Demartek

Marvell FastLinQ 41000 Series 10GbE Performance, iSCSI Offload Competitive Evaluation and Storage Spaces Direct Use Cases

The Marvell® FastLinQ® 41000 Series is a high-performance 10GbE adapter with iSCSI hardware offload and Universal RDMA performance benefits.

Executive Summary

Non-Volatile Memory Express (NVMe) and Storage Class Memory (SCM) are offered in current generation servers with Intel Xeon Scalable Processors. The resulting server storage performance gains are driving an increase in network bandwidth: Virtual Machines (VMs) and containers are more densely deployed on servers, the internet Small Computer System Interface (iSCSI) is being used for high-bandwidth storage solutions, and Hyper Converged Infrastructure (HCI) needs extensive bandwidth for inter-node communications. A 10GbE network has become necessary to support this infrastructure.

It is also important to consider additional Network Interface Card (NIC) features that enhance overall performance that may be applicable to a deployment like iSCSI hardware offload or Remote Direct Memory Access (RDMA). Marvell FastLinQ 41000 Series adapters are equipped with one of the most extensive collections of these Ethernet features on the market.

Marvell commissioned Demartek to evaluate the benefits of the Marvell FastLinQ 41000 Series when used with latest generation servers. We tested for Layer 2 performance, compared the iSCSI hardware initiator offload performance to that of software initiator on a leading competitor, and evaluated Marvell FastLinQ 41000 Series use in a hyper-converged Storage Spaces Direct (S2D) cluster with SCM and NVMe storage.

We found that the Marvell FastLinQ 41000 achieved line-rate 10GbE Layer 2 performance. We also found that the iSCSI hardware offload increased processor effectiveness and the universal RDMA helped our S2D cluster achieve high bandwidth with little processor overhead.

Marvell FastLinQ 41000 Series 10GbE Performance, iSCSI Offload Competitive Evaluation and Storage Spaces Direct Use Cases

Demartek

July 2019

Key Findings

> The Marvell FastLinQ 41000 Series achieved line rate bidirectional performance for buffer sizes of 1KB up to 1MB.

> The Marvell FastLinQ 41000 Series hardware iSCSI initiator had an average of 7.2 times more IOPS than the Linux software iSCSI initiator on Intel for unidirectional workloads.

> The Marvell FastLinQ 41000 Series hardware iSCSI initiator achieved an average of 4.6 higher IOPS for 8KB block sizes and an average of 5.2 times higher IOPS for 32KB block sizes as compared to the Linux software initiator.

> The Marvell FastLinQ 41000 Series hardware iSCSI initiator used half the processor (51%) at 8KB block size that the Microsoft Windows software initiator did to deliver the same full bandwidth, doubling the processor effectiveness.

> For large block S2D read testing, the cluster utilizing the Marvell FastLinQ 41000 Series with Universal RDMA achieved a total average throughput of 10,470 MBPS while using on average 16% of available cluster processor.

> For large block S2D write testing, the cluster utilizing the Marvell FastLinQ 41000 Series with Universal RDMA achieved a total average throughput of 1,227 MBPS over RoCE and 2,360 MBPS over iWARP, while using on average 10% of available cluster processor.

Marvell FastLinQ 41000 Series

The Marvell FastLinQ 41000 Series supports:

> Available in 10GBASE-T, 10GbE DAC and 10GbE SR connectivity types

> Bandwidth provisioning to aid migration from 1GbE to 10GbE

> Single Root Input/Output Virtualization (SR-IOV) and NIC Partitioning (NPAR)

> Full iSCSI hardware offload

> Universal RDMA support for all current protocols:

- RDMA over Converged Ethernet (RoCE)
- RoCE version 2 (RoCEv2)
- Internet Wide-area RDMA Protocol (iWARP)

> Tunnel Offload – NVGRE, VXLAN, GENEVE

> DPDK Small Packet Acceleration

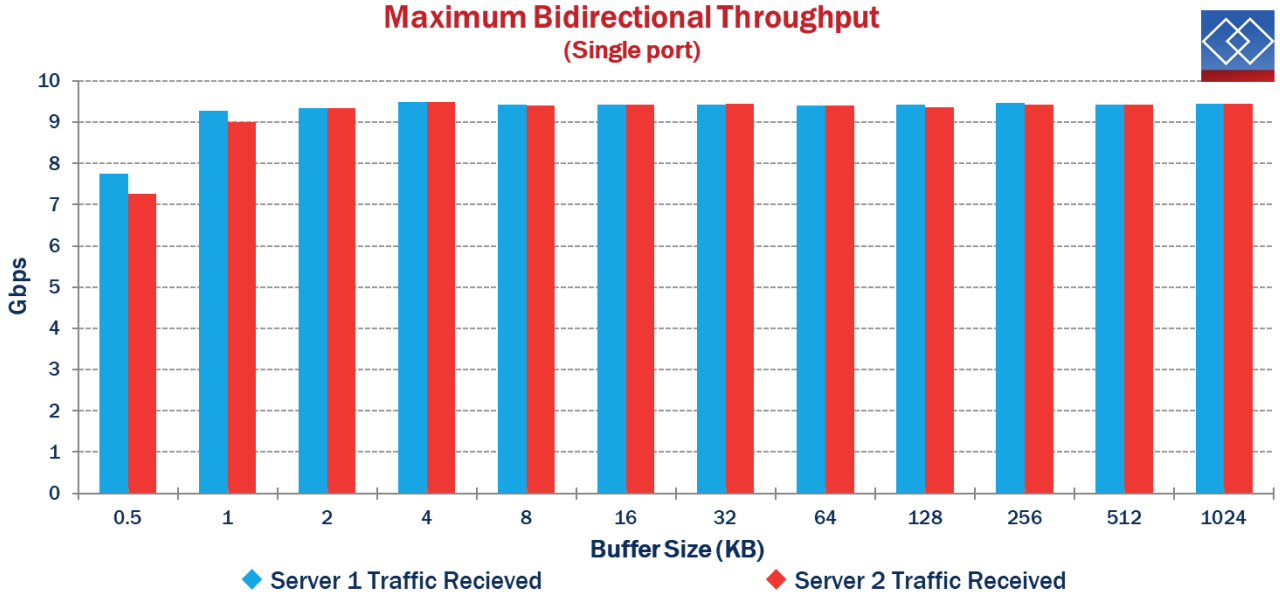
Marvell adapters have excellent performance and support the widest arrays of features on the market.

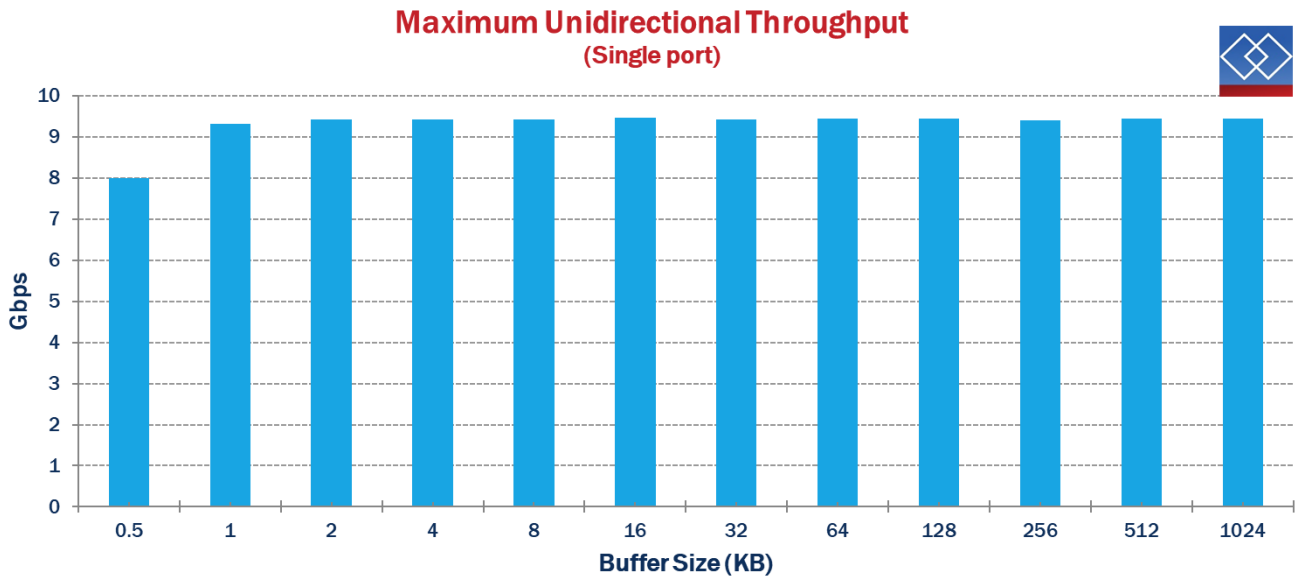
Marvell FastLinQ 41000 Series 10GbE Performance, iSCSI Offload Competitive Evaluation and Storage Spaces Direct Use Cases

Marvell FastLinQ 41000 Series Layer 2 Performance

Two Marvell FastLinQ 41000 Series adapters (each with two ports of 10GbE) were installed in two current generation servers with Intel Xeon Scalable processors, formerly known as "Purley" platforms. One port from each server was connected to an Ethernet switch and iperf, a network performance measurement tool, was used to test bandwidth between the two servers. Buffer sizes from 512B up to 1MB were tested at varying thread counts.

Both bidirectional and unidirectional bandwidth was measured. In all but the smallest buffer sizes, we saw line rate performance.





Marvell FastLinQ 41000 Series 10GbE Performance, iSCSI Offload Competitive Evaluation and Storage Spaces Direct Use Cases

Demartek

July 2019

Marvell FastLinQ 41000 Series iSCSI Hardware Offload

Software versus Hardware Initiators

The iSCSI initiator can be implemented in software or hardware. Software initiators are usually provided by the operating system and use system resources and any Network Interface Card (NIC) available. Hardware offloaded Initiators are available on select NICs. Hardware Initiators usually have their own TCP/IP stack and offload iSCSI processing to the NIC, conserving valuable system resources. Hardware Initiators are most valuable in deployments with limited host processor resources. Hardware initiators will reduce system processor utilization and, in some cases, can improve IOPS and throughput. The Intel Converged Network Adapter X710 does not offer hardware iSCSI initiator, however a hardware offloaded initiator is available on the Marvell FastLinQ 41000 Series adapter.

