ESG Lab Validation

Performance and Cost Efficiency of Intel and Microsoft Hyperconverged Infrastructure

By Mike Leone and Kerry Dolan, Senior Validation Analysts

March 2017

This ESG Lab Report was commissioned by Intel and Microsoft and is distributed under license from ESG.

Contents

ESG Lab Reports

The goal of ESG Lab reports is to educate IT professionals about data center technology products for companies of all types and sizes. ESG Lab reports are not meant to replace the evaluation process that should be conducted before making purchasing decisions, but rather to provide insight into these emerging technologies. Our objective is to go over some of the more valuable feature/functions of products, show how they can be used to solve real customer problems and identify any areas needing improvement. ESG Lab's expert third-party perspective is based on our own hands-on testing as well as on interviews with customers who use these products in production environments.

Introduction

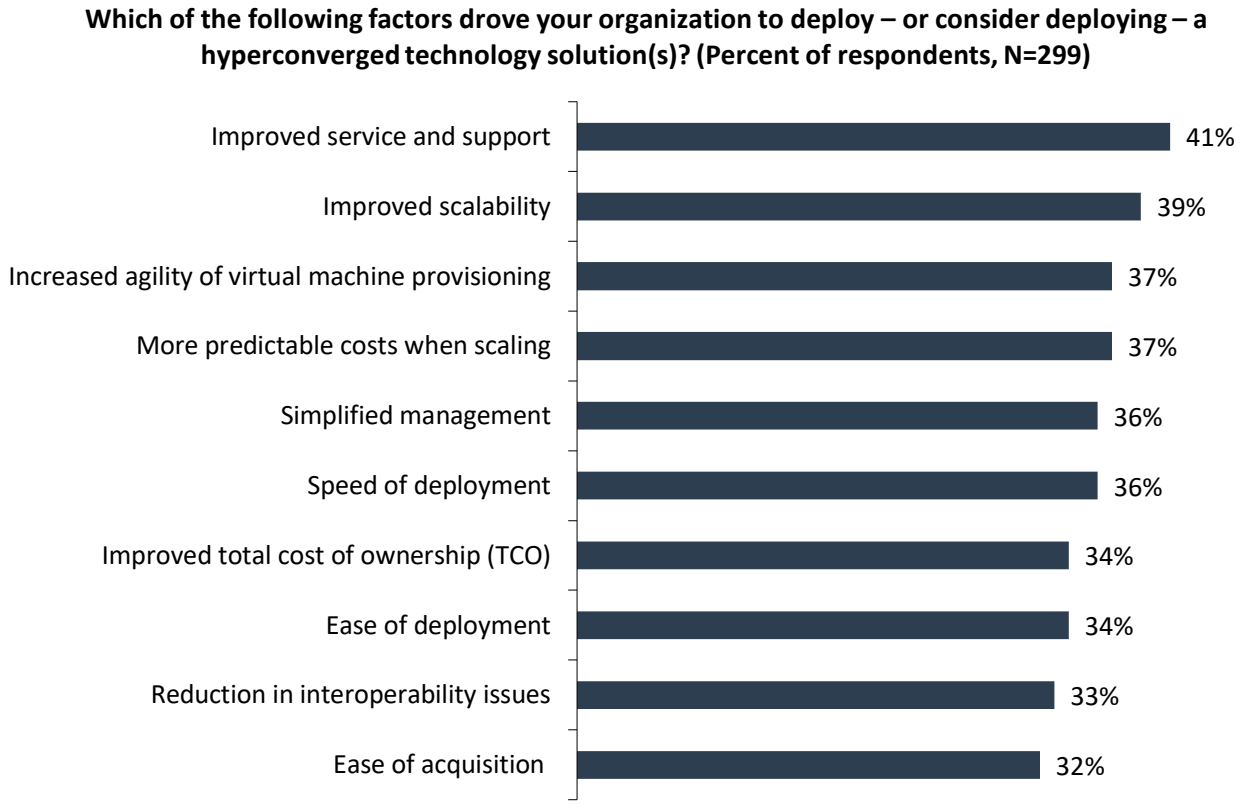
This ESG Lab Validation documents our audit of synthetic and real-world testing of hyperconverged infrastructure (HCI) reference architectures (RAs) delivered by Intel and Microsoft. The solutions include Intel-based servers and flash drives with Windows Server 2016 and Storage Spaces Direct. Testing focused on performance, scalability, and cost effectiveness.

Background

Businesses today must be agile while handling massive data growth. To accomplish this, they need easier management, simpler scalability, and cost efficiency. Traditional storage architectures limit organizations in all these areas; silos of dedicated components—servers, networks, and storage—are complex and costly to buy, manage, and grow. Hyperconverged infrastructures can solve these challenges, providing hybrid or private cloud services using a single, centrally managed appliance with software-defined data services.

In recent ESG research, 85% of respondents reported that they currently use or plan to use HCI solutions in the coming months.[1] This high percentage comes as no surprise when you look at the benefits that drove them to consider HCI. The list in Figure 1 covers a wide range of capabilities that are essential to success in today's business environment: improved service and support, scalability, agile VM provisioning, predictable costs, simplified management, fast deployment, better TCO, fewer interoperability problems, and ease of acquisition.[2] It sounds too good to be true.

Figure 1. Top Ten Factors Driving Deployment of Hyperconverged Technology Solutions



Source: Enterprise Strategy Group, 2017

[1] Source: ESG Research Report, *The Cloud Computing Spectrum: From Private to Hybrid*, March 2016.
[2] ibid.

And in some cases, it is too good to be true. One reason organizations have shied away from HCI is that solutions cannot always deliver the performance demanded by mission-critical workloads. While many are attracted to HCI because of the management simplicity, that's no longer the only priority. As more HCI options have come to market, the buying criteria have evolved to include performance, security, and resiliency.

Intel and Windows Server 2016 Hyperconverged Infrastructure

Intel and Microsoft engineers collaborated to develop performance-focused HCI reference architectures designed for various hybrid and private cloud workloads. These RAs consist of Intel or Intel-based servers, processors, and storage plus Windows Server 2016 Datacenter Edition with Storage Spaces Direct for software-defined compute, storage, and networking. These RAs have consolidated management and can be expanded easily and cost effectively. Three reference architectures are available to optimize performance, capacity, and cost as required. [3] A full description of the components involved is beyond the scope of this paper, but a brief description is provided below.

