A MICRON® WHITE PAPER

Micron® and Microsoft® Azure® Stack HCI: Cost-Effective, High-Performance Virtualized Infrastructure

Micron 7300: High-Performance Flash Benefits HCI

Microsoft® recently released Azure® Stack HCI, a hyperconverged infrastructure (HCI) platform that offers a cost-effective virtual server solution built on Microsoft Windows Server® 2019, Hyper-V™ and Storage Spaces Direct. This combination is well suited for organizations of all sizes with data center and remote office needs.

Organizations are increasingly adopting virtualized infrastructure platforms that provide benefits through server consolidation and cloud-centric management. These benefits can result in lower operational costs, better scalability and better resiliency for critical business applications. Micron offers a broad range of data center SSDs to enhance the benefits of these virtualized infrastructures, and this paper covers our newest NVMe™ SSD, the Micron 7300.

The Micron 7300 SSD with NVMe is Micron's latest generation of SSDs built to provide fast throughput and IOPS, low latency and consistent responsiveness in the data center. These characteristics make it well suited for Microsoft Azure Stack HCI.

This white paper highlights the benefits of a four-node Azure Stack HCI solution using Micron 7300 SSDs, second-generation AMD® EPYC™ CPU-based servers and high-performance networking by Marvell™. Test results highlight the value of the Micron 7300 to maximize cluster performance across multiple key metrics, in a solution designed to control the costs sometimes associated with all-flash HCI solutions.

Micron Benefits

1.4 million IOPS

Four nodes and standard components build a performance Azure Stack HCI cluster.[1,2]

Mainstream NVMe for HCI

The Micron 7300 SSD delivers mainstream NVMe performance for hyperconverged infrastructure.

NVMe performance, approachable price point

Get 6x the performance of data center SATA SSDs at comparable prices.[3]

micron.com/7300

1. In this document, we use the word "performance" to mean input/output operations per second (IOPS), MB/s or both
2. 100% read, 4 KiB (kibibytes)
3. The Micron 7300 PRO SSD 2TB U.2 with NVMe (3,000 MB/s sequential read) has six times higher performance than the Micron 5300 PRO SATA SSD 2TB (540 MB/s sequential read; 540 MB/s is the maximum bandwidth available to any SATA device). For pricing details, contact your Micron representative.

Tested Configuration

We used four standard x86 servers hosting Microsoft Azure Stack HCI (Azure Stack). Each server was a Dell EMC™ PowerEdge™ R7525 dual-socket AMD platform using two AMD EPYC 7502 CPUs that provided 32 physical cores per socket and a total of 256 physical cores for the cluster (512 logical cores). In addition, each server had 16 32GB Micron 3200 RDIMM memory modules, for a total of 512GB of DRAM per server (Table 1).

The storage configuration consisted of a single Dell BOSS (Boot Optimized Storage Solution) option that hosted a single Micron 5300 PRO 480GB M.2 SSD for the operating system, and eight Micron 7300 PRO 3.84TB SSDs per node for virtual machine storage, resulting in a total of 123TB of cluster capacity.

The network configuration for this cluster was built as a switchless design. The simplicity of switchless networks, combined with their reduced cost, makes them well suited for small deployments, without compromising overall cluster performance. Each node used two four-port 25 GbE Marvell QL41234HLCU network interface cards (NIC). For the first NIC in a server node, three ports were directly connected to a port on each of the corresponding NICs in the other nodes (Figure 1). This was repeated for the second four-port NIC to provide additional bandwidth and network redundancy. The final (fourth) port on each NIC was teamed with its counterpart NIC on the same server to provide a redundant interconnect to the data center client network to support access to virtual machines within the cluster.