Performance Test Setup

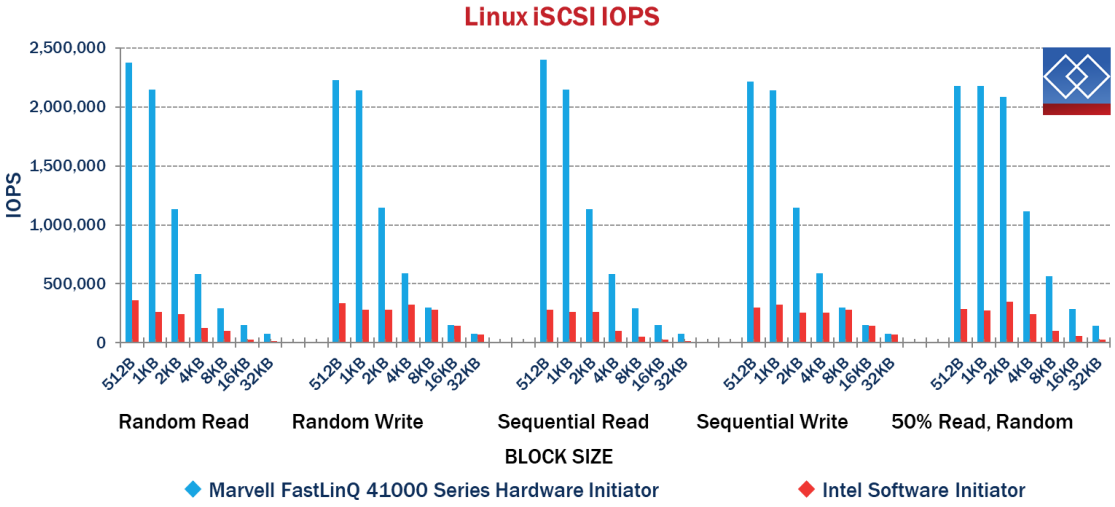
Tests comparing the Marvell FastLinQ 41000 Series hardware iSCSI initiator to the operating system software iSCSI initiator with the Intel Converged Network Adapter X710 were performed to validate the performance benefits offered by Marvell's hardware initiator. 32 iSCSI storage targets were deployed. A Marvell FastLinQ 41000 Series Adapter was deployed in the test server and the hardware initiator was used to connect to the iSCSI storage targets. Tests were run with fio on Linux and with Diskspd on Windows. The Marvell Adapter was replaced with an Intel Converged Network Adapter X710 and the tests were repeated using the operating system provided software initiator.

Marvell FastLinQ 41000 Series 10GbE Performance, iSCSI Offload Competitive Evaluation and Storage Spaces Direct Use Cases

Demartek
July 2019

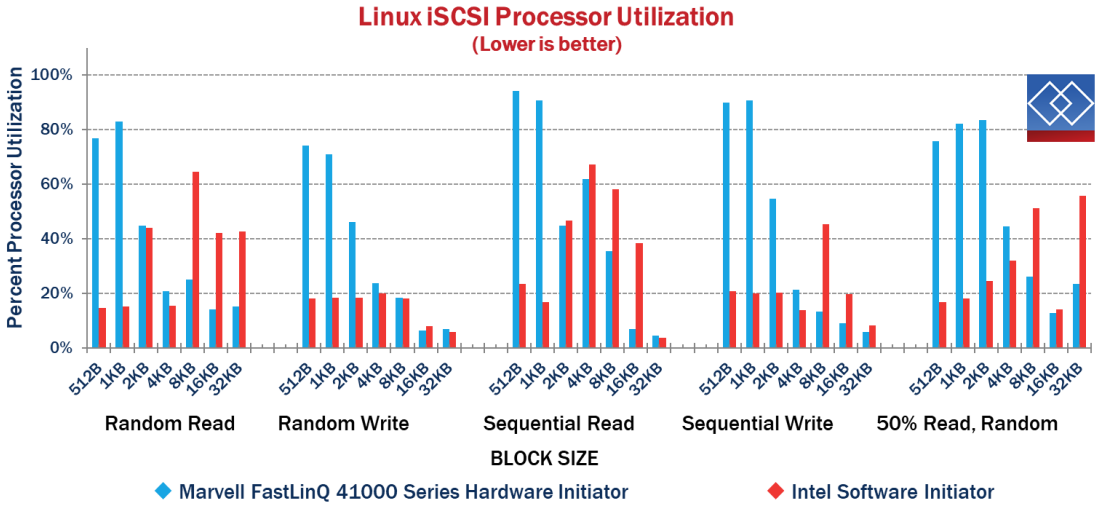
Marvell FastLinQ 41000 Series Hardware iSCSI Initiator versus Linux Software Initiator on Intel Converged Network Adapter X710

For 512B sequential reads, the Marvell FastLinQ 41000 Series hardware iSCSI initiator achieved 8.6 times the topline IOPS that the Linux software initiator with Intel software initiator, and 6.7 times higher overall.
For the read workloads, the Marvell hardware iSCSI initiator achieved an average of 4.6 higher IOPS for 8KB block sizes and an average of 5.2 times higher IOPS for 32KB block sizes.



Due to the Marvell FastLinQ 41000 Series hardware iSCSI initiator in general outpacing the Linux software initiator on Intel by such a wide margin, processor savings usually seen in hardware offload are not easily observed. To see the offload's effect on system resources, it is necessary to compare a workload where either IOPS or processor was the same for both.

For example, the Marvell FastLinQ 41000 Series hardware iSCSI initiator achieved 83% more IOPS than with Intel X710 while using approximately the same amount of processor for 4KB random writes. Without the offload, we would expect the processor usage to increase as the IOPS increased.



Marvell FastLinQ 41000 Series 10GbE Performance, iSCSI Offload Competitive Evaluation and Storage Spaces Direct Use Cases

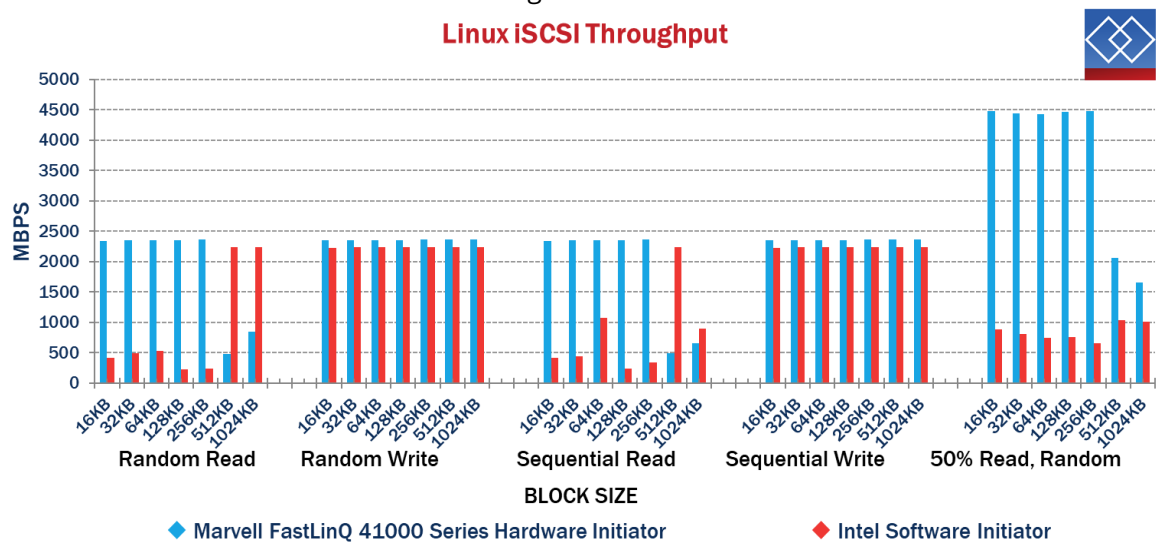
Calculating processor effectiveness can also help us to see the processor savings from hardware offload. We define processor effectiveness as the ratio of IOPS or throughput to percent processor utilization. This effectively tells us for each 1% of processor utilized, how much work can be achieved:

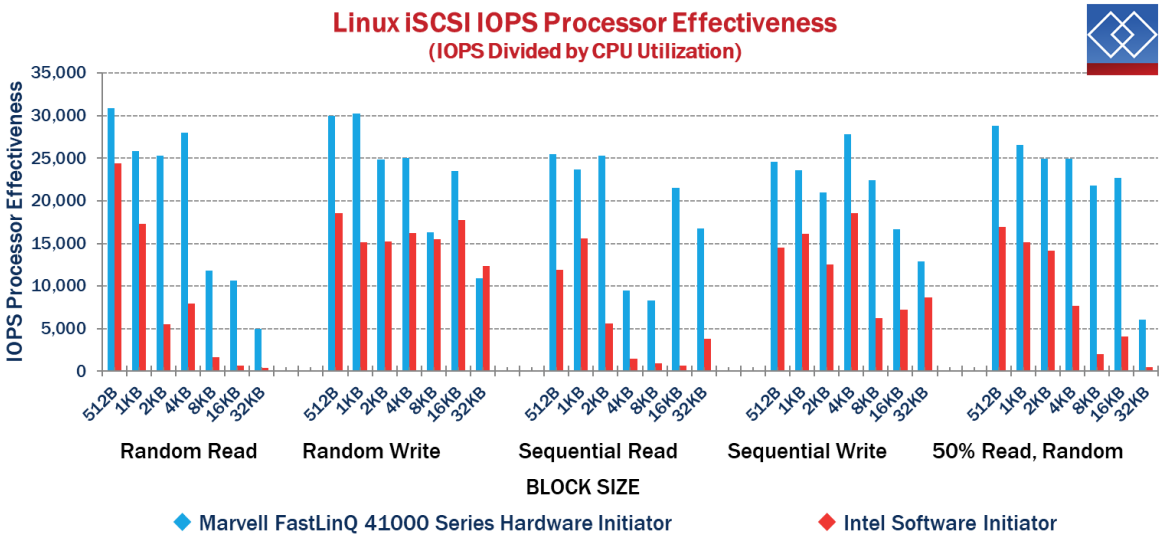
> The Marvell FastLinQ 41000 Series hardware iSCSI initiator used 30% of the processor that Linux software initiator on Intel did to achieve line rate for 8KB sequential writes, achieving 3.6 times the processor effectiveness of Linux software initiator on Intel.