Intel components include:

- Intel Xeon processor E5-2600 v4 family. These processors are designed for software-defined solutions, and include hardware enhancements that help monitor, secure, and orchestrate data center resources faster and with less IT effort. They are more efficient, enabling faster and less costly scaling for very large workloads.
- Intel SSD Data Center family. Designed for high performance and low latency, these SSD drives offer extended write endurance, end-to-end data protection, and 256-bit AES encryption. For highest performance and lowest latency, Intel PCIe SSDs with the NVMe interface are recommended for the cache tier. The three RAs leverage a combination of NVMe SSDs, SATA SSDs, and high-capacity HDDs to deliver the right levels of performance and cost.
- Intel server boards and chassis. These elements are preconfigured, pretested, and pre-certified to simplify and speed deployment based on required CPU, memory, network, I/O controllers, and storage capacity.
    - Intel Cloud Blocks for Microsoft are also available. These pre-validated systems deliver enterprise-class security, performance, and reliability. Available in all-flash and hybrid configurations, Cloud Blocks for Microsoft are configured to optimize performance for Windows Server 2016 Storage Spaces Direct.
- Security. These features include hardware-assisted keys, software-based key generation, and crypto acceleration for encryption without performance impact.

Microsoft components include:

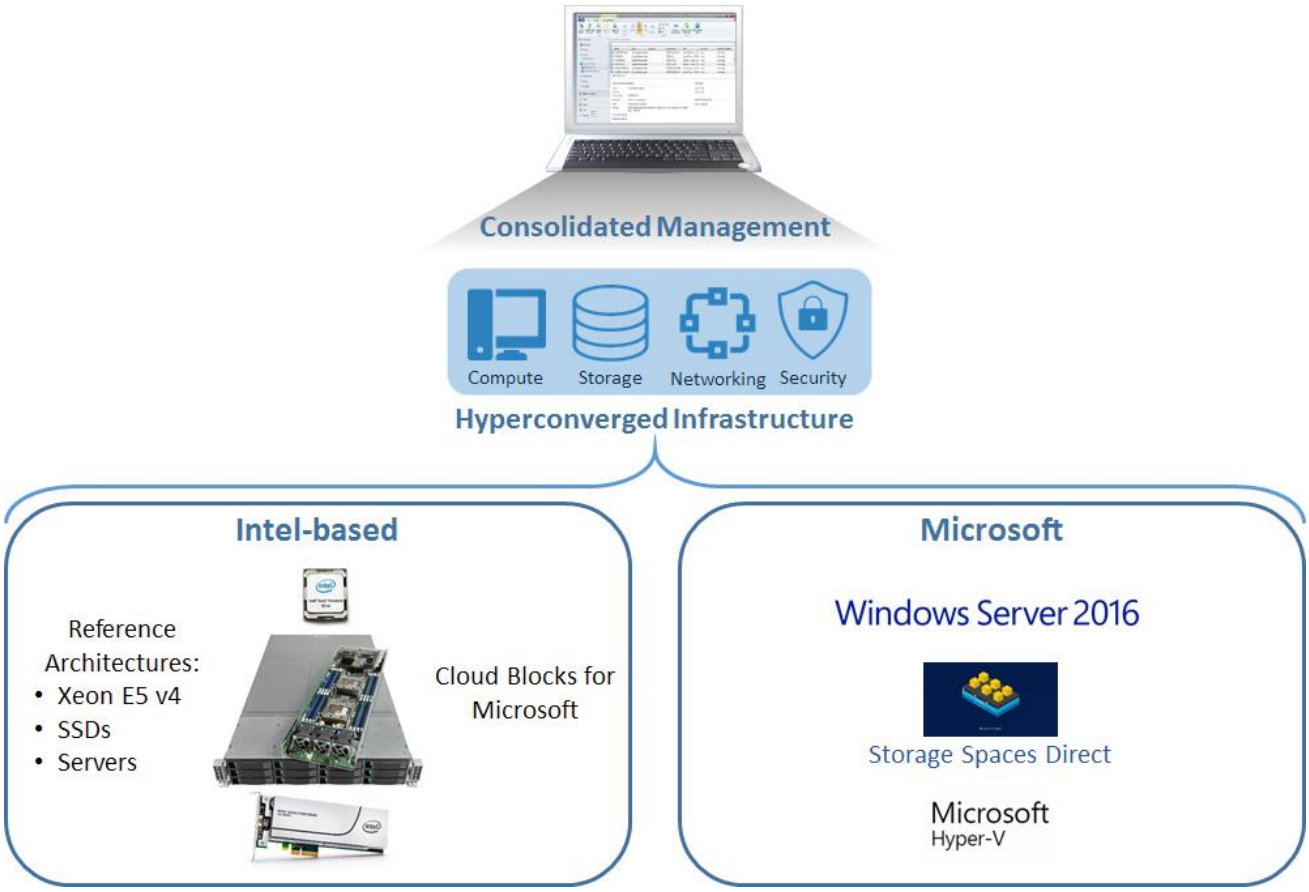
- Windows Server 2016 Datacenter Edition provides software-defined compute, storage, and networking, bringing resources that are scalable and cost effective in a smaller footprint. Hyper-V virtualization is built in, and support for containers enables application isolation on each VM. Because of the Intel and Microsoft collaboration, Windows Server 2016 Datacenter Edition is optimized for Intel compute, networking, and storage, delivering high performance, efficiency, and scalability that today's data centers and private clouds demand.
- Storage Spaces Direct for software-defined storage. Storage Spaces Direct pools local storage on Intel Xeon processor-based servers into highly scalable, available server clusters to provide storage for VMs. These clusters leverage RDMA-capable Ethernet for the storage fabric. The Intel processors provide the horsepower needed for advanced data services. NVMe SSDs are used for caching in all workloads, and for capacity in those workloads that require high IOPS,

[3] The reference architectures are described in detail on page 6.

low latency, and enhanced Storage QoS. Storage Spaces Direct detects the storage devices available and automatically tiers the devices based on performance and endurance levels. Storage Spaces Direct also integrates with Microsoft Scale-out File Server, Clustered Shared Volume File System, and Failover Clusters.

- Software-defined Networking for centrally managing virtual and physical networks, multi-tenant isolation, enhanced throughput, and NIC teaming for reliable performance.
- Shielded Virtual Machines increases VM security by delegating most administration to Windows PowerShell. Disk and VM states are encrypted, protecting them from accidental or malicious access.

