v1.2 | December 2021 | TN-2068

TECH NOTE

Infrastructure Resilience

NUTANIX
YOUR ENTERPRISE CLOUD

Copyright

Contents

1. Executive Summary

We designed Nutanix from the ground up to provide best-in-class resilience that keeps applications running regardless of underlying hardware and software failures. The Nutanix software runs as a virtual storage controller (Controller VM, or CVM) on each node in a cluster, forming a distributed system. All nodes work together to aggregate individual direct-attached storage resources into a single global namespace that all hosts can use. Nutanix distributed storage manages all storage resources to preserve data and system integrity in the event of node, disk, or software failure.

Document Version History

Version Number	Published	Notes
1.0	February 2017	Original publication.
1.1	July 2019	Updated the Availability Domains for Rack-Aware Fault Tolerance section.
1.2	December 2021	Updated the Availability Domains for Rack-Aware Fault Tolerance and Metadata Protection and Reliability sections.

2. Data Protection Through Tunable Redundancy

Nutanix storage uses a distributed operations log (oplog), analogous to a fault-tolerant journal on a local file system, as a staging area to absorb incoming writes onto a fast, low-latency SSD tier. The oplog coalesces the data and writes (or drains) it to back-end storage resources (the extent store) asynchronously. The extent store is made up of extents, the variable-sized contiguous regions of a vDisk.

Nutanix storage implements a fully distributed design that prevents oplog data loss in the event of a node failure. Before a host receives any write acknowledgement, the write synchronously replicates to an oplog on one or two other adjacent nodes. All nodes participate in this replication. The host only receives acknowledgment of a successful write after replication of the data—and its associated metadata—completes. This process ensures that data exists in at least two independent locations in the cluster and is fault tolerant. This design also means that the system doesn't rely on batteries for fault tolerance, so data is safe even through a complete datacenter outage.

For data resilience, you can have two or three copies of data, representing a data replication factor of 2 or 3, respectively.