> The Marvell FastLinQ 41000 Series hardware iSCSI initiator achieved 5.9 times the IOPS that Linux software initiator on Intel did while using 8% less processor for 4KB sequential reads, achieving 6.4 times the processor effectiveness of Linux software initiator on Intel.

> The Marvell FastLinQ 41000 Series hardware iSCSI initiator achieved 5.4 times the IOPS that Linux software initiator on Intel did while using 39% less processor for 8KB sequential reads, achieving 9 times the processor effectiveness of Intel.

The Marvell FastLinQ 41000 Series iSCSI hardware initiator achieved line rate throughput for all but the smallest block sizes. Linux software initiator on Intel did not.





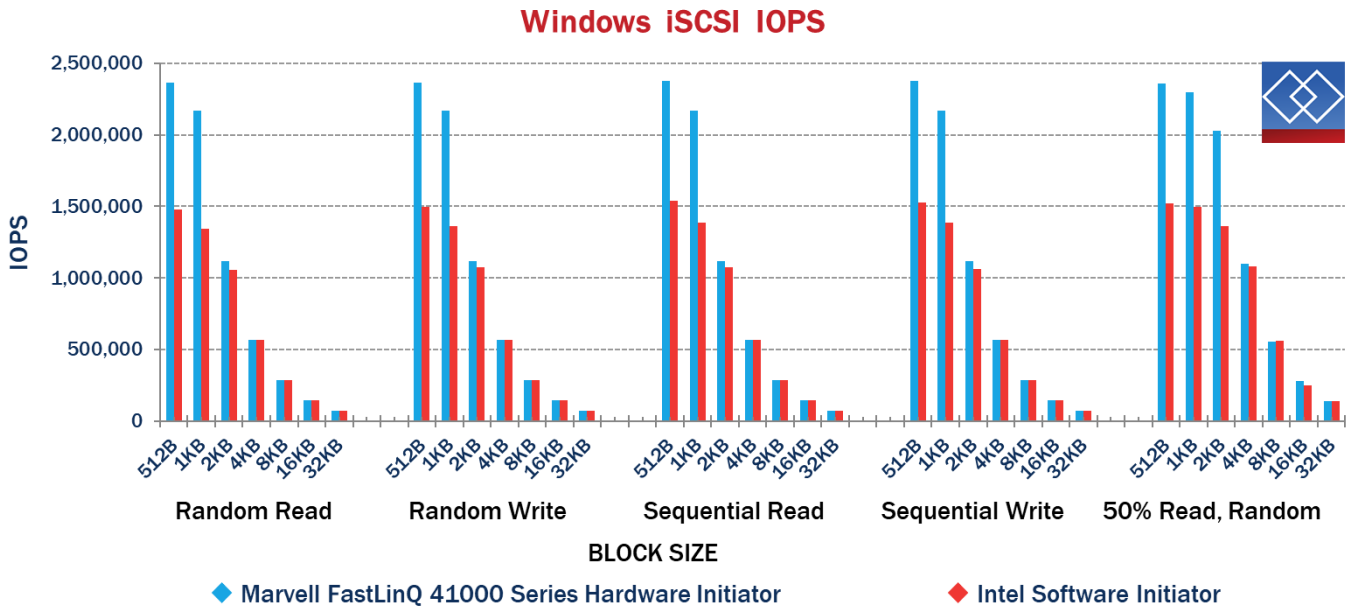
Marvell FastLinQ 41000 Series 10GbE Performance, iSCSI Offload Competitive Evaluation and Storage Spaces Direct Use Cases

Demartek

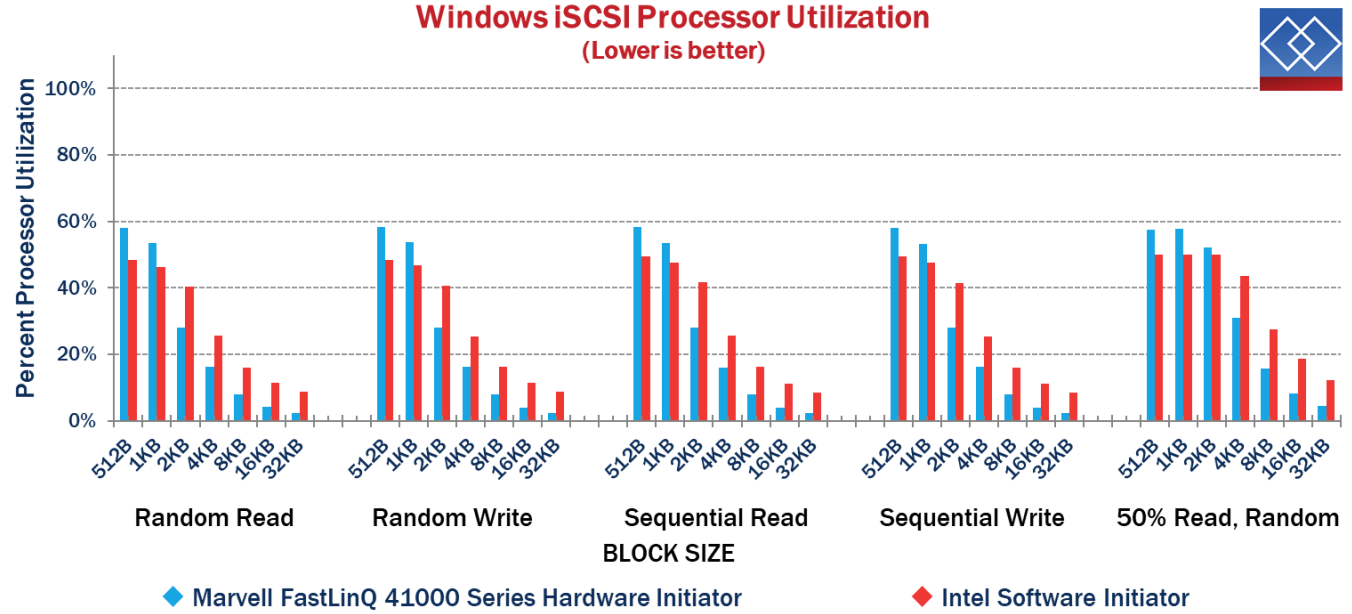
July 2019

Marvell FastLinQ 41000 Series Hardware iSCSI Initiator versus Windows Software Initiator on Intel Converged Network Adapter X710

For our smallest block sizes, the Marvell FastLinQ 41000 Series hardware iSCSI initiator achieved much higher IOPS with only small additional processor expense when compared to the Microsoft Windows software initiator on Intel. For example, for all 512B testing, the Marvell FastLinQ 41000 Series hardware initiator achieved on average 57% more IOPS.



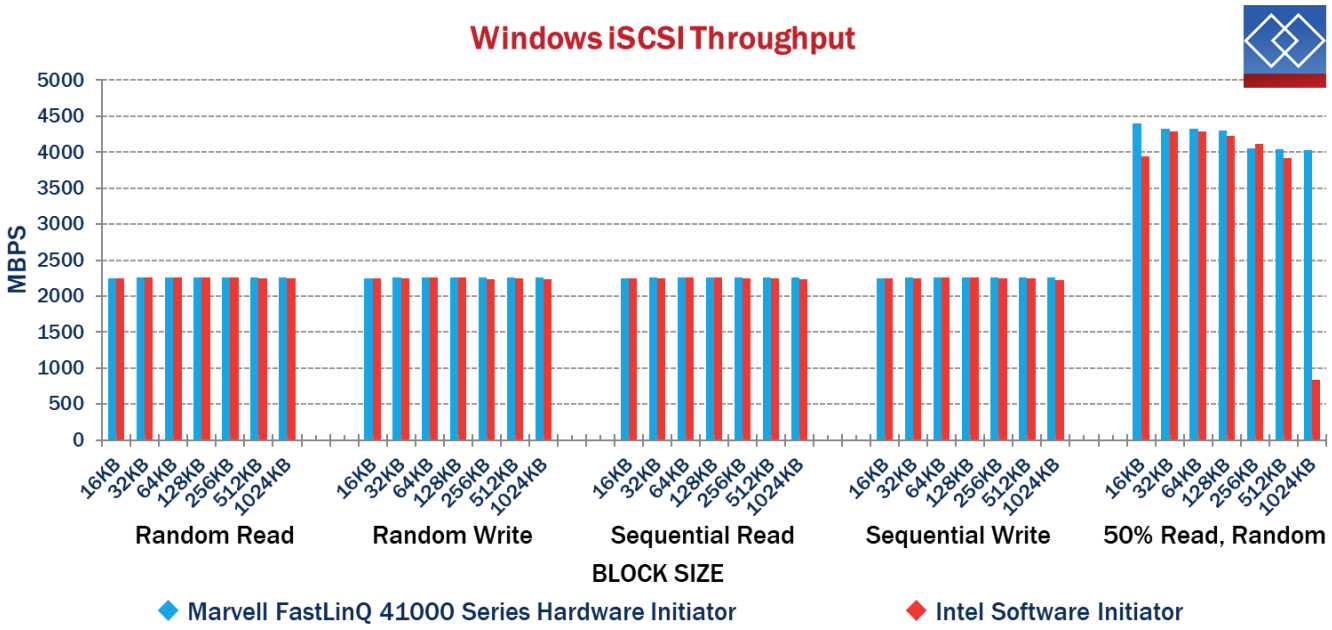
For the often used 8KB block size, the Marvell FastLinQ 41000 Series hardware iSCSI initiator used half the processor (51%) that the Microsoft Windows software initiator on Intel did to deliver the same full bandwidth, doubling the processor effectiveness.



Marvell FastLinQ 41000 Series 10GbE Performance, iSCSI Offload Competitive Evaluation and Storage Spaces Direct Use Cases

Demartek
July 2019

The Marvell FastLinQ 41000 Series hardware initiator achieved line rate throughput for all but the smallest block sizes.



Marvell FastLinQ 41000 Series 10GbE Performance, iSCSI Offload Competitive Evaluation and Storage Spaces Direct Use Cases

Demartek
July 2019

Marvell FastLinQ 41000 Series RDMA

RDMA is the remote memory management capability that allows server-to-server data movement directly between the application memory of each server without any processor involvement. Ethernet-based RDMA requires specialized NICs, sometimes called RNICs.

