Technical Validation

Mission-critical Workload Performance Testing of Different Hyperconverged Approaches on Cisco HyperFlex

By Alex Arcilla, Senior Validation Analyst

May 2022

This ESG Technical Validation was commissioned by Cisco Systems, Inc. and is distributed under license from TechTarget, Inc.

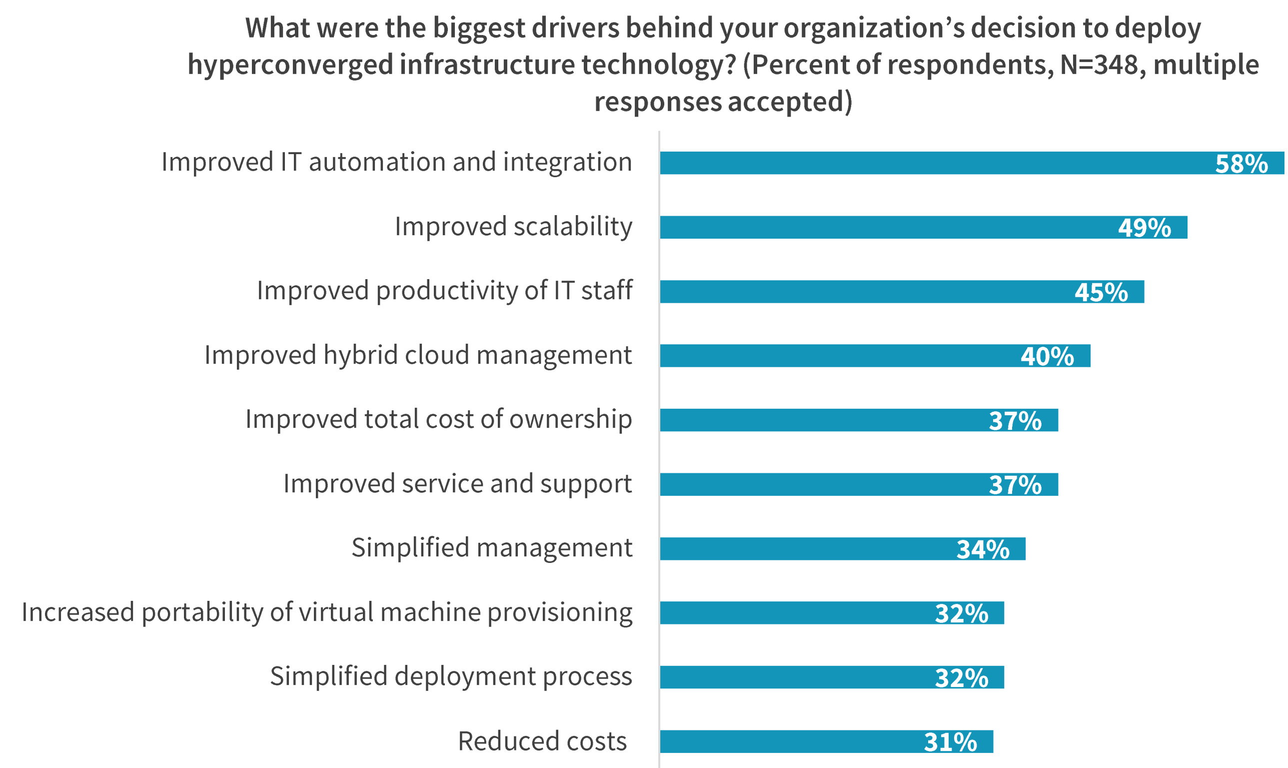
Introduction

This ESG Validation audits performance testing results for both scaling existing workloads and mixing different workloads on Cisco HyperFlex hyperconverged infrastructure (HCI). We compare the fully engineered appliance versions of the Cisco HyperFlex solution on Cisco UCS against two software-only HCI offerings from leading vendors. Both software-based offerings were independently validated to run on Cisco UCS hardware servicing mission-critical workloads.

Background

Every day, new workloads are spun up and existing workloads are scaled to meet evolving business needs. To support the additional work, many organizations have increased their use of hyperconverged infrastructure (HCI). In fact, ESG research has shown that 64% of organizations currently use HCI to support 21-50% of their production applications.[1] When asked what the biggest drivers were behind their organization's decision to deploy HCI, almost half (49%) of respondents noted that improved scalability was a top priority.

Figure 1. Top Ten Drivers for Deploying HCI Technology



Source: ESG, a division of TechTarget, Inc.

As organizations scale both the size and mix of workloads on HCI, the issue of performance arises. Powering complex, business-critical workloads can expose architectural deficiencies in an HCI solution not optimized to handle the workload requirements. An HCI platform deployed to support a tier-1 workload needs to not only provide high IOPS and low read/write latency, but it also needs to do so in a consistent, predictable manner. Predictable performance and low VM performance variability are critical to maximize end-user productivity across an organization.

[1] Source: ESG Survey Results, *Hyperconverged Infrastructure 2.0*, October 2021. All ESG research references and charts in this ESG Technical Validation have been taken from this survey results set, unless otherwise noted.

Key Metrics to Consider when Evaluating HCI Solutions

Simplicity is no longer the only priority; as adoption of latency-sensitive mission-critical workloads continues to grow, performance needs to be included as a key buying criterion for HCI solutions to enable the next generation of HCI-powered workloads. While first-generation HCI architecture—consisting of software running on x86 servers connected through commodity grade switches—worked for early use cases, the mission-critical nature of tier-1 workloads requires a solution that can deliver trusted performance.

Input/output operations per second (IOPS)—Adoption of flash-based storage has greatly reduced I/O challenges in traditional shared-storage environments, but in a clustered environment like HCI, total IOPS can vary depending on the network connection between nodes as well as the software layer powering the HCI solution. For HCI deployments, it is important to evaluate both the total number of IOPS delivered by the cluster as well as the IOPS consistency that is delivered. Consistent VM performance has been a challenge since the beginning of virtualized computing, but “noisy neighbor” VM performance can be even more pronounced with HCI deployments based on how the software layer writes data across the cluster.

Latency—While IOPS are an important performance indicator, latency as it relates to the application should also be considered when purchasing an HCI solution. Clustered environments like HCI can have multiple bottlenecks like storage performance, responsiveness, and network throughput, all of which can contribute to application latency. Increased latency means decreased responsiveness of applications for users.

- **Read latency—**This is the time required for the storage controller to find and deliver the requested user data. For flash storage, as evaluated in this paper, this includes the time for the flash subsystem to find the required data blocks and prepare to transfer them, and the transit time through the network.
- **Write latency—**This is the time it takes for the storage controller to perform all the activities required to write data blocks, including determination of the proper location for the data and performance of overhead activities (block erase, copy, and “garbage collection”), then creating persistent copies and acknowledging the write back to the host, VM, or container.
- **Overall latency—**Total latency is simply a combination of the read and write latencies calculated using the ratio of reads and writes used by the application. For example, for a workload that consists of 70% reads and 30% writes, the total latency is the average of the read and write results, weighted according to the percentage of each.