Figure 2. Intel and Microsoft Hyperconverged Infrastructure Components



Source: Enterprise Strategy Group, 2017

Performance Validation

ESG audited synthetic and real-world performance tests run by Intel. Testing leveraged industry-standard workload generation tools and was designed to demonstrate the speed, scalability, and cost effectiveness of Microsoft Storage Spaces Direct running Intel Xeon processors and SSDs. The three joint hyperconverged reference architectures were used for testing to highlight the configuration flexibility of the solution to meet the evolving performance needs of all types of organizations.

1. *Capacity Optimized (CO)* – A hybrid configuration that leverages NVMe or SATA SSDs as a data cache with data volumes being fully allocated on high-capacity HDDs. This configuration serves as a baseline option to house applications with relatively cold datasets, such as file servers or workloads with modest performance requirements. The CO configuration was made up of four 2U Intel servers, each containing two Intel Xeon processors E5-2650 v4 with 12 cores running at 2.2 GHz and 256 GB of RAM. The storage cache tier consisted of two 2TB Intel DC P3700 Series SSDs, while the capacity tier had eight 6TB 3.5" Seagate ST6000NM0024 HDDs. A 1x10 GbE dual-port Chelsio T520 adapter and 1x10GbE Extreme Networks Summit X670-48x switch were leveraged as the underlying networking infrastructure.
2. *Performance/Capacity Optimized (PCO)* – An all-flash configuration that leverages NVMe as a data cache to absorb large write bursts and deliver the highest levels of IOPS performance, while high-capacity SATA SSDs continue delivering flash performance for workloads that fall outside of cache. Ideal applications to be run on this configuration include business-critical OLTP database workloads, VDI, infrastructure-as-a-service (IaaS), and some data warehouse workloads. The PCO configuration was made up of four 2U Intel servers, each containing two Intel Xeon processors E5-2695 v4 with 18 cores running at 2.1 GHz and 256 GB of RAM. The storage cache tier consisted of four 800GB Intel DC P3700 Series SSDs, while the capacity tier had twenty 1.6TB Intel DC S3610 Series SSDs. A 1x10 GbE dual-port Chelsio T520 adapter and 1x10GbE Extreme Networks Summit X670-48x switch were leveraged as the underlying networking infrastructure.
3. *Performance Optimized (PO)* – An all-flash configuration that leverages all NVMe to deliver ultra-high levels of predictable performance. This type of configuration is ideal for mission-critical applications that serve as a lifeline to the business with strict performance SLAs, including OLTP databases, VDI, and IaaS. The PO configuration was made up of four 2U Supermicro SuperServers, each containing two Intel Xeon processors E5-2699 v4 with 22 cores running at 2.2 GHz and 384 GB of RAM. The storage cache tier consisted of two 800GB Intel DC P3700 Series SSDs, while the capacity tier had eight 2TB Intel DC P3500 Series SSDs. A 1x40 GbE dual-port Chelsio T580 adapter and 1x40 GbE Extreme Networks Summit X770-32x switch were leveraged as the underlying networking infrastructure.

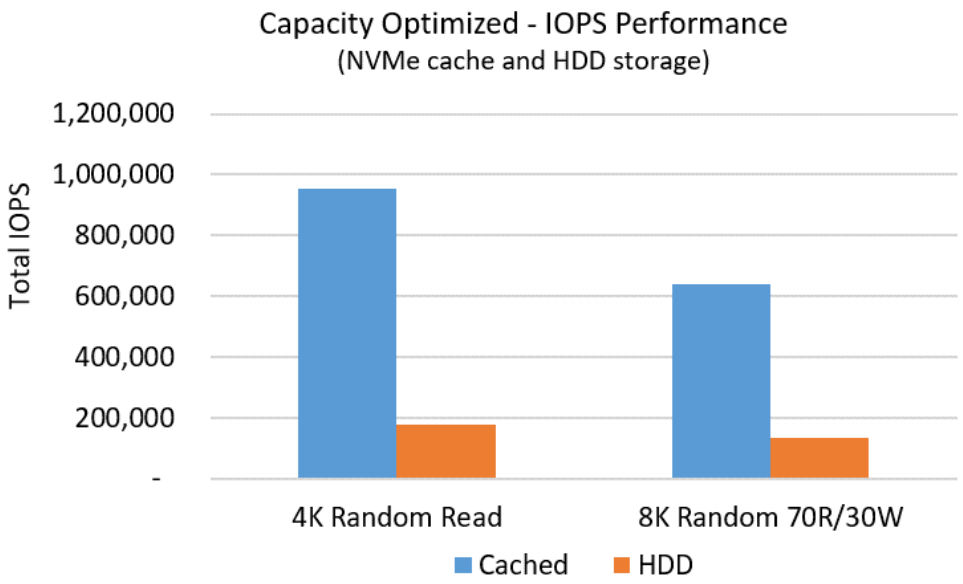
Synthetic Workload Performance Analysis

The first phase of analysis focused on the underlying speed of Storage Spaces Direct across all three reference architectures. The industry-standard Diskspd Utility was leveraged to generate synthetic I/O workloads (4Kb random reads; 8Kb random 70% reads/30% writes) with a goal of producing a high level of IOPS to stress the underlying storage and identify the scenarios in which each configuration is ideally suited.

The CO test focused on the ability of the solution to deliver high levels of performance for working sets with a predictable size that can fully fit in the SSD cache. As such, a comparison was made to a similarly running workload that falls outside of the CO's SSD cache by increasing the working set to consume 78% of the total available storage of the cluster, meaning a small portion of the workload would be serviced by cache, while a majority would be serviced by the underlying HDDs. Twenty-four Azure-like VMs were configured per node (96 VMs total across the cluster) with 2 vCPUs and 3.5 GB of RAM, which mapped directly to the number of processing cores available per node in the cluster. Each VM contained a 60GB VHD for the operating system and a 500GB VHD for data. Within the data VHD, two 10GB Diskspd files were used for the cached dataset, while four 98GB Diskspd files were serviced by the underlying HDDs. Both workloads were driven by four threads and 32 outstanding I/Os per VM.

The results are shown in Figure 3. When the working set fit entirely in the SSD cache, the 4Kb 100% read test yielded impressive total IOPS of 954,200, while the 8K 70/30 read/write mixed test hit 642,000 IOPS. As expected, when the working set was configured to consume 78% of the total available storage, the workload was primarily serviced by the HDDs, which yielded a large reduction in IOPS. The pure read test was reduced by nearly 5.5x, yielding just 176,500 IOPS. The mixed workload performance dropped 4.7x, hitting 135,600 IOPS.

