Technical Review

Nutanix Architecture and Performance Optimization

Date: September 2020 **Author:** Tony Palmer, Senior Validation Analyst

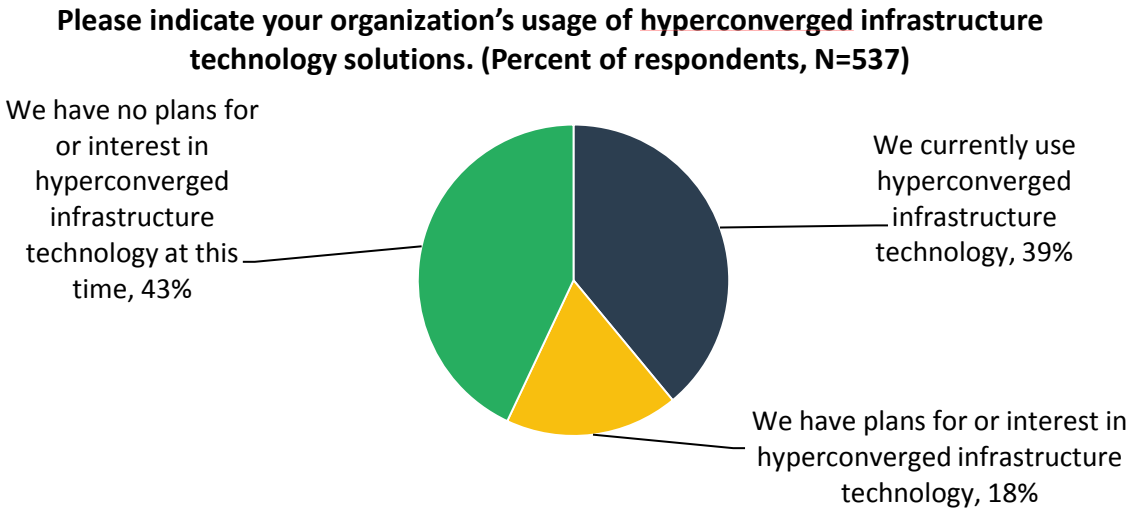
Abstract

This ESG Technical Review documents the results of recent Nutanix performance testing that focused on real-world performance scalability and sustainability.

The Challenges

As hyperconverged technologies continue to replace legacy technology solutions, organizations' buying criteria have continued to expand. They're no longer just looking for simplicity and cost savings; organizations are also prioritizing requirements such as performance, scalability, and reliability—recognizing that technologies like the cloud and software-defined storage will be far less complex and more cost-effective than a traditional siloed approach. In an ESG research study, 57% of respondents reported that they were using or planning to use HCI solutions. This is not surprising, given the factors driving them to consider HCI. Deployment drivers most cited by respondents include improved scalability (31%), total cost of ownership (28%), ease of deployment (26%), and simplified systems management (24%).[1]

Figure 1. Hyperconverged Infrastructure Usage Trends



Source: Enterprise Strategy Group

ESG research also reveals that flash storage continues to be adopted across IT, delivering not only improved performance levels, but also other key data center benefits such as reliability and TCO savings.[2] Eighty-four percent of respondents report that their organization is currently leveraging flash storage to some extent, with nearly half (47%) of organizations using the technology for more than 30% of their applications and workloads. In the same survey 71% of organizations indicated that they're either already using or plan to use NVMe-based storage technology. The top objectives organizations identified as driving their use of or interest in on-premises NVMe-based flash storage were to future proof their environment, improve performance, consolidate storage infrastructure, and reduce the operational cost of storage tuning and optimization.

[1] Source: ESG Master Survey Results, *Converged and Hyperconverged Infrastructure Trends*, October 2017.
[2] Source: ESG Research Report, *Data Storage Trends in an Increasingly Hybrid Cloud World*, March 2020.