RoCE and iWARP

RDMA has two main implementations: RoCE and iWARP. For best results, RoCE deployments typically need a lossless fabric supplied by a switch that supports Data Center Bridging (DCB). A combination of Flow Control, Priority Flow Control (PFC), Enhanced Transmission Selection (ETS) and Explicit Congestion Notification (ECN) may be used to improve network performance and guarantee losslessness for RoCE traffic. Care must be taken to determine the optimal configuration for deployment and all switches and adapters must be configured identically as a single dropped packet can be extremely detrimental to RoCE performance. However, when deployed correctly RoCE can provide the most efficient performance.

iWARP does not require a lossless fabric, making deployment simpler as the standard TCP/IP stack is used. However, latency delivered by iWARP could be higher than RoCE, but most deployments like S2D would be unaffected due to this difference.

Both types of RDMA may be used in many use cases, including iSCSI Extensions for RDMA (iSER), Network File System over RDMA (NFSoRDMA), NVMe over Fabrics (NVMe-oF), and Server Message Block Direct (SMB Direct). Microsoft uses SMB Direct in an HCI environment in Windows Server 2016's and 2019's Storage Spaces Direct (S2D), where RDMA traffic is used for cluster inter-node communications. S2D is positioned as the top use case for RDMA – both RoCE and iWARP.

Performance Test Setup

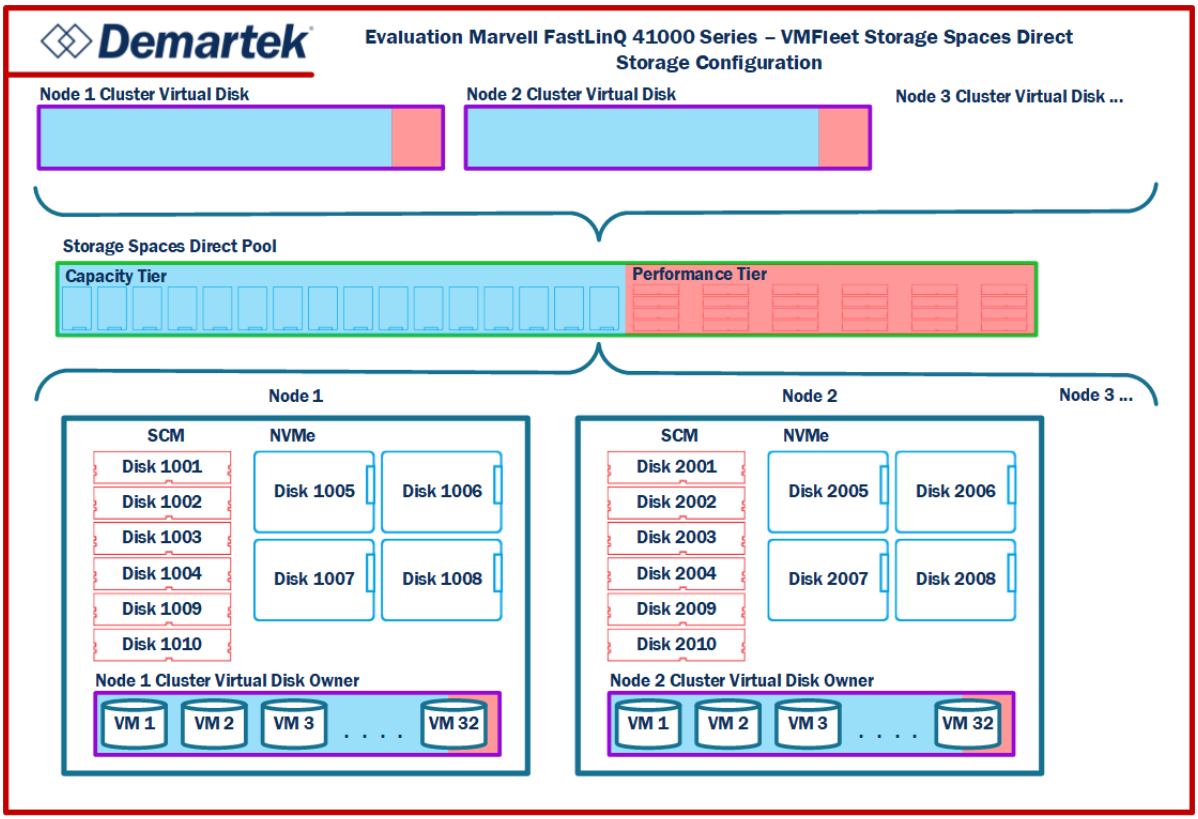
A four node Storage Spaces Direct Cluster was created. Each server node had the following hardware:

> 1xMarvell FastLinQ 41000 Series Adapter.

> 4xNVMe drives.

> 6x16GB NVDIMMs (used as SCM).

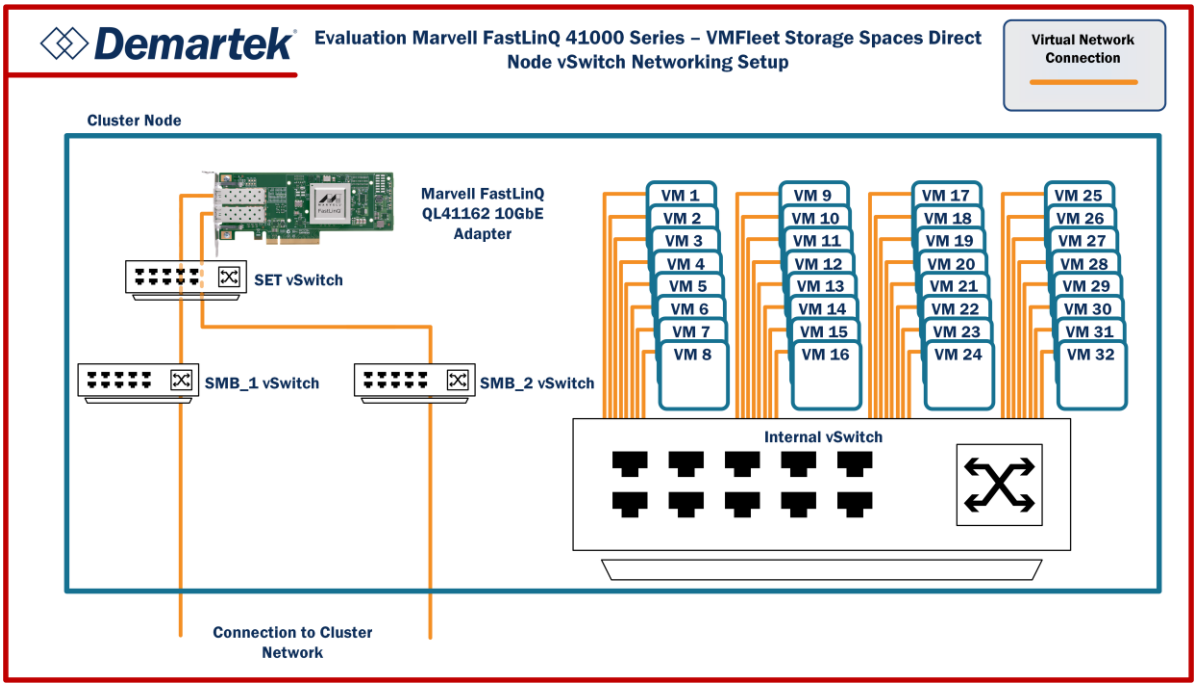
Marvell FastLinQ 41000 Series 10GbE Performance, iSCSI Offload Competitive Evaluation and Storage Spaces Direct Use Cases



An S2D cluster with 3-way mirroring was created from these servers with the SCM NVDIMMs as the performance tier and NVMe as the capacity tier.

Switch Embedded Teaming (SET) with RDMA was used for inter-node communication.

Marvell FastLinQ 41000 Series 10GbE Performance, iSCSI Offload Competitive Evaluation and Storage Spaces Direct Use Cases



VMfleet is an open source tool developed by Microsoft and is used to load and stress servers, typically in order to characterize S2D performance. It consists of VMs deployed on the S2D Storage, and scripts to run coordinated DiskSpd test runs across the VMs, was deployed on the cluster and the total cluster throughput was measured.

VMfleet generates inter-node traffic that is typical of hyper-converged environments, producing ***high-performance on read*** and lower performance on write. As with any environment, the underlying storage media influences the behavior of the environment. SSDs typically have poorer write performance than read performance. However, in addition to this, ***S2D's 3-way mirroring will make additional work when a cluster writes.*** Upon write, three copies of the item must be written to three different nodes. This generates quite a bit of inter-node traffic, and greatly increases the work of each write.
By contrast, upon read, only one copy of the item must be read. In most of these cases, because there are so many copies of the data in the 3-way mirror, there is a high chance the data will be local, and limited network traffic is generated. Lastly, cluster performance will be affected by our S2D storage tier structure.

As a result of differences between read and write performance on the storage media, S2D's 3-way mirror, and S2D's caching, read throughput can be as much as ten times the write throughput in some cases.

