Shared NVMe™ Becomes Mainstream With NVMe Over TCP Software-Defined Storage

Micron® 9300 drives high-performance Lightbits Labs™ LightOS® deployments

Overview

Modern, scale-out storage solutions have recently flooded the enterprise cloud/data center market. They all have one thing in common. They focus on flash. But while they all deliver flash storage — the fastest mainstream storage solution available today — not all offerings provide the same set of capabilities.

Micron understands that fast flash, like the Micron® 9300 family of NVMe™ solid-state drives (SSDs) enabling fast data analytics and scalable, low-latency transaction processing, is an essential part of a complete enterprise storage solution. Very early on, Micron focused on advanced storage access options for SSDs such as NVMe and NVMe over Fabrics™ (NVMe-oF™).

NVMe-oF has captured attention because it enables storage administrators and storage vendors alike to have easy access to flexible options for deploying high-performance flash storage. By extending NVMe — historically only available in server-local deployments — beyond the server via networking infrastructures, NVMe-oF supports separating high-performance storage services from the application servers using them.

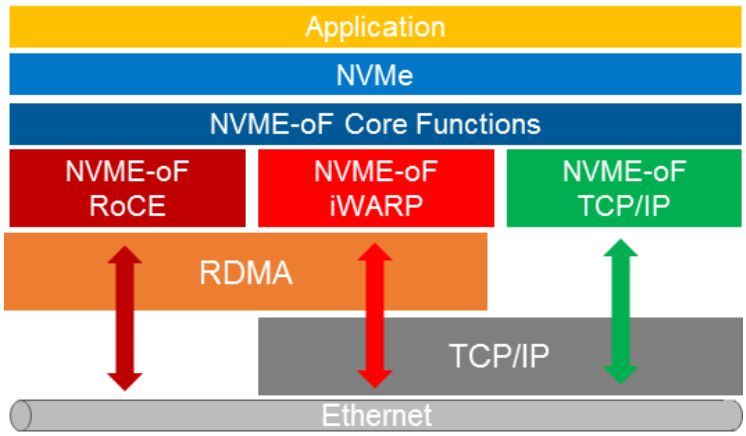


Figure 1: Comparison of NVMe over Fabrics Ethernet Implementations

NVMe over TCP

NVMe over TCP (NVMe/TCP) is the latest development in shared NVMe. NVMe/TCP differs from other NVMe-oF implementations. Unlike NVMe over Converged Ethernet (RoCE), NVMe/TCP does not require complex, potentially more costly networking infrastructures that depend on remote direct memory access (RDMA) (see Figure 1), which itself depends on advanced Ethernet functionality called data center bridging (DCB).

Deploying an advanced networked storage option for NVMe across well-understood Ethernet infrastructures using TCP/IP can dramatically reduce the cost and complexity of sharing NVMe while still offering a high-performance storage solution. To learn more about this new standard, visit the NVM Express™ website at https://nvmexpress.org/welcome-nvme-tcp-to-the-nvme-of-family-of-transports/.

NVMe/TCP is a new way to share NVMe storage. To test NVMe/TCP capabilities, we selected Lightbits Labs™, which led the development of the NVMe/TCP standard and was one of the first to market with a commercial NVMe/TCP solution.

Lightbits Labs offers a software-defined hardware-accelerated solution. The first component is the Lightbits LightOS® NVMe/TCP software-defined storage (SDS). LightOS, running on one or more standard x86 servers populated with NVMe SSDs, becomes an NVMe target that application servers can access for storage resources such as databases, webpages or traditional file services.

The second component is the optional Lightbits Lightfield™ storage acceleration PCIe add-in card that enhances the performance and efficiency of the storage resources by relieving the CPU from managing the storage. Lightfield is a reprogrammable device that performs several advanced storage functions for the solution, including NVMe/TCP offload, erasure encoding offload, inline data compression and compaction, and flash storage management.

The Test

We compared the performance and latency of two database configurations:

1. Cassandra servers using local NVMe SSDs (Figure 2a)

   For local NVMe configuration tests, each Cassandra server contained 2x Micron 3.84TB 9300 PRO NVMe SSDs. Configured as a single, striped 4.9TB logical volume using native LVM (logical volume manager), this logical volume stored the local database on each server. The four servers connected using 100GbE to a single switch.

