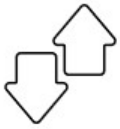
Micron® HSE Enhances Shared Storage Using Lightbits Labs™

Micron® testing has proven that SSDs combined with storage software optimized to take advantage of flash are the cornerstones of data-centric application performance. Whether deployed as server-local storage or in shared storage implementations, our data center SSDs with NVMe™ are key enablers for a successful solution. Using the recently released open-source Micron Heterogeneous-Memory Storage Engine (HSE) for Linux® servers, users benefit from faster reads and writes via a data interface optimized for advanced non-volatile storage technologies. HSE is not constrained exclusively to server-local storage. By using HSE with shared storage solutions, such as Lightbits Labs™ LightOS®, Micron offers not only faster data access, but higher endurance and more cost-effective ways to maximize your NVMe investment.

This white paper highlights the benefits of using Micron HSE with shared NVMe storage deployments for a common NoSQL database solution, MongoDB. MongoDB uses a modular architecture that supports the replacement of its default storage engine, called WiredTiger, with HSE to manage data I/O between the database and storage device resources. HSE, when used in place of WiredTiger, dramatically improves overall database performance.

LightOS + HSE Benefits

In benchmark tests HSE provides up to:



6x Better Ops/second[1]

30x Better QoS Latency[2]

5x Better Power Efficiency[3]

1-YCSB Workload A, HSE vs. WiredTiger
2-YCSB Workload F, HSE vs. WiredTiger.
3-YCSB Workload A, HSE vs. WiredTiger.

Test Configuration

For this comparison, we used standard, x86 servers hosting MongoDB. Each database server used HSE version 1.7.1, as illustrated in Figure 1. Tables 1a – 1d list the specific server configuration for each infrastructure role, along with the operating system and installed relevant software. We list specific optimizations for various software components in the section “How We Test” at the end of this white paper.

For database storage, we used a LightOS cluster consisting of two Dell® EMC PowerEdge R7525 servers. LightOS is a software-defined storage (SDS) solution that enables higher efficiency by creating a centralized, high-performance NVMe-over-TCP™ platform for data storage. While LightOS performs like local NVMe storage, it provides additional features commonly associated with traditional, hardware-based storage area networks, such as logical volumes with data protection, compression and thin-provisioning services.

Protection for all database content uses 2x replication between storage nodes and erasure coding across drives within each node through LightOS Elastic RAID. On each node, data is stored on 12x 15.36TB Micron 9300 NVMe SSDs. Four 8TB logical volumes span across the 12 SSDs in each node with a logical volume assigned to each MongoDB database server.