Figure 3. Core Performance of Intel's Capacity Optimized Solution

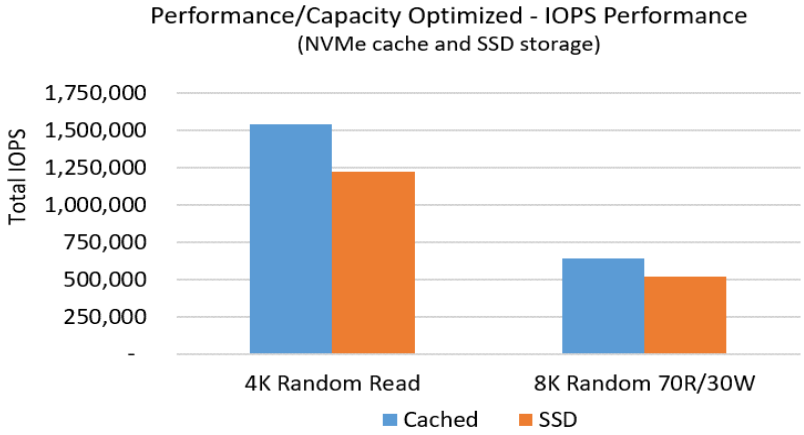


Source: Enterprise Strategy Group, 2017

Next, the PCO test was run to show how dynamic workloads with unpredictable working set sizes can reap the performance benefits of SATA SSDs when workloads are serviced by the underlying storage to continue delivering high, predictable levels of performance. Again, a comparison was made between the workload fitting entirely in cache and the workload being serviced by the SATA SSDs by increasing the working set to consume 75% of the total available storage of the cluster. Thirty-six Azure-like VMs were configured per node (144 VMs total across the cluster) with 2vCPUs and 3.5 GB of RAM, which mapped directly to the number of processing cores available per node in the cluster. Each VM contained a 60GB VHD for the operating system and a 150GB VHD for data. Within the data VHD, one 70GB Diskspd file was used for the cached dataset, while two 70GB Diskspd files were serviced by the underlying SSDs. Both workloads were driven by four threads and 32 outstanding I/Os per VM.

As shown in Figure 4, when the workload fell entirely in the caching tier, performance was quite impressive, achieving 1,544,000 IOPS for pure reads and 642,000 IOPS for the 70R/30W mixed workload. When the workload was serviced by the SATA SSDs, the performance was impacted, but significantly less than we had witnessed with HDDs. Consistent levels of high IOPS performance were still achieved, with a slight measured decrease of 25%. The pure read test produced 1,221,000 IOPS and the mixed workload yielded 518,000.

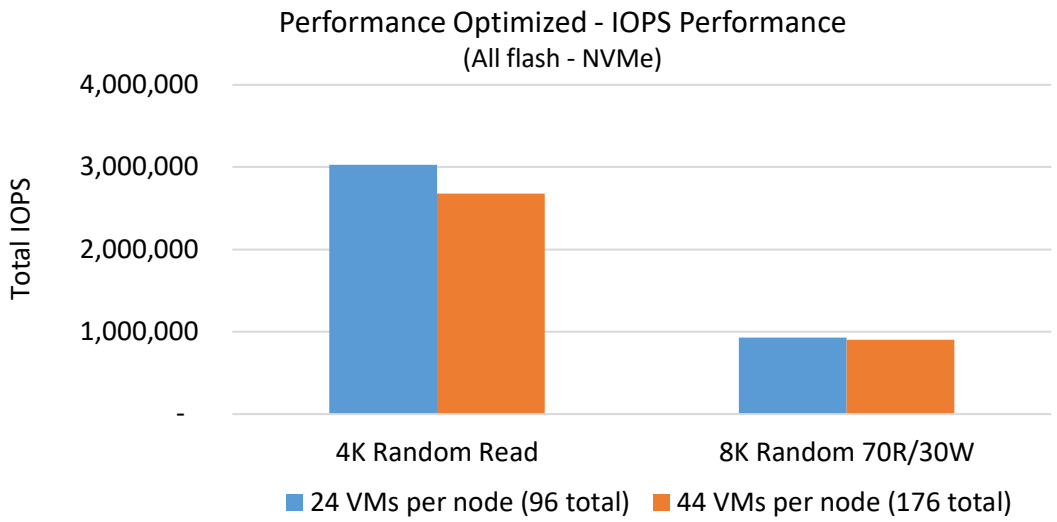
Figure 4. Core Performance of the Performance/Capacity Optimized Solution



Source: Enterprise Strategy Group, 2017

The last configuration tested was the PO configuration with all NVMe drives. The goal of this testing differed slightly from the previous two; instead of comparing cached with disk performance, this comparison looked at optimal performance versus equal physical to virtual CPU subscription performance. For the optimal performance scenario, 24 Azure-like VMs were configured per node (96 VMs total across the cluster) with 2vCPUs and 3.5 GB of RAM. Each VM contained a 60GB VHD for the operating system and a 60GB VHD for data. Within the data VHD, four 10GB Diskspd files were used. For the equal physical to virtual CPU subscription test scenario, the only difference was the number of Azure-like VMs, which increased from 24 to 44 (176 VMs total across the cluster), making the physical to virtual core ratio perfectly aligned. Again, both pure read and mixed workload tests were run and driven by four threads and 32 outstanding I/Os per VM.

Figure 5. Core Performance of the Performance Optimized Solution



Source: Enterprise Strategy Group, 2017

The results of the PO tests are shown in Figure 5. The PO configuration produced a massive 3.01 million IOPS for the pure read test with 24 VMs per node, while the mixed workload achieved an equally impressive 930,800 IOPS. When adjusting the configuration to consume the exact number of available physical cores, performance was impacted minimally, with the read test decreasing by just 13% to 2.68 million IOPS, and the mixed workload decreasing by a negligible 3% to 905,600 IOPS. These high levels of consistent performance position the PO configuration to easily handle the storage requirements of mission-critical applications where IOPS and latency must be guaranteed.

Why This Matters

As the buying criteria for hyperconverged offerings shifts from simplicity and cost savings to enterprise capabilities such as delivering consistently high levels of performance, understanding how different configurations meet specific organizational performance requirements is essential for making the best buying decision. One wrong decision can not only drain a budget, but also, worse, impact an application and the end-user experience, hurting customer loyalty and brand recognition.

