

TECHNICAL VALIDATION

VMware vSAN 8 Express Storage Architecture on Intel Fourth Generation Xeon Scalable Processors

Modernizing Infrastructure to Support Business-critical and Emerging Artificial Intelligence Workloads

By Alex Arcilla, Senior Analyst – Validation Services
Enterprise Strategy Group

Contents

Introduction	3
Background	3
VMware vSAN 8 ESA on Intel Fourth Generation Xeon Scalable Processors	4
Enterprise Strategy Group Technical Validation	5
Maintaining High Application Performance When Consolidating Business-critical Workloads	5
Maintaining High and Consistent Performance When Scaling Out Traditional and Modern Workloads.....	7
Achieving Performance Needed for AI-enabled Applications	10
Conclusion	13

Introduction

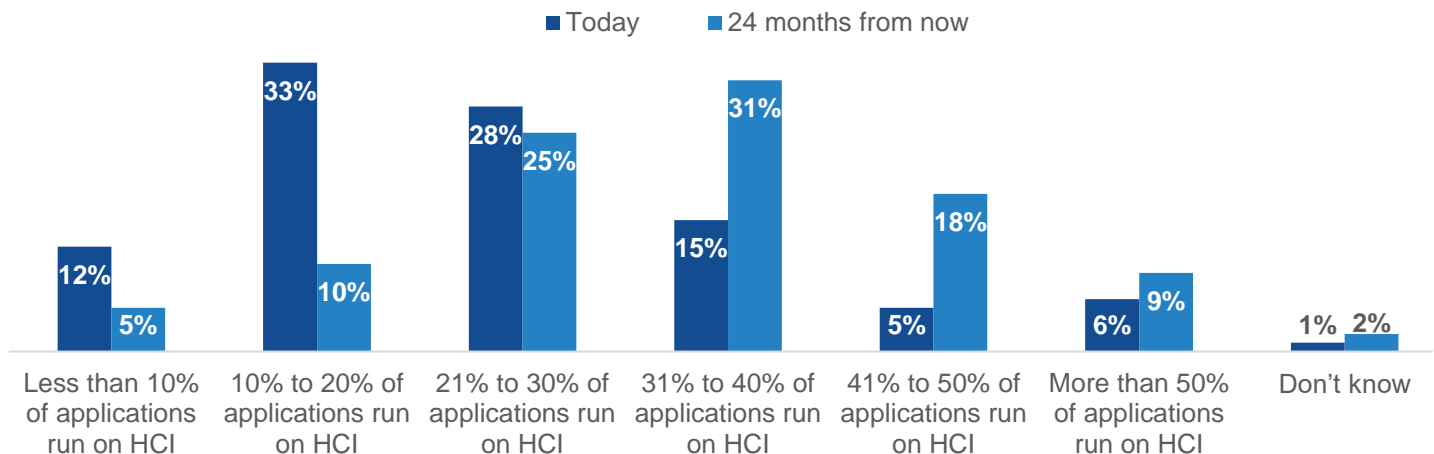
This Technical Validation from TechTarget’s Enterprise Strategy Group documents our evaluation of VMware vSAN 8 Express Storage Architecture (ESA) on fourth generation Intel Xeon Scalable Processors. We reviewed how this latest release of hyperconverged infrastructure (HCI) software, supported by the latest generation of Intel Xeon processors, can help organizations modernize IT infrastructure by facilitating server consolidation while maintaining the performance organizations expect from a wide variety of workloads, particularly artificial intelligence.

Background

The deployment of hyperconverged infrastructure in organizations’ private clouds shows no indication of slowing down. According to research from TechTarget’s Enterprise Strategy, 74% of respondents anticipate running between 21% and 50% of their business applications and workloads on HCI, up from 48% today.¹

Figure 1. Production Applications Running on HCI Now and in Two Years

Approximately what percentage of your organization’s production business applications/workloads are run on HCI today? Approximately what percentage of your organization’s production business applications/workloads do you expect will run on HCI



Source: Enterprise Strategy Group, a division of TechTarget, Inc.

Driving the increase of applications deployed on HCI is the constant pressure of IT in organizations to “do more with less.” Ever-shrinking IT budgets result in fewer funds available for capital and operational expenses. HCI has proven to be instrumental in server consolidation. Indeed, modernizing infrastructure via consolidation leads to numerous benefits, as a resulting smaller footprint requires less power and cooling and less cost to manage and administer the infrastructure.

However, organizations cannot risk sacrificing performance, as end users will not tolerate any degradation of performance of business-critical workloads that they rely on to complete their daily tasks, especially as organizations begin to deploy modern workloads, specifically those leveraging AI. Scalability also becomes a

¹ Source: Enterprise Strategy Group Research Report, [Navigating the Cloud and AI Revolution: The State of Enterprise Storage and HCI](#), March 2024. All Enterprise Strategy Group research references and charts in this Technical Validation have been taken from this research report, unless otherwise noted.