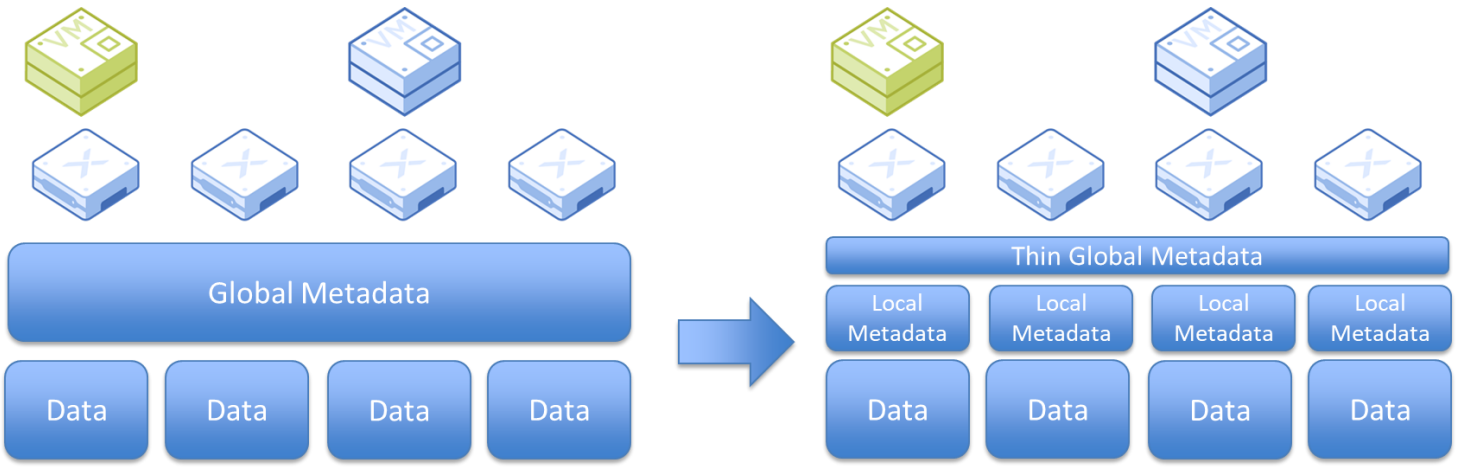
Nutanix AOS Platform

The Nutanix AOS Platform is designed to deliver a complete, software-driven IT infrastructure stack with the agility, scalability, and simplicity of the cloud combined with the security, performance, and cost predictability of a traditional on-premises infrastructure. The architecture is a scale-out fully distributed software platform leveraging web-scale engineering principles innovated by leading cloud companies such as Google, Facebook, and Amazon. The software integrates the compute, virtualization, and storage environments into a single solution. This integration eliminates the complexity of traditional SAN and NAS environments, costly, special-purpose hardware, and the specialized skill sets they require.

Next-generation Nutanix AOS with new Blockstore and SPDK technology—combined with other technologies like Autonomous Extent Store (AES), which was introduced in a prior version of AOS—brings HCI performance to the next level by capitalizing on the optimized architecture of AOS. These innovations optimize for high throughput and low latency applications and they are uniquely designed to deliver maximum benefits of new media such as NVMe and Storage-class memory.

Historically, storage devices have been slower than CPU and RAM by orders of magnitude, which meant that applications would interact with storage via OS using interrupts. The OS could do this efficiently because the latency of the storage medium was much slower than the rest of the system. This changed with the adoption of solid-stage storage (SSDs). With very low latency flash technology, the performance bottleneck is no longer the storage medium, but the software itself. System calls (i.e., interrupts) magnify the inherent overhead of general purpose in-kernel filesystems, with CPU cycles devoted to buffer cache management and metadata operations. With a goal of devising a more intelligent approach, Nutanix has engineered a simple free-space manager, which manages the space on the physical device at the block level, where the block is the minimum unit of allocation and deallocation. Nutanix calls this Blockstore. Blockstore includes a file-system layer built on top of the block free-space management layer, which adds directory structure and metadata management.