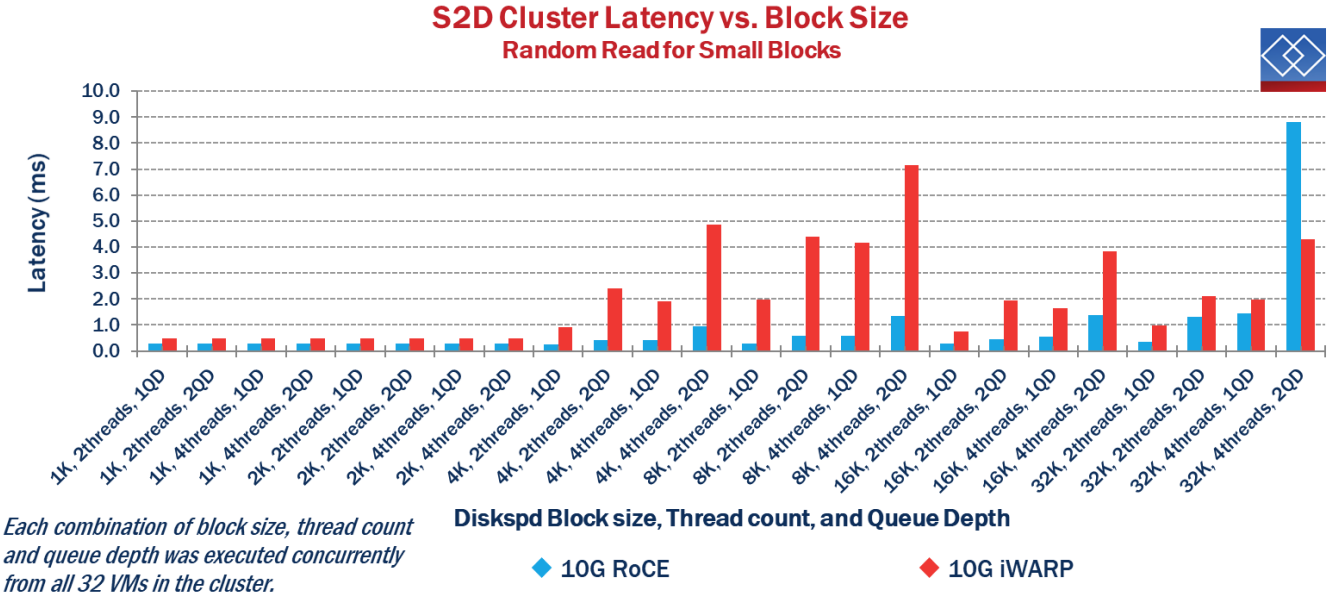
Workloads with both random and sequential access patterns were run, as is standard in most storage performance tests, however, the results showed very little performance difference between the access patterns. For this reason, only random performance is shown here.

The test results displayed on the following pages are across all four nodes of the S2D cluster. This includes the processor utilization charts, which are the average processor utilization across all the nodes. The IOPS and throughput metrics are the S2D-specific metrics and do not include general network traffic.

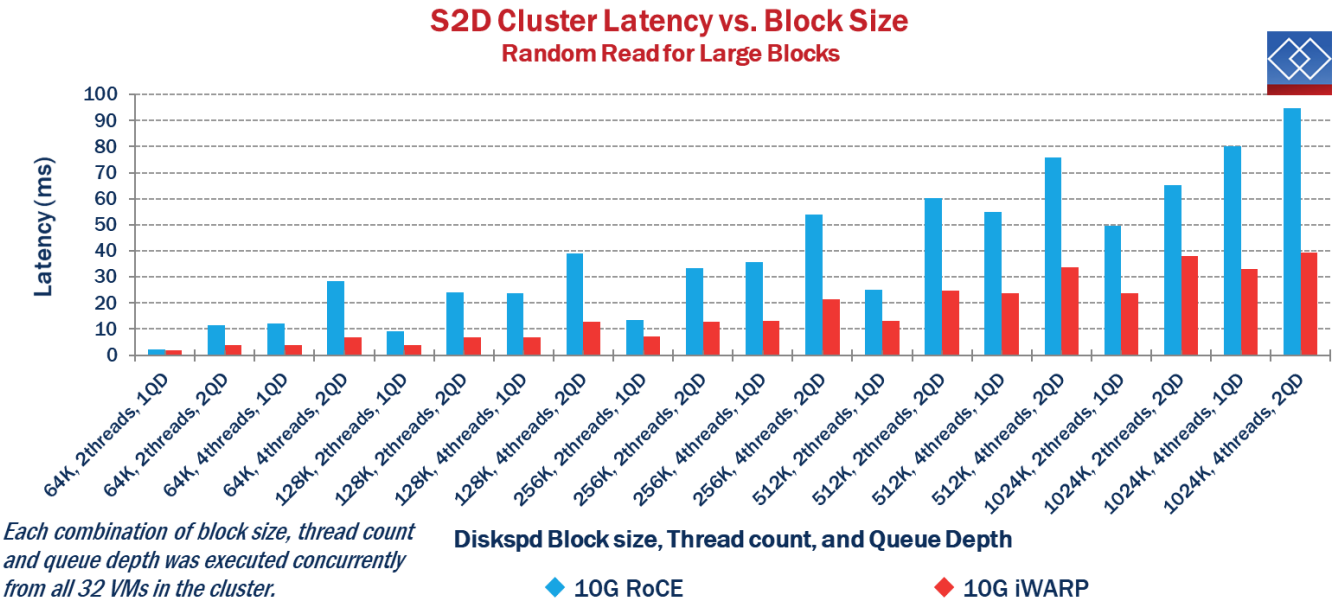
Marvell FastLinQ 41000 Series 10GbE Performance, iSCSI Offload Competitive Evaluation and Storage Spaces Direct Use Cases

Read Performance

For small block reads, iWARP had a higher latency than RoCE. But when block size increased, RoCE had higher latency. However, the differences between them could be considered in the realm of noise.



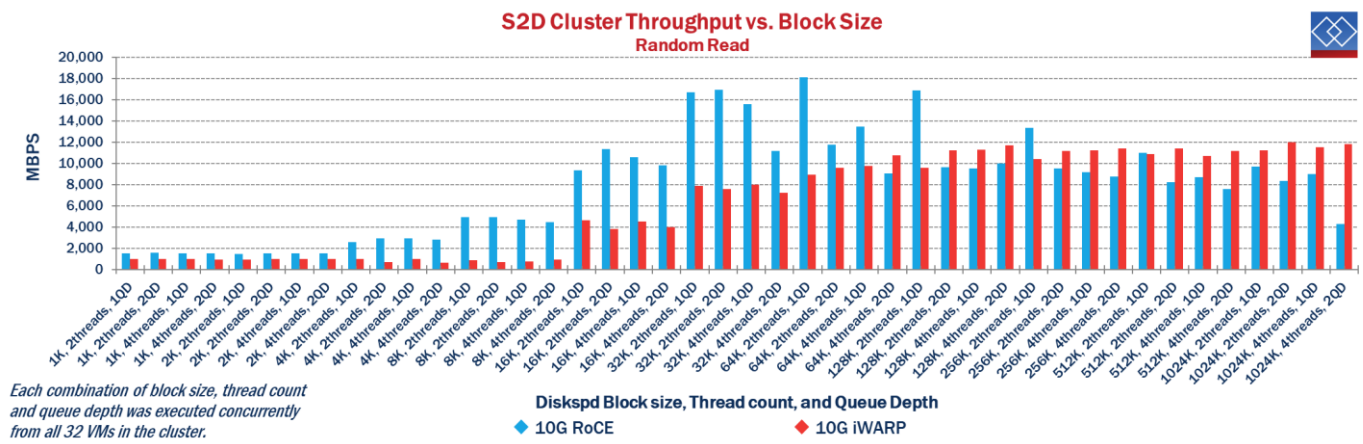
Each combination of block size, thread count and queue depth was executed concurrently from all 32 VMs in the cluster.



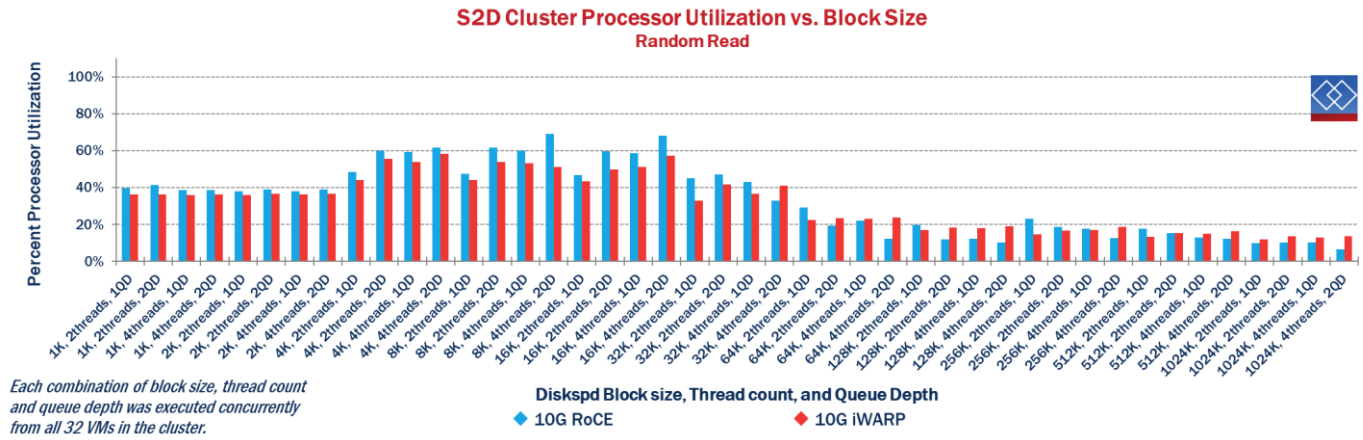
Each combination of block size, thread count and queue depth was executed concurrently from all 32 VMs in the cluster.

Marvell FastLinQ 41000 Series 10GbE Performance, iSCSI Offload Competitive Evaluation and Storage Spaces Direct Use Cases

RoCE yielded higher total cluster throughput than iWARP for the smaller block workloads, but at larger block sizes iWARP delivered slightly higher throughput.