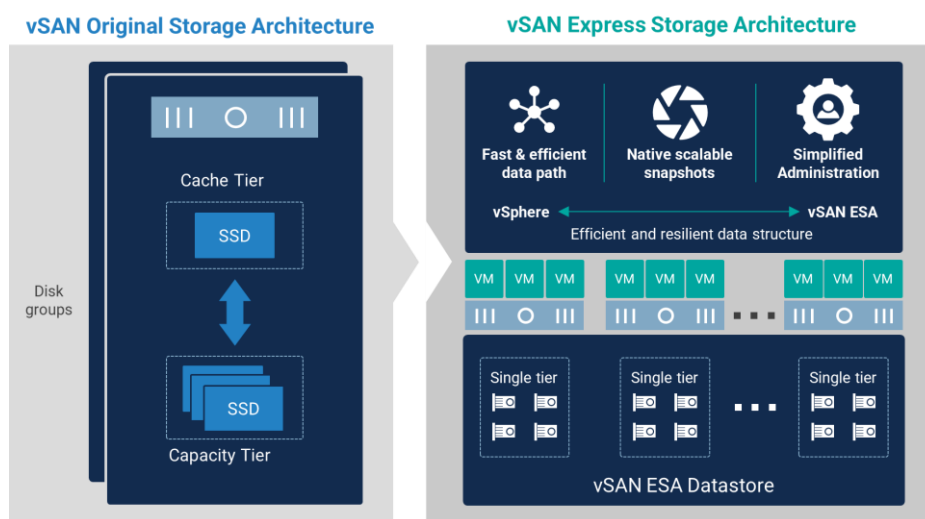
challenge, as organizations must scale workloads when business demands emerge, yet they must accomplish this under existing budgetary constraints. Tackling these challenges requires organizations to select the right combination of hardware and software that will deliver the benefits of server consolidation without sacrificing performance and scalability.

VMware vSAN 8 ESA on Intel Fourth Generation Xeon Scalable Processors

Built on VMware’s ESA, VMware vSAN 8 has been designed to help organizations better scale their infrastructure to run both traditional and modern workloads on fewer servers configured with Intel fourth generation Xeon Scalable Processors, whether they operate in an enterprise data center, a private cloud, or at the edge.

When compared to VMware vSAN Original Storage Architecture (OSA), VMware vSAN 8 takes advantage of hardware advances in storage so that organizations can maximize available storage capacity without sacrificing storage efficiency. Instead of the OSA’s two-tier architecture (cache and capacity tiers) that can work with a range of older storage devices, ESA leverages high-performance NVMe-based TLC flash devices to create a single-tier architecture so that all NVMe storage devices form a single storage pool to be utilized by the vSAN data store. ESA is also designed to maximize the storage capacity of a vSphere cluster using RAID-6 while achieving RAID-1 performance. With improved data compression and a native snapshot engine, organizations can also increase storage efficiency. To ensure data security, VMware vSAN 8 provides FIPS-validated data-at-rest and data-in-transit encryption.

Figure 2. Comparison of VMware vSAN 8’s OSA and ESA



Source: VMware by Broadcom and Enterprise Strategy Group, a division of TechTarget, Inc.

VMware vSAN 8 can also support high application performance requirements by taking advantage of Intel’s fourth generation Xeon Scalable Processor design. This latest generation of processors features Intel Accelerator Engines designed to accelerate performance across a wide variety of both traditional and modern workloads, especially AI-powered applications. With the processor’s higher core count per socket, servers using Intel fourth generation Xeon Scalable Processors can increase physical and virtual CPU utilization, thus increasing VM density. Combined with the processor’s power efficiency, organizations can decrease their server footprint, thereby reducing power consumption and related costs.

For emerging workloads, particularly those enabled by AI, the fourth generation Intel Xeon Scalable Processor is equipped with Intel Advanced Matrix Extensions (AMX) for both training and inference, building upon the

acceleration achieved with other accelerator technologies in the third generation Xeon processors, such as Intel Advanced Vector Extensions 512 (AVX-512) and Intel Deep Learning (DL) Boost.

The combination of VMware vSAN 8 ESA with servers using Intel fourth generation Xeon Scalable Processors ultimately helps organizations reduce server and storage costs associated with running both traditional and modern workloads using vSAN clusters, without sacrificing the expected performance and scalability.

Enterprise Strategy Group Technical Validation

Enterprise Strategy Group validated how the combination of VMware vSAN 8 ESA and servers powered by Intel fourth generation Xeon Scalable Processors can help organizations satisfy the performance requirements of both traditional and modern workloads by reviewing performance testing results. (Note that testbed configurations for all performance results audited in the Validation are located in the Appendix.)

Maintaining High Application Performance When Consolidating Business-critical Workloads

Consolidating business-critical workloads—namely those that support daily business operations such as databases and VDI—requires the right combination of updated hardware and software that can adequately support mixed workloads while maintaining the performance end users expect from each workload individually. Operating VMware vSAN 8 on servers powered by Intel fourth generation Xeon Scalable Processors can help organizations achieve those results.