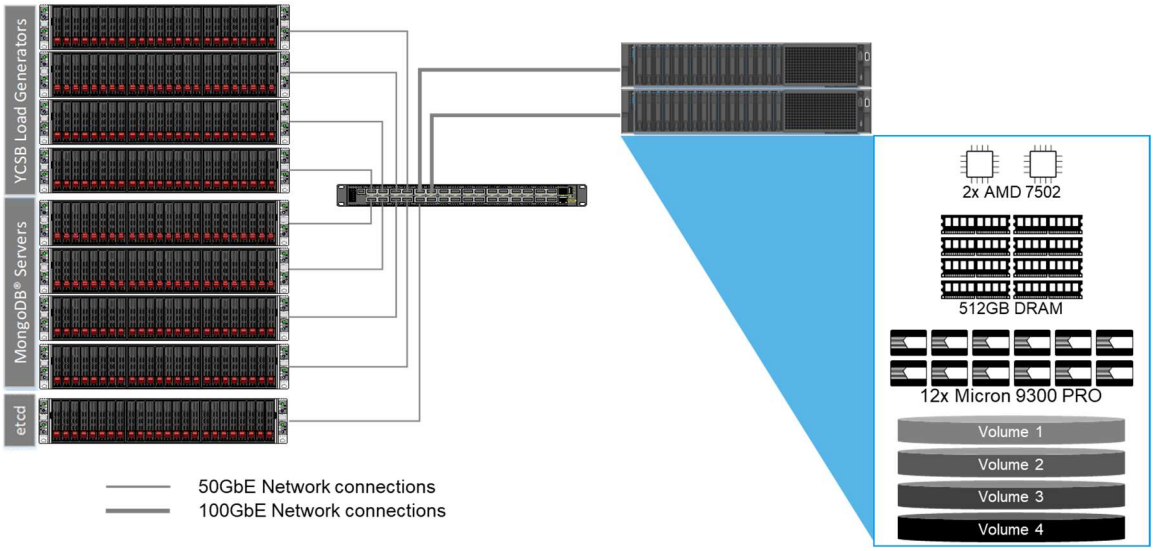


Figure 1: Test Configuration Using Lightbits LightOS Remote NVMe Storage

To generate transactions for this test, we used the Yahoo® Cloud Server Benchmark (YCSB) tool to simulate five workload profiles, defined in YCSB workloads A, B, C, D and F[1] (see "YCSB workload definitions" in this white paper).

LightOS Storage Server (2x)	
Vendor/Model	Dell R7525
Processor (2x)	AMD EPYC™ 7502 (32-core, 3.3GHz)
Memory (DRAM)	512 GB (Micron DDR4-3200)
Network	Mellanox® ConnectX-5 MCX516A 100GbE (x16 PCIe Gen 4)
Storage	Boot (1x): Micron 240 GB 5100 SATA SSD Data (12x): Micron 15.36 TB 9300 PRO NVMe SSD
Operating System	CentOS® 7.8.1810 (custom kernel 4.14.175)
LightOS	Version 2.0

YCSB Load-Generation Servers (4x)	
Vendor/Model	Supermicro® SYS-2028U-TNRT+
Processor (2x)	Intel® Xeon™ E5-2960 v4 (14-core, 2.6 GHz)
Memory	256 GB (Micron DDR4-2666)
Network	Mellanox ConnectX-4 50GbE
Storage	Boot (1x): Micron 256 GB M510DC SATA SSD
Operating System	CentOS 7.6
YCSB	0.17.0

MongoDB® Database Servers (4x)	
Vendor/Model	Supermicro SYS-2028U-TNRT+
Processor (2x)	Intel Xeon E5-2690 v4 (14-core, 2.6 GHz)
Memory	256 GB (Micron DDR4-2666)
Network	Mellanox ConnectX-4 CX413A 50GbE
Storage	Boot (1x): Micron 256 GB M510DC SATA SSD Database storage: 1x 8TB NVMe remote logical volume
Operating System	Red Hat® Enterprise Linux version 7.8 (kernel 5.7.0-1.el7)
MongoDB	WiredTiger Storage Engine v4.2.8 HSE Storage Engine v3.4.17.2 (HSE version 1.7.1)

Quorum (etcd) Server (1x)	
Vendor/Model	Supermicro SYS-2028U-TNRT+
Processor (2x)	Intel Xeon E5-2690 v4 (14-core, 2.6 GHz)
Memory	256 GB (Micron DDR4-2666)
Network	Mellanox ConnectX-4 CX413A 50GbE
Storage	Boot (1x): Micron 256 GB M510DC SATA SSD etcd (1x): Micron 1.92 TB 7300 PRO
Operating System	Red Hat Enterprise Linux version 7.8

Table 1a-d: Server Configuration by Role

[1] YCSB also defines a Workload E. This workload focuses on index scan only and is not impacted by performance of storage technology, therefore this workload is not executed during performance analysis of storage solutions.

Results

Test results focus on four aspects of performance that may impact the application: solution performance, tail latency (consistency), power efficiency and endurance. All data reflects metrics using 2x data replication.

Solution Performance

An application is heavily impacted by how fast data can be moved into and out of the storage. Solution performance focuses on the number of database operations completed per second for each storage engine. Figure 2 compares the performance for the five YCSB benchmark workloads.