ESG confirmed that the Intel and Microsoft joint hyperconverged reference architectures with Microsoft's Storage Spaces Direct offer configuration flexibility, with three clear options to meet varying capacity and performance requirements, all while delivering the core benefits of hyperconverged (i.e., simplicity and cost efficiency). For working sets with predictable capacity requirements, the capacity optimized offering easily delivers high levels of IOPS performance; the performance/capacity optimized offering provides underlying SSDs to cover workloads that fall outside of cache to continue to deliver high levels of performance for dynamically sized workloads. For latency-sensitive workloads, the performance optimized offering consistently delivers millions of IOPS, even when all resources are consumed, providing organizations with peace of mind knowing they're covered even when unexpected demands are put on the system.

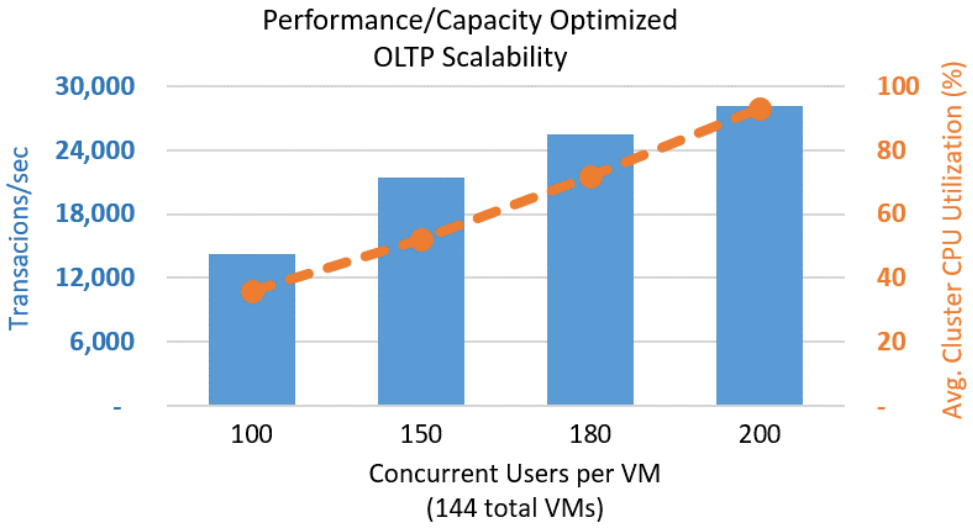
Real-world Workload Performance Analysis

Hyperconverged offerings combine storage and compute into scalable building blocks. To validate performance, ESG shifted from a synthetic analysis of the underlying storage to simulating a real-world workload that exercised both the storage and compute. A real-world OLTP workload was used to simulate the activity of Microsoft SQL Server 2016 users. The workload emulated the database activity of users in a typical online brokerage firm as they generated trades, serviced account inquiries, and performed market research. The workload was composed of nine transaction types covering database inserts, updates, deletes, and pure reads. The workload generated a high level of I/O activity with small access sizes and spent significant execution time at the operating-system-kernel level, combining with a large cache resident working set to effectively evaluate the expected performance levels of the PCO and PO reference architectures. The workload was driven by two independent servers outside of the cluster, each containing two engines capable of driving a predefined number of user streams to each SQL Server instance in the cluster. The four total drivers mapped exactly to the number of nodes in the cluster.

The PCO configuration was tested to show high levels of sustained concurrency across the cluster to deliver scalable OLTP performance. Thirty-six Windows Server 2016 VMs were configured per node (144 VMs total across the cluster) with 2vCPUs and 3.5 GB of memory. Each VM contained a 60GB VHD for the operating system and a 150GB VHD for a Microsoft SQL Server 2016 database instance with a scale factor of 4,000—104 GB for the database, 10 GB for the SQL Server log, and 20 GB for the tempdb file. This sizing was selected to ensure the workload was serviced outside of the cache. ESG audited four thread counts: 100, 150, 180, and 200. These numbers represent the number of active customers per VM database, meaning that the 200-thread count represented 28,800 simultaneous database customers (144 VMs at 200 thread counts each). ESG analyzed transactions/sec, average transaction response time, and CPU utilization.

Figure 6 highlights the transactions/sec performance scalability as the thread count increased from 100 to 200. Also shown is the average CPU utilization of the cluster. As expected, performance scaled near linearly, while CPU usage increased to peak levels. At the 200-concurrent user count, ESG witnessed a cumulative total of 28,823 transactions across the cluster. Though not shown in the chart, average transaction response time remained extremely low through the scalability points. This is critical for high-performing OLTP database workloads. The average 95th percentile response time, which is deemed as the premium class of service (CLOS), for all transactions measured on a per SQL VM basis remained under two seconds, and when looking at the scaled back data point of 180 concurrent users, remained under one second.

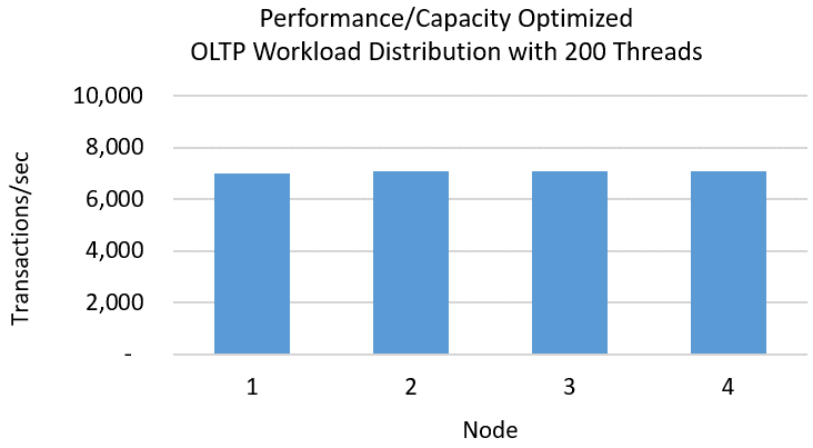
Figure 6. OLTP Performance Scalability of the Performance/Capacity Optimized Configuration



Source: Enterprise Strategy Group, 2017

Since hyperconverged solutions leverage a shared-resource architecture, an even workload distribution across the cluster is essential to delivering an optimally performing environment. This is especially true at peak performance levels. Therefore, ESG focused on the 200-concurrent user count and dug deeper into the results on a per node level. As shown in Figure 7, the workload was distributed almost equally across each of the four nodes. Of the 28,823 transactions, the lowest performing node contributed 6,977 transactions/sec, while the highest performing node delivered 7,098 transactions/sec.