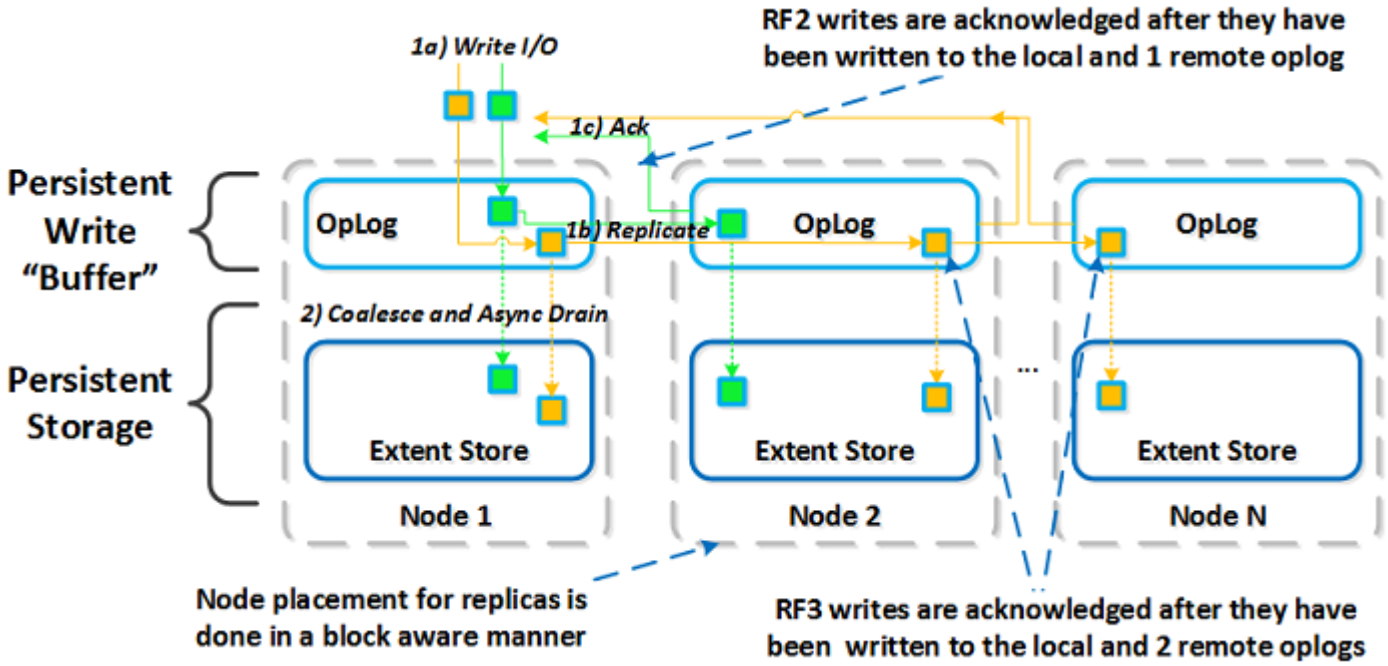

Figure 1: Tunable Redundancy

3. Data Path Redundancy

Nutanix distributed storage withstands a variety of hardware failures and builds strong redundancy into the software stack. Nutanix software processes are designed to fail fast when they encounter a serious error, a design principle that quickly restarts normal operation instead of waiting for a potentially faulty process to complete. AOS continuously monitors components and, in the event of an error, stops and restarts them so they can recover as quickly as possible rather than linger in a nonresponsive state.

As an example of this principle in action, let's look at how the storage fabric tracks the health of all CVMs in the cluster. Each host relies on its local CVM to service all storage requests. If an unrecoverable error occurs on a particular CVM, Nutanix pathing automatically reroutes requests from the host to a healthy CVM on another node, providing data path redundancy.

This redirection continues until the local CVM failure issue is resolved. Because the cluster has a global namespace and access to replicas for all the data on that node, it services requests immediately. This structure provides a high degree of fault tolerance and failover capability for all VMs in a Nutanix cluster. If the node's CVM remains unavailable for a prolonged period, data automatically replicates again to maintain the necessary replication factor.

4. Failure Recovery Is Easy

We've discussed what happens when a storage controller fails in a distributed system, but that's a fairly easy use case. The harder problem to solve is when a service just crawls along. Slow services can cause real headaches in a distributed system.

If a remote CVM isn't performing well, it can affect the acknowledgement of writes coming from other hosts. Several factors may affect performance, including:

- Significant network bandwidth reduction.
- Network packet drops.
- CPU soft lockups.
- Partially bad disks.
- Hardware issues.

Even as-yet unknown issues can affect performance, so Nutanix Engineering has developed a scoring system that uses votes to compare all running services fairly. We call the voting system degraded node detection.

Services running on each node of the cluster publish scores, or votes, for services running on other nodes. Peer health scores are computed based on various metrics like RPC latency, RPC failures or timeouts, network latency, and so on. If services running on one node consistently receive bad scores for an extended period (approximately 10 minutes), its peers convict it as a degraded node, unless the current cluster fault tolerance level is still within the accepted range.

Upgrades and break-fix actions aren't allowed while a node is in the degraded state. By default, degraded node logging is enabled, but degraded node action is disabled. You can set an action policy to take over when AOS detects a degraded node. Action policy options include no action, reboot CVM, and shut down.

Degraded node detection alerts you to potential failure situations and allows for quick recovery, preventing outages before they happen.

5. Availability Domains for Rack-Aware Fault Tolerance

Availability domains offer robust protection from catastrophic hardware failures by allowing Nutanix systems to survive losing multiple servers in a physical enclosure without data or service loss. Combining powerful data redundancy, configured at the Nutanix storage container level, with intelligent data placement, the system can automatically ensure availability of both data and data access if a physical block or rack fails. Availability domains provide greater system-level resilience without increasing the required storage capacity.

Disk and node awareness is always on. Block awareness is best effort and requires a minimum of three blocks. For rack awareness, you need at least three racks, and the administrator must define which racks the blocks should be placed in.