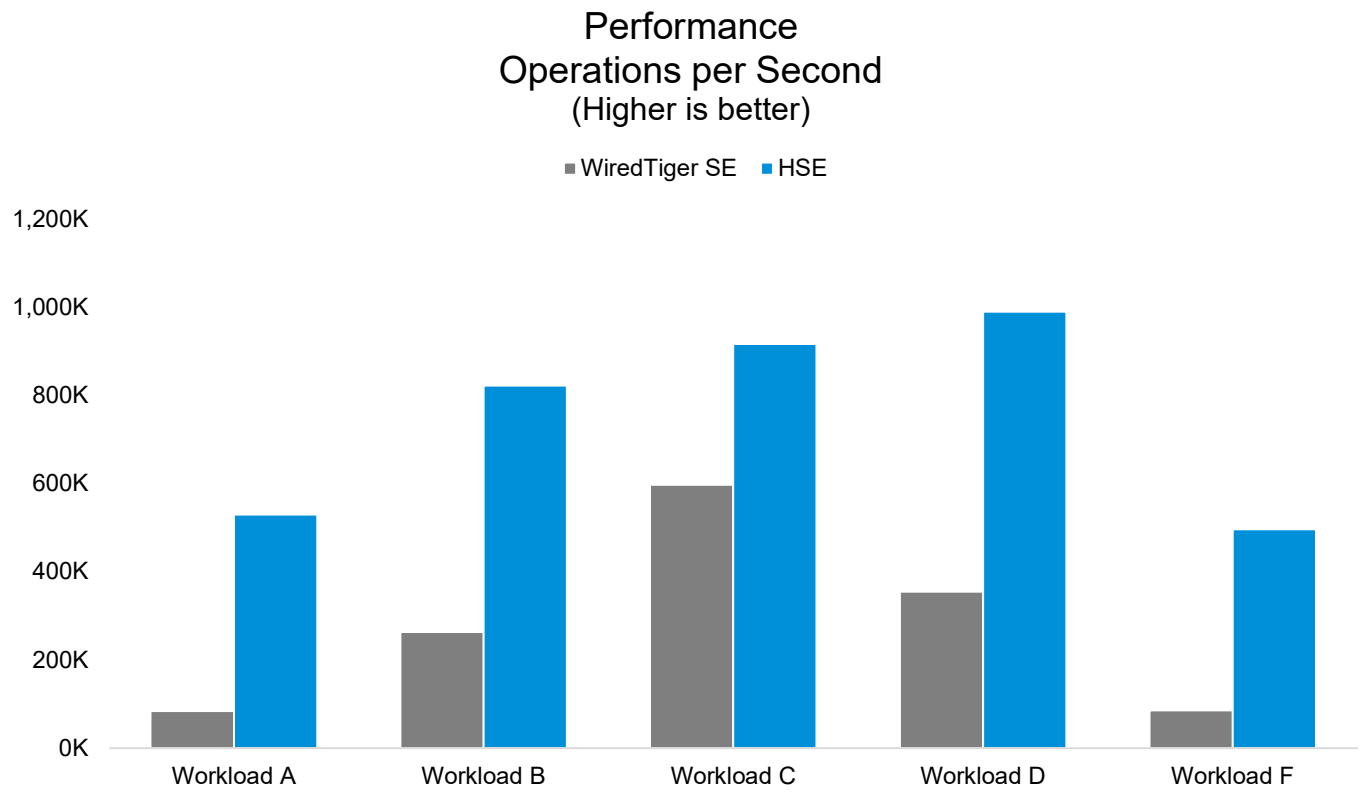


Figure 2: Performance Comparison of HSE vs. WiredTiger Storage Engines

For all YCSB workloads, HSE offered improvements in database operations per second over the default WiredTiger storage engine. Workloads that have a high percentage of data write operations, such as Workload A and Workload F, experience a larger relative performance increase – up to six times better than WiredTiger. While this result was previously documented for MongoDB solutions leveraging SSDs installed directly into the database server (see Open Source Storage Performance With Micron® HSE and MongoDB™), experiencing such significant performance gains using HSE with a networked, shared-storage implementation means that HSE can benefit many open-architecture, data-centric applications regardless of where the NVMe devices are located.

Latency

Total operations per second performance is important to many workloads, but latency affects modern analytics and cloud application performance levels. Timely storage response means more transactions completed in a shorter time.