Table 1. Server configuration by role

LightOS Storage Server (4x)	
Vendor/Model	Dell EMC PowerEdge R7525
Processor (2x)	AMD 7502 (32-core, 3.3GHz)
Memory (DRAM)	512GB (Micron DDR4-3200)
Network	2x Marvell QL41234HLCU 4-port 25 GbE
Storage	Boot (1x): Micron 5300 PRO 480GB SATA M.2 SSD Data (8x): Micron 7300 PRO 3.84TB NVMe SSD
Operating System	Microsoft Windows Server 2019 Data Center Edition

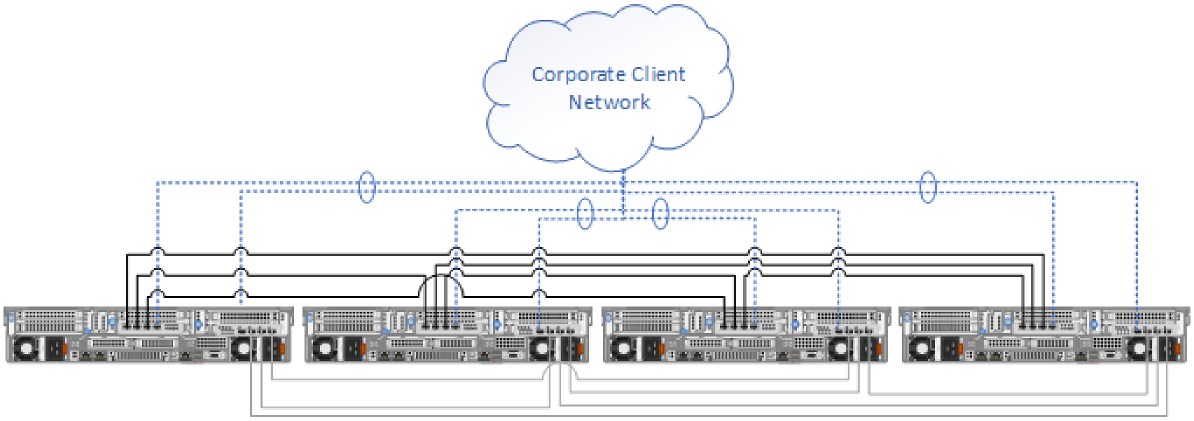

Figure 1. Azure Stack HCI test configuration with hypothetical client network connections

The Microsoft Azure Stack configuration used a definition for each virtual machine as an Azure A4v2 virtual machine (VM) consisting of four virtual CPUs and 8GB of DRAM. The total number of virtual cores running on each node ranged from four vCPUs (one VM per node) to 96 vCPUs (24x VMs per node) to ensure even distribution of workload. Based on prior experience, this VM count would provide optimal results. (The number of VMs is a administrator-adjustable value.)

All VM data used three-times (3x) replication for data protection by ensuring that each data block was hosted on two additional cluster nodes.

Virtualized Infrastructure Performance Results

To highlight the results obtained during testing, we performed two analyses. First, we analyzed the overall cluster performance using IOPS performed by VMs at various VM counts along with average latency and 99.99% tail latency. We focused on small block (4 KiB) random I/O for 100% read, 100% write, 70% read/30% write, and 90% read/10% write workloads. Second, we analyzed the performance for a single node in the cluster to gain a deeper understanding of how the SSDs, CPU and network were used.

Cluster Performance

The cluster consisted of four servers working cooperatively to provide shared, redundant resources to hosted VMs. VMs may reside on any node, but VM affinity ensures that VMs run on the node on which they were created, unless a node failure occurs. This ensures a balanced VM load across the cluster. Data presented in this section represents the performance of the cluster with an equal number of VMs hosted on each node, for a total VM count illustrated in each chart from four VMs (one VM per node) to 96 VMs (24x VMs per node).

In addition, we used two per-VM queue depths to simulate “light” (queue depth 1, or QD1) and “heavy” (queue depth 8, or QD8) resource consumption per VM, highlighting different use cases for the virtualized infrastructure.

For 100% read workloads with 96 active VMs, the system reached 1.9 million IOPS with low latencies (Figure 2). At QD1, representing a “light” application, we experienced a range of average latencies between 224 and 338 microseconds (µs) and 99.99% latency, also known as “tail latency," in the range of 5 to 10 milliseconds (ms). At QD8, a “heavy” application, average latencies reached 1.6 ms, while tail latency reached 50 ms. These are relatively low values, making the solution suitable for many VM configurations.