Enterprise Strategy Group Analysis

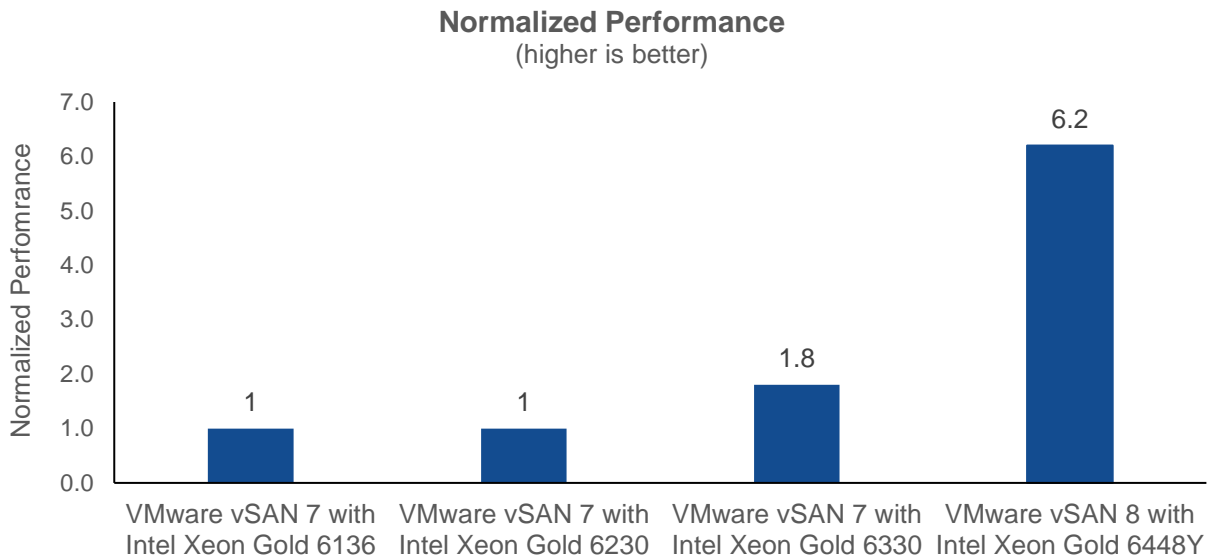
Enterprise Strategy Group reviewed the results of performance testing designed to compare improvements in performance and latency across the following configurations:

- VMware vSphere and vSAN 7 with first generation Intel Xeon Scalable Processors.
- VMware vSphere and vSAN 7 with second generation Intel Xeon Scalable Processors.
- VMware vSphere and vSAN 7 with third generation Intel Xeon Scalable Processors.
- VMware vSphere 8 and vSAN 8 with fourth generation Intel Xeon Scalable Processors.

Tests were conducted using HClBench, an industry-standard tool designed to test the performance of HCI clusters running virtual machines. HClBench leverages the industry-standard Vdbench storage benchmark tool to automate the end-to-end process that includes deploying test VMs, coordinating workload runs, aggregating test results, and collecting data.

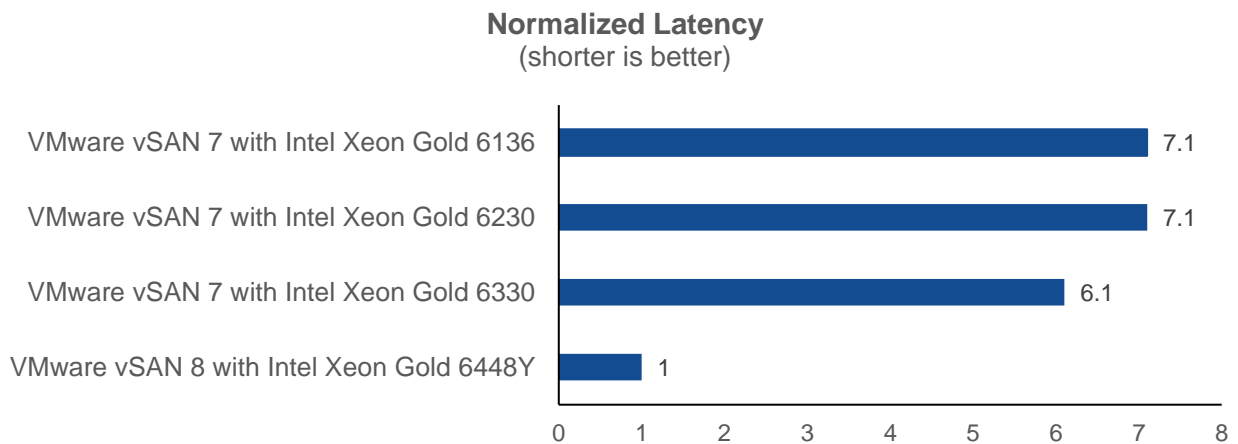
Test scenarios simulated a 70/30 workload with an 8KB block size and 100% random access. Scenarios reflect database (SQL/Oracle) and VDI workloads. Figures 3 and 4 show improvements based on normalized performance and latency results.