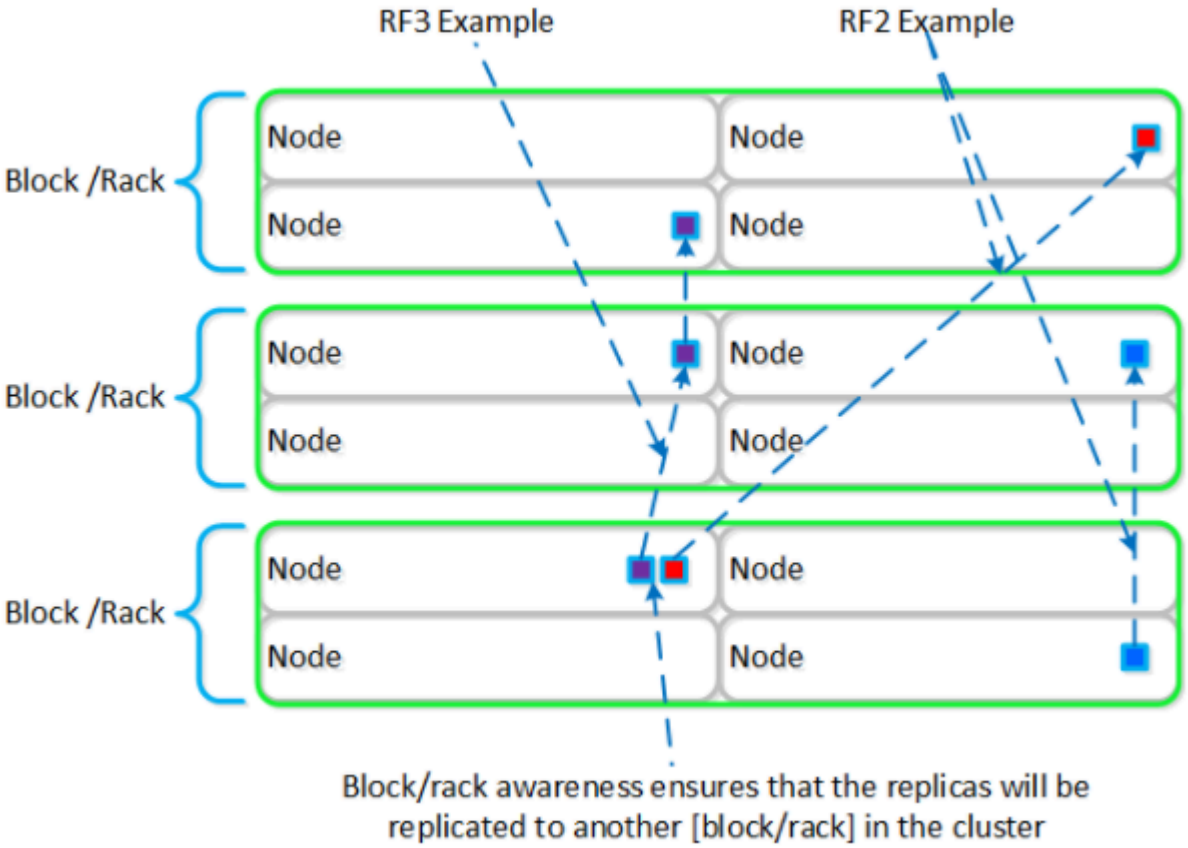

Figure 2: Block and Rack Awareness

6. Metadata Protection and Reliability

Nutanix maintains metadata in two separate databases: global and local. Global metadata is maintained in a strictly consistent, distributed, high-performance NoSQL database called Cassandra, and local metadata is maintained in a database called AES DB that's based on RocksDB. This design optimizes performance for metadata and provides a highly available and incrementally scalable platform for storage metadata, keeping it separate and protected. By default, it replicates metadata to two other nodes in the cluster; you can choose to set the system to keep up to five copies, providing significant metadata redundancy in case of a node failure.

Once metadata has been safely updated, the related extent is written to disk. As data is written to the extent store, each extent synchronously replicates to extent stores on other nodes in the cluster. This process (which is separate from the oplog described earlier) ensures that there are always multiple copies of all data, providing fault tolerance.

Additionally, the system computes a checksum on the write and includes it in the metadata record. Thus, when Nutanix storage reads an extent, it computes a new checksum and compares it with the checksum stored in the metadata. If the two checksums match, the data is deemed consistent and valid. If the checksums don't match, the system marks the extent as corrupt and fetches a replica from a remote node for all subsequent requests. The fabric then initiates a replication task to restore the data to the desired replication factor and remediate any corrupt extents.

When the system maintains three metadata copies, the three copies are mapped to three consecutive nodes in a ring. In the unlikely event that two nodes fail, as long as the nodes lost are separated by at least two other nodes, metadata availability is unaffected. Even in the rare case of two consecutive nodes failing, the system can seamlessly make the remaining copy active.

7. Preventing Errors Due to Network Partitions

Distributed storage uses the Nutanix Paxos algorithm to avoid split-brain scenarios. Paxos is a proven protocol for reaching consensus (or quorum) among several participants in a distributed system. Before any file system metadata is written to Cassandra, Paxos ensures that all nodes in the system agree on the same value. If the nodes don't reach quorum, any write operation in progress fails to eliminate potential corruption or data inconsistency.
This design protects against events such as network partitioning, in which communication between nodes may experience failure or packets may become corrupt, leading to a scenario where nodes may disagree on values. The fabric also uses timestamps to ensure that updates are applied in the proper order.

8. Proactively Resolving Bad Disk Resources

Nutanix distributed storage incorporates a process called Curator, which performs background housekeeping tasks to keep the entire cluster running smoothly. Curator's responsibilities include ensuring file system metadata consistency and combing the extent store for corrupt and under-replicated data.