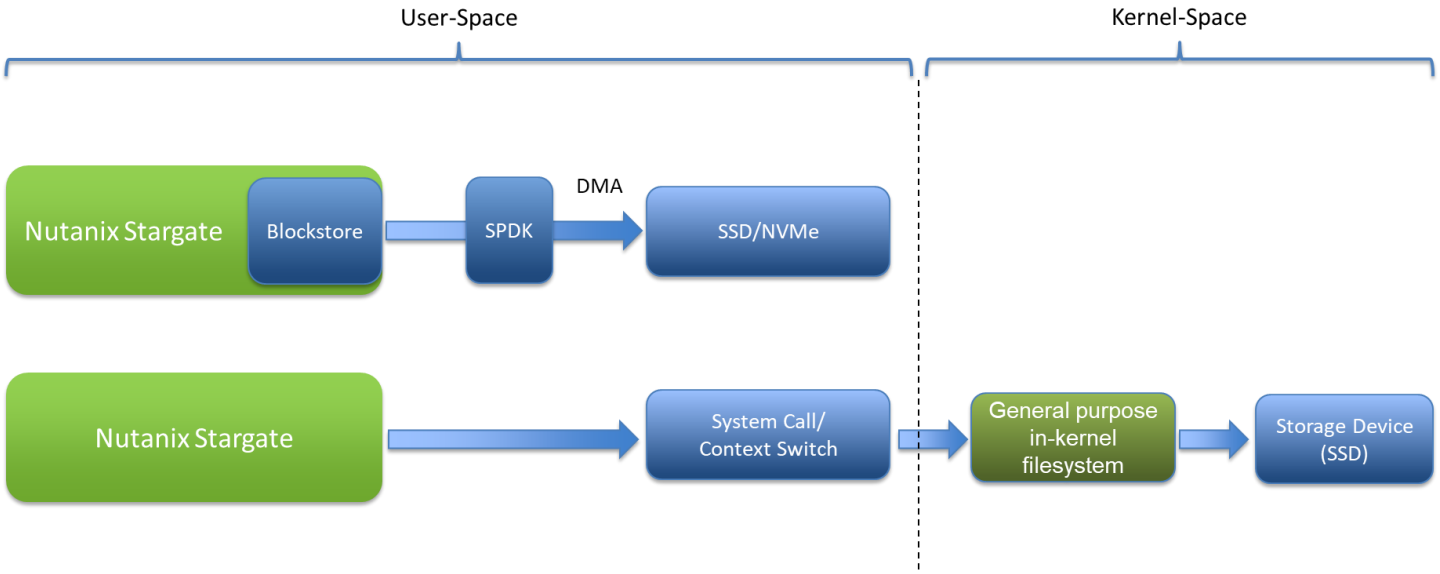
Figure 2. Autonomous Extent Store (AES) and Metadata Locality



Source: Enterprise Strategy Group

Modern storage systems typically manage their own metadata. Stargate, the main Nutanix component that manages data on the cluster, is no different. Global metadata is stored in a distributed key-value store (Cassandra) and local metadata in a local high-performance key-value store (RocksDB). This means that the Extent Store—where the data resides—doesn't need the full capabilities of an in-kernel filesystem. Eliminating the in-kernel filesystem and handling this in the Nutanix stack reduces overhead and lowers latency.

Figure 3. Optimizing Data Access with Blockstore



Source: Enterprise Strategy Group

New storage devices (like NVMe-based devices) come with user space tools and libraries that allow access to storage devices directly from user space, eliminating the need to make expensive system calls. This removes the overhead of user/kernel context switching. The Storage Performance Development Kit (SPDK) is an open source library developed by Intel for accessing raw NVMe devices from user space. SPDK achieves extremely high performance by avoiding interrupts, system calls, and locks. It addresses the issue of interrupt latency by polling the storage device. And since it is a user space library, it bypasses the context switch entirely, further reducing latency.