Figure 7. OLTP Workload Distribution on the Performance/Capacity Optimized Configuration

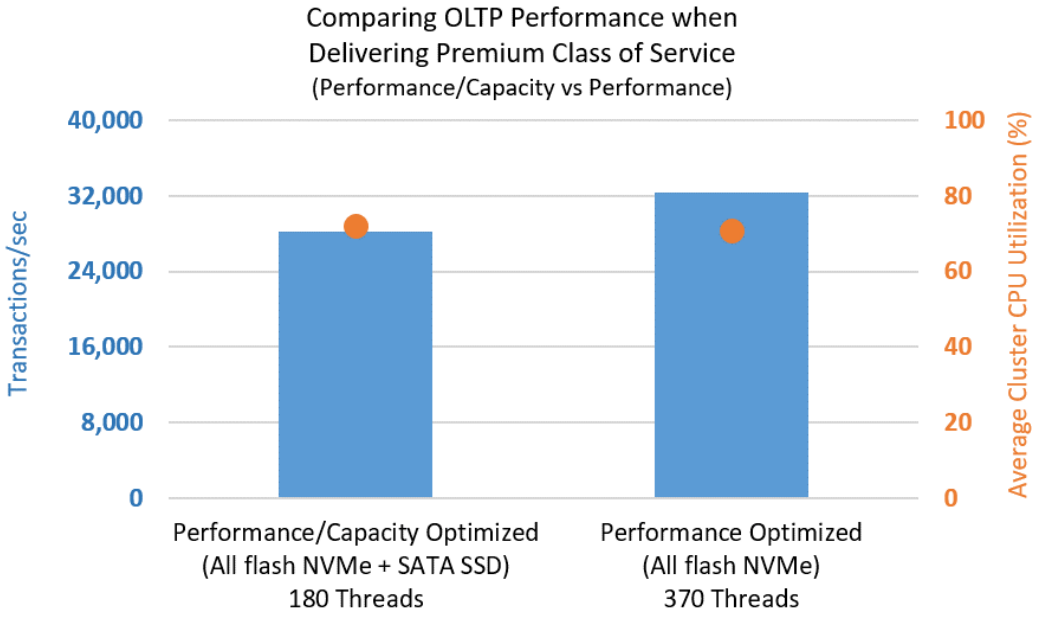


Source: Enterprise Strategy Group, 2017

For even higher levels of performance, ESG shifted to the PO configuration to push the thread count higher. The virtual infrastructure differed slightly, in that only 22 VMs were provisioned per node, totaling 88 across the cluster with 4 vCPUs and 3.5 GB of memory. Each VM contained a 60GB VHD for the operating system and a 100GB VHD for a Microsoft SQL Server 2016 database instance, with a scale factor of 2,000—50 GB for the database, 10 GB for the SQL Server log, and 20 GB for the tempdb file. Based on preliminary testing, a thread count of 370 (8,140 total users across the cluster) was selected to ensure delivery of a premium CLOS, as witnessed with the PC0 configuration at the 180-thread count.

ESG compared the PO's number of cumulative transactions/sec and CPU utilization while delivering a premium CLOS with the PCO configuration. The results are shown in Figure 8. The number of transactions on the PO achieved a 27% boost in performance compared with the PCO configuration, going from 28,223 on the PCO to 32,356 on the PO configuration. It should be noted that in both cases, CPU utilization remained steady in the 70-73% range and the average 95th percentile transaction response time remained under one second.

Figure 8. Comparing PCO with PO for Delivering Premium Class of Service



Source: Enterprise Strategy Group, 2017

Why This Matters

OLTP databases drive critical client activities such as bank, retail, airline, and other online transactions. Performance disruptions can have catastrophic impacts. Organizations expect premium levels of service and predictable, scalable performance, all while meeting their simplicity and budgetary requirements. Hyperconverged solutions have become increasingly popular for these activities, but real-world performance has been difficult to validate from a customer's standpoint.

ESG validated the real-world performance capabilities of Intel/Microsoft all-flash, hyperconverged reference architectures to deliver consistent levels of premium class service. The number of transactions/sec scaled on the PCO configuration as the concurrent user count increased from 100 to 200 per VM. The PO configuration yielded a large boost of 27% to transaction performance when compared with the PCO configuration, while continuing to deliver a premium class of service, with average transaction response times remaining under one second and CPU utilization remaining around 70%.