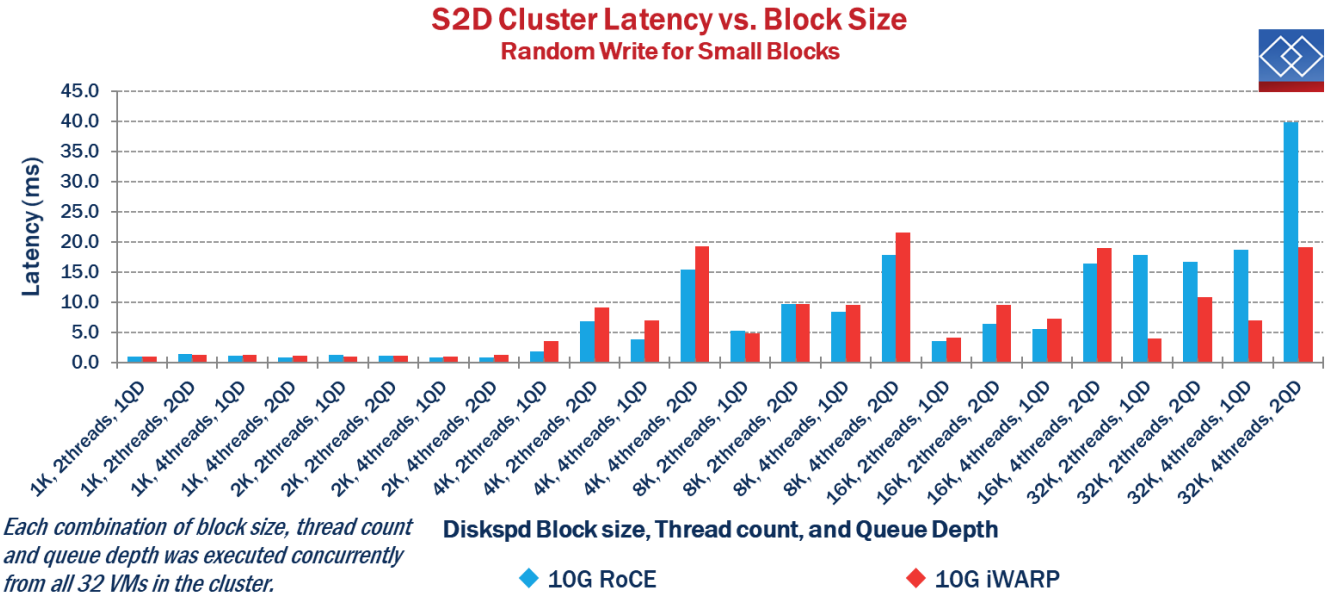
We can see that for both RoCE and iWARP, RDMA pays off, resulting in low processor utilization.



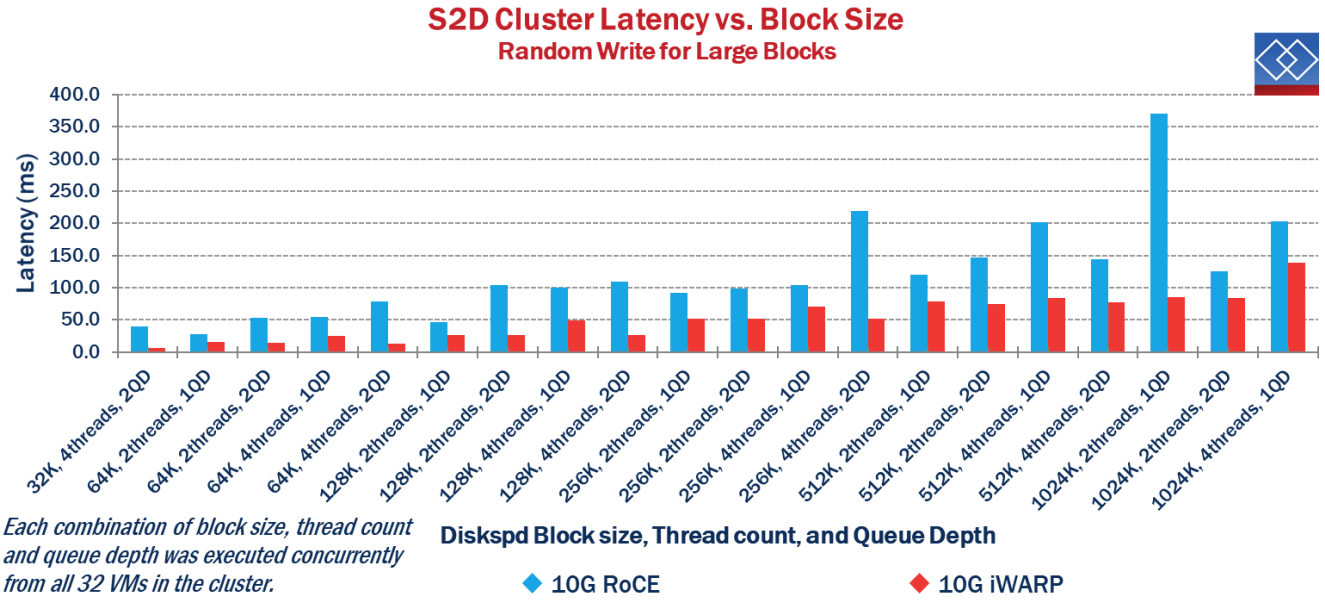
Marvell FastLinQ 41000 Series 10GbE Performance, iSCSI Offload Competitive Evaluation and Storage Spaces Direct Use Cases

Write Performance

Write performance will have higher latency than read due to the underlying storage media and traffic generated by the writes in a 3-way mirror.



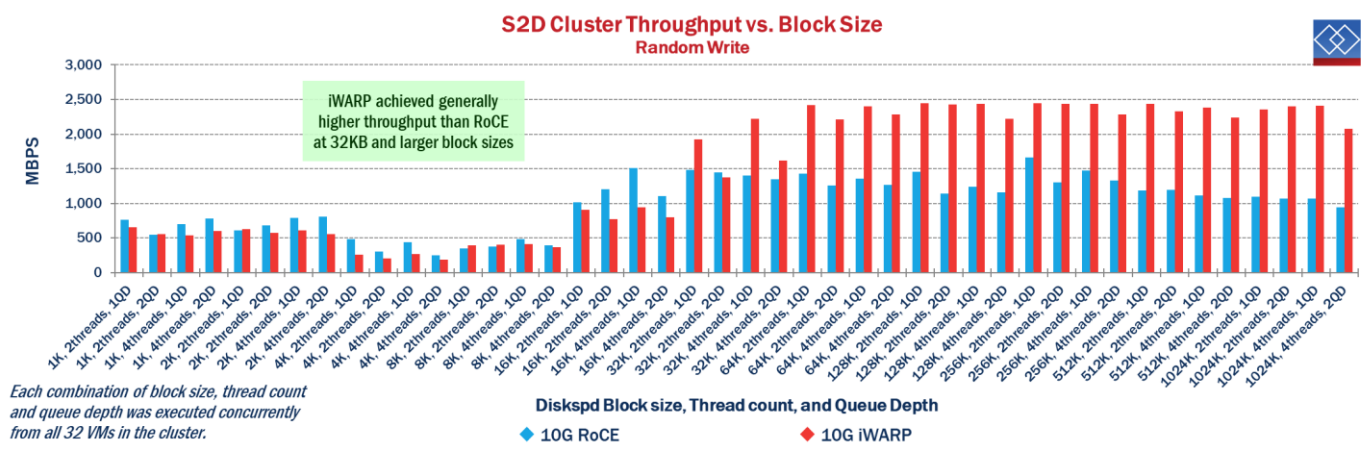
Each combination of block size, thread count and queue depth was executed concurrently from all 32 VMs in the cluster.



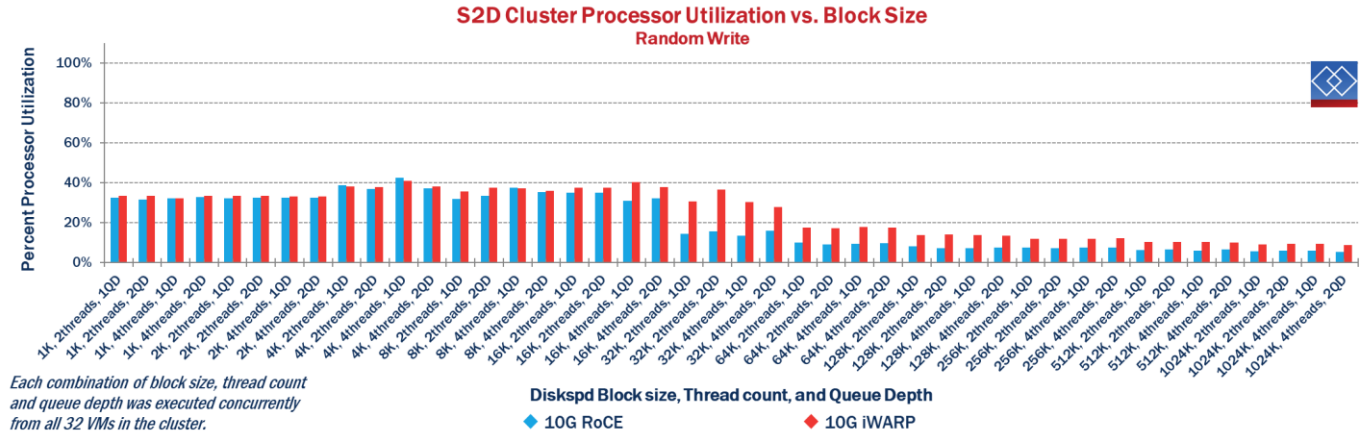
Each combination of block size, thread count and queue depth was executed concurrently from all 32 VMs in the cluster.

Marvell FastLinQ 41000 Series 10GbE Performance, iSCSI Offload Competitive Evaluation and Storage Spaces Direct Use Cases

For small block sizes, the 10GbE random write throughput for RoCE and iWARP was similar. However, at block sizes of 32KB or larger, iWARP achieved generally higher thoughput than RoCE for random writes.