2. Cassandra servers using shared NVMe storage over a 100GbE network (Figure 2b)

   Remote NVMe configuration tests used 8x Micron 3.84TB 9300 PRO NVMe SSDs in a single Linux server running Lightbits LightOS NVMe/TCP software-defined storage solution. To further minimize the impact of the network for writes, the Lightbits server used two Micron 32GB NVDIMMs — nonvolatile DRAM — to help absorb data writes from the Cassandra application server nodes. The Lightbits LightOS SDS server had a Lightfield PCIe acceleration card installed to provide hardware offload support. All remote NVMe tests used the Lightfield acceleration card for data compression, TCP offload and NVMe/TCP offload. Four logical volumes were striped across the eight SSDs. We assigned one logical volume to each Cassandra server node to host each Cassandra server's database.

All tests used Apache Cassandra™, a popular open-source, cloud-ready web data platform, with the Yahoo!™ Cloud Servicing Benchmark (YCSB) suite executing YCSB Workload "A," a 50% read/50% update I/O profile to provide insights into both read and write comparisons.

Four load-generation servers ran the YCSB benchmark for all tests. All load-generation servers used 50GbE connections to the switch.

YCSB allows a test run to limit the maximum workload placed on the database server through a `-target <ops/sec>` parameter. Throttling enables a comparison of how each storage configuration compares as workload levels increase. Tests ran with target workload levels of 10K, 20K, 40K, 60K and unthrottled operations per second.

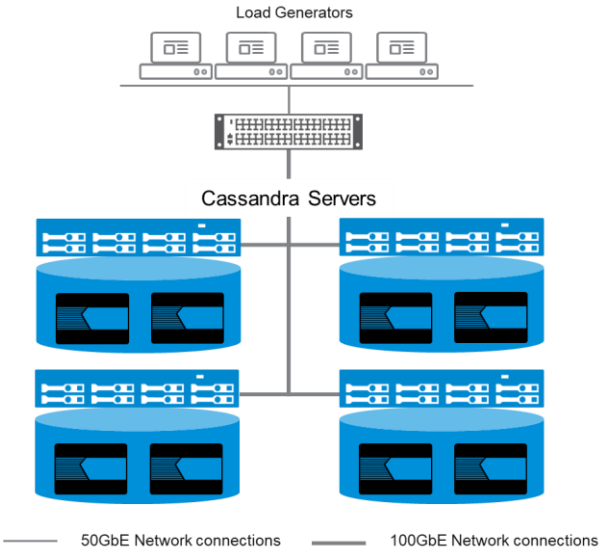


Figure 2a: Cassandra test configuration using local NVMe

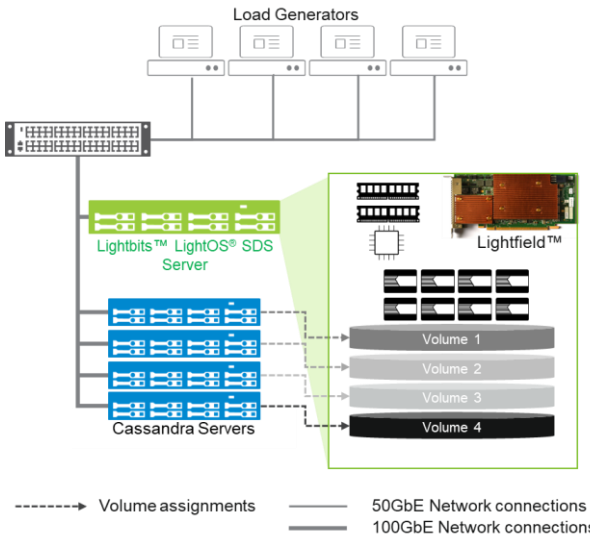


Figure 2b: Cassandra test configuration using remote NVMe

Each server's configuration is listed below:

