the mobility driver from Nutanix Guest Tools.
- Avoid migrating VMs with delta disks (hypervisor-based snapshots) or SATA disks.
- Ensure that protected VMs have an empty IDE CD-ROM attached.
- Run AOS 5.10 or later in both clusters.
- Ensure that network mapping is complete.

Sizing

- Use the application's change rate to size local and remote snapshot usage.

Bandwidth

- Seed locally for replication if WAN bandwidth is limited.
- Set a high initial retention time for the first replication when you seed.

Single-Node Backup

- Keep all protection domains, combined, under 30 VMs total.
- Limit backup retention to a three-month policy. We recommend seven daily, four weekly, and three monthly backups as a policy.
- Map an NX-1155 to one physical cluster only.
- Set the snapshot schedule to at least six hours.
- Turn off deduplication.

One- and Two-Node Backup

- Keep all protection domains under five VMs per node.
- Limit backup retention to a three-month policy. We recommend seven daily, four weekly, and three monthly backups as a policy.

- Map each backup target to one cluster only.
- Set the snapshot schedule to six hours or longer.
- Turn off deduplication.

Cloud Connect

- Limit each protection domain to one VM to speed up restores. This approach also saves money, as it limits the amount of data going across the WAN.
- Set the RPO to a minimum of four hours.
- Turn off deduplication.
- Use Cloud Connect to protect workloads that have an average change rate of less than 0.5 percent.

About Nutanix

Nutanix is a global leader in cloud software and a pioneer in hyperconverged infrastructure solutions, making clouds invisible and freeing customers to focus on their business outcomes. Organizations around the world use Nutanix software to leverage a single platform to manage any app at any location for their hybrid multicloud environments. Learn more at www.nutanix.com or follow us on Twitter @nutanix.

List of Figures

Figure 1: Example NX-1175S Node Configuration....10
Figure 2: Using Labels for Management....17
Figure 3: Grouping Clusters Using Labels....18
Figure 4: Tagging Clusters....18
Figure 5: Cluster Health....20
Figure 6: Setup Options for a Remote Site....22
Figure 7: Maximum Bandwidth....24
Figure 8: vStore and Container Mappings for Replication....25
Figure 9: Multiple Schedules for a Protection Domain....27
Figure 10: Example Snapshot Schedule....29