Figure 3. Normalized Performance Observed Across Configurations



Source: Enterprise Strategy Group, a division of TechTarget, Inc.

Figure 4. Normalized Latency Reduction Achieved Between Configurations



Source: Enterprise Strategy Group, a division of TechTarget, Inc.

What the Numbers Mean

Performance was observed to be up to 6.2x higher with VMware vSAN 8 ESA with the fourth generation Intel Xeon Processor when compared with previous releases of VMware vSAN and Intel Xeon processor generations. Additionally, latency was observed to be up to 7.1x lower with VMware vSAN 8 ESA with the fourth generation Intel Xeon Processor when compared with previous releases of VMware vSAN and Intel Xeon processor generations. Based on these results, Enterprise Strategy Group concluded that multiple traditional workloads (such as the VDI workloads simulated) can be consolidated onto fewer servers without sacrificing performance or increasing latency.

Why This Matters

As organizations continue to deploy more applications, infrastructure modernization—particularly server consolidation—becomes a key issue under the pressure of IT budgetary constraints. Yet, application performance cannot be sacrificed or else end users cannot complete critical business tasks.

Enterprise Strategy Group validated that VMware vSAN 8 running on servers configured with Intel fourth generation Xeon Scalable Processors can support higher application performance and lower latency even when consolidating workloads. Based on performance tests simulating traditional workloads run across different configurations of previous and current generations of Intel Xeon Scalable Processors and VMware vSAN, we saw that the combination of VMware vSAN 8 and Intel fourth generation Intel Xeon Scalable processors achieved the highest performance and lowest latency. We concluded that this combination can achieve higher VM density and storage utilization efficiency, without sacrificing the performance that end users expect.

Maintaining High and Consistent Performance When Scaling Out Traditional and Modern Workloads

When scaling out traditional and modern workloads, organizations must ensure that expected performance is maintained. Operating VMware vSAN 8 on servers powered by Intel fourth generation Xeon Scalable Processors can help organizations achieve the desired scalability as business needs change.

Enterprise Strategy Group Analysis

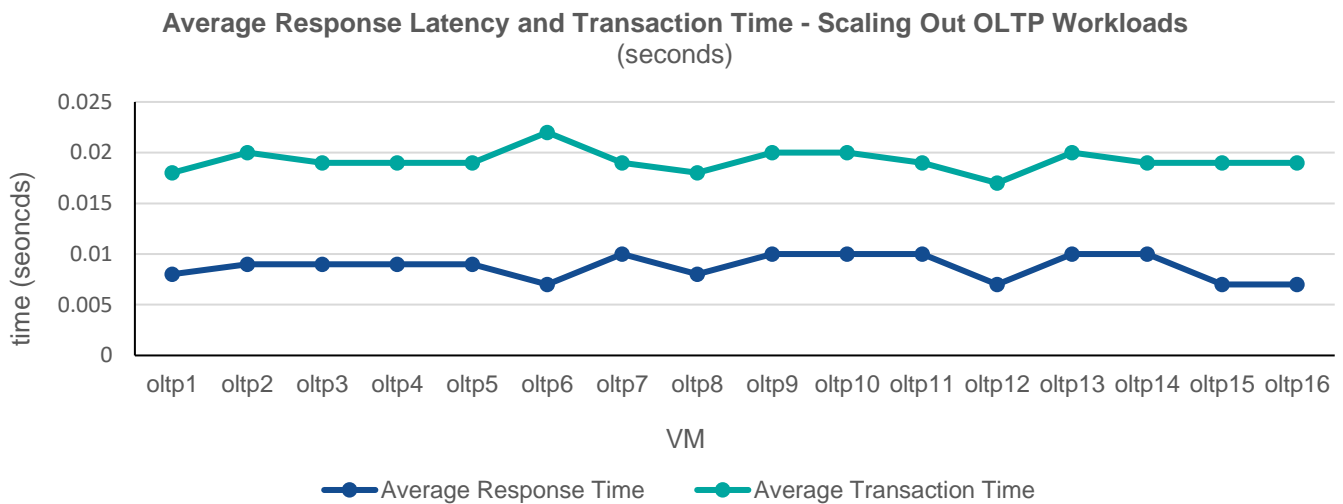
Enterprise Strategy Group proceeded to evaluate testing results that illustrated the performance consistency achieved when scaling out both a traditional online transaction processing (OLTP) workload (i.e., SQL database) and a modern database workload (i.e., NoSQL database) running on VMware vSAN 8 ESA.

To show how application performance was affected when scaling out a SQL database, performance testing was conducted using Benchmark Factory with a TPC-E-like benchmark.² Testing was designed to show performance consistency across multiple OLTP database workloads running on hosts configured with VMware vSAN 8 ESA.

Four hosts each ran four simulated OLTP workloads simultaneously, with one workload per VM, for a total of 16 VMs. Each host was configured with Intel fourth generation Xeon processors. During the testing period, 40 users simultaneously accessed each workload. A total of 33,300 transactions per second were generated during the testing period. Test results showing transaction latency and average response time for each OLTP workload are shown in Figure 5.