Lightbits LightOS Storage Server (1x)	
Vendor/Model	Dell® R740XD
Processor (2x)	Intel® Xeon™ Platinum 8168
Memory	768GB Micron DDR4-3200
Network	Mellanox® ConnectX-5 100GbE
Storage	Boot (1x): Micron 240GB 5100 SATA SSD Data (8x): Micron 3.84TB 9300 PRO NVMe SSD
Operating System	CentOS® 7.6.1810 (kernel 4.14.47)
Lightbits LightOS	Version 1.1.25
Cassandra Database Servers (4x)	
Vendor/Model	Supermicro SYS-1029-TN10RT
Processor (2x)	Intel® Xeon™ Platinum 8168
Memory	384GB Micron DDR4-2666
Network	Mellanox® ConnectX-4 100GbE
Storage	Boot (1x): Micron 240GB 5100 SATA SSD Local Data (2x): Micron 3.84TB 9300 PRO NVMe SSD
Operating System	CentOS® 7.6.1810 (kernel 4.14.47)
Cassandra	3.0.9 DataStax Cassandra Community Edition
Benchmark Load-Generation Servers (4x)	
Vendor/Model	Supermicro® SYS-2028U-TNRT+
Processor (2x)	Intel® Xeon™ E5-2960v4
Memory	256GB Micron DDR4-2666
Network	Mellanox® ConnectX-4 50GbE
Storage	Boot (1x): Micron 256GB M510 SATA SSD
Operating System	CentOS® 7.7
YCSB	0.16.0

The Results

When comparing NVMe/TCP storage performance to local in-server storage performance using a common enterprise application workload, the initial assumption is that local storage will naturally be faster than the remote, networked storage solution. As illustrated in the following charts, this is, in fact, not necessarily true. As shown in Figure 3, both storage configurations generate the same performance at lower workload levels. As workload levels increase above 40,000 database operations per second, the NVMe/TCP configuration generates 12% more operations per second than local NVMe. This performance comes from the fact that each database in the remote storage implementation can take advantage of all SSDs, instead of just two local SSDs, and from hardware acceleration of data services through the Lightfield add-in card.

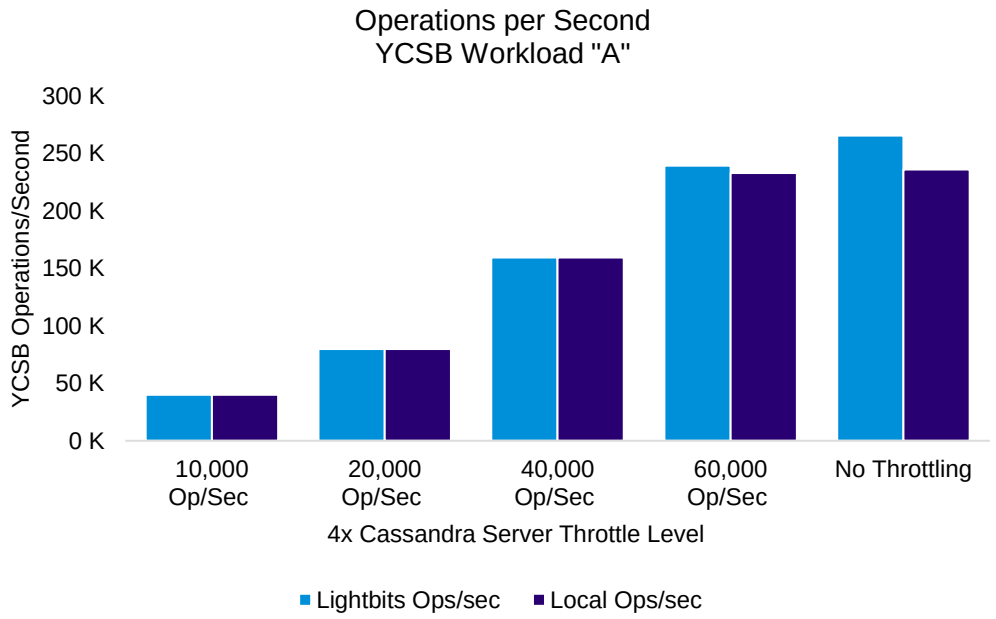
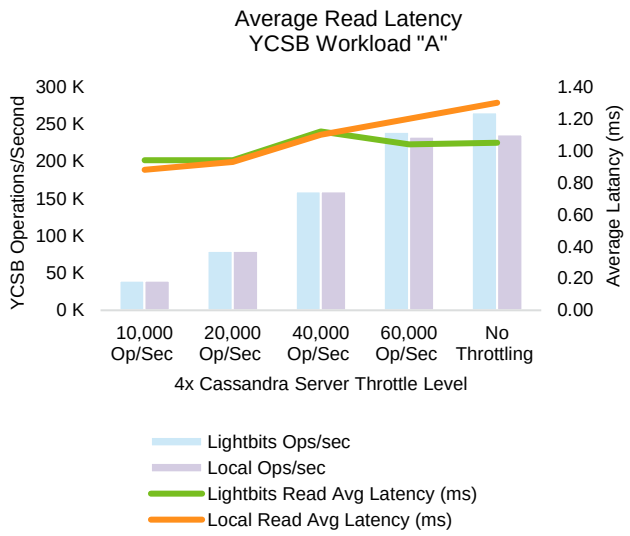