Industry Approaches to HCI—Software-validated versus Fully Engineered Appliance

HCI was conceived as the next step in the evolution of the modular data center concept. The goal was to simplify rack-level converged infrastructure (CI) to node-level deployments. Rather than three tiers of infrastructure managed through a common software platform, HCI combines virtualized compute and software-defined storage integrated through the software layer and deployed on a single chassis to create a node. Nodes are connected through network switches to form a shared pool of resources that can be scaled on demand by adding a new node to the cluster. Vendors have taken distinct approaches to bring HCI solutions to market, and all of them should be considered.

HCI Distribution Models:

Software-only HCI: This model focuses on the software layer used to integrate compute and storage into a single node. Users purchase the HCI software, which can either be installed in-house or by a third party on industry-standard servers. Initial HCI deployments tended to support tier-2 or tier-3 workloads, so it was common for the software to be deployed on off-the-shelf servers and connected through commodity switches to keep costs low. As HCI has matured and more critical workloads are deployed, organizations have been demanding that HCI run on trusted hardware platforms.

It is important to note that not all hardware manufacturers carry the same validations, so it would be wise for potential users to read the fine print for these types of deployments. Deploying software on a hardware platform that does not carry validation from all parties can open the door to finger pointing and add a degree of risk that might be more than some organizations are willing to accept.

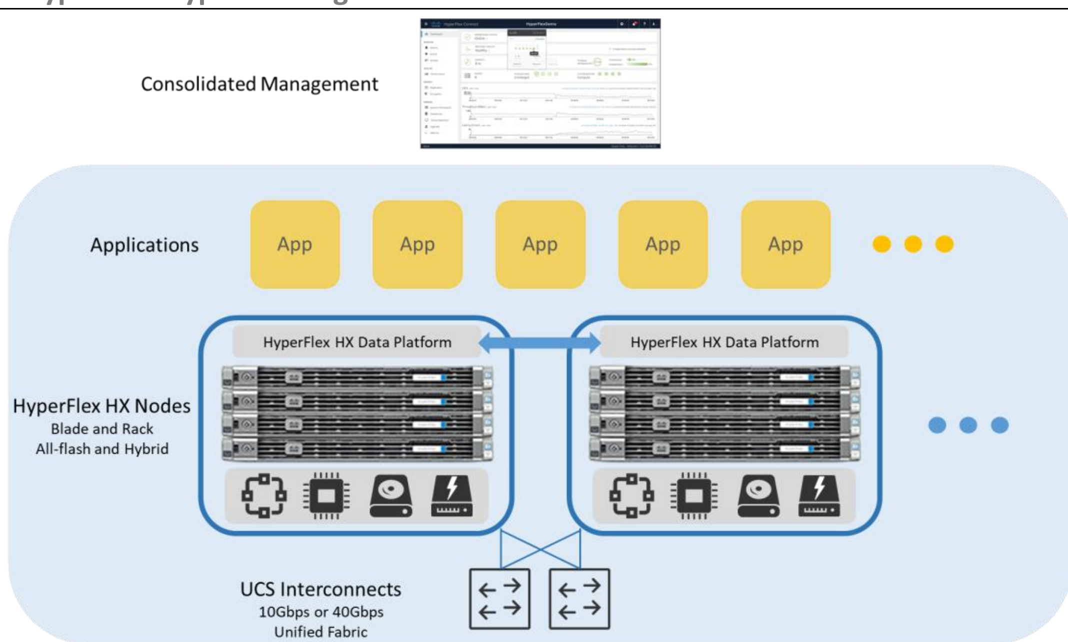
Fully engineered HCI: This model is designed to provide the simplest way to deploy an HCI solution. Predominantly sold by tier-1 vendors, users selecting this approach get appliances built on trusted hardware platforms that are shipped with the software preinstalled. Vendors choose this approach because it enables them to engineer and optimize every component of HCI across computing, networking, and storage into a simple appliance offering. Fully engineered HCI solutions remove a level of risk by ensuring users are delivered factory-validated configurations whose operation and support are guaranteed by a single source for computing, networking, and storage software. It is important to note that some appliances are created through partnerships between hardware and software vendors, and the level of optimization between the hardware and software layers can vary and affect overall performance.

Cisco's Fully Engineered Appliance Versions of HyperFlex on UCS

Cisco HyperFlex is a fully engineered hyperconverged system that combines compute and software-defined storage as well as fully integrated networking optimized for the east-west traffic flow between nodes in an HCI platform. This fully integrated platform is designed to scale resources independently and deliver consistent high performance. Cisco HyperFlex is engineered on Cisco UCS, combining the benefits of the UCS platform (such as policy-based automation for servers and networking) with those of the HX Data Platform's distributed file system for hyperconvergence.

It supports end-to-end workloads from mission-critical core data center applications to remote locations. HX offers a range of hybrid, all-flash, and all-NVMe nodes in order to support a broad range of mission-critical workload. HyperFlex deployments require a minimum two-node cluster in data center deployments for high availability, with data replicated across multiple nodes for data protection purposes.

Figure 2. Cisco HyperFlex Hyperconverged Infrastructure



Source: ESG, a division of TechTarget, inc.

HyperFlex HX-Series Nodes are engineered on the Cisco UCS platform and powered by the latest generation of Intel Xeon Scalable processors, as well as the third-generation AMD EPYC processors. Nodes were comprised of the following:

- **Cisco HyperFlex HX Data Platform.** The core of any HCI solution is the software platform, and the HX Data Platform was engineered specifically for HCI software-defined storage. Operating as a controller on each node, the HX Data Platform is a high-performance, distributed file system that combines all SSD and HDD capacity across the cluster into a distributed, multi-tier, object-based data store, striping data evenly across the cluster. It also delivers enterprise data services such as snapshots, thin provisioning, and instant clones. Policy-based data replication across the cluster ensures high availability. Dynamic data placement in memory, cache, and capacity tiers optimize application performance, while inline, always-on deduplication and compression optimize capacity.
    - The HX Data Platform handles all read and write requests for volumes accessed by the hypervisor. By striping data evenly across the cluster, network and storage hotspots are avoided, and VMs enjoy optimal I/O performance regardless of location. Writes go to local SSD or NVMe cache and are replicated to remote SSDs in parallel before the write is acknowledged. Reads are from cache if possible or retrieved from remote SSDs.
    - The log-structured file system is a distributed object store that uses a configurable SSD or NVMe cache to speed reads and writes, with capacity in HDD (hybrid), SSD (all-flash), or all-NVMe persistent tiers. When data is de-staged to persistent tiers, a single sequential operation writes data to enhance performance. Inline deduplication and compression occur when data is de-staged; data is moved after the write is acknowledged so there is no performance impact.
    - Cisco's partnerships with Cohesity and Veeam enable customers with data protection capabilities to address secondary data requirements, including target storage, backup, replication, and disaster recovery, so as to meet stringent RTOs and RPOs.
- **Cisco UCS compute-only nodes**. Both UCS blade and rack servers can be combined in the cluster, with a single network hop between any two nodes for maximum east-west bandwidth and low latency. HyperFlex lets you alter the ratio of CPU-intensive blades or servers—compute nodes—to storage-intensive capacity nodes—HX nodes—so users can optimize the system as application needs shift.
- **Cisco Unified Fabric—UCS 6200/6300/6400 Fabric Interconnects** enable software-defined networking. High bandwidth, low latency, and 10/25/40Gbps connectivity in the fabric enable high availability as data is securely distributed and replicated across the cluster. The network enables HX clusters to scale easily and securely. The single hop architecture is designed to maximize the efficiency of the storage software to enhance overall cluster performance.
- **Cisco Application Centric Infrastructure (ACI)** for automated provisioning. ACI enables automation of network deployment, application services, security polices, and workload placement per defined service profiles. This provides faster, more accurate, more secure, lower cost deployments. ACI automatically routes traffic to optimize performance and resource utilization and reroutes traffic around hotspots for optimal performance.
- **Choice of industry-leading hypervisors, including VMware ESXi and vCenter, Microsoft Hyper-V, and Cisco Intersight Workload Engine IWE).** Cisco IWE enables enabling container-native virtualization for Cisco HyperFlex.

To consolidate management of Cisco HyperFlex nodes, organizations can leverage Cisco Intersight. Delivered as SaaS, Cisco Intersight provides correlated visibility and management from bare metal servers, hypervisors, containers, and serverless and application components. Users can manage all their Cisco HyperFlex and Cisco Unified Computing System (Cisco UCS) infrastructure including traditional, hyperconverged, edge, and remote/branch offices through a single cloud-based GUI.

Cisco HyperFlex delivers numerous benefits, including:

- **High performance.** In addition to performance features mentioned above, HyperFlex Dynamic Data Distribution securely and evenly distributes data across all cluster nodes to reduce bottlenecks.
- **Fast, easy deployment**. This pre-integrated cluster can be deployed just by plugging into the network and applying power. Node configuration and connection is handled through Cisco UCS service profiles. Cisco says that customers report typical deployment times of less than one hour.
- **Consolidated management**. Systems are monitored and managed through Cisco Intersight, VMware vCenter plugin, or Cisco HyperFlex Connect, which eliminates separate management silos for compute and storage. HyperFlex Connect lets organizations manage and monitor clusters from anywhere and at any time with metrics and trends to support the entire management lifecycle.
- **Independent scaling**. Different from other HCI systems, HyperFlex can independently scale compute and storage resources without the need to add full nodes to the cluster. Users can easily incorporate compute-only nodes with bare UCS servers through the Fabric Interconnects to add additional compute to the cluster, or, if more storage is needed, add individual drives to each node; data is automatically rebalanced. This provides the right resources for different application needs, instead of scaling in predefined node increments that also add additional software licensing costs.

ESG Technical Validation

Testing was conducted using industry-standard tools and methodologies and was focused on comparing the performance of Cisco's fully engineered HyperFlex HCI appliance configurations against software-based HCI solutions. The bulk of the testing used HCIBench and HXBench, tools designed to test the performance of HCI clusters running virtual machines. Both tools leverage the industry standard Vdbench storage benchmark tool[2] and automate the end-to-end process that includes deploying test VMs, coordinating workload runs, aggregating test results, and collecting data.

This extensive testing was executed using a stringent methodology, including many months of baselining and iterative testing. While it is often easier to generate good performance numbers with a short test, benchmarks were run for long periods of time to observe performance as it would occur in a customer's environment.

In addition, tests were run many times, never back to back but separated by days and weeks, and the results averaged. These efforts add credibility by reducing the chances that results were influenced by chance circumstances. Also, testing was conducted using data sets large enough to ensure that data did not remain in cache but leveraged the back-end storage across each cluster.[3]

Mission-critical Workload Testing

The test bed was comprised of a four-node HyperFlex HX220C M5 version 4.5.2a cluster and two additional four-node Cisco UCS C220 M5 rack servers for the other two HCI solutions. Configuration details are listed in Table 1.

[2] For more information about Vdbench, please go to this link.
[3] When evaluating technology solutions, customers would be wise to understand the details behind vendor testing. Timing of test runs, volumes of data, and other details will impact performance results; these results may or may not be relevant to the customer environment.

Table 1. Tested HCI Configurations[4]

Platform	Nodes	Processors/Cores Per Node	RAM Per Node	Cache Per Node	Storage Capacity Per Node	Hypervisor
Cisco HyperFlex – Fully Engineered HCI with Cisco UCS	Four	2x CPU 6138, 20 Cores	256GB	1 X 1.6TB NVMe drive	8 X 960GB SSD Samsung Enterprise Value	VMware vSphere 7.0U2a
Vendor A Software-only HCI Validated on Cisco UCS	Four	2x CPU 6138, 20 Cores	384GB	2 X 1.6 TB NVMe drive	8 X 960GB SSD Samsung Enterprise Value	VMware vSphere 7.0U2a
Vendor B Software-only HCI Validated on Cisco UCS	Four	2x CPU 6138, 20 Cores	384GB[5]	Note[6]	8 X 800GB SSD Toshiba Enterprise Performance	VMware vSphere 7.0U2a

The mixed workload tests were run with 140 total VMs (35 VMs per node), each with 4 vCPUs, 4 GB RAM, and one 40GB disk and running RHEL version 7.2. The working set size was 5.6 TB. Tests were run for a minimum of one hour and up to five hours, with a five-minute ramp-up before each test and a minimum one-hour cool-down between tests.

Before every test was run, each VM was primed with written data by the test tool. This ensures that the test is reading "real" data and writing over existing blocks and not simply returning null or zero values directly from memory. This happens when data is not primed, so it is an important step to ensure that the test accurately reflects how data is read and written in an application environment. Priming of this large working set can take many hours to complete but is a wise investment in time to get more accurate performance results.

Testing was performed using the I/O profiles of benchmarks designed to emulate standard virtualization workloads, with 100% random data access. VMs by nature generate random I/O by combining I/O from multiple applications and workloads. It is important to note that all tests were run with compression and deduplication active on the Cisco HX cluster and the software-only HCI clusters.