There are a broad range of latency measures available, but most important in applications that make use of multiple services (such as MongoDB) are tail latencies. In this case, we focused on the 1/1000 (99.9%) measurements that can make services inconsistent and diminish the overall user experience. Ideally, we want all transactions processed in the shortest time possible, but in real-world situations, consistency is often more important. As large applications have many pieces working together, slow response outliers tend to affect overall system responsiveness and hence, not just low latency, but consistent low latency is the goal of application developers and service providers.

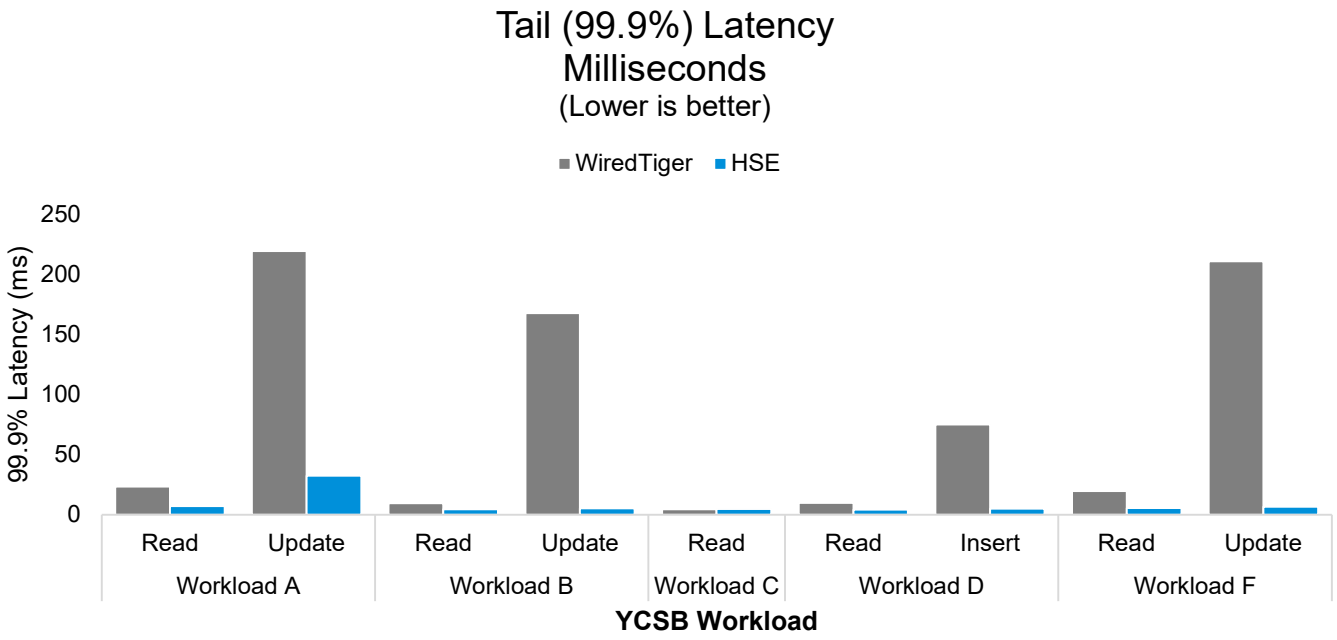


Figure 3: Tail Latency Comparison of HSE vs. WiredTiger Storage Engines

For this white paper, we compared tail latency levels of 99.9% of transactions for each workload's read and write (update or insert) operations. For reads, the net gain provided by HSE can be as high as four times better, while write operations see up to 20 times better 99.9% latency than the WiredTiger storage engine. One outlier is Workload C where the latency delta shows a 7% decrease for HSE. At such low latency levels (4.5 ms), the difference is statistically similar. The advantage of HSE comes in improved latency for workloads performing higher write operations (update and insert operations in the chart) as shown in Figure 3.

Power Efficiency

As data centers expand to support ever increasing demand, power efficiency becomes a factor in solution design. Small increases in the amount of work realized from a fixed power budget can result in dramatic cost savings for a company. HSE can significantly improve the efficiency of your MongoDB solution. Figure 4 shows HSE generating more transactions per watt of energy consumed where we see HSE generating up to five times more transactions per watt consumed (Workload C). This means we get more work when dealing with a fixed power budget.

