v3.1 | May 2022 | TN-2096

TECH NOTE

Nutanix Core Performance

NUTANIX
YOUR ENTERPRISE CLOUD

Copyright

Contents

1. Executive Summary

Storage system performance has traditionally been one of the biggest pain points in deployments. In a typical SAN or NAS architecture, the storage controller can become a performance bottleneck, especially when it uses fast media like SSDs. The Nutanix architecture provides performance in a scaled-out, distributed way by providing every node in the cluster with its own Controller Virtual Machine (CVM).

With Nutanix, you don’t have to design your deployments around future performance needs. You can increase performance and scale linearly by adding nodes to the cluster. Because the Nutanix solution is completely software-defined, you only need to upgrade the software to get new performance features; you don’t need any special hardware or custom gear.

Nutanix also enables you to visualize performance through Prism. Prism provides users and administrators with a complete, detailed view of performance and event correlation data for everything from VMs to disk drives.

2. Introduction

Audience

This tech note is part of the Nutanix Solutions Library. We wrote it for architects and administrators responsible for managing performance. Readers should already be familiar with basic virtualization and performance concepts.

Purpose

In this document, we cover the following topics:

- AOS storage performance.
- Data tiering with Nutanix.
- I/O path for reads and writes.
- Data locality in the Nutanix architecture.
- Blockstore and SPDK.
- vDisk sharding in AOS.
- Snapshot and clone performance in Nutanix.

Document Version History

Version Number	Published	Notes
1.0	February 2019	Original publication.
1.1	March 2020	Content refresh.
2.0	December 2020	Updated for AOS 5.10.
3.0	May 2022	Major updates throughout.
3.1	May 2022	Updated vDisk Sharding section.

3. AOS Storage

AOS storage drives high performance for guest VMs by providing storage resources to VMs locally on the same host. This method enables the local storage controller (one per Nutanix node) to devote its resources to handling I/O requests made by VMs running on the same physical node. Other controllers running in the cluster are then free to serve I/O requests made by their own local guest VMs. This architecture contrasts with traditional storage arrays that have remote storage controllers and resources located across a network (SAN or NAS).

The Nutanix architecture has several important performance benefits. Because storage resources are local, read requests don't traverse the network, which decreases latency because it eliminates the physical network from the I/O path. As you add new Nutanix nodes to the cluster, CVMs are added at the same rate, providing predictable, scalable, linear performance. The scale-out architecture allows for predictable high-performance storage. In addition, AOS I/O path has recently been optimized to make full use of advancements in hardware technologies. The tech note will highlight some of these enhancements.

4. Effective Data Classification and Tiering

AOS monitors I/O patterns and treats various data types differently to optimize performance for each guest VM. The following diagram provides an overview of Nutanix I/O path.