ESG Tested

ESG audited complete and detailed results from performance tests using a four-node Nutanix NX-8170-G7 cluster populated with eight 2TB NVMe devices per node that compared both synthetic and realistic application workloads. The testing used industry-standard application workload generation tools that exercised the Nutanix AOS Platform to compare the performance of Blockstore and SPDK with the previous version of AOS and show software improvements. The test configurations were identical, as was the hardware, in all cases. The workloads we look at in this report include:

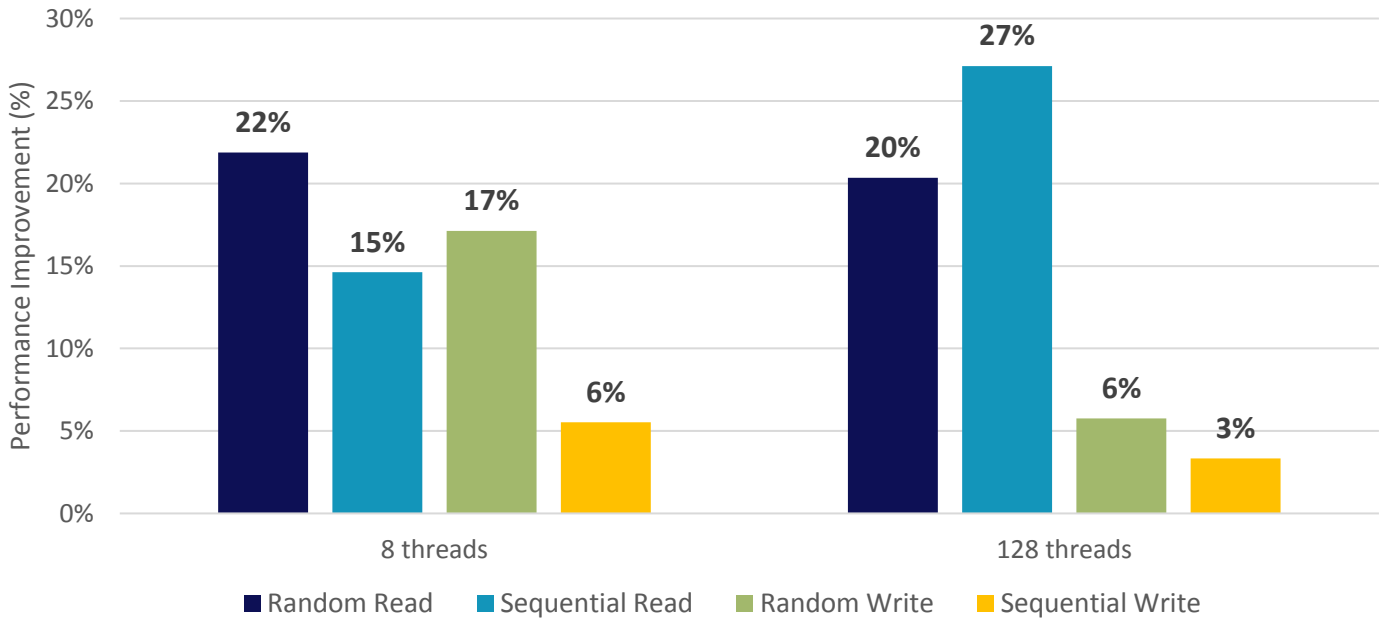
- Four corners — Synthetic workload to compare relative performance between Blockstore and pre Blockstore.
  - I/O Profile — 8KB random reads and writes, 1MB sequential reads and writes.
- High-performance database simulation—
  - I/O Profile — database: 100% random, 80% read, 90% 8KB, 10% 32KB, eight outstanding I/O.
  - Logs 100% 32KB writes, 90% sequential, one outstanding I/O.
  - One user VM per node.
- Multiple database online transactional processing (OLTP)—
  - Tests run with one to four databases active at a time, each with their own instance driving traffic.
  - 24 schemas per database, 32GB per schema.
  - I/O Profile — 100% random, 80% read.
- Postgres Analytics —
  - Pgbench was used to run a query set against Postgres DB.
  - A total of 12 database VMs were deployed across the cluster.
  - Buffer cache was kept small so the working set size (WSS) was larger than hosts' cache. The WSS was three times larger than the memory on each database VM.