² The TPC-E-like benchmark was designed using a simulated database within a brokerage firm that supports transactions related to trades, account inquiries, and market research.

Figure 5. Average Transaction Time When Scaling Out OLTP Workloads



Source: Enterprise Strategy Group, a division of TechTarget, Inc.

What the Numbers Mean

The average response time was estimated to be eight milliseconds, with response time ranging from seven to 10 milliseconds. Additionally, the average transaction time was estimated to be 19.2 milliseconds with over 33,300 transactions processed per second, with transaction time ranging from 17 to 22 milliseconds. Based on these test results, Enterprise Strategy Group observed that scaling out workloads on hosts running VMware vSAN 8 ESA can provide consistently high performance.

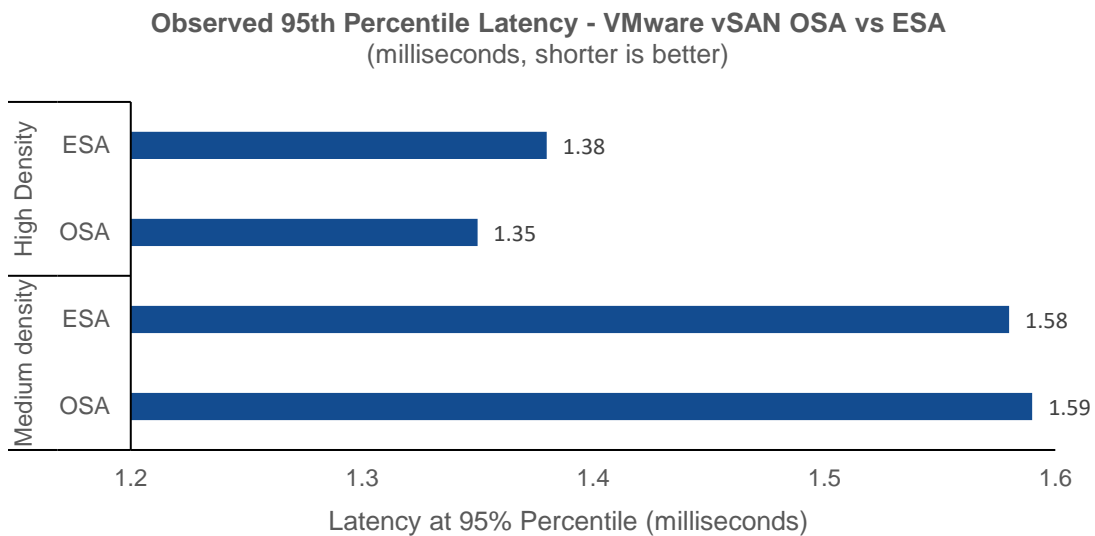
Enterprise Strategy Group then proceeded to review testing results that revealed the performance consistency achieved when scaling out a NoSQL database on VMware vSAN 8 ESA. We specifically reviewed the performance achieved with Apache Cassandra, an open source, NoSQL distributed database. Its distributed nature enables easier and faster horizontal scalability.

Performance testing was conducted using NoSQLBench, a tool designed for performance and scale testing for NoSQL databases. Testing was designed to compare performance of a 10% read/90% write Cassandra workload running on both VMware vSAN ESA and OSA.

To show the benefits of scaling out a workload, we observed Cassandra workloads running on a medium-density (two VMs per host) and high-density (four VMs per host) scenario. Each VM was configured with 16 vCPUs, 64GB RAM, a 256GB OS disk, a 2TB data disk, and a 100GB log disk. A data set of 4 billion records was preloaded for the medium-density scenario. For the high-density scenario, a data set of 8 billion records was preloaded.

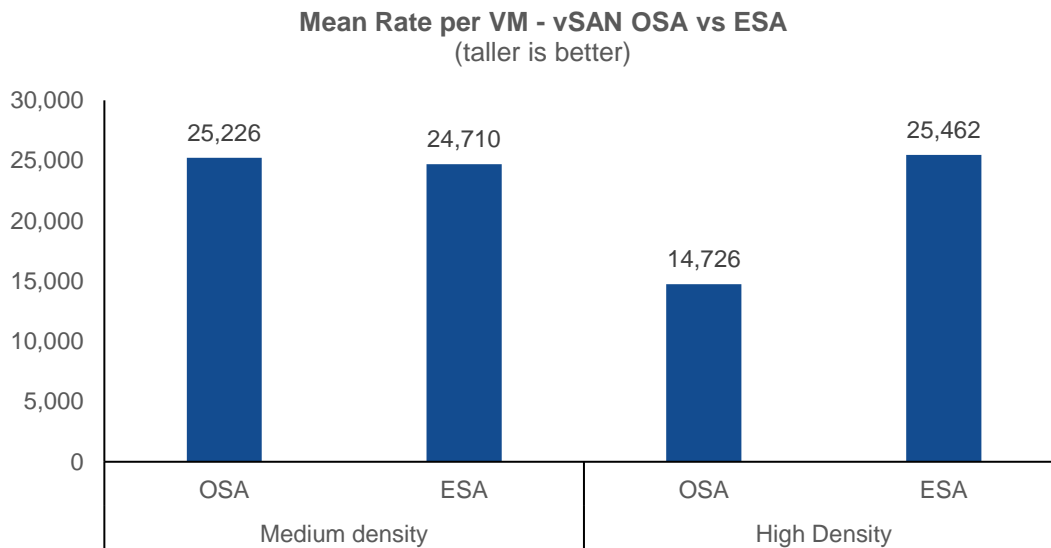
Test results showing transaction latency and mean rate per VM for both scenarios are shown in Figures 6 and 7.