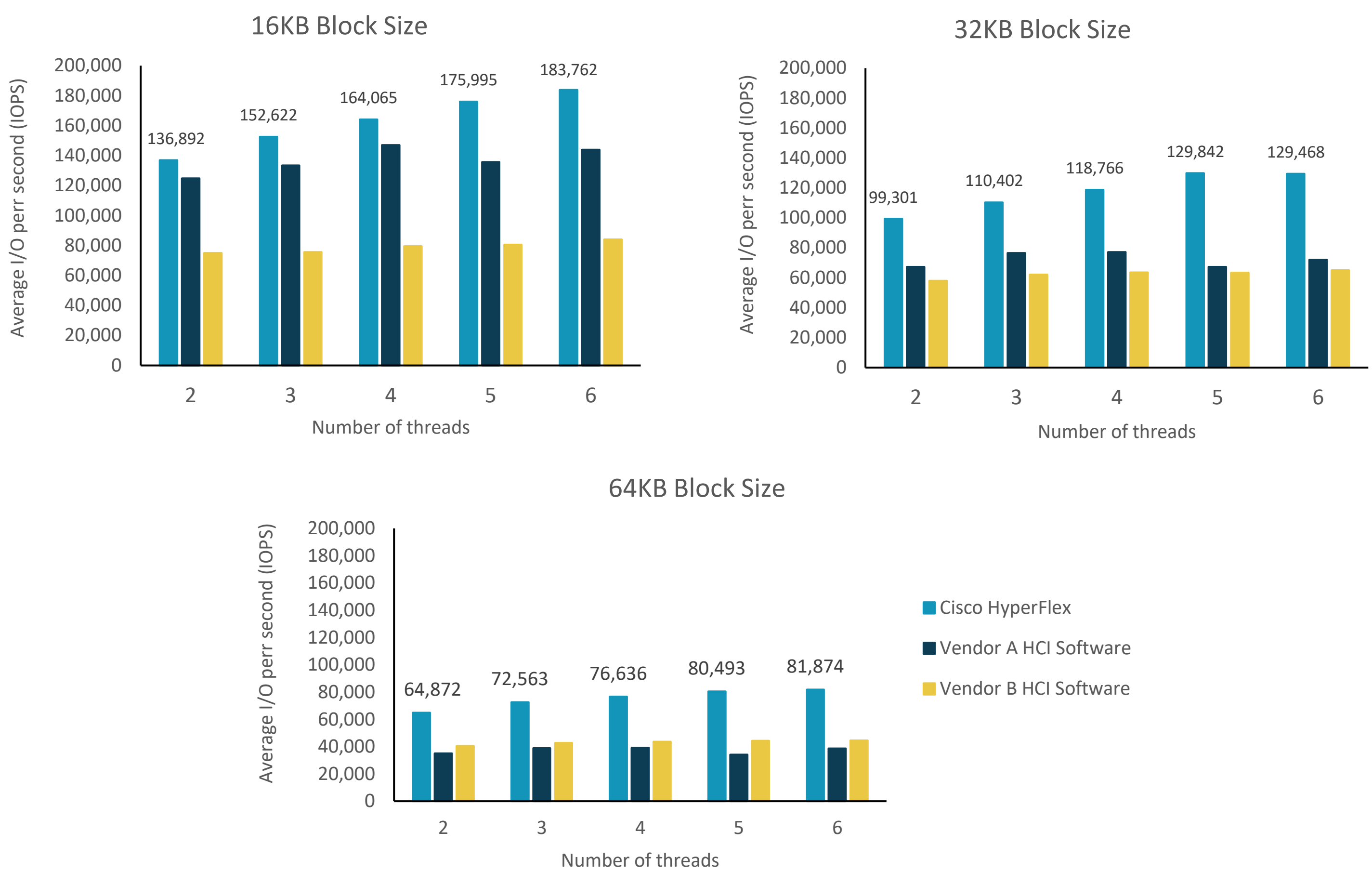
ESG Testing

First, ESG looked at a workload designed to emulate an application in which block size doubled over time. Vdbench was used to create a workload that exercised different transfer sizes and read/write ratios. In the Vdbench profile, the deduplication ratio was set to 2 with a unit size of 4 KB and the compressibility ration also set to 2. The test was run with a total of 140 virtual machines.

We first audited test results documenting average IOPS and latency for simulated 70% read/30% write workloads. During this test, we doubled the block size—from 8 KB to 64 KB. For each block size, we also increased the number of threads from two to six per VM to simulate an increase in cluster utilization. Total duration of the test was four hours, including one hour testing with 1.5 - 3 hours priming, depending on vendor. Results are displayed in Figure 3.

[4] All vendor configurations were equipped with the most recent software releases at the time testing occurred.
[5] The amount of CPU resources and memory available had no measurable impact on performance across vendors. CPU and memory resource utilization on each node for all vendors was far below available capacity.
[6] Note: Vendor validated configuration requires all Enterprise Performance SSDs only, no cache.

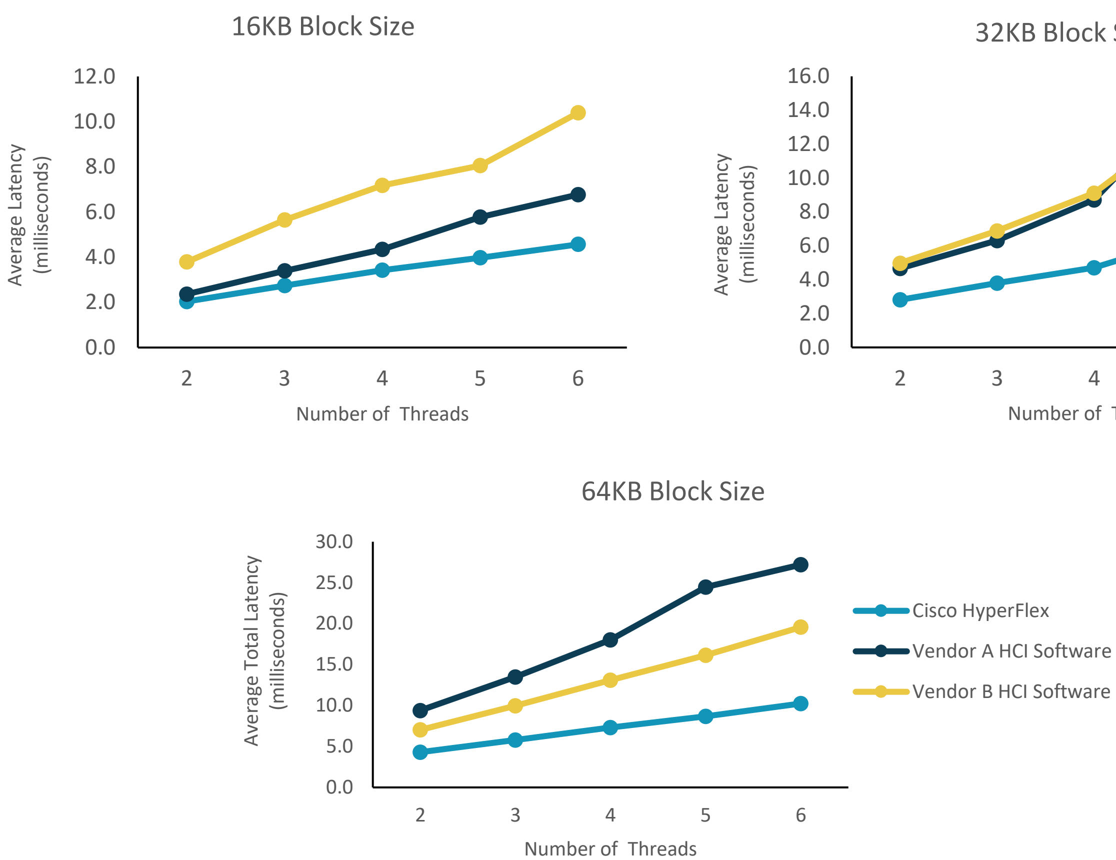
Figure 3. Average IOPS for 70R/30W Workload as Number of Threads Increased (Taller Is Better)