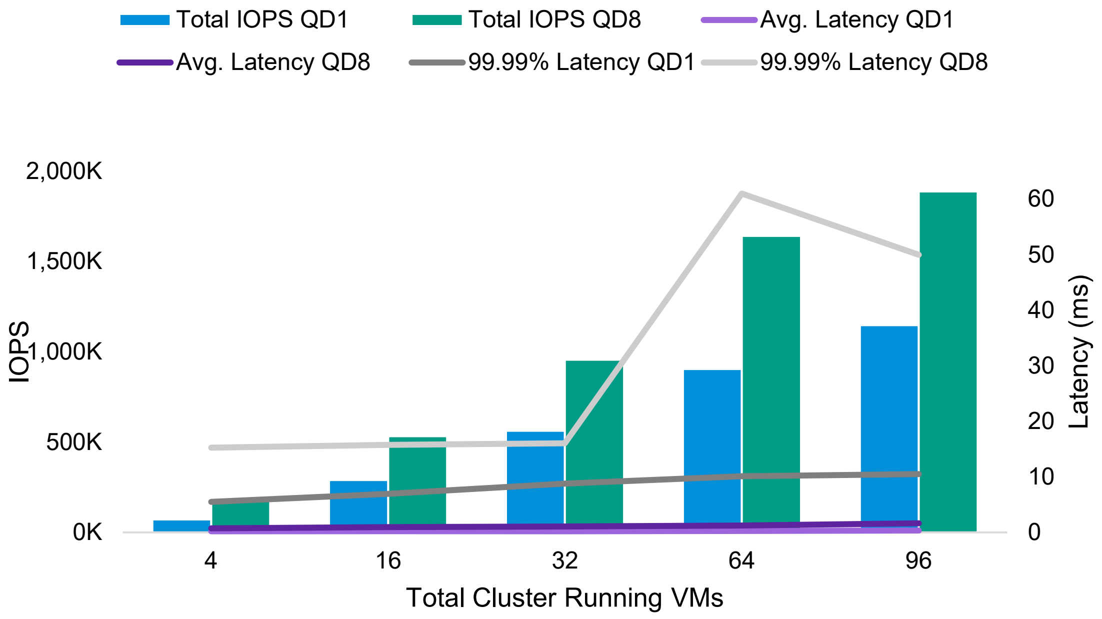

Figure 2. Cluster performance for various VM loads with 100% read I/O

For 100% write workloads and 96 active VMs, the Azure Stack cluster reached 531,000 IOPS. As discussed earlier, each write operation was replicated to two additional cluster nodes before the write could be acknowledged as completed. At QD1, average I/O latency ranged from 330 µs to 1.4 ms, and at QD8, average latencies reached 5.8 ms. Tail latencies ranged from 15.5 ms to 32.8 ms and from 17.4 ms to 52.7 ms at respective queue depths (Figure 3). Replication strategy might increase latencies based on several factors such as network utilization, queue depth and ongoing VM management being performed by Azure Stack HCI. Balancing data protection with performance will vary based on individual requirements.

