Lenovo - Mellanox - Excelero NVMesh® Reference Architecture

How adding a dash of software and intelligent interconnect to your server platform turns DAS into a high performance shared storage solution.

Introduction

Following the example of Tech Giants like Google, Facebook and Amazon, enterprises and service providers are increasingly using industry standard servers paired with high speed networking and intelligent software to run their scale-out applications. This approach is sometimes coined as the software-defined data center (SDDC). The promise of SDDC is to provide flexibility and reliability - without compromises - for environments of any scale. Customers need the ability to deploy storage and compute infrastructures starting with a few servers and scale out limitlessly with guaranteed reliability, predictable performance and seamless integration with their infrastructure and applications while maintaining operational simplicity and good ROI.

Current generations of servers are flexible systems, designed to enable the SDDC. They incorporate strong CPU power, high-speed networking, local high-performance flash storage and memory. Advanced compute and clustering methods such as parallelization enable pooling of CPU resources across multiple systems to tackle large compute jobs. Pooling local high-performance storage resources such as NVMe across servers has proven to be more challenging as you either compromise on storage performance, use valuable CPU for storage (vs. applications) or both. Excelero NVMesh was designed to solve exactly this problem: it enables efficient pooling of NVMe flash resources across multiple servers without impacting target CPU and thus saving the most powerful (and expensive) resource for its main purpose, the applications.

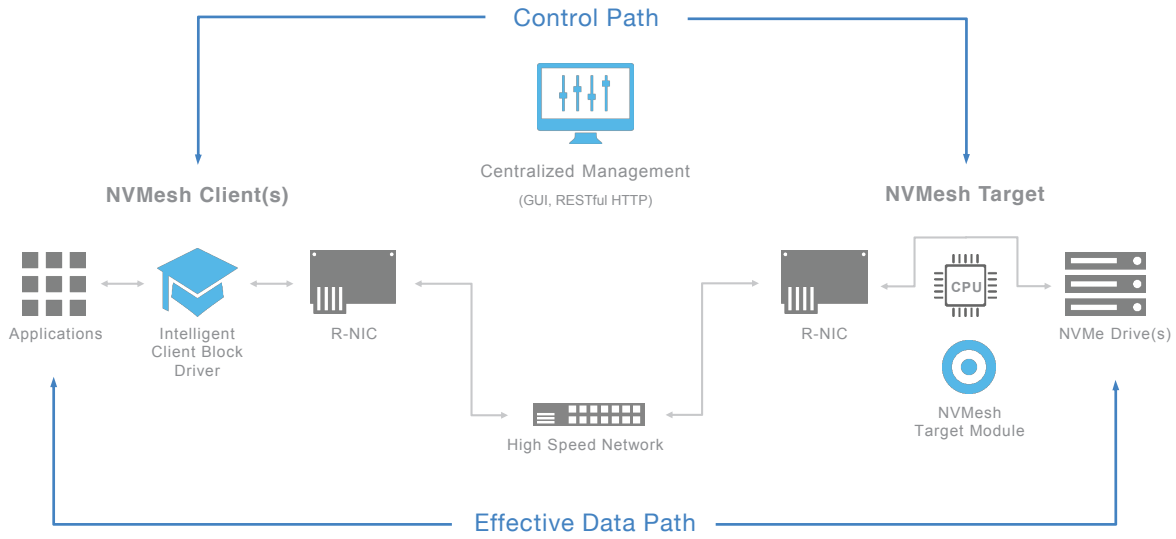
This paper describes how you can leverage Lenovo advanced server designs, Mellanox ConnectX-4 high performance network adapters and Excelero NVMesh Virtual SAN software to take the hardware you would normally purchase for an application cluster and turn it into both a compute AND a high-performance storage solution. You can do this without sacrificing valuable application CPU. With this "just add software" approach, you can remove storage bottlenecks without the need to purchase external storage arrays or purchase additional servers dedicated to storage purposes.