Source: ESG, a division of TechTarget, Inc.

We also examined test results showing the average total latency achieved with each solution as the number of threads increased for our 70% read/30% write workload. Results are displayed in Figure 4.

Figure 4. Average Total Latency for 70R/30W Workload as Number of Threads Increased (Lower Is Better)



Source: ESG, a division of TechTarget, Inc.

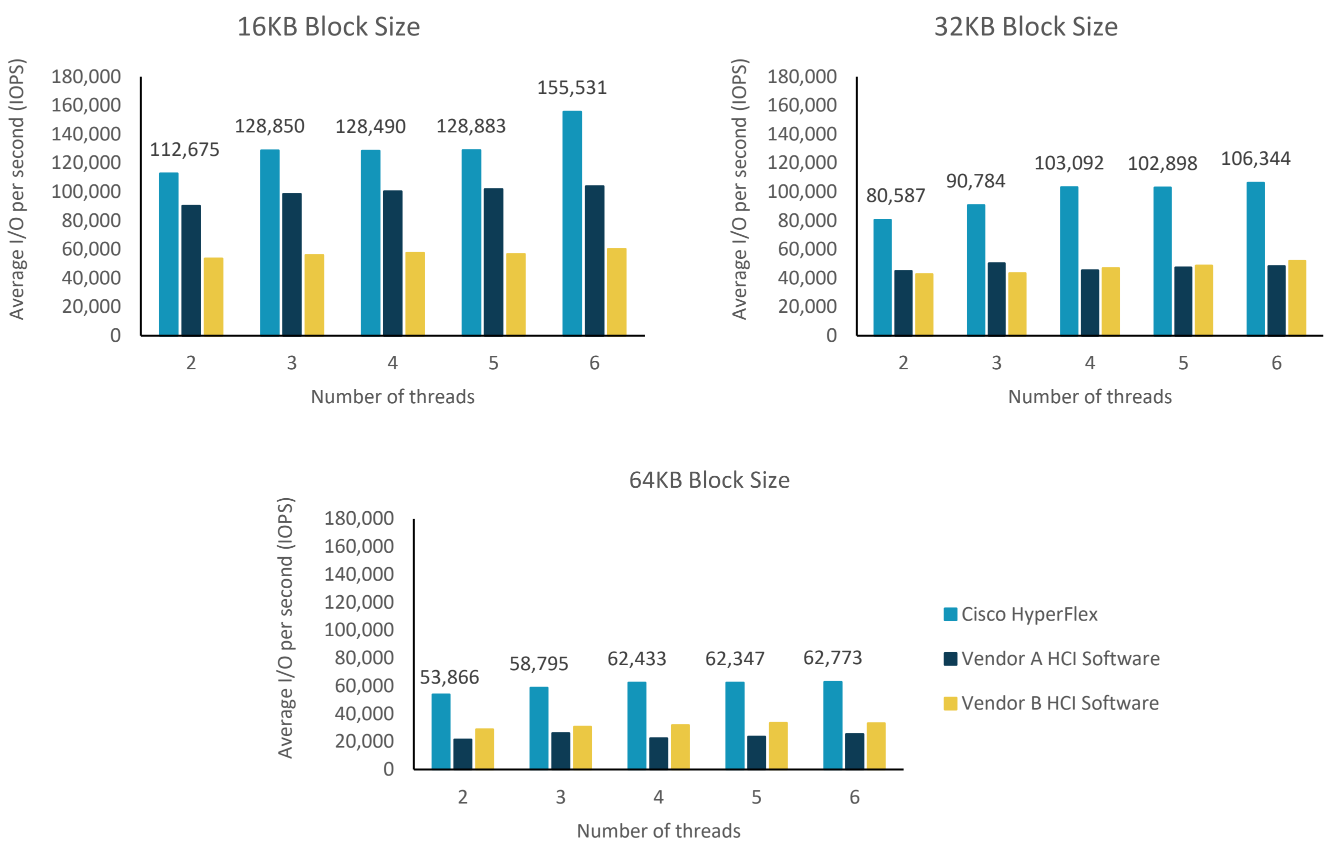
What the Numbers Mean

- Over the course of this four-hour test, ESG observed that average IOPS achieved by Cisco HyperFlex were consistently higher for 70% read/30% write workloads, across 16 – 64 KB block sizes, than those achieved by the two software-based HCI solutions. We also noted that average IOPs with Cisco HyperFlex increased as the number of threads increased for 16, 32, and 64 KB block sizes.
- Regardless of HCI solution, ESG also noted that as block size increased, average IOPS over all thread counts decreased. That is expected, as the amount of data being retrieved increased as block size doubled. However, Cisco Hyperflex continued to outperform the software-only HCI solutions. In fact, as the block size increased, the improvement in average IOPs achieved by Cisco HyperFlex increased. Improvements ranged from 19 - 58%.
- ESG also reviewed test results of average IOPs achieved with an 8KB workload as thread count increased. For this block size, the improvement in average IOPS achieved by Cisco HyperFlex over the software-based HCI solutions was not observed over all thread counts.
- Conversely, ESG also observed that average latencies achieved with Cisco HyperFlex were consistently lower, over all thread counts, than those achieved by either software-only HCI stack.

- While average total latency increased as block size doubled, ESG still observed that Cisco HyperFlex consistently achieved lower latencies than those observed by the software-only HCI solutions. Reduction in average latency ranged from 16% – 62% across 16 – 64 KB block sizes.

ESG then audited test results documenting average IOPS and latency for simulated 50% read/50% write workloads. During this test, we doubled the block size—from 8 KB to 64 KB. For each block size, we also increased the number of threads from two to six to simulate an increase in cluster utilization. Results are displayed in Figure 5.