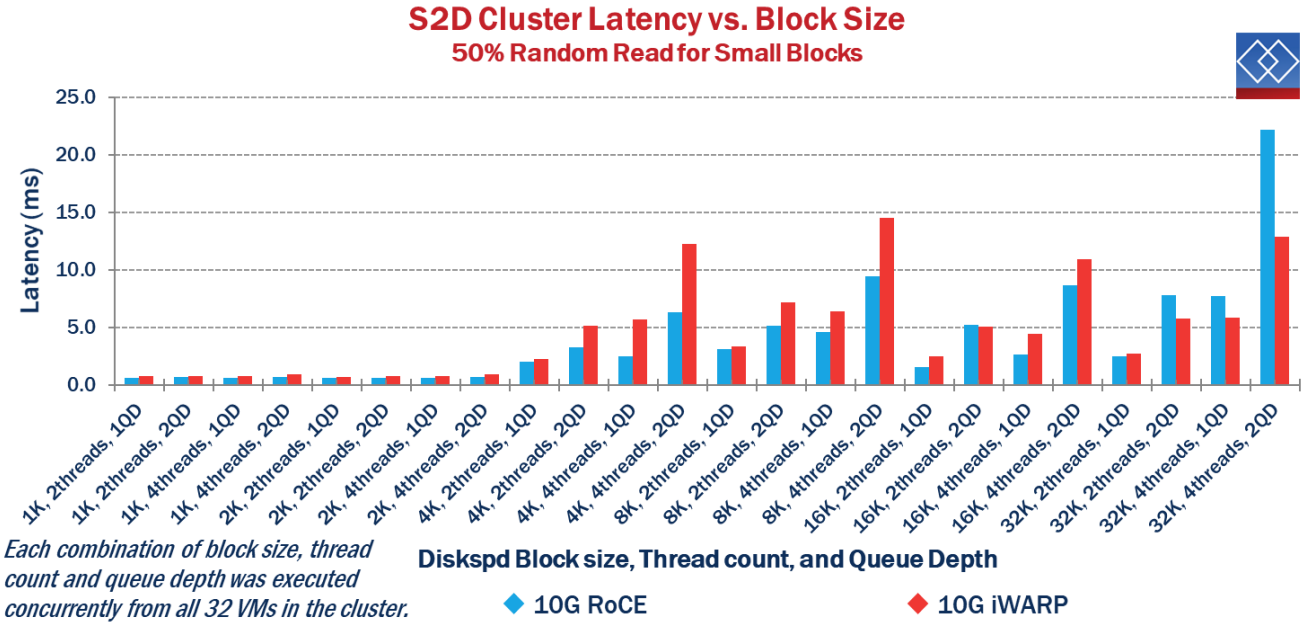
For both types of RDMA, total processor utilization is low. We see more processor utilization for iWARP for large block writes, most likely due to supporting the higher throughput that iWARP has in this case.



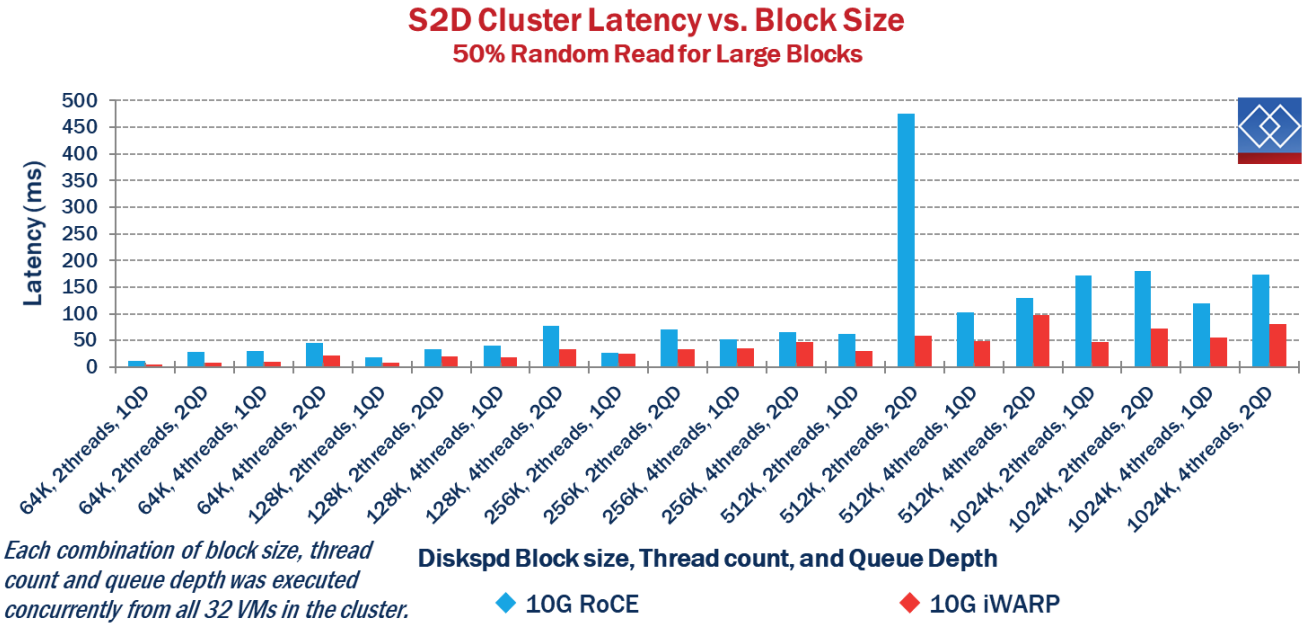
Marvell FastLinQ 41000 Series 10GbE Performance, iSCSI Offload Competitive Evaluation and Storage Spaces Direct Use Cases

50% Read Performance

When performing 50% read, 50% write, the performance is a blend of what we have observed from our 100% read and 100% write testing. Latencies were generally lower for RoCE at the smaller block sizes and lower for iWARP at the larger block sizes. For example, at 4KB block sizes, the 10GbE RoCE latencies were on average about 38% lower than for 10GbE iWARP. At 8KB, RoCE latencies were about 24% lower than iWARP latencies. By contrast, at block sizes of 64KB and higher, the iWARP latencies were, on average, about 51% lower than RoCE.



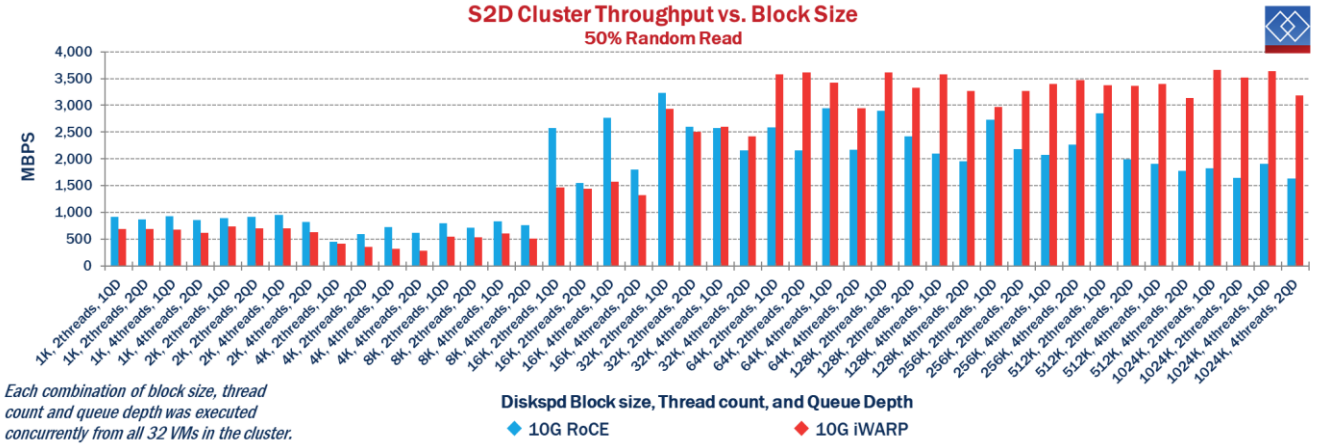
Each combination of block size, thread count and queue depth was executed concurrently from all 32 VMs in the cluster.



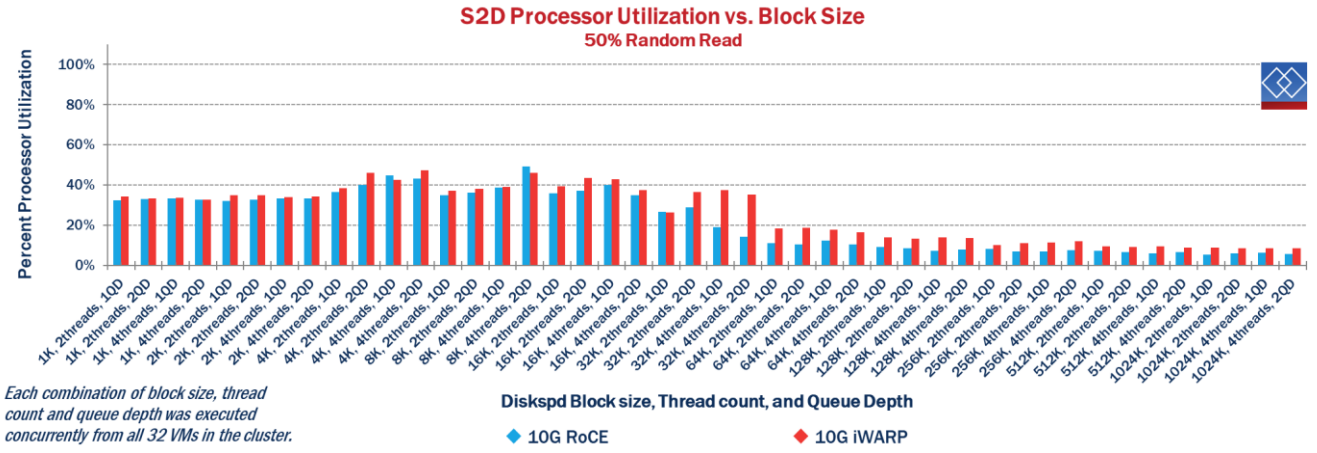
Each combination of block size, thread count and queue depth was executed concurrently from all 32 VMs in the cluster.

Marvell FastLinQ 41000 Series 10GbE Performance, iSCSI Offload Competitive Evaluation and Storage Spaces Direct Use Cases

Similar to what we found in our 100% write testing, at large block sizes iWARP can handle more cluster throughput and uses a larger amount of processor to support that throughput. For this testing, we found that, on average, for 64KB block sizes and higher, iWARP delivered approximately 58% higher throughput than RoCE.