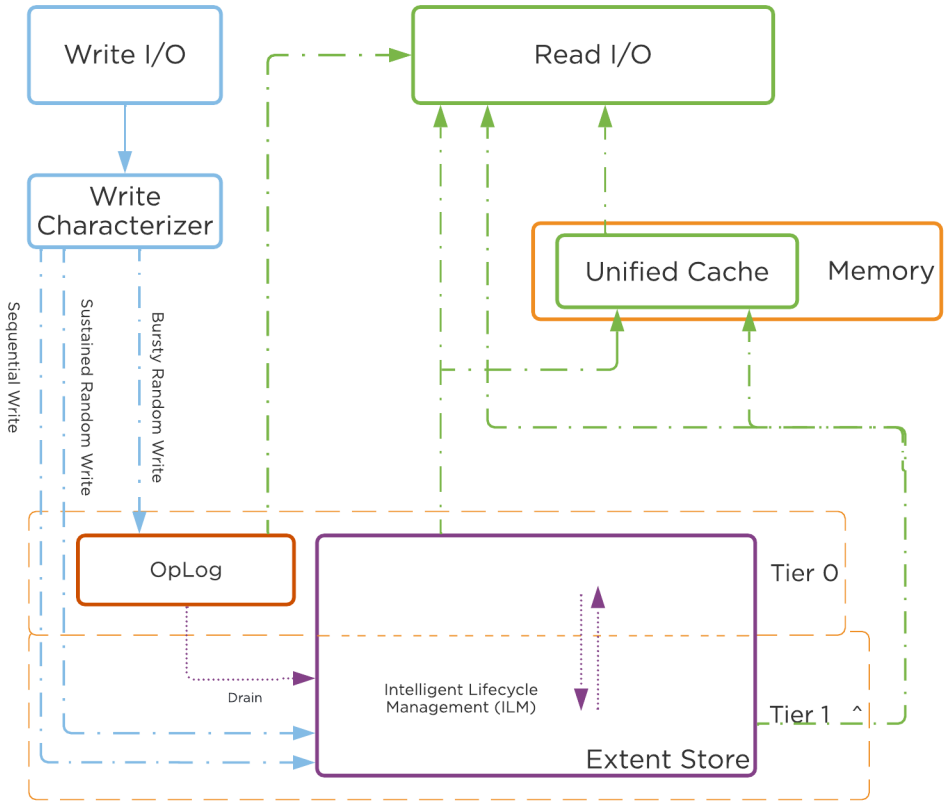

Figure 1: Nutanix I/O Path

When write I/O is received by AOS, Stargate, a process that handles I/O requests from user VMs, classifies the I/O as random or sequential and determines where that I/O will land.

Bursty Random I/O

Bursty random I/O is written to a dedicated area on the fastest tier called the oplog. The oplog is a persistent write buffer that is made up of SSDs providing the lowest latency for I/O.

During a random write burst, AOS writes data to the local oplog and simultaneously sends the data across the network to the CVMs on one or more other nodes in the cluster, where it is replicated to those nodes' oplogs. After data persists on at least two different nodes, a successful write is acknowledged back to the guest VM.

The data in oplog is then coalesced and sequentially drained to the extent store, the bulk persistent storage in AOS.

Sequential I/O

Sequential writes skip the oplog entirely and go directly to the extent store because sequential data is continuous and can be efficiently written to disks in large blocks, driving high throughput without a performance impact. Additionally, sequential data requires fewer metadata updates, so users spend less time performing metadata updates. Write I/O is classified as sequential when there is more than 1.5 MB of outstanding I/O to a vDisk of a user VM. Large write I/O with low outstanding I/O will still go to the oplog.

Autonomous Extent Store

Starting with AOS 5.10, if you meet certain conditions, the extent store, rather than the oplog, handles sustained random write workloads with a feature called Autonomous Extent Store (AES). For sustained random writes, AES writes and stores data in the extent store directly. Before, AOS stored all metadata globally, but now metadata has two parts: one stored locally to the node and another stored globally. Local metadata storage provides metadata locality in addition to data locality. Nodes don't need to know where each piece of physical metadata resides, which optimizes metadata lookups and allows you to achieve efficient sustained random write performance without using the oplog. For random write bursts, AOS still uses the oplog and drains to the extent store, using AES where possible.

For reads, I/O is serviced from the oplog or extent store depending on where the data is residing when it is requested. If there is a read request for data that is in extent store, that data is read from the extent store and is also placed into unified cache, which is a read cache stored in CVM's memory. All subsequent read requests will then be serviced from the read cache.

From a tiering perspective:

1. In all-flash node configurations (all NVMe SSD, all SATA/SAS SSD, NVMe+SATA/SAS SSD), the extent store consists of only SSD devices and no tier intelligent life-cycle management (ILM) occurs as only a single flash tier exists.
2. For hybrid, non all-flash scenarios, the flash is Tier 0 and HDD is Tier 1.
3. Oplog is always on SSDs in hybrid clusters and on both Optane and NVMe SSDs in Optane+NVMe clusters.
4. Data is moved between tiers by ILM based on access patterns.