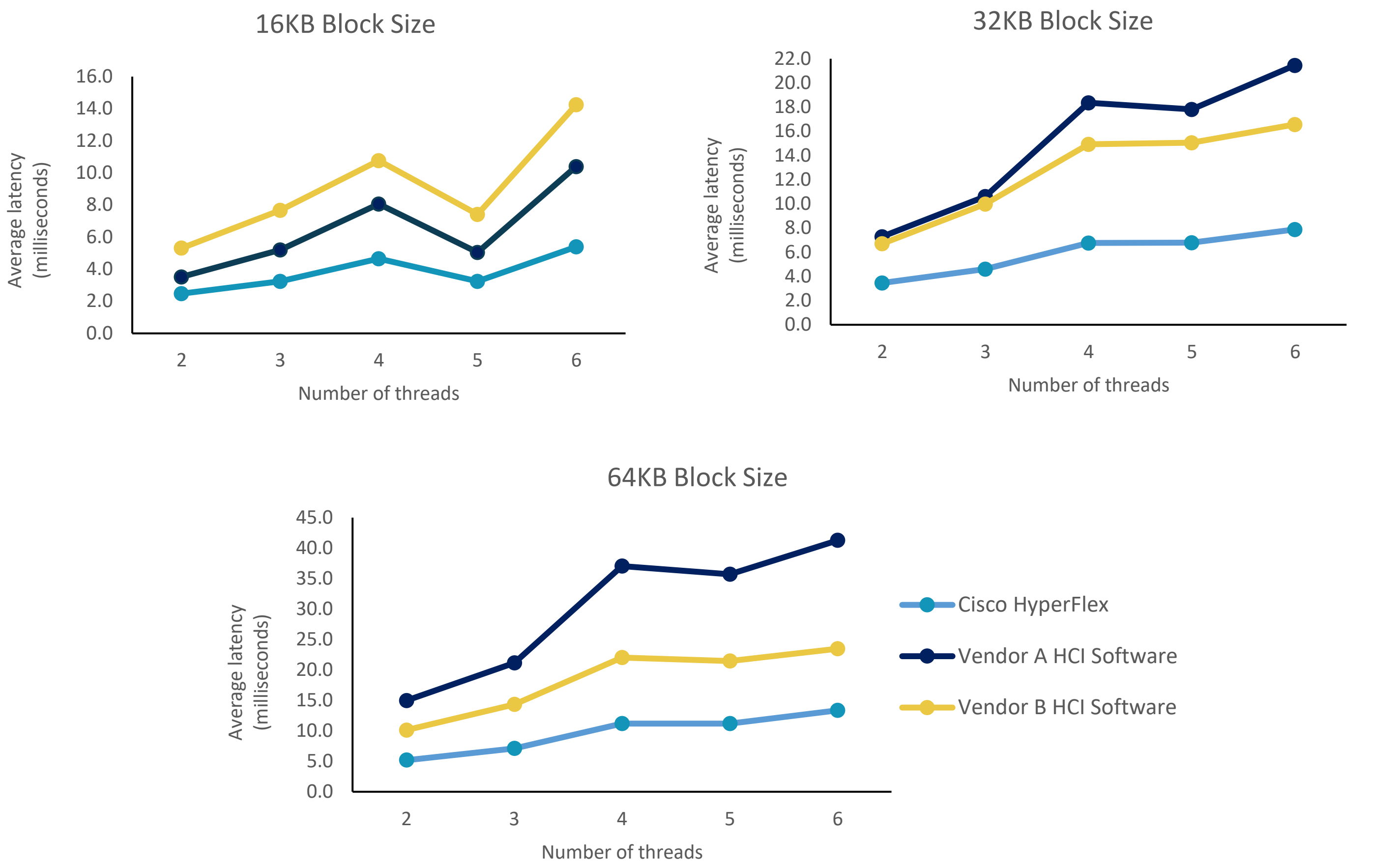
Figure 5. Average IOPS for 50R/50W Workload as Number of Threads Increased (Taller Is Better)



Source: ESG, a division of TechTarget, Inc.

We also examined test results showing the average total latency achieved with each solution as the number of threads increased for our 50% read/50% write workload. Results are displayed in Figure 6.

Figure 6. Average Total Latency for 50R/50W Workload as Number of Threads Increased (Lower Is Better)



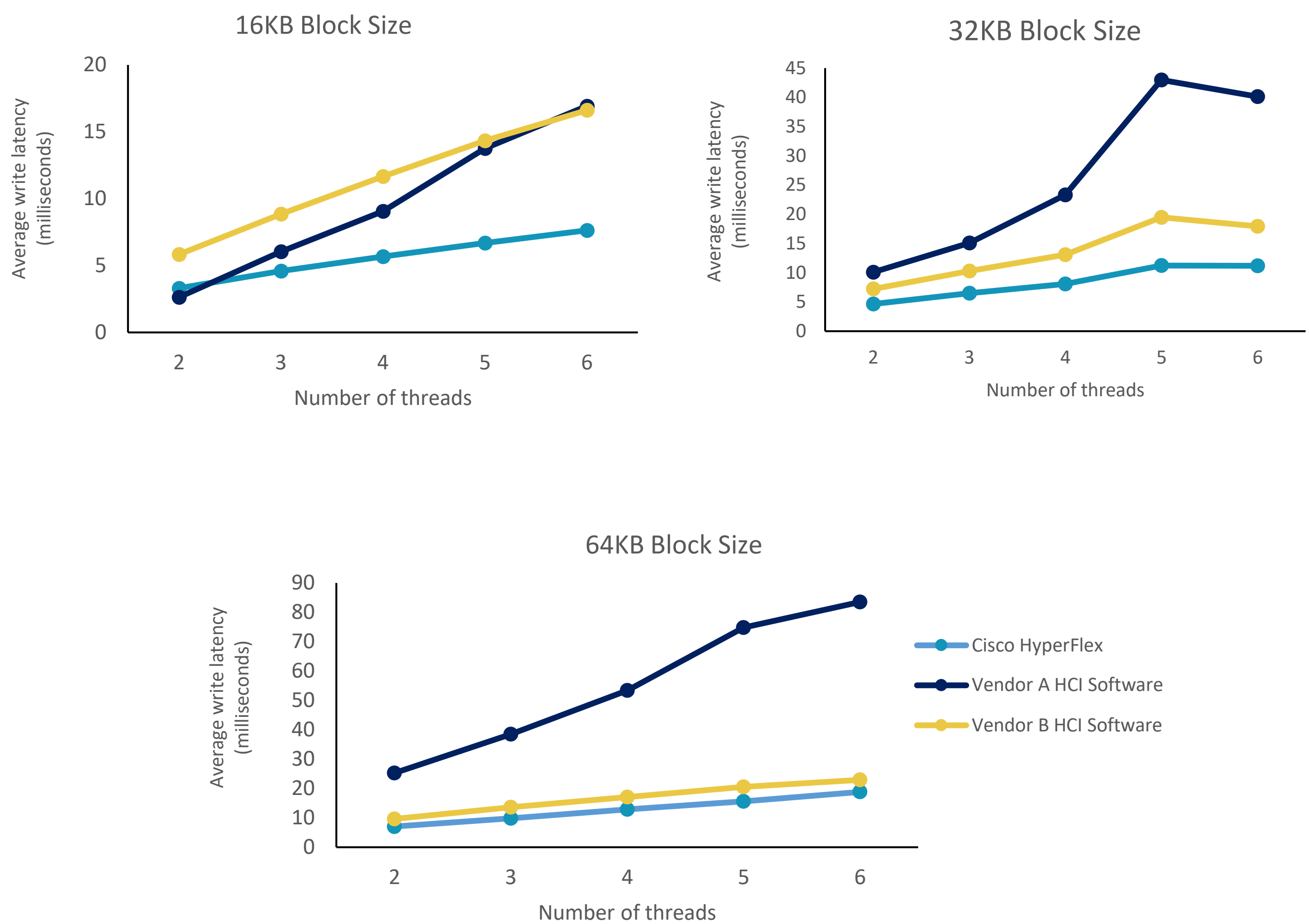
Source: ESG, a division of TechTarget, Inc.