Figure 6. Latency at 95% Percentile – VMware vSAN OSA versus ESA



Source: Enterprise Strategy Group, a division of TechTarget, Inc.

Figure 7. Mean Rate per VM – OSA versus ESA



Source: Enterprise Strategy Group, a division of TechTarget, Inc.

What the Numbers Mean

While the 95% percentile latency for ESA and OSA are terribly similar, the performance improvement can be clearly seen when considering the mean rate, especially in the high-density scenario. The results indicated that scaling out Cassandra workloads on hosts running VMware vSAN 8 ESA can exhibit higher performance (i.e., read and write throughput increases) as more VMs are added to the host. As the VM density of Cassandra workloads increased

from two to four, we observed that VMware vSAN 8 ESA can maintain, if not increase, mean rate per VM, while exhibiting a low and steady 95% latency.

Why This Matters

42% of respondents to Enterprise Strategy Group research indicated that their organization has realized improved scalability from its use of HCI. To see this as a benefit implies that application performance must be maintained, if not improved, as workloads are scaled out to meet business demands.

Enterprise Strategy Group validated that VMware vSAN 8 ESA can deliver better application performance when scaling out both traditional and modern workloads on hosts running on Intel fourth generation Xeon Scalable Processors. After reviewing performance testing results comparing performance and latency of simulated SQL and NoSQL database workloads on VMware vSAN 8 OSA and ESA, we observed that performance was maintained, if not improved, as VM density increased on hosts running VMware vSAN 8 ESA. Simultaneously, we also saw that the 95% latency remained low and consistent.

Achieving Performance Needed for AI-enabled Applications

As AI adoption continues, organizations must prepare their IT infrastructure to support AI-enabled applications to quickly deliver the insights needed to deliver customer value. However, organizations must provide the processing power needed to analyze larger amounts of data to gain those insights, while controlling hardware and software costs. Server infrastructure using VMware vSAN 8 ESA and Intel fourth generation Xeon Scalable processors equipped with Intel AMX can provide the performance necessary for AI-enabled workloads.