Excelero NVMesh Virtual SAN

NVMesh is a Software-Defined Block Storage solution that forms a Non-Volatile Mesh, a distributed block layer that allows unmodified applications to utilize pooled NVMe storage devices across a network at local speeds and latencies. Distributed NVMe storage resources are pooled with the ability to create arbitrary, dynamic block volumes that can be utilized by any host running the NVMesh block client. These virtual volumes can be striped, mirrored or both while enjoying centralized management, monitoring and administration. In short, applications can enjoy the latency, throughput and IOPs of a local NVMe device while at the same time getting the benefits of centralized, redundant storage.

A key component of Excelero's NVMesh is the patented Remote Direct Drive Access (RDDA) functionality, which leverages RDMA to bypass the target CPU and, thus, avoids the noisy neighbor effect for the application. The shift of data services from centralized CPU to client side distribution enables unlimited linear scalability, provides deterministic performance for applications and enables customers to maximize the utilization of their flash drives. Elastic Virtual SAN is deployed as a virtual, distributed non-volatile array and supports both converged and disaggregated architectures, giving customers full freedom in their architectural design.

The architecture is simple and comprised of three components: An intelligent block client (a Linux block driver) installed on any host that wants to utilize NVMesh logical volumes. Any hosts with an NVMe device load the target enablement module. Lastly, the management module is loaded on at least one host on the network to provide centralized management, monitoring and configuration. While the components are pictured below on separate physical hosts, it should be noted that it's common for the NVMesh client and target to be present on the same physical host in converged environments.



Lenovo Servers

Lenovo rack servers feature innovative hardware, software and services that solve customer challenges today and deliver an evolutionary fit-for-purpose, modular design approach to address tomorrow's challenges. These servers capitalize on best-in-class, industry-standard technologies coupled with differentiated Lenovo innovations to provide the greatest possible flexibility in x86 servers.

x3650 M5 2U 2 Socket Server

Key advantages of deploying Lenovo rack servers include:

- #1 in Reliability – An independent survey of 550 companies shows Lenovo servers have the industry's highest availability[1].
- Multiple world records for performance[2].
- High-efficiency power supplies with 80 PLUS Platinum and Titanium certifications.
- Lenovo servers use hexagonal ventilation holes, a part of Calibrated Vectored Cooling™ technology. Hexagonal holes can be grouped more densely than round holes, providing more efficient airflow through the system.
- Expansive storage capacity and flexible storage configurations for optimized workloads
- The System x3650 M5 server supports up to 8 SFF front hot swap NVMe drives from 400GB to up to 16TB of total capacity.

For cloud deployments, database, or virtualization workloads, trust Lenovo rack servers for world-class performance, power-efficient designs and extensive standard features at an affordable price.

The NVMesh architecture can take advantage of locally attached NVMe drives to increase utilization since excess capacity is not orphaned in remote nodes. Given the server's flexibility the customer only needs to buy the minimum capacity and can scale at a later date.

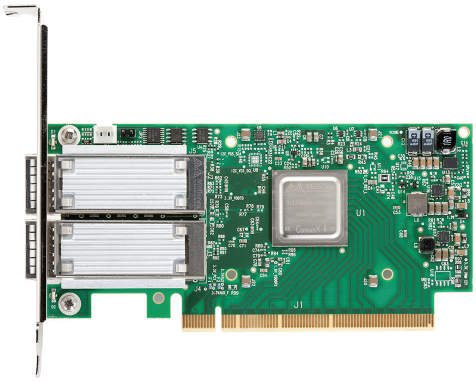
1. ITIC Global Server HW: http://www.lenovo.com/images/products/system-x/pdfs/whitepapers/ itic_2015_reliability_wp.pdf
2. Based on #1 in x86 2P SPECvirt_sc2013PPW, SPECvirt_sc*2013, SPECvirt_sc2013 ServerPPW, TPC-E, SPECfp*_rate_base2006, SPECfp

Mellanox ConnectX-4

The Mellanox ConnectX-4 adapter cards with Virtual Protocol Interconnect (VPI), supporting EDR 100Gb/s InfiniBand and 100Gb/s Ethernet connectivity, provide the highest performance and most flexible solution for high-performance, Web 2.0, Cloud, data analytics, database, and storage platforms.

With the exponential growth of data being shared and stored by applications and social networks, the need for high-speed and high performance compute and storage data centers is skyrocketing.

ConnectX-4 provides exceptional high performance for the most demanding data centers, public and private clouds, Web2.0 and Big Data applications, as well as High-Performance Computing (HPC) and Storage systems, enabling today's corporations to meet the demands of the data explosion.