ESG's observations of both average IOPS and total latency for 50% read/50% write workloads were similar to those made with 70% read/30% workloads.

ESG then specifically reviewed write latencies associated with these tests (see Figure 7). Unlike times associated with read latencies, ESG saw that the differences between write latencies observed on Cisco HyperFlex and the competitive offerings were larger. We specifically observed that these differences actually increased as block size increased, across all thread counts.[7] Write latencies achieved with Cisco HyperFlex were observed to be between 24% and 79% lower compared to Vendor A and between 18% and 54% compared to Vendor B. For organizations that are deploying more write-intensive workloads (e.g., video content creation), ESG noted that Cisco HyperFlex can offer the write performance these workloads demand.

[7] For tests run with an 8KB block size, Cisco HyperFlex did not consistently exhibit lower write latencies compared with Vendor A and Vendor B's software-only HCI stacks as thread count increased.

Figure 7. Average Write Latency for 50R/50W Workload as Number of Threads Increased (Lower Is Better)



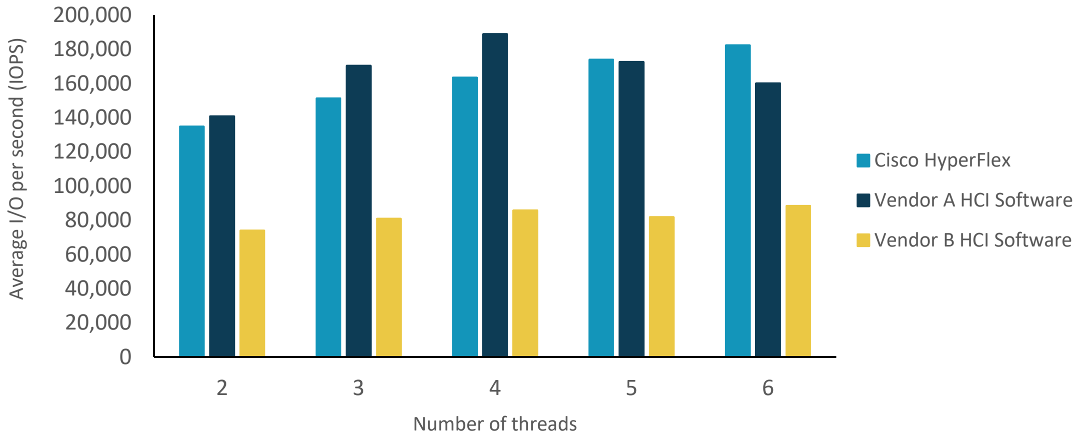
Source: ESG, a division of TechTarget, Inc.

Next, we simulated a mixed workload designed to emulate a virtualized environment with multiple VMs running different applications. Vdbench was used to create a 70% read/30% write workload that emulated the "blender effect" of running different workloads with varying transfer sizes, resulting in an average block size of 13 KB.

Tests were run using HCIBench against 140 VMs in each cluster—35 per node, emulating a mixed workload environment with virtual machines running a variety of applications. In the Vdbench profile, the deduplication ratio was set to 2, with a unit size of 4 KB and the compressibility ratio also set to 2.

For these tests, ESG also considered the effect of increasing the number of threads as well as limiting the cluster utilization. Multiple test runs were completed using two to six threads at cluster utilization ranging from 20 – 100% (increasing in increments of 20%). We then averaged the results obtained (from multiple runs for each combination of thread count and cluster utilization) to show both IOPS and latency for each thread count as cluster utilization increased. Results for average IOPs for 70% read/30% write workloads, at 100% cluster utilization, are shown in Figure 8.

Figure 8. Average IOPS per Thread Count for 70R/30W Workload at 100% Cluster Utilization[8]



Source: ESG, a division of TechTarget, Inc.

While ESG noted that Cisco Hyperflex performed as well or better than the software-only HCI solutions at 100% utilization, organizations will typically run HCI at less than 100% utilization in production networks. However, comparing how vendors perform at lower utilization rates is not a fair comparison. The interaction of hardware and software creates many variables that affect performance. To better understand how well each solution can perform at lower cluster utilization rates, ESG calculated the percentage increase between latencies as rates increased from 20% to 100% (see Figure 9).

[8] ESG observed similar results as cluster utilization increased from 20% to 80%.

Figure 9. Average Total Latency for 70R/30W Workload per Thread Count as Utilization Rate Increases

Source: ESG, a division of TechTarget, Inc.

What the Numbers Mean

- For 70% read/30% write workloads with an average block size of 13 KB, ESG observed that average IOPS achieved by Cisco HyperFlex were consistently on par or slightly lower compared to Vendor A's HCI software, regardless of thread count and utilization rate. On the other hand, average IOPS achieved by Cisco HyperFlex were consistently higher than those achieved by Vendor B's HCI software.

- ESG also noted that the average latency observed for our 70% read/30% write workload at 100% cluster utilization was either approximately the same or higher for lower threads (two to four), than latencies observed with Vendor A's HCI software. Again, average latency achieved by Cisco HyperFlex was consistently lower than that achieved by Vendor B's HCI software.