Additionally, Curator scans extents in successive passes, computing each extent's checksum and comparing it with the stored metadata checksum to validate data consistency. When the checksums don't match, the storage fabric replaces the corrupted extent with a valid extent from another node. This proactive data analysis protects against data loss and identifies bad sectors you can use to flag disks that are about to fail.

9. Maintaining Availability in the Event of Disk Failure

When Nutanix storage detects an accumulation of errors for a particular disk (for example, I/O errors or bad sectors), it deploys the Hades service running on the CVM. Hades 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 uses a unified component called Stargate to manage receiving and processing data. All read and write requests go to the Stargate process running on the node holding the data involved. Once Stargate sees delays in responses to I/O requests to a disk, it marks that disk offline. Hades then automatically removes the offline disk from the data path and runs smartctl checks against it. If the checks pass, Hades marks the disk online again and returns it to service. If the Hades 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 procedure ensues:

1. The cluster's configuration marks the disk for removal.
2. The disk is unmounted.
3. The disk's red LED turns on to provide a visual indication of the failure.

The cluster automatically begins to create new replicas of any data stored on the failed disk. Stargate marks the failed disk as a tombstoned disk to prevent it from being used again without manual intervention.

An alert activates when Stargate marks a disk offline and the system immediately removes the offline disk from the storage pool. Curator identifies all extents stored on the failed disk and prompts the storage fabric to make copies of the associated replicas to restore the desired replication factor. By the time the Nutanix administrators learn of the disk failure through Prism, SNMP trap, or email notification, the storage fabric is already on its way to healing the cluster.

Compared to traditional RAID data protection schemes, the Nutanix data rebuild architecture provides faster restoration times with no performance impact to workloads supported by the cluster. RAID groups or sets are usually made of a small number of drives. Typically, when a RAID set performs a rebuild operation, it selects one disk as the rebuild target. The other disks in the RAID set must divert enough resources to rebuild the data quickly on the failed disk. Diverting resources can lead to performance penalties for workloads served by the degraded RAID set.

With its distributed design, Nutanix storage can disperse remote copies found on any individual disk among the remaining disks in the cluster. Therefore, storage replication operations can work as background processes, with no impact to cluster operations or performance. The distributed storage can access all disks in the cluster at any given time as a single, unified pool of storage resources. This architecture provides a key Nutanix advantage: when cluster size increases, the length of time needed to recover from a disk failure decreases, because every node in the cluster participates in the replication. Because the data needed to rebuild a disk is distributed throughout the cluster, more disks are involved in the rebuild process. This broad participation increases the speed at which the system can replicate affected extents.

It's important to note that Nutanix also maintains consistent performance during rebuild operations. In hybrid systems, Nutanix rebuilds cold data to cold data, so large hard drives don't flood the SSDs. In all-flash systems, Nutanix implements quality of service for back-end I/O to minimize the impact on user I/O.

In addition to a many-to-many rebuild approach to data availability, the Nutanix data rebuild architecture ensures that all healthy disks are available for use all the time. Unlike most traditional storage arrays, Nutanix clusters don't need hot-spare or standby drives. Because the system can rebuild data to any of the remaining healthy disks, you don't need to reserve physical resources for failures.

10. Reliable and Redundant Hardware

The Nutanix cloud is an integrated hardware and software appliance. Nutanix conducts rigorous validation and testing for all hardware platforms and components. By performing testing up front, Nutanix increases hardware reliability and longevity. Every Nutanix node has fully redundant components, including CPUs, memory, and power supplies. A full node failure automatically triggers a high availability event, and VMs fail over to other hosts in the cluster. Curator then migrates data to the VM's local CVM to localize I/O operations. Simultaneously, the system copies data as needed to maintain replication factor and overall availability.

11. Holistic Proactive Problem Detection

Each service running in the AOS software package has robust data consistency, failure mode detection, and recovery built into the core design. Detecting issues that may affect services, VMs, or network performance requires independent software that can investigate and identify several problem areas simultaneously to find a clear root cause. Nutanix Cluster Check (NCC) is a distributed code base that scrutinizes all aspects of cluster health. NCC is an extensible framework that runs multiple checks on the Nutanix cluster, hypervisor, VMs, and network. This sophisticated framework provides error detection that is unmatched among datacenter infrastructure solutions. Administrators can view the data NCC collects from the Prism Cluster Health page, which allows rapid visual diagnosis.