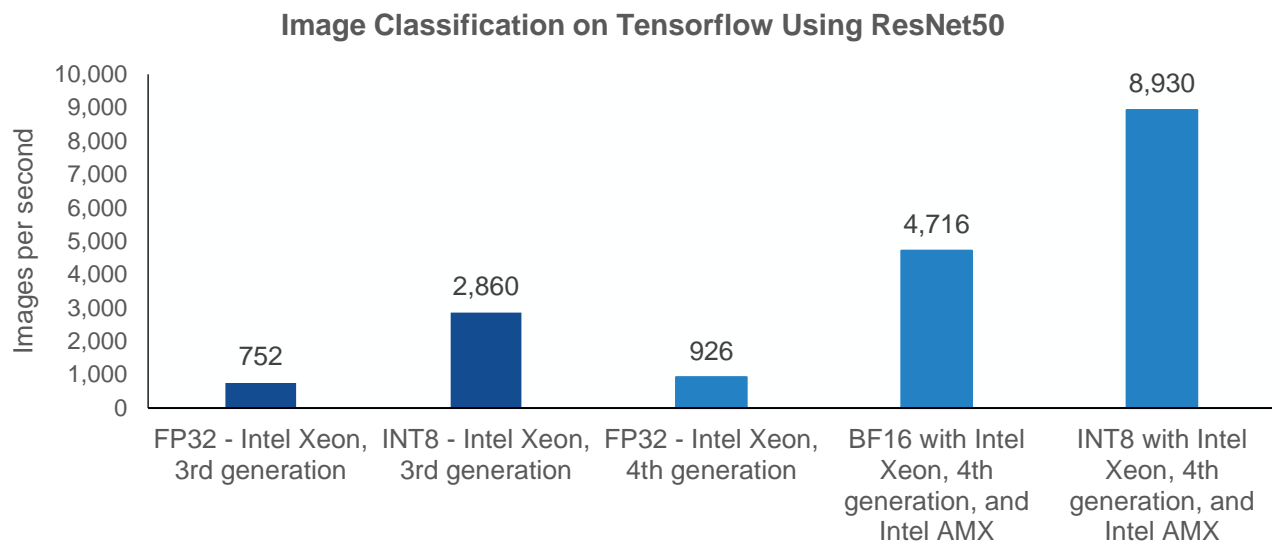
Enterprise Strategy Group Analysis

Enterprise Strategy Group then reviewed the results of performance testing designed to compare improvements in performance between Intel third and fourth generation Xeon processors, as Intel AMX was available starting with the fourth generation. We first evaluated the results of a test running an image-classification workload using deep learning. This workload ran on a testbed consisting of hosts configured with Intel third and fourth generation Xeon processors and VMware vSAN 8 ESA. (The Appendix contains these hardware and software configurations.) The test used the ResNet-50³ benchmark, a benchmark for measuring image classification workloads. Performance (in terms of images processed per second) used the following data formats that are supported by Intel third and fourth generation Xeon processors:

- **FP32** – Floating point data format for deep learning where data is represented as a 32-bit floating point number. This format is widely used for training deep learning models and inferencing.
- **Bfloat16 (BF16)** – A truncated version of 32-bit floating point, used for both training and inference. While this offers accuracy similar to FP32, BF16 enables faster computation than FP32.
- **INT8** – A data type that can significantly boost performance with minimal impact on accuracy.

Testing employed Tensorflow 2.11, an open source software library for machine learning and AI, with a focus on training and inference of deep neural networks. Results are illustrated in Figure 8.

³ ResNet-50 is an image classification model that can be trained on large data sets. Using residual connections, this convolutional neural network uses residual connections, enabling the network to quickly learn residual functions that map the input to the desired output.

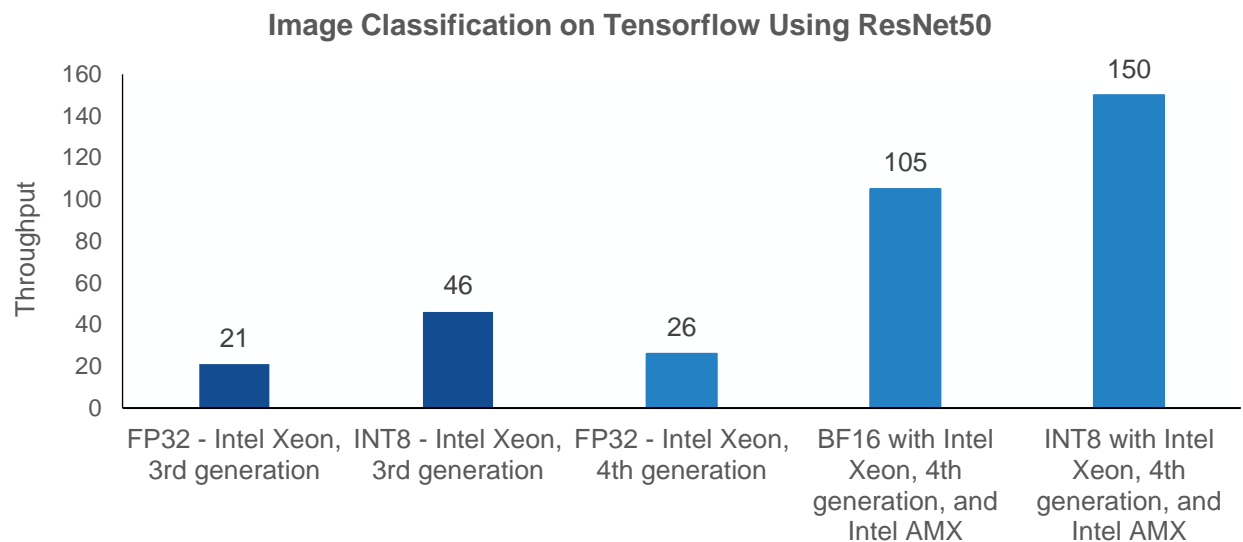
Figure 8. Performance of AI-enabled Image Classification

Source: Enterprise Strategy Group, a division of TechTarget, Inc.

What the Numbers Mean

- When using the FP32 data type, 1.23x higher performance was observed on Intel fourth generation processors (752 versus 926).
- When using the INT8 data type, 3x higher performance was observed on Intel fourth generation processors (2,960 versus 8,930).
- When comparing FP32 and BF16 data types used in workloads running on Intel fourth generation processors, 5x higher performance was observed when running Intel AMX, with practically no loss in accuracy (926 versus 4,716).
- When comparing FP32 and INT8 data types used in workloads running on Intel fourth generation processors, 9x higher performance was observed when running Intel AMX, with minimal loss in accuracy (926 versus 8,930).

Enterprise Strategy Group then evaluated test results illustrating the performance of a natural language processing model running on the testbed used previously. This test specifically used the BERT-Large pretrained model. BERT was one of the first LLMs for baselining NLP tasks, including general language understanding, question and answer, and named entity recognition. The test used the same data types as in the previous test, comparing performance improvements between third and fourth generation processors. Tests were run using Tensorflow 2.11. Results are shown in Figure 9.