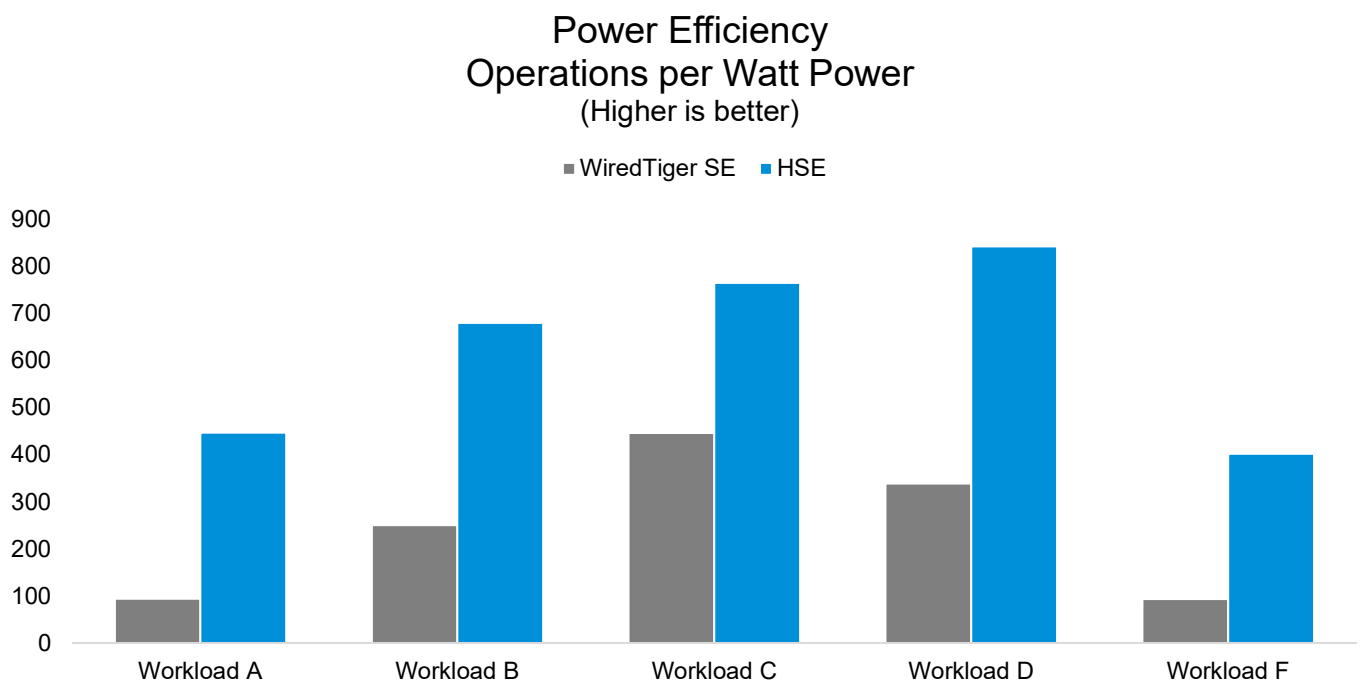


Figure 4: Power Efficiency Comparison of HSE vs. WiredTiger Storage Engines

Durability: I/O Amplification

Many decisions are made focusing on only the measurements just discussed. But a commonly overlooked aspect of selecting solutions such as an all-flash, shared-storage solution, is **I/O amplification**, or **the number of additional data read or written for each database transaction**. I/O amplification affects a shared storage solution in two ways: network utilization and SSD endurance. Each of the YCSB workloads has varying levels of read and write database-level operations. Reducing the number of I/O operations performed at the storage device can result in significant gains in performance and efficiency from existing data center infrastructure.

Network Utilization

Network utilization is impacted by I/O amplification for both read and write operations. For every 1KB database request, I/O amplification can result in the need to transfer multiple kilobytes of data across the network. Sources of the write amplification can be operating system file system page size, network frame size and data replication, among others. Micron optimized HSE to understand these low-level aspects of data management and use I/O coalescing for more efficient data transmission. HSE not only has advantages for writes but can offer data efficiencies for reads of up to 6 times better than the native WiredTiger data engine. Figure 5 highlights the data sent and received by network interfaces for each database I/O operation.