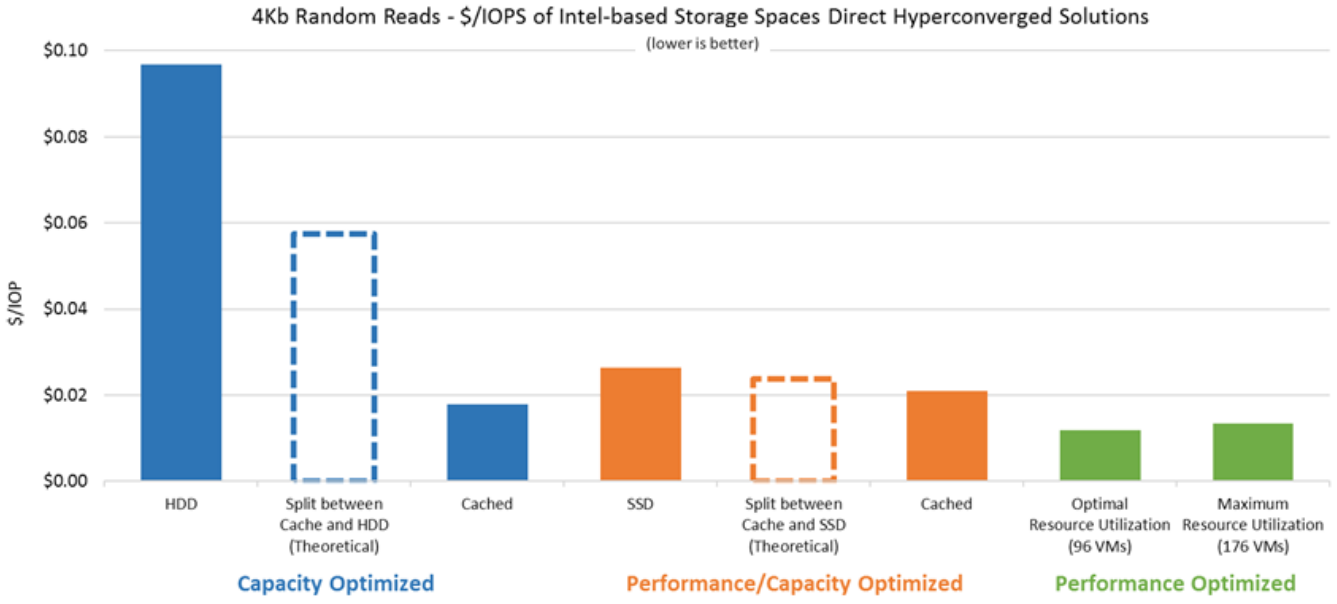
Price/Performance Analysis

Achieving high levels of performance is important to meeting the requirements of organizations looking to adopt hyperconverged technologies, and achieving that performance cost effectively is essential. With that in mind, ESG completed a price/performance analysis of Intel/Microsoft reference architectures with Storage Spaces Direct to highlight the cost flexibility in relation to an organization's performance requirements. The price/performance analysis was based solely on hardware cost of acquisition.

Considering the components of each Intel reference architecture, ESG combined Intel's pricing with that of other industry-leading vendors and resellers that offer similar components to calculate average pricing for each offering. The average price was then used to calculate price/performance based on the performance results highlighted in the previous sections of this report. The price/performance points were calculated using street pricing, as opposed to list, to set a reasonable expectation for potential customers and to serve as a guideline for evaluating hyperconverged solutions.

Figure 9 shows $/IOPS based on the Diskspd test results for the 4Kb random read test.

Figure 9. $/IOPS for 4Kb Random Reads with Intel-based Storage Spaces Direct Hyperconverged Solutions



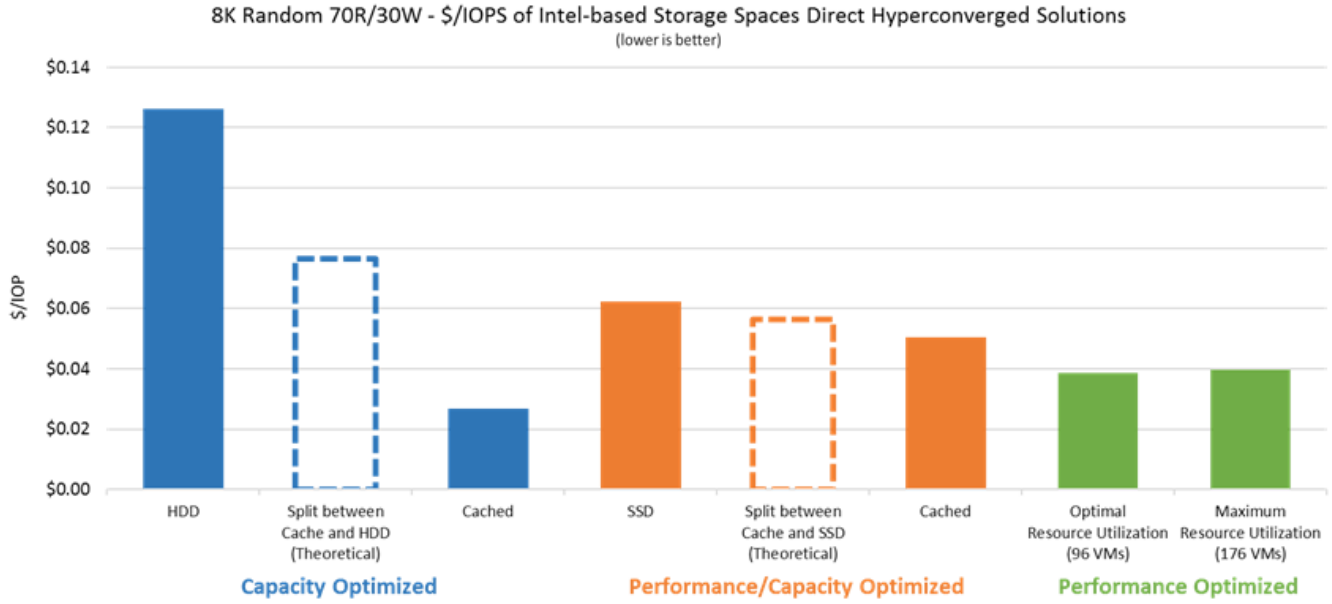
Source: Enterprise Strategy Group, 2017

What the Numbers Mean

- As expected, the test serviced outside of the cache on the capacity optimized configuration produced the highest $/IOPS since the workload was mostly serviced by HDDs, producing much lower performance than the NVMe or SATA SSD drives in every other data point on the chart. ESG witnessed the lowest $/IOPS on the all-NVME performance optimized configuration.
- Two theoretical data points were charted to show likely results when cache serviced most of the workload, but the remaining workload was serviced by the underlying HDDs or SSDs.
- For capacity optimized configurations, the results highlight a very important reality. Though the cost of HDDs is significantly less than SSDs on a price sheet, the use of HDDs creates $/IOPS price points that are far more expensive than the all-flash configurations. The only scenario where the CO configuration provides compelling $/IOPS value is when an organization can ensure its active dataset fits entirely in cache.
- For performance/capacity optimized configurations, there is less variation between the three data points since the underlying storage consists of flash (SATA SSDs), while cached is serviced by NVMe flash. The $/IOPS ranges from $.026 for performance serviced by the underlying storage to $.021 for cached workloads.
- For performance optimized configurations, the $/IOPS are extremely low. The all-NVMe architecture eliminated the impact of requiring underlying storage to service the requests. In both test scenarios, $/IOPS remained less than $.02. Optimal utilization produced a $/IOPS that was nearly half of any other calculated $/IOPS, and even the maximum resource utilized test yielded a 33% savings over the next closest $/IOPS measured.

ESG witnessed a similar pattern when analyzing the $/IOPS for the 8Kb random read/write mixed workload, but when writes were factored in, all $/IOPS increased, as expected (see Figure 10).