The database simulation was generated with Nutanix X-Ray, a tool that uses FIO, while the multiple database workload leveraged the widely adopted and publicly available Silly Little Oracle Benchmark (SLOB), and the Postgres workload was generated using pgbench with Nutanix X-Ray. It should be noted that the tests were not designed to exercise the Nutanix platform to its maximum specified performance, but to illustrate the relative performance improvement provided by Blockstore and SPDK.

Four Corners

First, we tested the cluster to find the "four corners" raw performance, a common assessment of basic horsepower of the system. As shown in Figure 4, Nutanix with Blockstore and SPDK showed distinct performance improvement over a range of different queue depth values. Sequential performance was measured in GB/sec and random performance was measured in IOPS.

Figure 4. Four Corners Performance Improvement



Source: Enterprise Strategy Group

With Blockstore and SPDK:

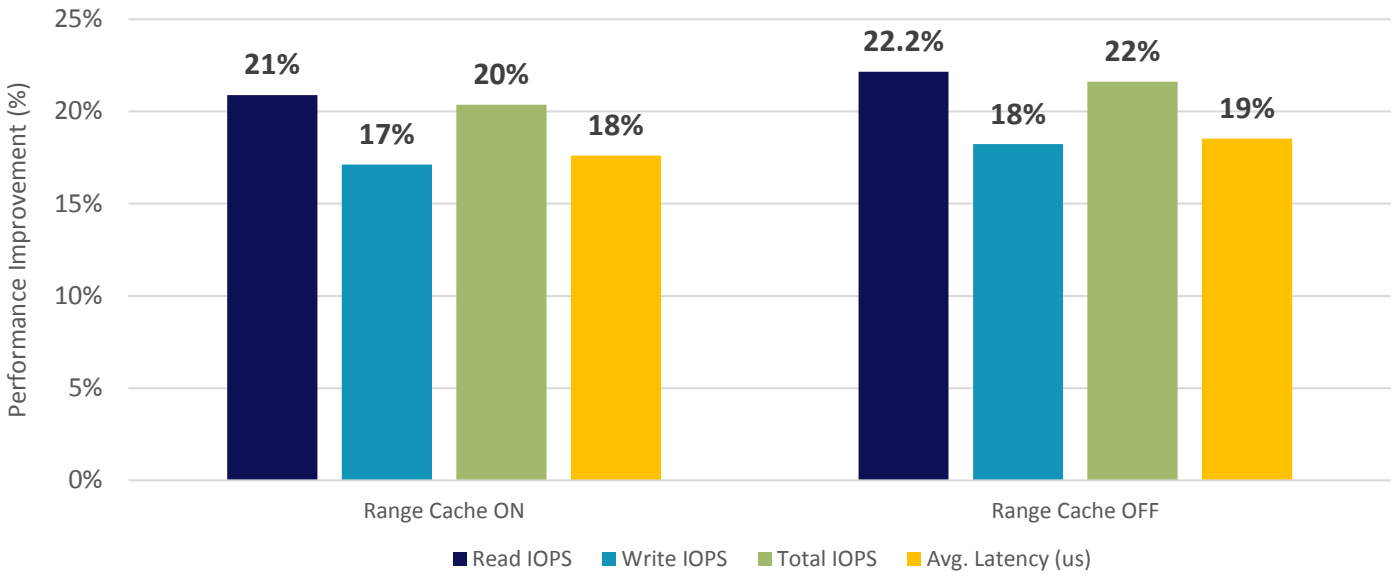
- Read workloads performed 15-27% better across various thread counts and 21% better on average.
- Write workloads improved by up to 17%.

Nutanix can also use the RDMA protocol for inter-node traffic with different hypervisors, which helps in acceleration of write throughput.