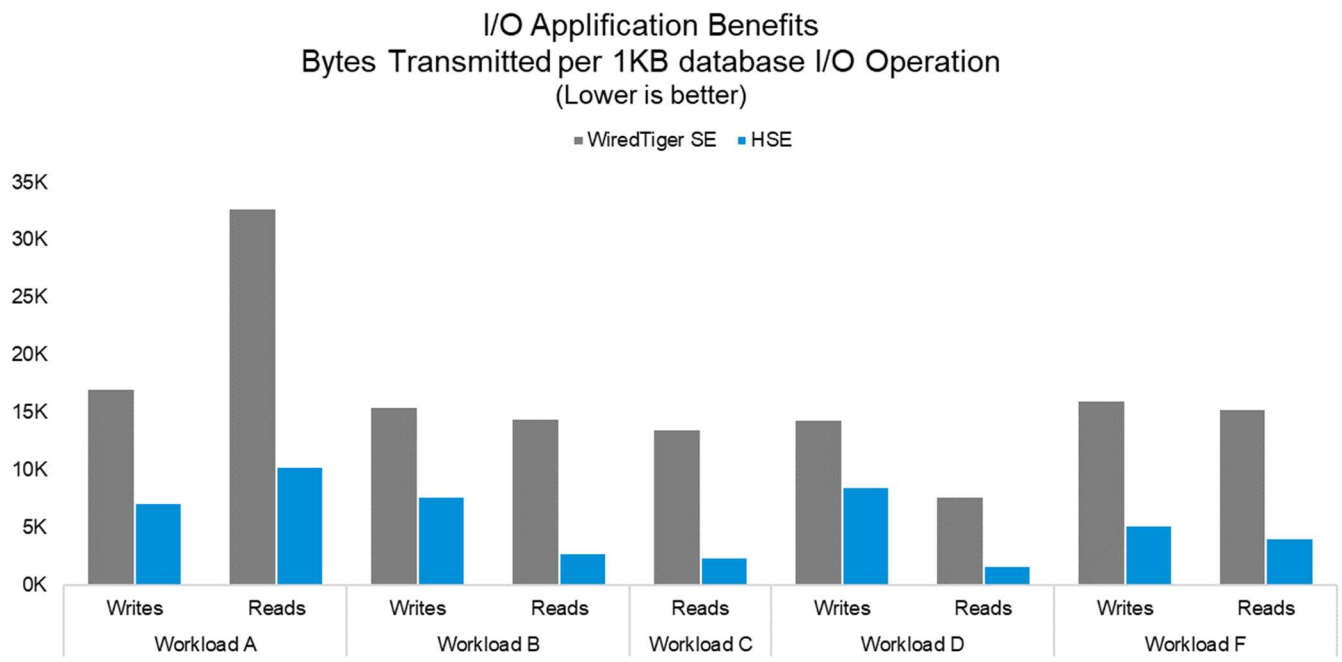


Figure 5: I/O Amplification Comparison of HSE vs. WiredTiger Storage Engines

SSD Endurance

The need to perform multiple SSD write operations for each database write request is known as write amplification. Write amplification can directly affect the SSD life, as well as the actual production life, of an SSD. If there is a way to reduce write amplification, then the SSD can have a longer life even in higher-write-percentage workloads.

One of the drivers for Micron's development of HSE is a reduction in SSD write amplification. As just discussed, the amount of data transmitted over the network for both reads and writes is improved with HSE. SSD endurance is only concerned with the data sent to the array that must be written to the SSDs. As shown in Figure 5, HSE reduces the number of bytes sent to the LightOS cluster (writes), and subsequently written to the SSDs, for each database write operation compared to WiredTiger. Fewer writes to SSDs results in fewer writes over the life of each SSD. This enables solution designers to consider SSDs for two reasons. One, the life of SSDs will likely be prolonged using HSE due to the smaller number of writes to them. Two, SSDs with lower endurance and lower costs, even in environments that have higher data-write requirements, can be used.