Figure 3: Cassandra Workload "A" performance comparison

Total operations per second performance is important to many workloads, but latency affects modern analytics and cloud application performance levels. Comparing both average and quality of service (tail, or 99.9%) latency levels for both read and update operations, the NVMe/TCP networked storage solution performed each operation faster than the local NVMe at higher transaction loads, as shown in Figure 4 and Figure 5 below. It should be noted that YCSB throttles workload by injecting a "sleep" counter between each transaction that is based on historic average of transaction times prior to the throttle operation, a process that could be a cause for the inconsistent latency values recorded.

Average Read Latency



Read Tail Latency

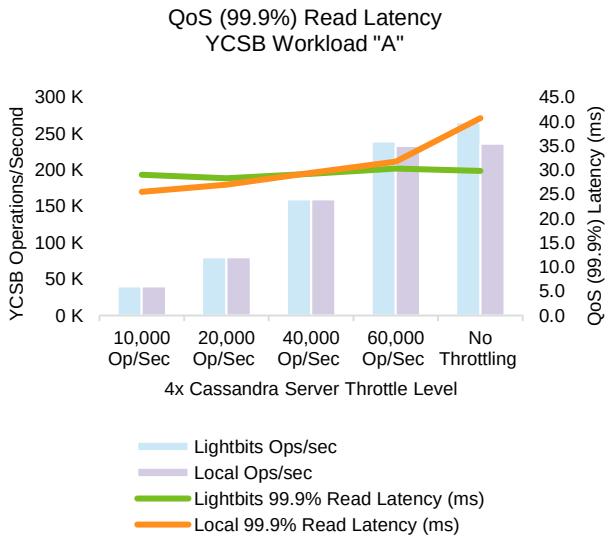
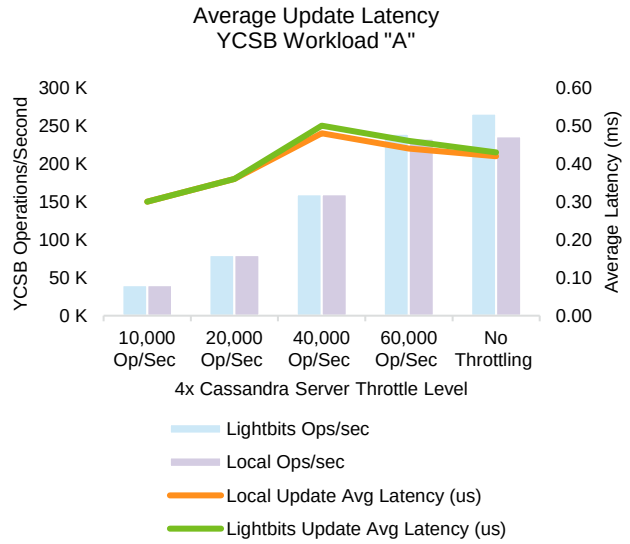