Database Simulation

Next, we tested a database simulation using Nutanix X-Ray. This workload was designed to simulate the database and logging I/O of active high-performance databases driving more than 200,000 IOPS at sub-millisecond response times across the cluster. Testing was executed twice for each configuration, with and without range cache enabled. Range cache caches user data in the Nutanix Controller VM's memory. The reason for this was to show a more complete picture of the improvement irrespective of caching.

Figure 5. Database Performance—IOPS and Latency Improvement



Source: Enterprise Strategy Group

As Figure 5 shows, Blockstore and SPDK improved:

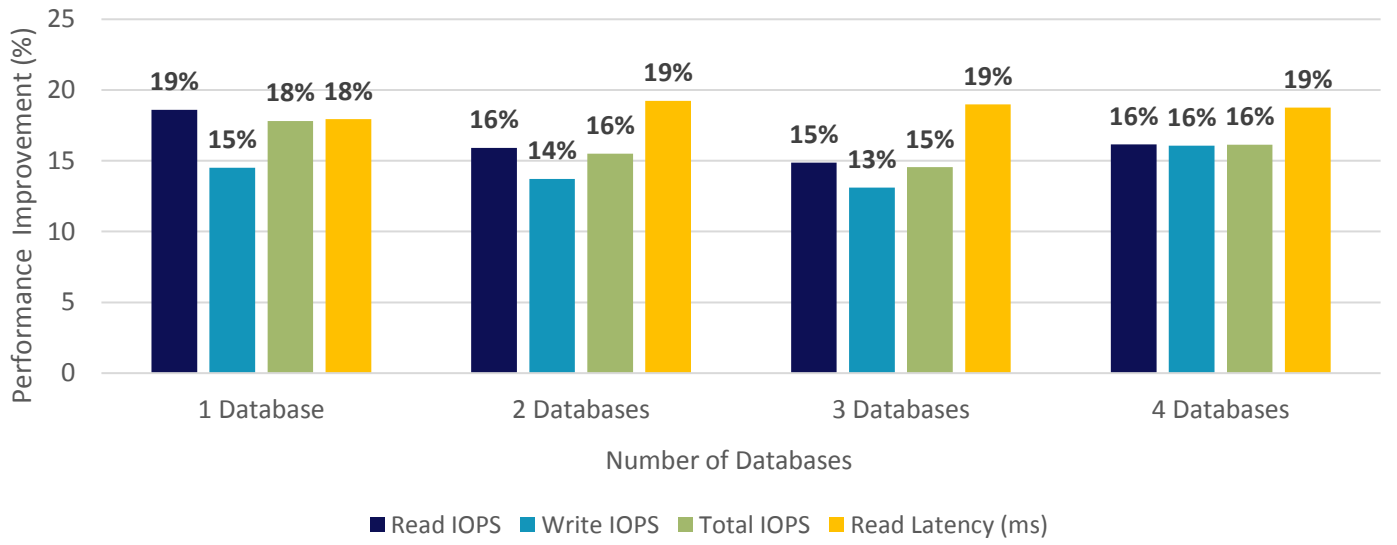
- Total IOPS by over 20% regardless of whether cache was enabled or not.
- Average latency by about 18%, also completely irrespective of whether cache was enabled or not.

Multiple Database OLTP Simulation

Next, we analyzed results of an emulated real-world transactional workload running on multiple Oracle databases and the SLOB benchmark utility. SLOB was used to generate realistic system-wide, random, single-block, and application-independent SQL queries. The tool exercised all components of the server and storage subsystems by stressing the physical I/O layer of Oracle through SGA-buffered random I/O, without being limited to a specific load-generating application. Four database VMs were created—one per node—and tests were run with one, two, three, and four databases active at a time, each with their own SLOB instance driving traffic. SLOB was run with 24 schemas per DB, 32GB per schema to ensure the working set size was significant. In these tests, range cache was disabled to prevent user data caching in the Nutanix controller VM (CVM).