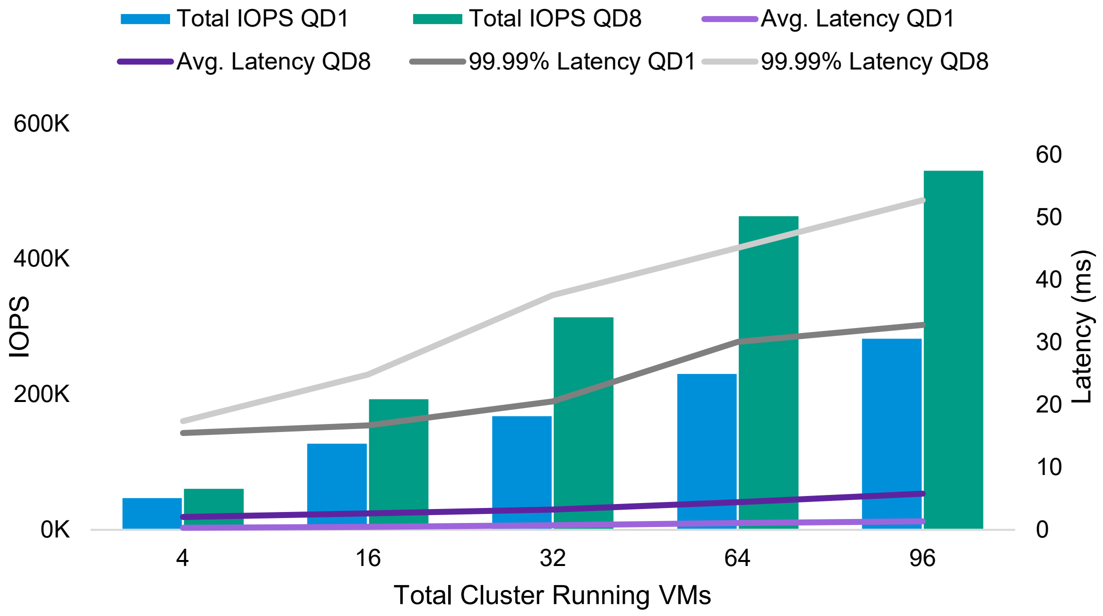

Figure 3. Cluster performance for various VM loads with 100% write I/O

Mixed read-write workloads using random 4 KiB I/O at 90% read and 10% write showed the cluster was able to reach up to 1.5 million IOPS with very low average latency. At QD1, average I/O latency showed a maximum of 437 µs for 96 VMs and 2.1 ms at QD8. Tail latencies showed a significant increase for QD8 above 32 VMs, ultimately reaching 288.5 ms for 96 VMs (Figure 4).

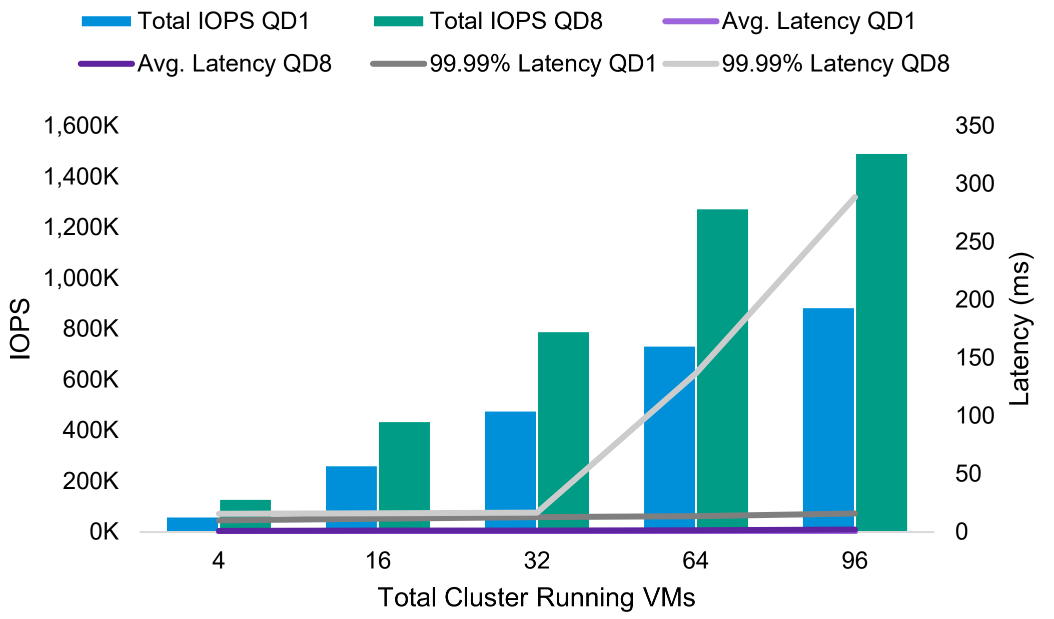

Figure 4. Cluster performance for various VM loads with 90% read/10% write I/O

Mixed read-write workloads using random 4KiB I/O at 70% read and 30% write showed up to 1.1 million IOPS with low average latency. At QD1, average IO latency showed a maximum of 650 μs for 96 virtual machines and 2.8 ms at QD8. Tail latencies showed a maximum of 10.4 ms at QD1 and 17.9 ms for QD8 and 96 virtual machines (Figure 5). Reduced tail latency at lower write percentage levels was due to the use of DRAM cache on the 7300 SSDs, which enabled the configuration to buffer a small number of writes.

