NVMe™ SSDs Future-Proof Apache Cassandra®

Get More Insight from Datasets Too Large to Fit into Memory

Overview

Some of the most compelling Cassandra cluster deployments have been built with SSDs — typically SATA SSDs due to their high capacity and better-than-hard drive performance,[1] which are both paramount when the entire data set does not fit into memory (DRAM).

The success of SATA SSD-based Cassandra deployments has helped generate increased demand for Cassandra databases and greater demand for still higher performance.

This technical brief highlights the next logical step for performance Cassandra applications: SSDs with NVMe, like the Micron® 9300 PRO.

To demonstrate the 9300 PRO's effectiveness, we compared two 4-node, same-capacity, same-SSD-count Cassandra clusters — one built using Micron's 9300 PRO NVMe SSD and a second using enterprise SATA SSDs[2] — and found startling results in both higher performance and lower latency.

Due to the broad range of Cassandra deployments, we tested multiple workloads and multiple worker counts. You may find some results more relevant than others for your deployment. The terms "performance" and "database operations per second (or OPS)" are used interchangeably.

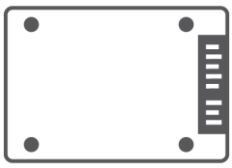
Figure 1: Micron 9300 PRO NVMe SSD

Fast Facts

4-Node Cluster Results (2X SSDs/node)

Micron 9300 NVMe

Enterprise SATA



Workload	9300 Advantage[3]
Session Recording	3.4x
Meta data Tagging	3.6x
User / Credentials Authentication	3.5x
Read Latest Updates	5.8x
Recording User Activity	3.1x

1. We use the terms database operations per second (OPS) and performance interchangeably in this paper.
2. Capacity, GB/s, and IOPS vary by SSD. This paper focuses on the Micron 9300 3.84TB U.2 NVMe SSDs and enterprise-class 3.84TB SATA SSDs. Other SSD models, capacities, and/or interfaces may give different results.
3. Calculated values: (9300 maximum performance/SATA maximum performance) by workload, all tested worker counts.

NVMe SSDs Meet Growing Demands

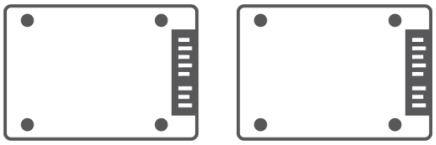
High-capacity, lightning-quick NVMe SSDs are changing the NoSQL cluster design rules. We can expect far, far greater results with no additional footprint as we show in this example. High-capacity NVMe SSDs enable new all-SSD design opportunities and performance thresholds.

We used the Yahoo! Cloud Serving Benchmark (YCSB) workloads A–D and F[4] to compare two 4-node Cassandra test clusters: One built with Micron NVMe SSDs and the other built with SATA SSDs.

Note: Due to the broad range of Cassandra deployments, we tested multiple thread counts from 64 to 1024. We used two SSDs per node — a pair of 9300 PRO 3.84TB SSDs in one and a pair of enterprise SATA SSDs, also 3.84TB each, in the other. Each node in our SSD test cluster stores about 7.68TB and uses four nodes per cluster.

Figure 2: Test Clusters

NVMe Test Cluster: 2x 3.84TB Micron 9300 PRO SSDs per node (4 cluster nodes)



SATA Test Cluster: 2x 3.84TB Enterprise SATA SSDs per node (4 cluster nodes)

The Micron 9300 is the Flash Leader

The test results are organized by YCSB workload. The figures show database operations per second (higher is better) and average read latency (lower is better). The number of workers tested increases from 64 to the maximum number that each 4-node configuration could reasonably support.

These figures show that the SATA-based cluster could not effectively support as many workers. As we added workers, the SATA cluster's average read latency line continued to rise, reaching extremely high values. We stopped testing the SATA cluster beyond 512 workers due to the SATA cluster's unreasonably high read latency for all workloads.

A brief workload description and use case accompanies each workload's test results. These descriptions are based on the YCSB descriptions from GitHub. Where appropriate we've used the definitions from GitHub as written. In other cases, we've added clarification. For additional details, use cases, and examples, please see the GitHub link.

4. We did not test YCSB workload E because it is not universally supported.

Workload A

This is an update-heavy workload with about 50% of all storage IO being written. Examples of this workload can be seen when users sessions are recorded.