Stargate moves data from a lower tier to a higher tier when access to that data increases. Curator, a background process that performs file system operations like tiering, rebalancing, and fixing errors with data redundancy, moves data from a higher tier to a lower tier when access to the data become less frequent. This ensures AOS makes optimal use of higher tier space and only the most frequently accessed data that needs lower latency stays there.

Optane Tiering

Intel Optane SSDs have a significantly better read latency performance than non-Optane NVMe SSDs and SATA SSDs. Starting with AOS 6.1, AOS groups Optane SSDs in a node into a separate storage tier and keeps non-Optane NVMe SSDs in a tier lower than that. ILM manages the movement of data in this hybrid flash scenario.

5. Data Locality

In a traditional shared storage environment, users access data over the network, so a VM's data stays in the same place (that is, on the central array) even if the VM migrates throughout the cluster. In AOS, every host/node has user VMs residing on it and data is stored on clusters formed by CVMs. As part of storing the data, AOS always writes one copy of the data local to where the user VM is residing and the other copy or copies on other nodes. This ensures that data is always local and accessed quickly without the need to traverse the network. This provides the fastest performance and minimizes both cross talk and network utilization.

In VM clusters, VMs migrate from host to host within the cluster throughout the day and over time in order to optimize CPU and memory resources. Because AOS storage serves data locally to guest VMs, the VM's data must follow when it moves between hosts.

After a VM completes its migration to another host, the CVM on the destination host takes ownership of the migrated VM's files (vDisks) and begins to serve all I/O requests for these vDisks. Accordingly, the writes also go to the local CVM on local storage to ensure that write performance remains as fast as it was before the VM migration event.

The Nutanix platform serves all read requests for newly written data locally and forwards previously written data to the source host's CVM. In the background, Curator dynamically moves the VM's remote data to the local Nutanix node so that all future read I/O is now performed locally and doesn't traverse the network.

6. Blockstore and SPDK

Technological advancements have made physical storage much faster with the advent of NVMe and Optane SSDs. AOS runs inside CVM, which is a Linux based virtual machine. Stargate, which processes and manages data, needs to invoke system calls to kernel filesystem and block sub-system in the CVM. Blockstore efficiently manages storage media and I/O with lower CPU overhead and latency. With Blockstore, the filesystem inside the CVM is removed from kernel completely into application user space. This allows Stargate to avoid the overhead of system calls and interrupts when accessing and storing data on storage devices, especially flash devices.

Blockstore also gives AOS the ability to leverage libraries and APIs that allow access to storage devices directly from user space. One such library is Intel Storage Performance Development Kit (SPDK). SPDK allows AOS to have zero-copy, direct parallel access to NVMe and Optane devices, unlocking its complete potential. Blockstore combined with SPDK makes the AOS data path inside CVM very lean, delivering performance benefits to applications.