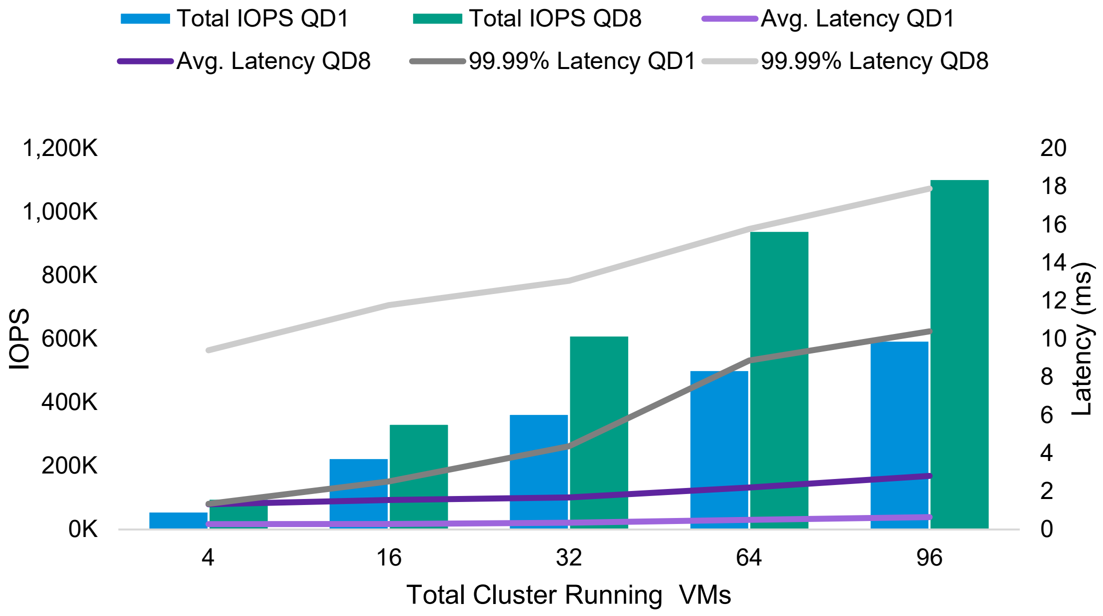

Figure 5. Cluster performance for various VM loads with 70% read/30% write I/O

Single-Node Performance

To understand more about the latency behavior of the Azure Stack HCI cluster, we analyzed an individual node's results. Based on our comparison of the four nodes, each node behaves similarly across all measured areas. Therefore, we used the data from one cluster node as representative of the behavior of each node in the cluster. Also, we focused on one profile that most closely represented common, general purpose, virtual infrastructure use by analyzing the 70% read and 30% write workload profile with a queue depth of eight (QD8).

Storage Performance

Each Azure Stack HCI node hosted eight Micron 7300 NVMe SSDs. Storage performance across all eight SSDs was consistent, generating between 50,000 and 60,000 IOPS (Figure 6). Each node contributed approximately 440,000 IOPS to the cluster. Differences in IOPS across the eight SSDs in a single node were due to node management tasks and roles assigned to each disk from Azure Stack HCI, which uses the SMB 3.0 protocol for storage access management.

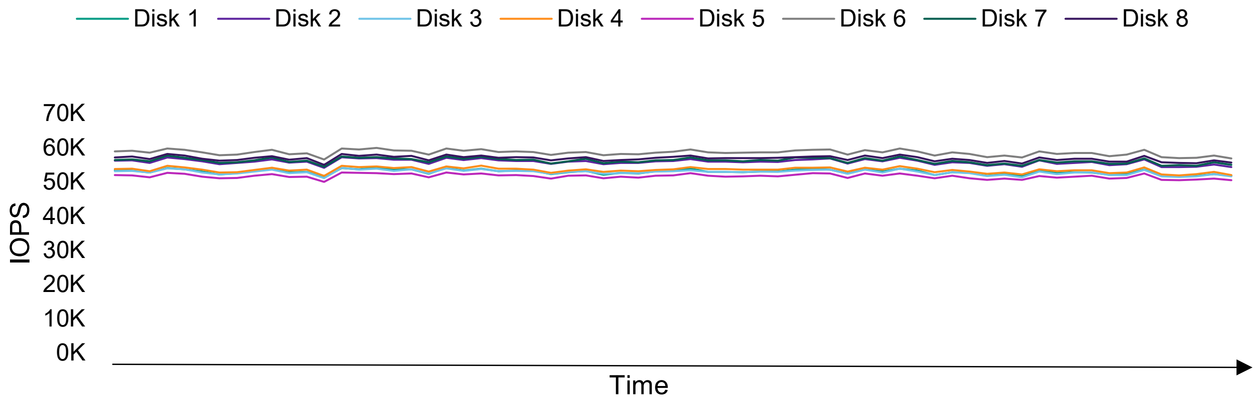

Figure 6. Single-host-per-disk IOPS over time for 4 KiB random 70% read/30% write I/O