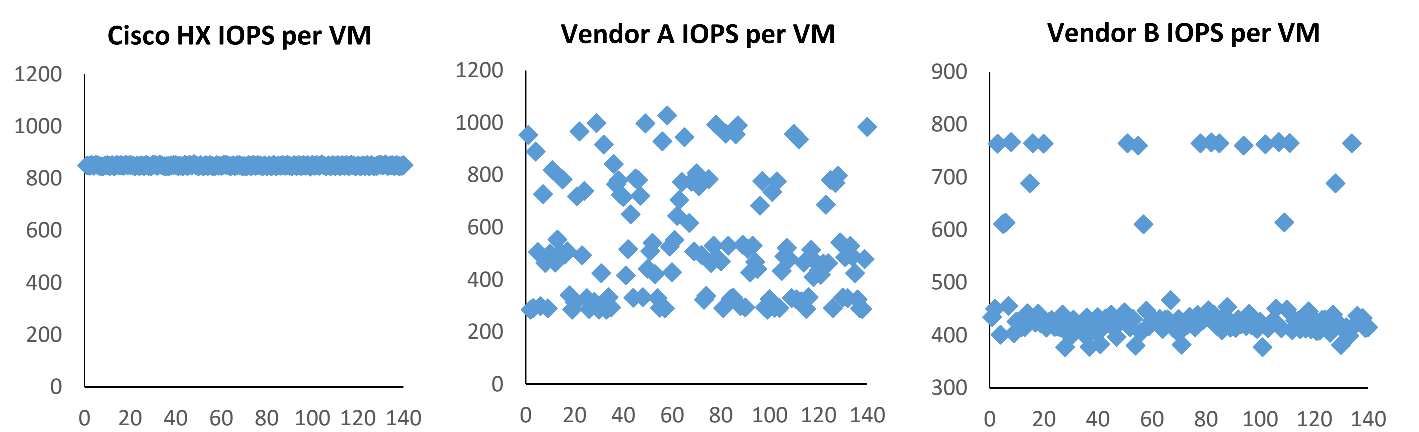
- While average latency observed on Cisco HyperFlex at 100% utilization was approximately equal or lower compared to that observed with Vendor A at up to four threads, latency achieved with Vendor A increased compared to Cisco HyperFlex at five and six threads.

- As we reviewed the percentage changes in total latency as cluster utilization increased, ESG found that the change decreased steadily compared to the software-based HCI stacks. Vendors A and B experienced much higher percentage changes in latency as the workload utilization increased from 60% to 100%. ESG noted that running mixed workloads using solutions by Vendors A and B would experience dramatic changes in latency as cluster utilization increased. On the other hand, percentage changes in latency observed with Cisco HyperFlex steadily decreased when we increased cluster utilization. We could attribute this to the design of the Cisco distributed log structured file subsystem.

- ESG also noted that the average block size would be greater than 13KB as organizations add more larger workloads to Cisco Hyperflex. Based on observed results, we expect that performance achieved with Cisco HyperFlex would continue to increase in a linear fashion as more pressure is placed on the file subsystem.

We finally reviewed the variability in performance from VM to VM for a 70% read workload, increasing both block size and number of threads. Software-only HCI vendors A and B both showed considerable variability in performance from VM to VM. To illustrate, Figure 10 compares the IOPS variability associated with a 32KB workload and four threads. While Cisco HyperFlex showed minor variation across all 140 VMs—aggregate testing IOPS stayed close to the target of 800. However, both Vendors A and B exhibited variability over a wide range.

Figure 10. 32KB Workload, 70% Read, 100% Random, Four Threads—140 Virtual Machines



It is important to note that this variability was observed in every iteration of testing, across all block sizes and thread counts. no form of storage QoS was used during these test runs on any of the clusters. Network QoS was used for all systems.

Inconsistency like this could be quite problematic for administrators, who would likely need to use some form of QoS (if available from the HCI vendor) to attempt to control the VMs that are consuming more than their share of resources so others are not starved.

Why This Matters

To simplify their IT environments and lower overall costs within their data center infrastructure, organizations continue to deploy HCI. However, as with other data center purchases, they are also considering its scalability to handle larger and more complex workloads. In fact, ESG research uncovered that 57% of survey respondents consider scalability an important, if not critical, data center infrastructure purchase criterion.[9] Yet, organizations must also consider how adding these workloads will affect overall performance and their ability to fulfill business needs without unnecessary delay.

ESG validated that Cisco HyperFlex delivered higher performance than other similarly configured HCI solutions. In the majority of cases, ESG verified that Cisco HyperFlex outpaced competitors in terms of IOPS and latency when we simulated both workloads with increasing block size and mixed workloads in light of increasing cluster utilization. We saw that it also offered more consistent, predictable performance per VM and per node than both software-based systems. This translates directly to lower upfront and ongoing costs because a given workload can potentially be serviced by a smaller number of Cisco HyperFlex nodes.

The Bigger Truth

Organizations are deploying more complex workloads that are accessing and processing larger amounts of data alongside traditional OLTP workloads. As the number of these applications grow, organizations face the challenge of deploying sufficient infrastructure without incurring additional capital and operational expenses. While HCI solutions can help to simplify IT environments and reduce costs, workload performance cannot be sacrificed.

Cisco HyperFlex provides the typical benefits of HCI: It is cost-effective and simple to manage and lets organizations start small and scale. Cisco HyperFlex provides the high performance and low latency that mission-critical, virtualized workloads demand. The *consistency* of performance over time and across all VMs in a cluster was particularly notable. In addition, its independent resource scalability enables organizations to adapt quickly to changing requirements, which today's environments demand.

Cisco HyperFlex HCI solutions have highly integrated, fully engineered appliance configurations, powered by the latest generation of Intel Xeon Scalable processors and third-generation AMD EPYC processors, that provide pre-integrated clusters that include the network fabric, data optimization, unified servers, and choice of hypervisor, including VMware ESXi/vSphere, Microsoft Hyper-V, and Cisco Intersight Workload Engine, enabling fast deployment. This makes them simple to manage and scale. ESG has previously validated that HyperFlex provides consistent high performance for VMware environments running mission-critical workloads, outpacing multiple competitive solutions with higher IOPS, lower latency, and better consistency over time and across VMs.

The test results presented in this report are based on applications and benchmarks deployed in a controlled environment with industry-standard testing tools. Due to the many variables in each production data center environment, capacity planning and testing in your own environment are recommended. While the methodology in these tests was more stringent than most, customers are well advised to always explore the details behind any vendor testing to understand the relevance to their environment.

When market evolution changes the buying criteria in an industry, there is often a mismatch between what customers want and what they can get. Vendors that can see what is missing and fill the void gain an advantage. Cisco delivers an HCI solution that provides the essential simplicity and cost-efficiency features of HCI, but also the consistent high performance

[9] Source: ESG Survey Results, *2021 Data Infrastructure Trends*, September 2021.

that has been missing—and that customers need for mission-critical workloads. HyperFlex supports VMware and Microsoft on-premises virtualized environments, and expansion to bare metal, containerized, and multi-cloud environments.

HCI solutions have been focused on second tier workloads, but the consistent high performance offered by Cisco HyperFlex All NVMe further validates HyperFlex as extremely well-suited to tier-1 production workloads. Organizations seeking cost-effective, scalable, high-performance infrastructure solutions for mission-critical workloads would be smart to take a close look at Cisco HyperFlex.