Key Benefits of Mellanox ConnectX-4

- Highest performing silicon for applications requiring high bandwidth, low latency and high message rate
- World-class cluster, network, and storage performance
- Smart interconnect for x86, Power, ARM, and GPU-based compute and storage platforms
- Cutting-edge performance in virtualized overlay networks (VXLAN and NVGRE)
- Efficient I/O consolidation, lowering data center costs and complexity
- Virtualization acceleration
- Power efficiency
- Scalability to tens-of-thousands of nodes

The Mellanox ConnectX-4 adapters provide an unmatched combination of bandwidth, the lowest available latency, highest messages per second rate and application hardware offloads, addressing both today's server needs as well as the next generation of compute and storage data center demands. It includes native hardware support for RDMA over Converged Ethernet and Ethernet stateless offload engines which are key for RDDA performance, which makes them a perfect match for Excelero NVMesh.

Excelero Virtual SAN Software on Lenovo Servers Connected with Mellanox Networking

By combining the Lenovo advanced server designs and Mellanox network adapters with Excelero NVMesh software, customers can use the hardware they deployed to run their application cluster to run a virtual, high-performance Server SAN. With this software-only approach, customers remove storage bottlenecks without the need to purchase external storage arrays or additional servers dedicated to storage purposes.

Benefits of the Joint Solution

- Build a high-performance Virtual SAN without needing to purchase additional or dedicated servers.
- Utilize pooled NVMe storage across a network at local speeds and latencies to maximize NVMe utilization and avoid data locality issues for applications.
- Logical disaggregation of storage and compute allows applications to leverage the full capability of CPU.
- Scale performance linearly at near 100% efficiency by shifting data services from centralized CPU to client side distribution.

Reference, Tested Architecture

Hardware

Lenovo Servers

	Type	Details
4 X 3650 M5 Servers	CPU	Dual Intel Xeon E5-2699v4
	Memory	512GB
	NIC Card	2 x Mellanox 40GbE – only 1 port used for testing
	NVMe Card	2 x Intel P3700 1.6TB NVMe 2.5" G3HS

Mellanox/Lenovo Network (Lenovo branded)

NIC: 2x Mellanox ConnectX-4 100GbE/EDR IB QSFP28 VPI Adapter (running at 40GbE for this testing) with its Industry-leading throughput and latency performance.

Switch

The Lenovo RackSwitch™ G8332 provides low latency, lossless performance and a feature-rich design with key virtualization features, such as Converged Enhanced Ethernet (CEE)/Data Center Bridging (DCB), high availability, and enterprise class Layer 2 and Layer 3 functions. The RackSwitch G8332 also delivers excellent cost savings as you consider acquisition costs, energy costs, operational expenses, and ease of use and management for a 40 Gbps class switch. The RackSwitch G8332 has 32 QSFP+ ports and is suitable for clients who use 10 Gigabit Ethernet or 40 Gigabit Ethernet connectivity (or both).

Software

TCentOS 7.2.1511, kernel 3.10.0-327.36.1.el7.x86_64
Mellanox OFED version 3.3-1.0.4
Excelero NVMesh 1.1

Scope, Goal & Success Criteria

Excelero NVMesh is the only solution that enables a 100% converged infrastructure with no CPU penalty by offering full logical disaggregation of storage and compute. Additionally, NVMesh software allows for local flash resources to be pooled and then logically distributed over a compute cluster – allowing maximum utilization of capacity, IOPs and bandwidth of this precious investment. Together with Lenovo, it's possible to build a joint solution that delivers the best $/GB/s or $/IOPs for high-performance block storage.

A team of Lenovo, Mellanox and Excelero engineers joined forces to deploy an Excelero NVMesh on Lenovo hardware and to run CPU and storage performance tests on the environment.
There were four key objectives:

- To validate functional operation of the joint solution (interoperability)
- To demonstrate the low-to-zero impact of the Excelero software on the CPU of the servers.
- To measure IO and throughput capabilities of the Virtual SAN
- To demonstrate the architectural flexibility of the Excelero NVMesh architecture and ease-of-use

Success would be measured by how close NVMesh testing, distributed over shared logical volumes, could come to the published (local) SSD performance numbers.