The SSD industry typically measures SSD life by the amount of data written to the SSD, called Total Bytes Written or Terabytes Written (TBW). SSD read operations do not affect SSD life calculations.

The Bottom Line

As customers look for faster, easily manageable storage solutions, they are turning to scalable, shared storage solutions built using high-performance flash. These solutions offer the flexibility and performance customers need to meet their applications' unique needs.

Data applications such as MongoDB offer an open-source, scalable database platform for a wide variety of use cases. While MongoDB provides a complete data solution out of the box, it has a modular architecture that supports the use of alternative components, such as the storage engine, which is critical to efficient data access.

Micron HSE is a storage engine alternative that is compatible with the MongoDB application. HSE incorporates Micron's deep understanding of data management with our combined understanding of memory and storage to provide a highly efficient storage engine. The result is MongoDB performance improvements across multiple design criteria, offering gains of up to six times in operations per second while improving quality of service by up to 30 times for writes (Workload B and F) and up to four times for reads (Workload F). HSE is a key factor in extending the life of SSDs by reducing the read and write transactions being executed against the SSDs. This supports the flexibility to consider lower endurance SSDs or extend the effective life of read-centric solutions. Finally, HSE offers dramatic improvements in overall efficiency by supporting up to 5 times the database transactions per watt of consumed power (Workload A).

Micron believes that the NVMe protocol is the future of data center storage. LightOS greatly improves the use rates of NVMe via disaggregation with rich data services. Micron HSE greatly improves the performance of LightOS logical volumes while reducing write amplification, power consumption and network bandwidth. Micron NVMe SSDs, along with advanced software-defined storage solutions like Lightbits Labs' LightOS, can be a critical part of your successful migration to NVMe. By offering a next-generation storage engine that takes advantage of flash and tiered storage architecture, Micron is the right SSD provider for your application needs.

Learn More

To learn more about Micron Heterogenous-Memory Storage Engine, visit us at www.micron.com/hse.

To download the latest HSE source code visit the HSE repository on github.com.

The release of MongoDB supporting Micron HSE can be found here.

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

How We Tested

Test Methodology

The Linux operating system on each server was configured to use XFS with mount parameters `noatime`, `nodiratime` and `discard`.

A 2TB database was created on each MongoDB server using the YCSB Workload C with a load parameter.

After completing the load process, a test pass consisting of a single execution of each YCSB workload (A, B, C, D and F) executed using the parameters listed in Table 2. Three test passes for each YCSB workload executed and the results documented in this white paper are the average results of the three test passes for each YCSB workload.

Table 3 lists the characteristics and I/O profile of each YCSB workload.

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

Table 2: YCSB Benchmark Parameters

Parameters	Value	Description
Number of threads	96	Number of YCSB threads to generate
Fieldcount	10	Standard 1KB record size
Recordcount	2 billion	Number of records in the database
Operationcount	2 billion	Dataset size within database
ExecutionTime	30 minutes	Duration of the test

Table 3: YCSB Workload Definitions

YCSB Workload	Type	**Ratio**
A	Update heavy	50% Read, 50% Write
B	Read mostly	95% Read, 5% Write
C	Read only	100% Read, 0% Write
D	Read latest	95% Read, 5% Insert
F	Read-modify-write	50% Read, 50% Read-Modify-Write

Configuration Parameters

Table 3 defines the configuration parameters used for various aspects of the test environment.

Table 3: Configuration Parameters Used

Server	Object	Configuration Description
LightOS Storage Node	Read Ahead	Set to 8 (Default is 256): blockdev –setra 8 /dev/nvme0n1
MongoDB Application Nodes	Storage Engine	WiredTiger: XFS with mount options: noatime, nodiratime, discard HSE: mpool mode=0600 capsz=32
	Transparent Hugepages	Disabled
	/etc/sysctl.conf	vm.swapiness = 1 vm.max_map_count = 131072 vm.zone_reclaim_mode = 0 /etc/security/limits.d/99-mongod.conf: mongod - fsize unlimited mongod - cpu unlimited mongod - as unlimited mongod - memlock unlimited mongod - nofile 64000 mongod - rss unlimited mongod - nproc 64000

micron.com

Rev. A 9/2020 CCM004-676576390-11494