Figure 9. Performance of AI-enabled NLP

Source: Enterprise Strategy Group, a division of TechTarget, Inc.

What the Numbers Mean

- When using the FP32 data type, Enterprise Strategy Group observed performance was 1.24x higher on Intel fourth generation processors (21 versus 26).
- When using the INT8 data type, we observed performance was 3.2x higher on Intel fourth generation processors (46 versus 150).
- When comparing the FP32 and BF16 data types used in workloads running on Intel fourth generation processors, we observed performance was 4x higher on Intel fourth generation processors running Intel AMX, with practically no loss in accuracy (26 versus 105).
- When comparing the FP32 and INT8 data types used in workloads running on Intel fourth generation processors, we observed performance was 5x higher on Intel fourth generation processors running Intel AMX, with minimal loss in accuracy (26 versus 150).

Why This Matters

According to Enterprise Strategy Group research, 41% of respondents indicated that they have generative or predictive AI initiatives driving their multivendor technology evaluation project for HCI. Yet, AI workloads demand highly performant infrastructure that enables organizations to ultimately extract value from data as quickly as needed to meet business demands faster.

Enterprise Strategy Group validated that AI workloads on VMware vSAN 8 ESA running on hosts equipped with Intel fourth generation Xeon processors performs better overall compared with Intel third generation processors. When comparing the use of different data types that can exhibit increased performance with minor impact on accuracy, we observed that AI workload performance was maximized when supported by VMware vSAN 8 ESA running on hosts configured with Intel further generation Xeon processors and Intel AMX.

Conclusion

HCI is here to stay, as organizations have realized numerous benefits, specifically reduced hardware footprints via server consolidation, leading to reduced capital and operational costs. Yet, the pressure remains on HCI to support increasing performance and scalability demands as more applications—traditional and modern—are deployed across data centers, private clouds, and the edge. The combination of VMware vSAN 8 ESA and Intel fourth generation Xeon Scalable Processors can help organizations address this challenge.

VMware vSAN 8 ESA has been designed to leverage hardware advances in both processing power and storage to meet higher performance and scalability requirements. As the adoption of NVMe-based TLC flash devices increases, organizations can leverage VMware vSAN ESA to maximize available storage capacity without sacrificing storage efficiency. Simultaneously, VMware vSAN ESA can take advantage of the processor's higher core count per socket in order to increase density. When combined with the processor's power efficiency, the combination of VMware vSAN 8 and Intel fourth generation Xeon Scalable Processors not only can help organizations decrease their server footprint and related power consumption and operational expenses, but also deliver the performance and scalability needed for both traditional and modern workloads.

Throughout our evaluation of performance test results, Enterprise Strategy Group validated that VMware vSAN 8 ESA operating on hosts configured with Intel fourth generation Xeon Scalable Processors can:

- Achieve higher application performance for traditional workloads compared to combinations of previous generations of VMware vSAN and Intel Xeon processors.
- Increase workload scalability without sacrificing performance and latency of both traditional and modern workloads.
- Support the higher performance requirements of AI-enabled applications.

In light of ever-shrinking budgets, organizations are constantly facing the challenge of delivering an IT infrastructure that will support the performance and scalability requirements across an increasing mix of traditional and modern workloads and applications. With VMware vSAN 8 ESA operating on hosts running Intel further generation Xeon Scalable processors, organizations can continue to reap the benefits of HCI while meeting the application performance and scalability end users expect. With that in mind, Enterprise Strategy Group suggests that organizations look closely at this technology combination if facing the aforementioned challenges.



©TechTarget, Inc. or its subsidiaries. All rights reserved. TechTarget, and the TechTarget logo, are trademarks or registered trademarks of TechTarget, Inc. and are registered in jurisdictions worldwide. Other product and service names and logos, including for BrightTALK, Xtelligent, and the Enterprise Strategy Group might be trademarks of TechTarget or its subsidiaries. All other trademarks, logos and brand names are the property of their respective owners.

Information contained in this publication has been obtained by sources TechTarget considers to be reliable but is not warranted by TechTarget. This publication may contain opinions of TechTarget, which are subject to change. This publication may include forecasts, projections, and other predictive statements that represent TechTarget's assumptions and expectations in light of currently available information. These forecasts are based on industry trends and involve variables and uncertainties. Consequently, TechTarget makes no warranty as to the accuracy of specific forecasts, projections or predictive statements contained herein.

Any reproduction or redistribution of this publication, in whole or in part, whether in hard-copy format, electronically, or otherwise to persons not authorized to receive it, without the express consent of TechTarget, is in violation of U.S. copyright law and will be subject to an action for civil damages and, if applicable, criminal prosecution. Should you have any questions, please contact Client Relations at cr@esg-global.com.

About Enterprise Strategy Group

TechTarget's Enterprise Strategy Group provides focused and actionable market intelligence, demand-side research, analyst advisory services, GTM strategy guidance, solution validations, and custom content supporting enterprise technology buying and selling.

 contact@esg-global.com
 www.esg-global.com