Write Latency

Storage latency across all eight SSDs was consistent, with average latency in the range of 87 µs (Figure 7). Disk-level latency contributed only 3% of the overall 2.8 ms (Figure 5, QD8, 96 VMs) average latency on a per-VM basis across the cluster. Data writes were replicated onto two additional nodes for redundancy. Based on this factor, most of the latency likely results from network latency and Azure Stack HCI storage management.

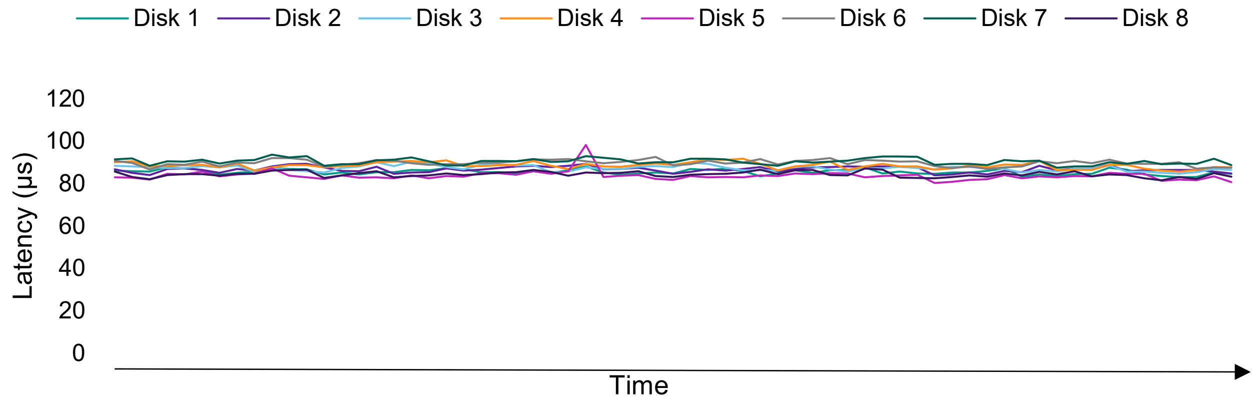

Figure 7. Single-host average write latency over time for 4 KiB random 70% read/30% write I/O

CPU Utilization

CPU workload partly determines overall performance of the applications running within the VMs, as well as how many VMs can be supported by a single server before additional servers are required. AMD EPYC 7002 generation CPUs support modifying the number of active NUMA (non-uniform memory access) complexes within a server from one to 16 per socket, depending on the CPU model. Based on our previous experiences with altering the NUMA per socket (NPS) settings and our use of 32-core-per-socket CPUs, we set the NPS at 4, resulting in eight total NUMA complexes per server, with each complex hosting eight cores. We measured CPU utilization at a NUMA level. By measuring CPU utilization for storage-related processing, we gained an understanding of the CPU resources available for dedication to VM-hosted applications.

CPU utilization was consistent across seven of the eight NUMA complexes, at an average utilization of 21.5%. One outlier to this utilization level was NUMA complex 1, which averaged over 50% utilization (Figure 8). While we have no direct verification of why a single NUMA (or maybe even a single core within NUMA 1) would have a higher utilization, we speculate this was due to the architecture of Azure Stack HCI and how it manages VMs and storage resources by dedicating cores to these roles.