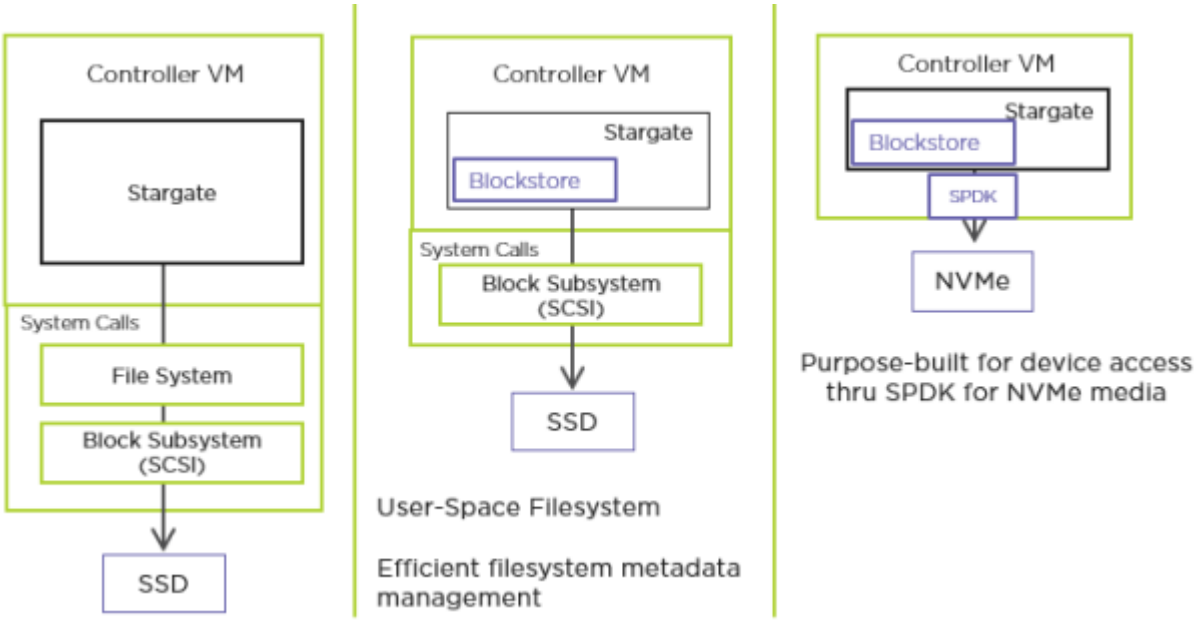

Figure 2: AOS Datapath with Blockstore and SPDK

7. vDisk Sharding

AOS architecture can deliver consistent high performance at scale to user VMs that have multiple vDisks due to the technologies mentioned above. However, there are still some traditional applications and workloads that have a big user VM with single vDisk. These VMs have not been able to leverage AOS capabilities to their full potential because, inside Stargate, I/O is processed by an entity called vDisk controller, which does single threaded access to the vDisk. With AOS 6.1, the vDisk controller was modified such that requests to the vDisk under its control are distributed across multiple shard vDisk controllers that are managed by a primary vDisk controller. The access to the single vDisk is now spread across multiple threads by shard controllers, effectively sharding the single vDisk. This results in consistent high performance being delivered by AOS for these traditional applications that use single vDisks.