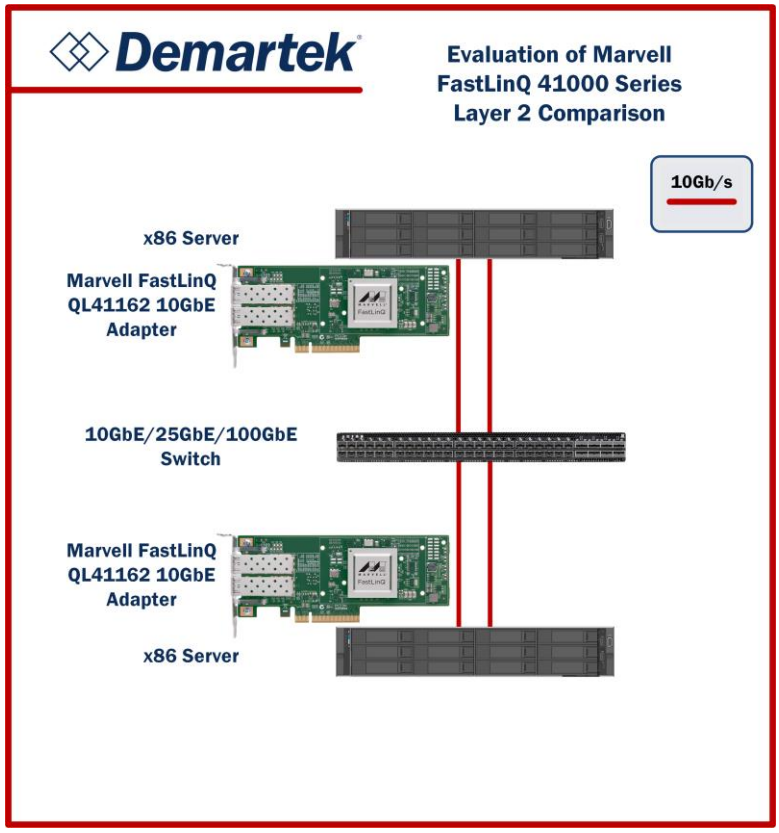
Again, processor utilization is very low due to our use of RDMA.



Marvell FastLinQ 41000 Series 10GbE Performance, iSCSI Offload Competitive Evaluation and Storage Spaces Direct Use Cases

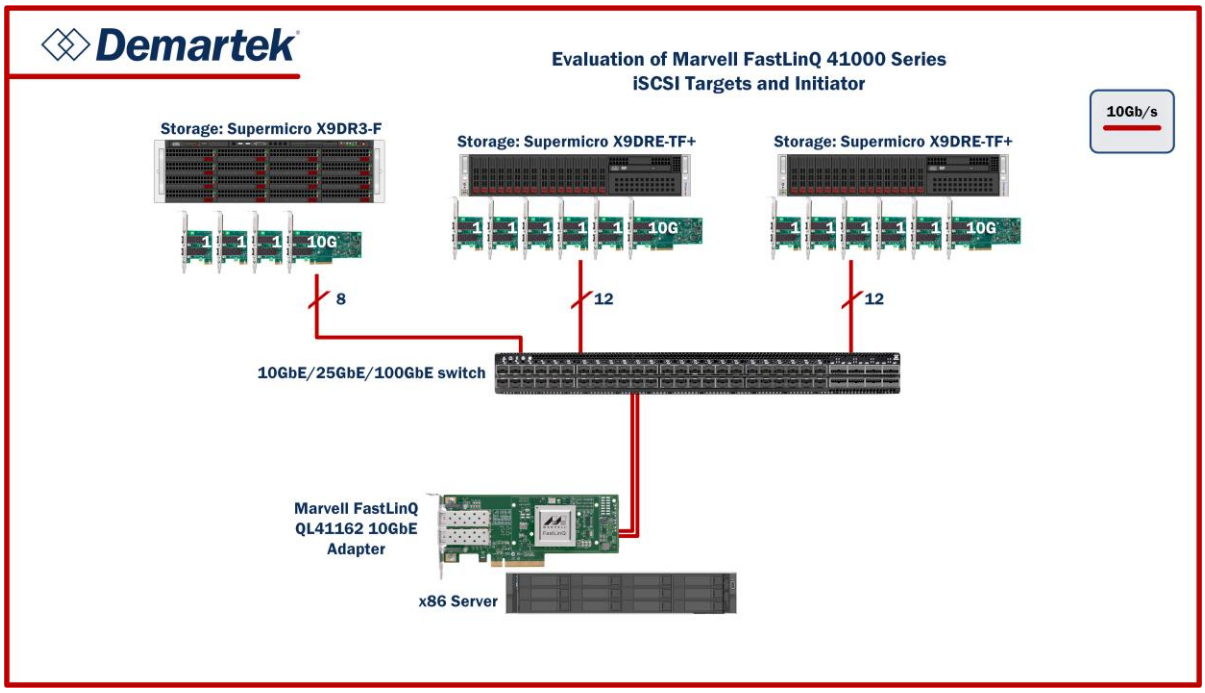
Test Environment

Marvell FastLinQ 41000 Series Level 2 Performance Tests

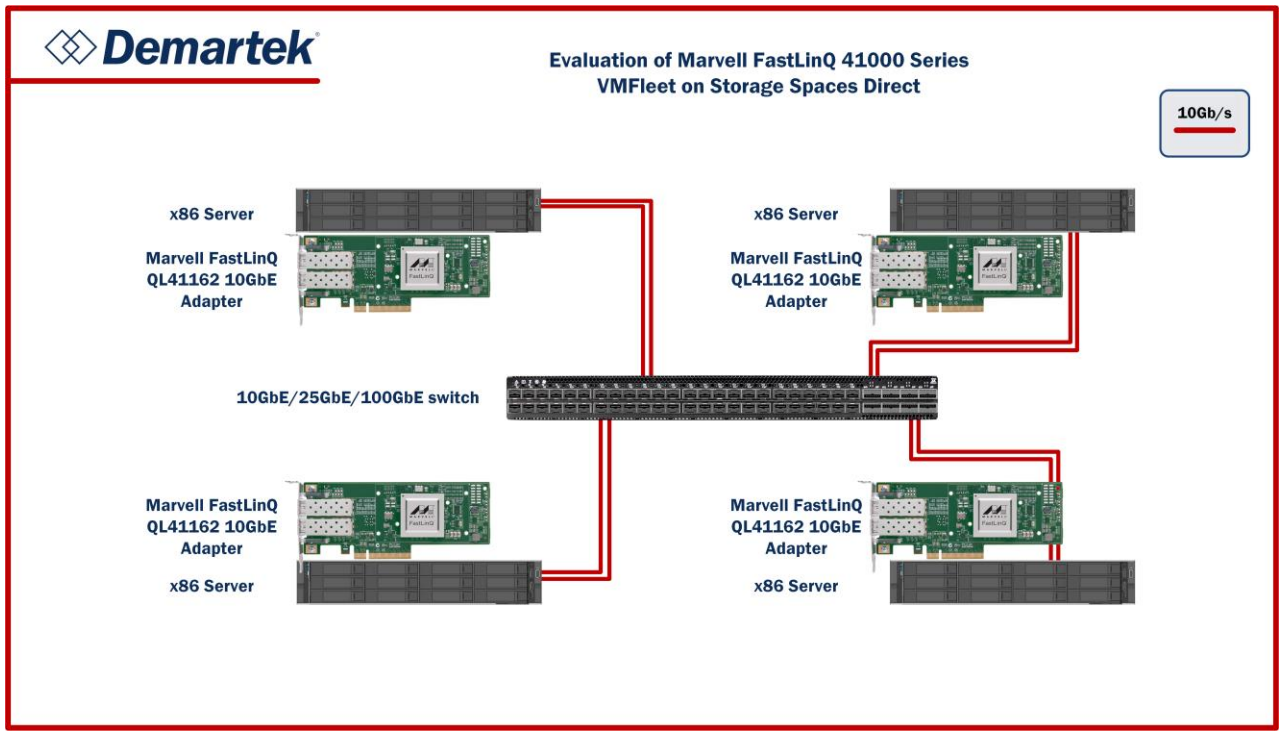


Marvell FastLinQ 41000 Series 10GbE Performance, iSCSI Offload Competitive Evaluation and Storage Spaces Direct Use Cases

Marvell FastLinQ 41000 Series iSCSI Offload vs Intel Converged Network Adapter X710 Software iSCSI Tests



Marvell FastLinQ 41000 Series Storage Spaces Direct Tests



Marvell FastLinQ 41000 Series 10GbE Performance, iSCSI Offload Competitive Evaluation and Storage Spaces Direct Use Cases

Demartek
July 2019

Servers

- 2x Intel Xeon Gold 6130, 2.1GHz, 32 total cores, 64 total threads
- 192 GB Memory
- Microsoft Windows Server 2019 Datacenter Build 17723 (Windows Insider Preview)

Adapters

- Marvell FastLinQ 41000 Series, Boot Code 8.30.10.1, MBI 8.30.13, driver qede 8.30.12.0

OR

- Intel Converged Network Adapter X710 when testing competitor software iSCSI, firmware version 4.25 0x8000143f 0.0.0, driver i40e 1.5.10-k

Switch

- Generic 25GbE/100GbE switch

iSCSI Storage Targets

2x Supermicro X9DRE-TF+

- 2x Intel Xeon E5-2690 v2, 3.0GHz, 20 total cores
- 256 GB Memory
- 6x Dual Port 10GbE NICs from multiple manufacturers
- RedHat Enterprise Linux 7.3
- Targetcli 2.1.fb41

1x Supermicro X9DR3-F

- 2x Intel Xeon E5-2690, 2.9GHz, 16 total cores
- 192 GB Memory
- 4x Dual Port 10GbE NICs from multiple manufacturers
- RedHat Enterprise Linux 7.3
- Targetcli 2.1.fb41

Marvell FastLinQ 41000 Series 10GbE Performance, iSCSI Offload Competitive Evaluation and Storage Spaces Direct Use Cases

Demartek
July 2019

Summary and Conclusion

The Marvell FastLinQ 41000 Series is an excellent choice for current generation servers with Intel Xeon Scalable processors. While offering a broad choice of offloads and Universal RDMA, Marvell also achieves great performance.

> The Marvell FastLinQ 41000 Series achieved line rate bidirectional performance for buffer sizes of 1KB up to 1MB.

> The Marvell FastLinQ 41000 Series hardware iSCSI initiator achieved an average of 7.2 times the IOPS of Linux Software Initiator with Intel for unidirectional workloads. Results were similar for the bidirectional workloads.

> The Marvell FastLinQ 41000 Series hardware iSCSI initiator achieved an average of 4.6 higher IOPS for 8KB block sizes and an average of 5.2 times higher IOPS for 32KB block sizes as compared to the Linux software initiator.

> The Marvell FastLinQ 41000 Series hardware iSCSI initiator used half the processor (51%) at 8KB block size that the Microsoft Windows software initiator did to deliver the same full bandwidth, doubling the processor effectiveness.

> Our S2D cluster built with Marvell FastLinQ 41000 Series Universal RDMA achieved on average 10,470 MBPS read throughput and 1,794 MBPS write throughput while using on average 16% or less of available cluster processor.