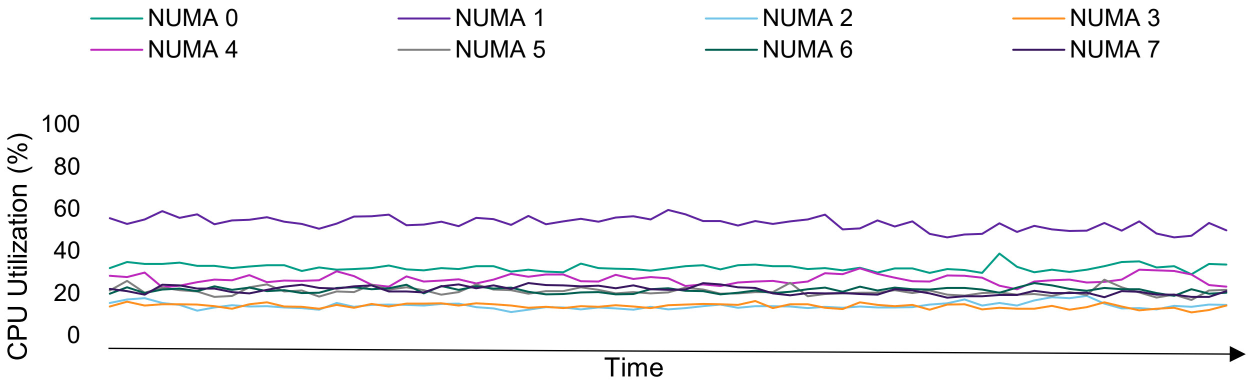

Figure 8. Single-host CPU utilization over time for 4 KiB random 70% read/30% write I/O

Network Utilization

Azure Stack HCI uses advanced Ethernet network functionality called remote direct memory access (RDMA) and server message block (SMB) for data management and transmission between servers. To measure network utilization, we measured the throughput into and out of each of the six Ethernet ports on the server, configured for internode communications within the Azure Stack HCI cluster.

Figure 9 shows the rate of data leaving a node in megabytes per second (MB/s). Data leaving a node can be described for either of two actions:

1. Data being transferred from a node to two additional nodes as part of the triple data replication to ensure data resiliency

2. Data being read by a remote VM

Since the layout of our VMs is designed such that a VM's primary data was hosted by the same server as the VM, most data leaving a node was due to the first reason above. Average transmit (write) throughput per port was 186 MB/s. There was a similar level of simultaneous receive throughput. This throughput level was approximately 6% of the available network bandwidth of each 25 GbE port, indicating that network bandwidth was not an issue.

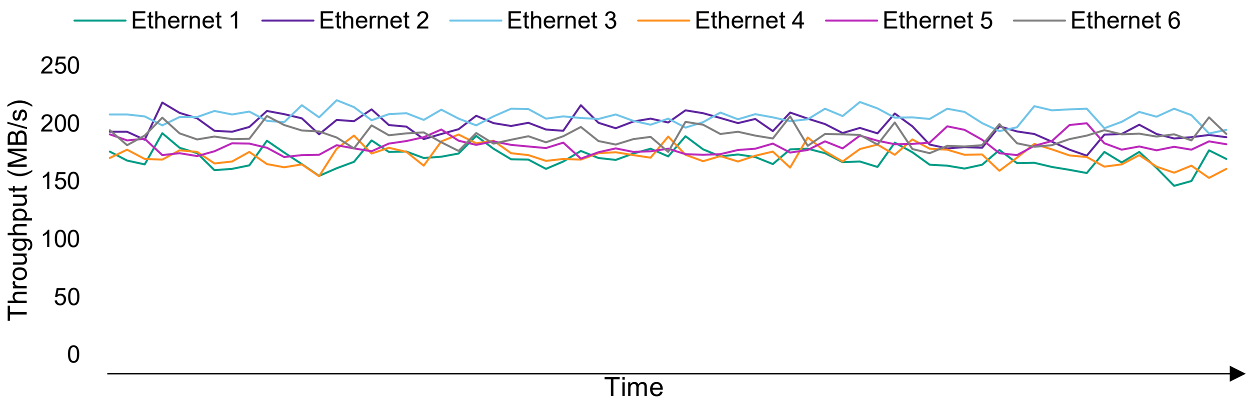

Figure 9. Single-host network outbound throughput per port over time for 4 KiB random 70% read/30% write I/O

A second measurement of network performance is the number of frames that can be sent or received by a port within the infrastructure. Each 25 GbE port could send and receive 9.5 million data frames per second. Figure 10 shows the number of frames being transmitted by each port within the node that was assigned to Azure Stack HCI cluster communications. An analysis of received frames is similar but not illustrated here. Each port was sending and receiving an average of 256,000 frames per second. This represented only 6% of capacity for each port, indicating that the network was not limiting overall performance.