Performance Test Methodology

All testing was performed using a random IO pattern, with a 70/30 Read/Write ratio. This was chosen as a representation of a mixed environment workload. Two tools were chosen to perform the benchmark testing. Linux fio (often used by SSD manufactures for their published benchmarks) was utilized to verify local SSD performance and to demonstrate the effect of increasing the number of outstanding IOs (threads x queue depths) against the remote logical volumes. Vdbench was subsequently used to run a controlled, simultaneous series of benchmarks more representative of enterprise-style storage testing.

The first step in the Lenovo - NVMesh benchmark consisted of characterizing the selected NVMe devices on a local host to validate the local latency, IOPS and throughput capabilities of the devices. This testing was performed with a general-purpose IO testing tool (Linux, fio). All eight devices on all four servers were found to meet or slightly exceed their published specification number of 240K 4K IOPs with a 70/30 R/W ratio.

NVMesh logical volumes were created using the same devices, but with the physical devices accessed over the network from the NVMesh client utilizing remote NVMesh targets. 4 logical striped (RAID-0) volumes were created, with each volume spanning ¼ all 8 drives. The 4 logical volumes where then attached to each server, with each server seeing all 4 logical volumes as block devices in /dev.

Modern low-latency Ethernet networks utilizing RDMA over Converged Ethernet (RoCE) protocol (v2), when combined with NVMesh software should exhibit only 5-6µs of additional overhead for RAID-0 volume operations vs. the locally measured latencies of the same devices. With adequate network bandwidth, IOPs and throughput should be virtually identical. Thus, if a local NVMe device (via the open source NVMe block driver) responded to read requests in 80µs, an NVMesh remote access utilizing the same device should respond in approximately 86µs. Thus, performance of (remote) SSDs in NVMesh logical volumes should be virtually indistinguishable from local SSDs.

Results

Software installation was completed in less than 30 minutes (from beginning to “ready for IO”), without any issues. The intuitive GUI allowed the administrator to inspect the system and prepare logical volumes for the real tests.

Remote IOs (via logical volumes), Single Outstanding IOs

All four servers, each attached to all 4 logical volumes ran a brief fio, 4K block size test with a single thread, and a single queue-depth to gauge the round-trip response time to one of the logical volumes. The results for reads from fio from each of the four servers follow:

clat (usec): min=0, max=7321, avg=85.03, stdev=105.59
clat (usec): min=0, max=7364, avg=85.55, stdev=106.30
clat (usec): min=0, max=8053, avg=85.89, stdev=105.31
clat (usec): min=0, max=9650, avg=86.14, stdev=104.79

This demonstrated that the remote drive reads for single IOs are completing in about 86µs; equivalent to local NVMe performance.

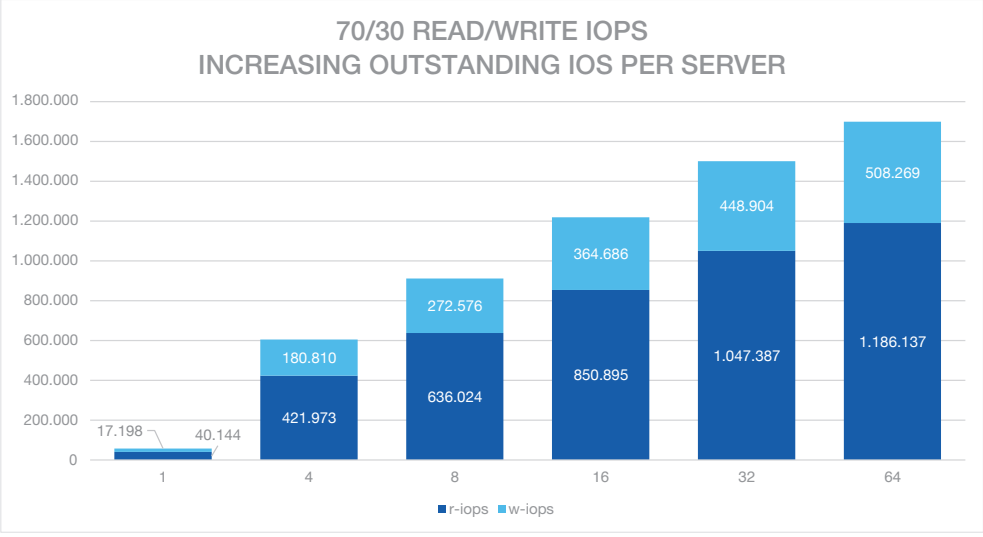
The results for writes from fio from each of the four servers follow:

clat (usec): min=0, max=6932, avg=19.02, stdev=13.67
clat (usec): min=0, max=8155, avg=19.55, stdev=13.99
clat (usec): min=0, max=6037, avg=19.92, stdev=11.18
clat (usec): min=0, max=7520, avg=20.16, stdev=15.82

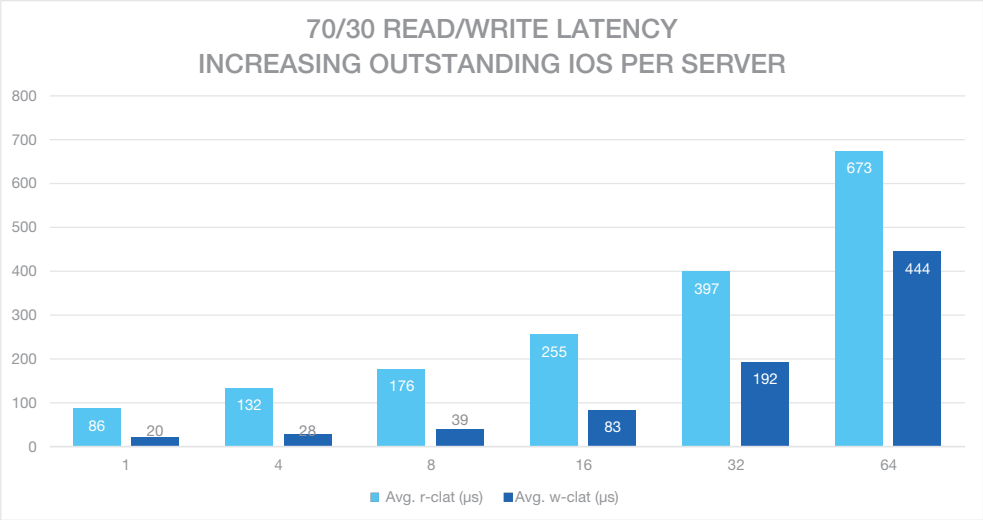
This demonstrated that the remote drive writes for single IOs are completing in about 20µs; equivalent to local NVMe performance.

Scaling Tests Using fio, Increasing Outstanding IOs

After verifying NVMesh overhead was negligible, fio ran simultaneously on all four servers to measure IOPs and response time over increasing numbers of outstanding IOs by utilizing 4 threads and incrementally increasing the queue depth. The results are shown in the following graphs depicting the aggregate results of all four servers:



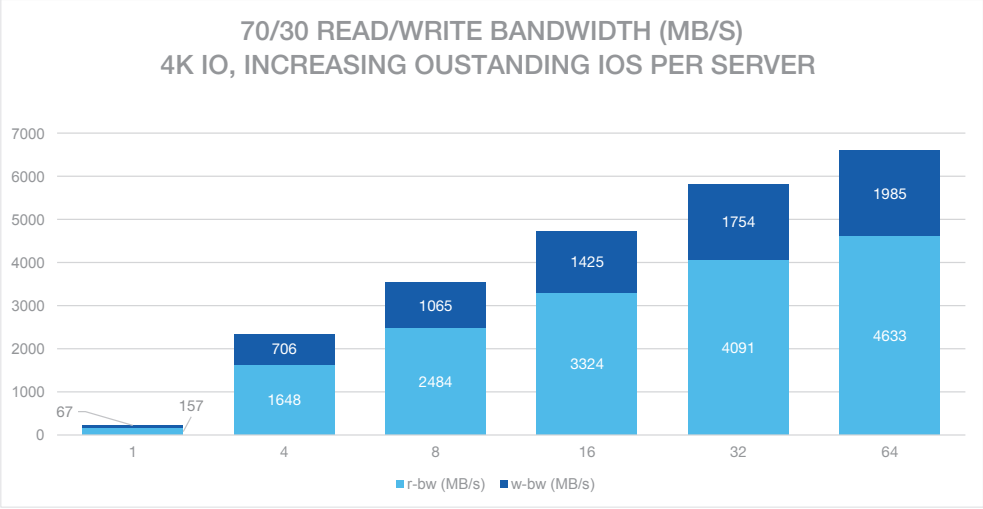
The graph demonstrates linear growth in both read and write IOPs with increases in queue depth per server. The average latency for those reads and writes is shown below:



The two graphs together demonstrate over 1.6 million random read/write IOPS at latencies far below the most performant all-flash-arrays. Further, at only about 1 million IOPs, latencies average below 200µs for reads and below 50µs for writes.

While IOPs could have been driven higher by further increasing threads average latencies would have approached 1ms and the decision was made to change focus to Vdbench.

Bandwidth scaling was equally impressive:



The graph demonstrates 4K IO totaling about 6.6GB/s of aggregate bandwidth.

Varying Block Size Testing Utilizing Vdbench

After running some simple scaling tests with fio, testing moved to the more enterprise-focused Vdbench test tool. The desire was to demonstrate IOPs and bandwidth over various block sizes in a converged fashion with the Vdbench Java processes emulating application IO.

The total summary results from the VDbench runs are below:

i/o rate	MB/sec 1024**2	bytes i/o	read pct	read resp	write resp	resp stddev	queue depth	cpu% sys+u	cpu% sys
Starting RD=4krand; I/O rate: Uncontrolled MAX; elapsed=3600; For loops: threads=256									
2122680.66	8291.72	4096	70.00	0.581	0.242	0.692	1018.3	7.1	4.5
Starting RD=8krand; I/O rate: Uncontrolled MAX; elapsed=3600; For loops: threads=128									
1117453.09	8730.10	8192	70.00	0.597	0.125	0.689	509.1	4.4	2.7
Starting RD=16krand; I/O rate: Uncontrolled MAX; elapsed=3600; For loops: threads=64									
535310.28	8364.22	16384	70.00	0.637	0.098	0.699	254.5	2.3	1.4
Starting RD=32krand; I/O rate: Uncontrolled MAX; elapsed=3600; For loops: threads=32									
248906.73	7778.34	32768	70.01	0.691	0.090	0.708	127.1	1.2	0.7
Starting RD=64krand; I/O rate: Uncontrolled MAX; elapsed=3600; For loops: threads=16									
110526.75	6907.92	65536	70.00	0.778	0.098	0.716	63.5	0.7	0.4

The above shows some very interesting results. From an IOPs point of view, the servers produced over 2.1M IOPs without impacting target CPU while serving over 500K IOPS per server. This is evidenced by the low CPU utilization reported by Vdbench in the last two columns.

Secondarily, the results demonstrate the ability to extract all the performance that the NVMe drives can provide. The tested drives are rated to perform at approximately 2 million random 70/30 R/W IOPS (240K IOPs per drive x 8 drives). The results above demonstrate complete drive saturation. Pushing the drives to their maximums increased average read response time to 581µs and average write response time to 242µs. As demonstrated in the fio results above, reducing the IOP load or increasing the number of drives can make these dramatically low latencies even lower.

For bandwidth, testing demonstrated a maximum aggregate bandwidth of 8.73GB/s utilizing an 8K block size. This also effectively hit the drives published maximum bandwidth at 8K IO of 8.75GB/s.

Conclusion

The results show that by simply adding an optimized software layer (NVMesh) it's possible to turn multiple servers with local drives into a high performance shared storage solution. It demonstrates the ability to extract full performance and utilization out of NVMe SSD resources by adding a "dash of software" to hardware platforms common in scale-out data center designs. Further, it's possible to achieve higher performance levels than proprietary all-flash-arrays at a much lower $/IOP, or $/GB/s. Marrying high-performance, reliable, standard servers from Lenovo with innovative and revolutionary software from Excelero enables what was previously unattainable: The cost savings of standard servers, the performance of local flash with the convenience and protection of centrally managed storage. Once again this is a white paper to show what is possible, and Lenovo has not made any commitments to bring this to market as a product or appliance.