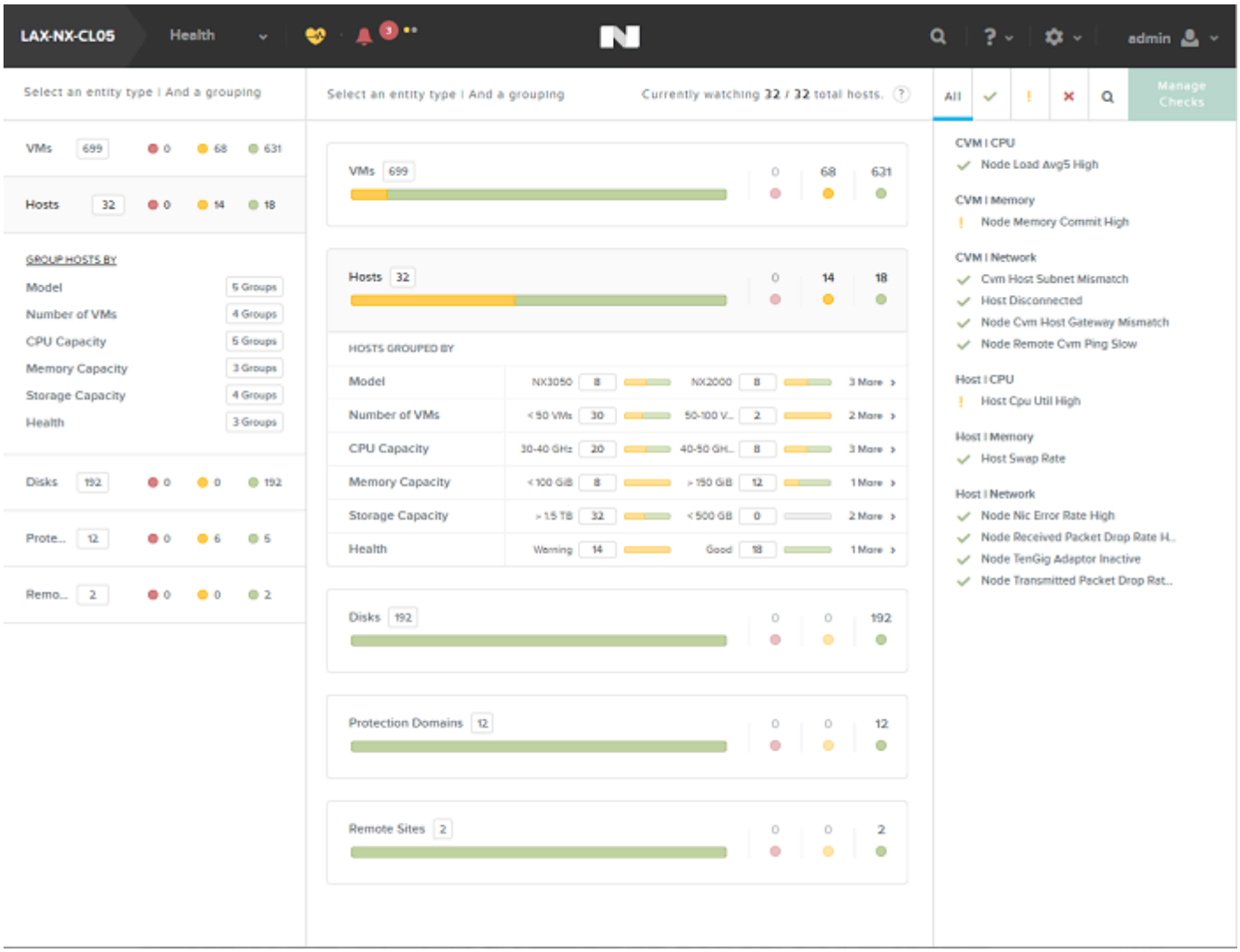

Figure 3: Prism Cluster Health Page

Administrators can also run NCC in interactive mode on an ad hoc basis, so they can collect the output and review it offline. You can send this output file to Nutanix Support for additional investigation as well. The file consists of two sections: a list of test results and the data evaluated for each test. You can diagnose problems much more quickly because NCC drastically reduces the need to manually parse log bundles offline. By the time you generate the output file, NCC has already analyzed data from multiple sources on the

local cluster in real time, so you don't need to collect as much data to send to support for analysis.

12. Conclusion

As you can see, Nutanix offers a comprehensive set of solutions across the stack to withstand hardware failures and software glitches and ensure that application availability and performance is never compromised. The best data protection helps prevent disasters from happening in the first place, and Nutanix resilience features, particularly early detection without human involvement and the ability to self-heal, provide stability. To learn more about the data protection and disaster recovery capabilities Nutanix offers, refer to the Data Protection and Disaster Recovery tech note.

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