Again, the suite of tests was run twice to compare performance of the exact same hardware running the exact same workload, with and without Blockstore and SPDK. The IOPS and latency results of the SLOB testing are shown in Figure 6.

Figure 6. OLTP Multi-database Performance—IOPS and Latency Improvement



Source: Enterprise Strategy Group

Again, the results painted a remarkably consistent picture. Blockstore and SPDK improved:

- Read IOPS by 15-19%
- Write IOPS by 13-16%.
- Read latency by nearly 20%
- Write latency—not shown—was similar with one database but improved by 5-6% with multiple databases.

As seen here, the increase in IOPS and reduction in latency were consistent as the load was scaled from one database to four databases.

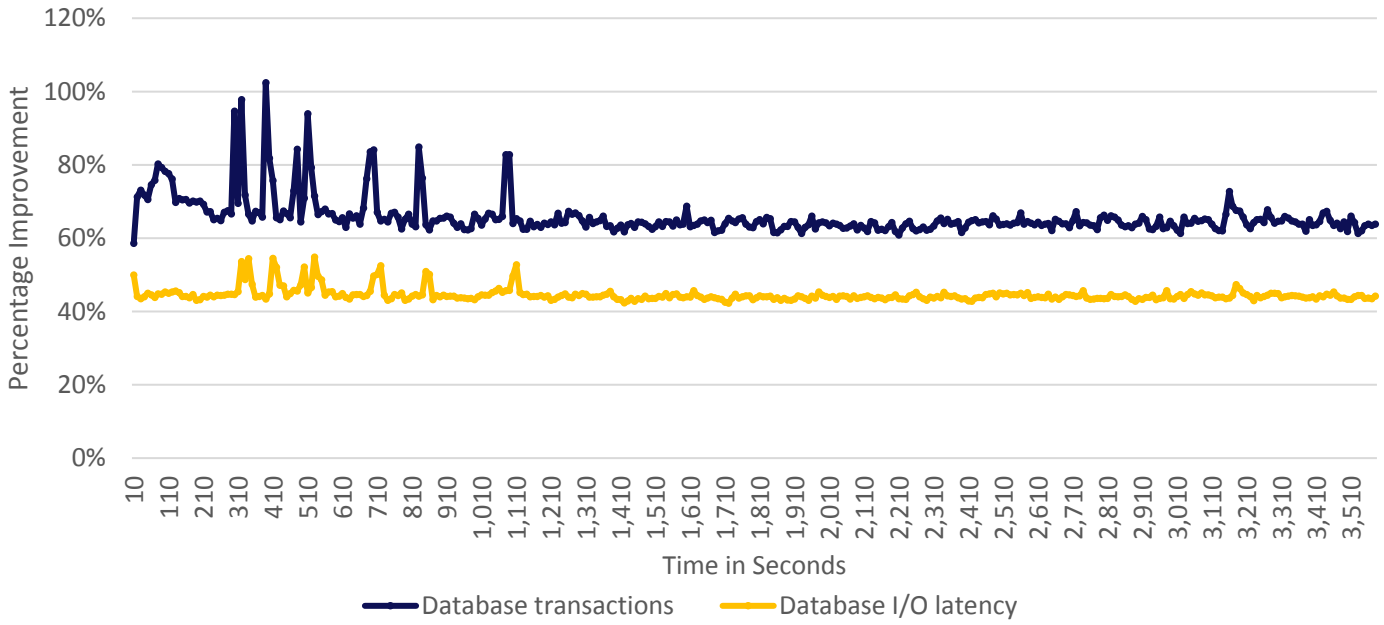
Postgres Performance

Finally, we used the pgbench utility to compare how applications might benefit by going from a typical all-flash cluster today to a next-generation cluster with Blockstore and SPDK leveraging Optane, NVMe drives, and other Nutanix technologies like Autonomous Extent Store (AES).