Figure 4 a-b: Cassandra read latency comparison for Workload "A"

Average Update Latency



Update Tail Latency

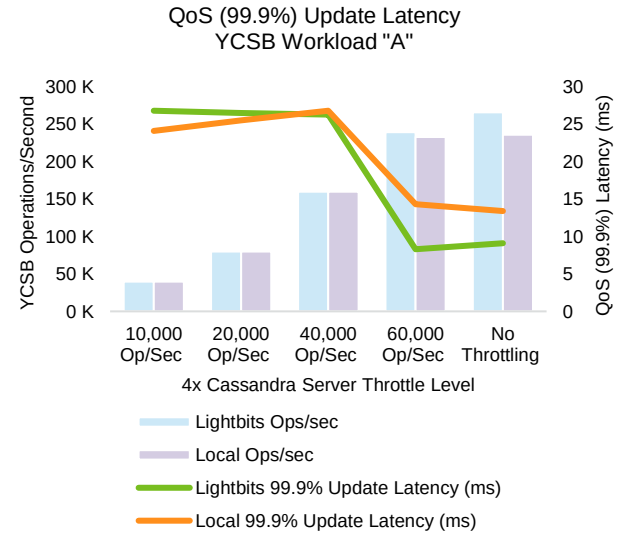


Figure 5 a-b: Cassandra update latency comparison for Workload "A"

The Bottom Line

As customers look for faster, easily manageable, scalable storage solutions, they can choose between purpose-built storage arrays or SDS solutions built on commodity, industry-standard server architectures. Modern SDS solutions can offer services similar to legacy array architectures, but they offer more control in terms of scaling, cost and performance options. Software-based SDS solutions such as Lightbits Labs' LightOS, along with its optional Lightfield PCIe add-in card, are easing the path to efficient, cost-effective, high-performance storage for cloud-enabled applications. By leveraging network infrastructures already deployed within your data center, Lightbits Labs' LightOS can lower the "cost of entry" and reduce the complexity of the storage infrastructure.

Micron believes that NVMe is the future of the data center. Advanced Micron NVMe SSDs, along with advanced SDS storage solutions like Lighbits Labs' LightOS, can be a critical part of your successful migration to NVMe. With the release of the Micron 9300 family of NVMe SSDs, Micron delivers industry-leading random and sequential read and write performance with the lowest average write latency on the market. Together, Micron and Lightbits Labs offer a compelling solution to meet the demands of your performance-critical cloud and enterprise workloads with the speed, performance and capacity your applications demand.

Learn More

To learn more about Micron's 9300 NVMe SSDs and ways they can contribute to your success, visit us at micron.com/9300.

To learn more about Lightbits Labs NVMe/TCP solutions, visit them at www.lightbitslabs.com.

Want to learn more about advanced NVMe over Fabrics solutions?

- Visit https://sniablog.org/category/nvme-over-fabrics/

or

- Visit https://nvmexpress.org/welcome-nvme-tcp-to-the-nvme-of-family-of-transports/

or

- Download the NVMe-oF white paper from nvmexpress.org (information required to download)

How We Tested

Our test methodology approximates real-world deployments and uses for a Cassandra database. The Cassandra configuration consists of four application servers operating independently. Modifications to the default Cassandra configuration include the following:

- Data distribution set to “uniform” from the default value “zipfian”
- Fieldcount set to “4” from the default value “10”
- Fieldlength set to “1024” from the default value “100”
- YCSB operations per second throttled using the -target <xxx> command line parameter where “xxx” was 10000, 20000, 40000, and 60000.
- Compression
    - For local data storage testing, Cassandra compression is configured as “enabled” and uses compaction class “LeveledCompactionStrategy”
    - For networked data storage testing, compression is configured as “disabled” and uses Lightbits compression services

Initial data load generated a dataset of approximately 1.36TB for each Cassandra server, exceeding the 384GB of available DRAM on each application server. The database is backed up to a separate location for quick reload of data between test runs. For each configuration under test, we restore the database from this backup, starting every test from a consistent state. Each test executes multiple times, and the average of each test result provides the documented results.

Tests execute using 50 threads and run for 20 minutes.