Figure 10. $/IOPS for 8Kb 70R/30W with Intel-based Storage Spaces Direct Hyperconverged Solution



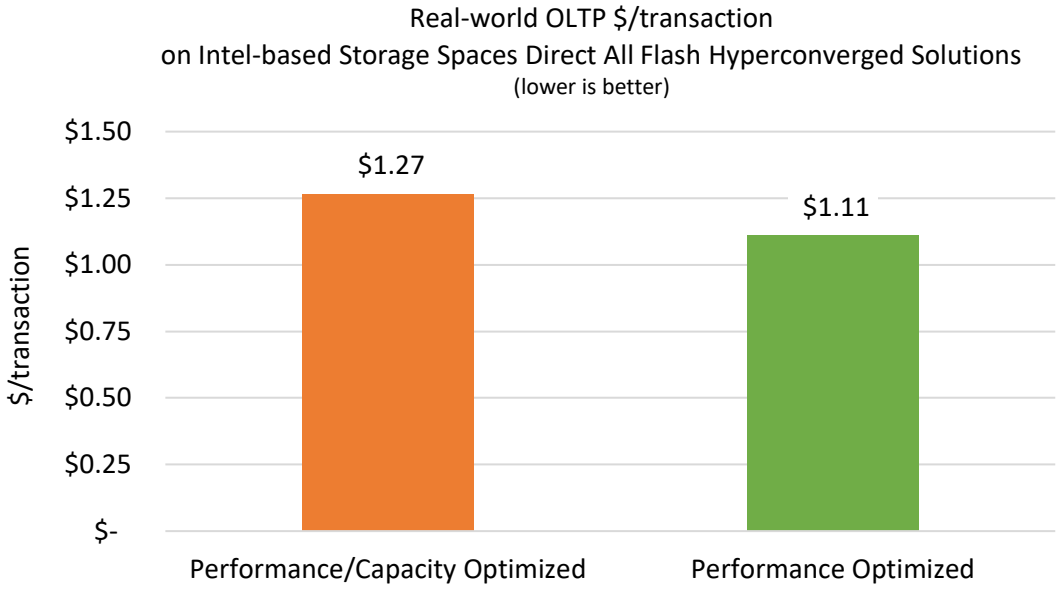
Source: Enterprise Strategy Group, 2017

What the Numbers Mean

- It is important to call out the $/IOPS for the cached result in the capacity optimized configuration. By leveraging HDDs as the underlying storage, an upfront cost for this type of configuration is significantly lower, but if an active dataset is small enough to fit entirely in cache, the performance penalty of the HDDs is never seen. This enables the configuration to deliver flash-like performance at the cost of an HDD configuration. For this test configuration and workload, the combination of a large cache (tested at 4 TB) and a low cost due to underlying HDDs produced a compelling $/IOPS that bested the all-flash configurations.

For the real-world OLTP workload, $/transaction was calculated using the performance results that delivered a premium class of service (average 95th percentile response time). As expected, $/transaction was higher than $/IOPS across the board; each transaction requires multiple I/Os. ESG was especially impressed with the ability to keep costs per transaction marginally low. The performance optimized solution offered nearly a 15% savings over the performance/capacity optimized configuration, as Figure 11 shows.

Figure 11. $/Transaction with Intel-based Storage Spaces Direct All-flash Hyperconverged Offerings



Source: Enterprise Strategy Group, 2017

The important takeaway from these price/performance analyses is that hyperconverged solutions based on Intel and Microsoft reference architectures that leverage Microsoft's Storage Spaces Direct provide highly cost-efficient options for organizations looking to modernize their infrastructures. By providing offerings that address both capacity and performance requirements at a low cost, Intel and Microsoft are setting themselves up to be a major contender in this space.

Why This Matters

Year after year, cost reduction is a top business priority, and it is a key factor when evaluating and selecting new technologies. Executives want to purchase new technologies to modernize their infrastructures and meet business requirements, but they prefer to not spend a lot to do so. This creates a difficult situation for IT personnel who are asked to find the perfect piece of technology that checks off every box on the requirements list.

It is nearly impossible to check every box while staying on budget, leading to tradeoffs. Enterprise-class performance and manageability are often traded away for fast ramp-up times and scalability, but these Intel/Microsoft-based reference architectures with Microsoft Storage Spaces Direct prove that performance and cost efficiency are not mutually exclusive. They offer configuration flexibility to meet dynamic performance and capacity requirements without breaking the bank.

The Bigger Truth

When market evolution changes the buying criteria in an industry, there is often a mismatch for a period of time between what customers want and what they can get. Vendors that can see what's missing and fill the void gain an advantage. The hyperconverged infrastructure market has begun to mature, and as a result, simplicity and cost efficiency, while still essential features of HCI, are no longer enough. Customers are demanding that HCI infrastructures meet higher performance requirements than ever before.

This is where industry leaders Intel and Microsoft are making a difference. They have collaborated to build HCI solutions that deliver the consistent high performance and scalability that organizations require today, while maintaining cost efficiency and management simplicity.

The combined Intel/Microsoft solution leverages Intel servers, processors, and flash drives and Microsoft Windows Server 2016 with Storage Spaces Direct. They offer configuration flexibility with three reference architectures that optimize performance and cost for various workloads: 1) a capacity optimized solution for workloads such as Exchange and SharePoint; 2) a performance/capacity optimized solution designed for workloads with additional performance needs such as VDI; and 3) a performance-optimized solution for mission-critical workloads with consistent high-performance SLAs such as OLTP, VDI, and IaaS.

ESG Lab validated synthetic and real-work testing on the three RAs, focusing on core IOPS performance as well as OLTP transaction performance. We validated that these RAs deliver high performance, up to millions of IOPS, while maintaining simplicity and cost efficiency. Real-world testing demonstrated linear performance scalability and low latency, as well as even distribution of workloads as thread counts scaled. The latter ensures that the shared resources continue to function optimally as workloads grow. In addition, ESG Lab's price/performance analysis showed that the three Intel-based reference architectures with Microsoft Storage Spaces Direct demonstrate that performance and cost efficiency are not mutually exclusive. Dynamic, scalable performance requirements were easily met, while costs remained low for both $/IOPS and $/transaction.

Poor application performance can cause serious business disruption, but organizations are sometimes forced to trade off performance for cost and simplicity. The key is to be able to address both, with a price/performance option such as the NVMe drives in this solution. ESG Lab validated that the Intel/Microsoft RAs tested achieve extremely high performance while maintaining the essential cost and simplicity objectives that drive HCI deployments. In our estimation, this is a premier HCI solution for organizations looking to improve performance, reliability, flexibility, and security for their private or public cloud deployments.

Enterprise Strategy Group is an IT analyst, research, validation, and strategy firm that provides market intelligence and actionable insight to the global IT community.