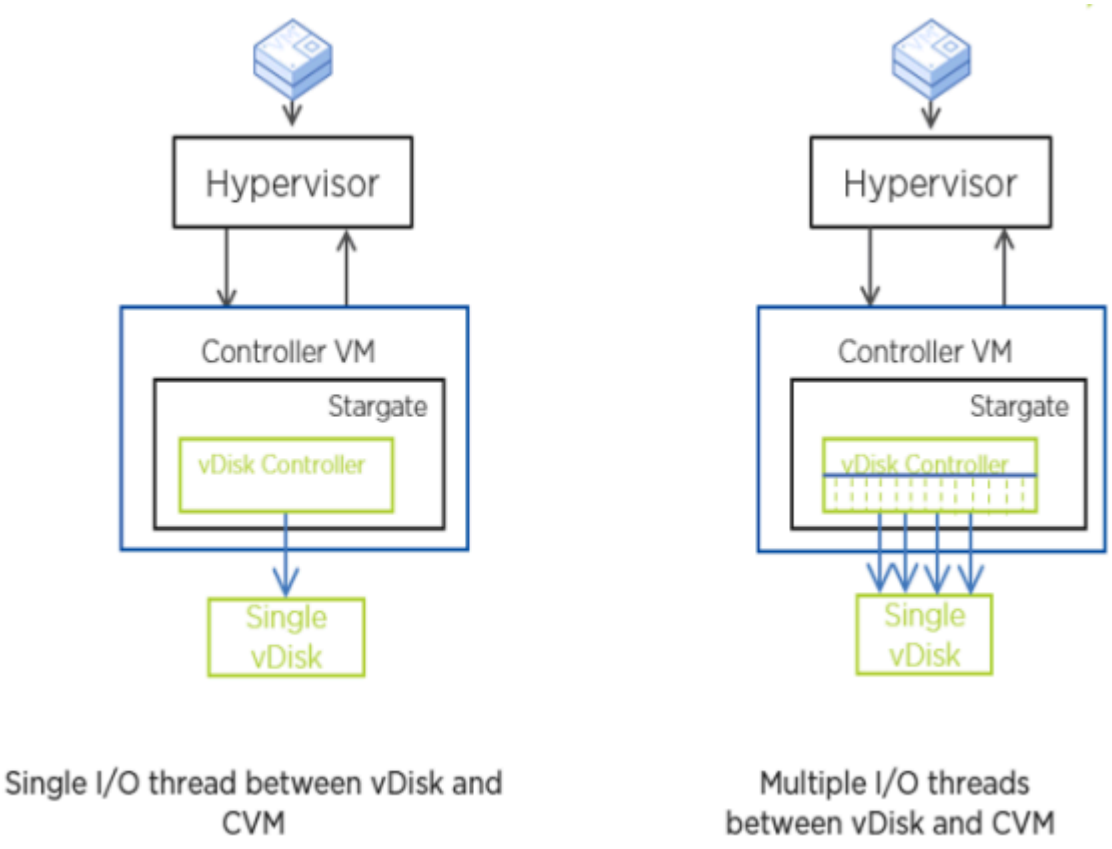

Figure 3: I/O Path without and with vDisk Sharding

8. High-Performance Snapshots and Clones

Traditional hypervisor-based snapshots degrade performance and typically aren't recommended or used for production environments. Performance degradation occurs because the hypervisor has little to no knowledge about the back-end storage. The performance challenge is that the hypervisor makes writes to delta files in the form of a change log, which records every change made to a file since the snapshot was taken. This process means that reads need to access the most recent delta file the block of data was written to as well as every change made to the original data. More snapshots result in more changed data, which in turn imposes a substantial performance penalty.

Nutanix AOS uses a redirect-on-write algorithm that dramatically improves system efficiency without sacrificing performance when taking snapshots because the snapshots have storage awareness and can provide enterprise-level data protection natively.

First, Nutanix snapshots allocate writes on a new block instead of writing them to a change log, which means AOS doesn't have to look up existing data when writing. Secondly, AOS storage intelligently handles the way snapshot trees are tracked in metadata to optimize performance and capacity, while simultaneously minimizing system overhead. When the child writes to any of the blocks inside the group, the child vDisk gets a copy of the parent metadata for an extent group. This process essentially eliminates any overhead as the snapshot chain grows because each clone has its own copy.

Additionally, the distributed metadata system allows users to request multiple vDisks in a single request, which greatly minimizes metadata lookup overhead for blocks that haven't been written or updated yet.

Clones (essentially writable snapshots) are closely related to snapshots. AOS storage uses the same underlying mechanism for cloning that it does for snapshots, so it benefits from the same metadata optimizations.

9. Conclusion

As discussed, although it's a cluster, Nutanix manages performance at the node level and, with the various enhancements that have been implemented in AOS recently, delivers the best possible performance for every guest VM. Data locality, which is unique to Nutanix, ensures that reads are served locally to VMs even after they move around in the cluster. Performance scales linearly as more nodes are added to the cluster.

Effective data tiering ensures that performance is optimized for every VM according to access patterns for I/O going to its vDisks. System processes make sure relevant data is hosted on the relevant tier to keep the cluster performing at the optimal level. With Optane tiering even on an all-flash system, AOS ensures it leverages the superior performance capabilities of Optane SSDs.

Blockstore and SPDK allows AOS to realize the full potential of storage hardware technology advancements by providing a lean data path.

vDisk sharding makes it possible to lift and shift traditional applications from SAN and NAS based storage to Nutanix without compromising on performance.

Finally, Prism provides a detailed end-to-end analysis of performance from the same management platform used to manage the cluster, so you don't need to install additional software specifically for managing performance.

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