The baseline test was run against a four-node NX-3060 cluster with dual 2.4 GHz Intel Xeon E5-2680 v4 CPUs, 256GiB of RAM per node, and four 500GB SSDs per node. The results were compared with the same workload run against a four-node NX-8170 cluster with dual 2.5GHz Intel Xeon Gold 6248 CPUs, 376GiB RAM per node, and six 2TB Intel P4510 NVMe drives per node and two 750GB Intel P4800 Optane SSDs per node.

Figure 7 shows the consistent performance improvement across the entire one-hour duration of the test.

Figure 7. Postgres Performance—Database Transactions and Latency Improvement



Source: Enterprise Strategy Group

The next-generation cluster with Blockstore and SPDK showed an average improvement in DB transactions for this workload of 66%. Latency was cut nearly in half, improving by an average of 45%. Latency is a much more important metric than raw IOPS, and reducing database transaction latency by half means that more response-time-sensitive tier-1 applications can be deployed on Nutanix HCI with confidence.

Why This Matters

Delivering high levels of performance is a requirement for IT environments that rely heavily on mission- and business-critical databases and applications. This is especially important in dynamic environments where data growth is constant and continuous accessibility is a requirement. The ability to easily meet these performance and scalability requirements is essential for anyone evaluating hyperconverged infrastructures. The challenge is that some organizations feel there is too much overhead between the virtualization and the essential underlying services that must always be running to not only ensure proper functionality of the hyperconverged infrastructure, but to also meet strict application performance SLAs.

ESG confirmed that Nutanix Blockstore significantly improves I/O efficiency and performance. Blockstore and SPDK improved application performance and reduced latency in synthetic and real-world testing. Our tests exercised both storage and compute to highlight the type of performance organizations can expect in their own OLTP database and complex application environments. While simply installing the latest version of AOS provided IOPS and latency improvements of up to 27%, organizations leveraging the latest AOS while moving from standard flash to all NVMe or Optane with NVMe could experience IOPS improvements of 66% while cutting latency in half. This can translate into support for significantly higher density and/or performance of supported databases and applications.

The Bigger Truth

Organizations want the simplicity and scalability of the cloud provided by HCI, but they need to be able to predict costs for mission-critical applications with high-performance SLAs. High levels of reliable and scalable enterprise-class performance have become less of a wish, and more of an expectation. Flash storage went a long way to move organizations closer to this goal but using the same storage interfaces as HDDs (SATA, SAS, and SCSI) has increasingly become a bottleneck. Using in-kernel filesystems introduces more layers and exacerbates the problem.

The NVMe protocol was developed to address the interface bottleneck by enabling SSDs to connect directly to the PCIe bus, increasing not only the potential performance of each drive individually but also the aggregate performance within a server. Memory-class storage technologies, like Intel Optane, utilize different underlying storage medium, can be connected either through NVMe or by directly plugging into memory (DIMM) slots on a server motherboard, and are capable of even greater performance than NVMe SSDs. Nutanix Blockstore streamlines the I/O stack, taking the filesystem out of the equation altogether.

ESG audited performance results and validated that the Nutanix platform has addressed these issues with its latest generation HCI clusters and Nutanix AOS with Blockstore. Testing confirmed that Nutanix meets the demanding performance requirements of dynamic, mission-critical applications. The Nutanix platform with NVMe, Optane, and Blockstore delivers significant IOPS and latency improvements with no changes to the applications. Real-world testing exercised both compute and storage resources to meet the high transactions and latency demands of scalable OLTP database deployments in both Oracle OLTP and Postgres analytics database environments.

As hyperconverged technologies continue to mature, Nutanix continues to expand the boundaries of what is possible by not only adopting cutting-edge technology but also providing software enhancements that take outstanding advantage of it, satisfying the performance requirements and expectations of the most performance-demanding mission- and business-critical applications. If you are looking to modernize your IT infrastructure to gain the benefits of today's most highly performant storage technology with the simplicity of HCI, you'd be smart to seriously consider Nutanix.