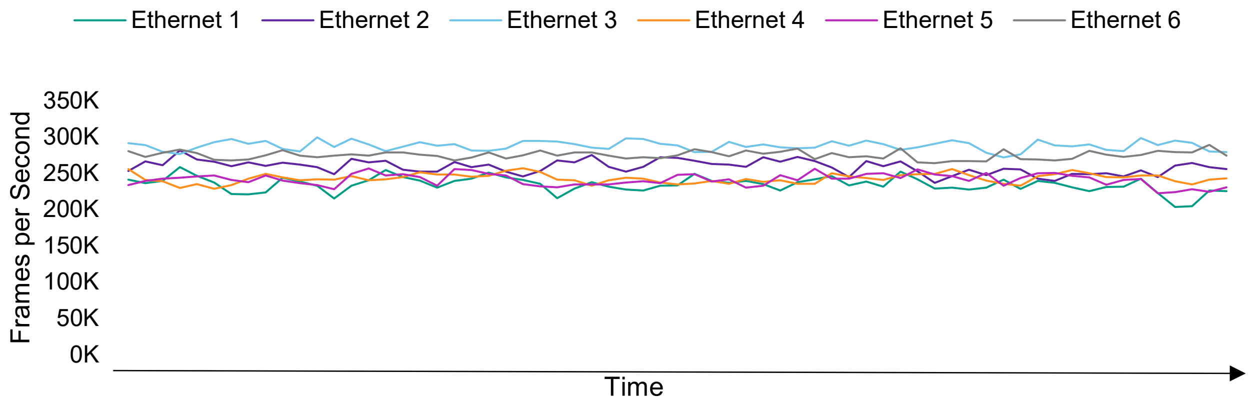

Figure 10. Single-host network outbound frames per second per port over time for 4KiB random 70% read/30% write I/O

The Bottom Line

Azure Stack HCI offers a highly robust virtualized infrastructure environment that emphasizes consistent performance, data protection and reliability. While it provides strong performance overall, Azure Stack HCI primarily focuses on providing a robust infrastructure for hosting virtual servers (VMs) that need balanced availability of all resources (compute, storage and network). Our testing shows that good performance is attainable using SSDs such as the Micron 7300 family of mainstream NVMe SSDs. In fact, the 7300 NVMe SSD is Micron's most cost-effective data center SSD and a great option for storage in this solution. Offering over 1 million IOPS for 70% read and 30% write workloads and over 1.4 million IOPS for 90% read and 10% write workloads, Azure Stack HCI using Micron 7300 SSDs offers great performance at a reasonable cost.

Learn More

To learn more about Micron data center SSDs, visit micron.com/products/ssd/usage/data-center-ssd
To learn more about Microsoft Azure Stack HCI, visit azure.microsoft.com/en-us/products/azure-stack/hci/

How We Tested

Test Methodology

The Microsoft VMFleet test tool was used to deploy a varying set of up to 96 VMs distributed evenly across the four Azure Stack HCI host systems (26 VMs per physical host). Within each VM, VMFleet executed `diskspd` with the required read/write workload profile. Each clusterwide test was executed three times and the results presented are the average of results from those three tests.

Four KiB blocks were used in the following read/write ratios:

- 100% read
- 90% read and 10% write
- 70% read and 30% write
- 100% write

Each workload profile sweep consisted of tests run by varying the number of active VMs from one to 24 per server node.

Each workload profile-VM count test also executed using a queue depth of one and eight to simulate a "light" and "heavy" use-case VM (Table 2).

For each result, tests executed for a specified time period as follows:

- Warmup: 3 minutes
- Test run: 10 minutes

The workload parameters are summarized in Table 2:

Table 2. VMFleet test parameters

Parameter	Value
Block size	4096 bytes
Threads/VM	4
Queue Depth/VM	1, 8
Write Percentages	0%,10%, 30%, 100%

Cluster Deployment

After cluster creation, a single 8TB clustered shared volume (CSV) was created per node. For this testing, CSV read cache was disabled. Each VM consisted of a 32GB VHDX file and was configured to the Microsoft Azure Standard_A4_v2 VM profile. Twenty-four VMs execute from each node, consuming a total of 112 logical cores and 224GB of DRAM per node. A total of 10.8TB of cluster storage was consumed, consisting of 3.6TB of VM images and two additional data replicas maintained due to the triple replication. All VMs ran from the same node that hosted its VHDX image.

micron.com/data-center-ssd