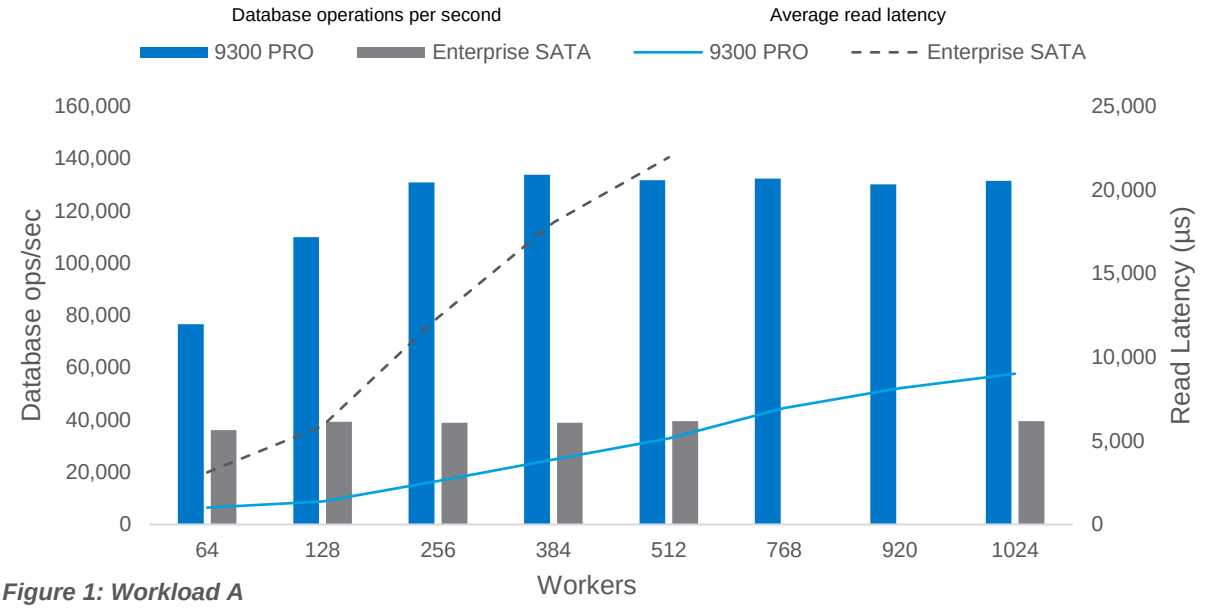


Figure 1: Workload A

The Micron 9300 cluster performance is clearly much greater than the enterprise SATA cluster, showing peak performance of 133,725 OPS at 384 workers. The 9300 cluster's average read latency starts low and increases slowly as additional workers are added with no marked read latency increases (spikes).

This suggests that the 9300-based cluster can be loaded above 384 workers without experiencing a sharp read latency increase or peak. This may be useful when the cluster must manage unforecasted (peak) demand.

The enterprise SATA cluster performance is consistent from 64 to 512 workers, remaining essentially flat with increased worker count and measuring much lower than the 9300 cluster. Its read latency showed a pronounced, steep increase at 128 workers and beyond. The worker count beyond 512 was not tested due to very high latencies.

Workload A	Cluster Type		Micron 9300 Advantage
#workers	9300 NVMe	Enterprise SATA	
64	76,642	36,078	2.1X
128	109,910	39,319	2.8X
256	130,866	38,946	3.4X
384	133,725	38,959	3.4X
512	131,702	39,488	3.3X
768	132,312	(not tested)	---
920	130,121	(not tested)	---
1024	131,505	(not tested)	---

Table 1: OPS/sec, Workload A

Workload B

Workload B is read-heavy with about 95% of all storage IO being read and only 5% being written. Examples of this workload include adding metadata to existing data (tagging). Most of the tags are read while few are written (or rewritten).

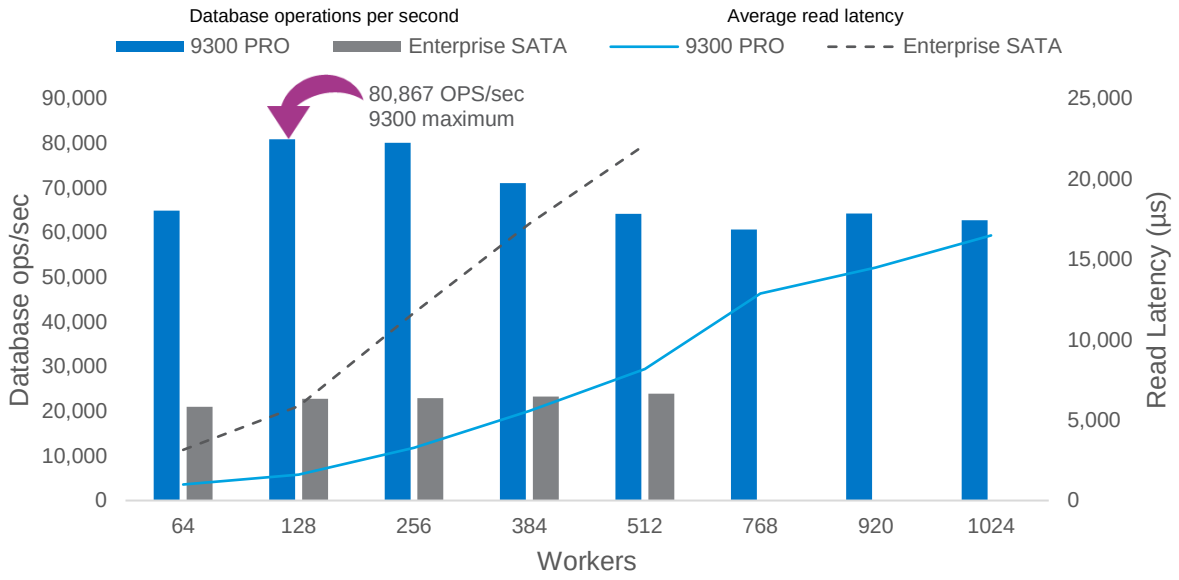


Figure 2: Workload B

The Micron 9300 cluster shows a much higher global maximum performance of 80,867 OPS at 128 workers. Its low average read latency increases smoothly to 512 workers, after which it shows an increased latency at 768 workers.

The enterprise SATA cluster performance ranges from a low of 20,998 OPS to a maximum of 23,943 OPS (at 512 workers), showing little performance variance from 64 to 512 workers. Its average read latency shows a pronounced, steep increase beyond 128 workers.

Workload B	Cluster Type		Micron 9300 Advantage
#workers	9300 NVMe	Enterprise SATA	
64	64,920	20,998	3.1X
128	80,867	22,770	3.6X
256	80,059	22,920	3.5X
384	71,051	23,246	3.1X
512	64,185	23,943	2.7X
768	60,689	(not tested)	---
920	64,276	(not tested)	---
1024	62,781	(not tested)	---

Table 2: OPS, Workload B

Workload C

Workload C is a 100% read workload (data does not change). Examples include reading immutable data for user authentication or reading a profile cache (when a user or system profile was created elsewhere).

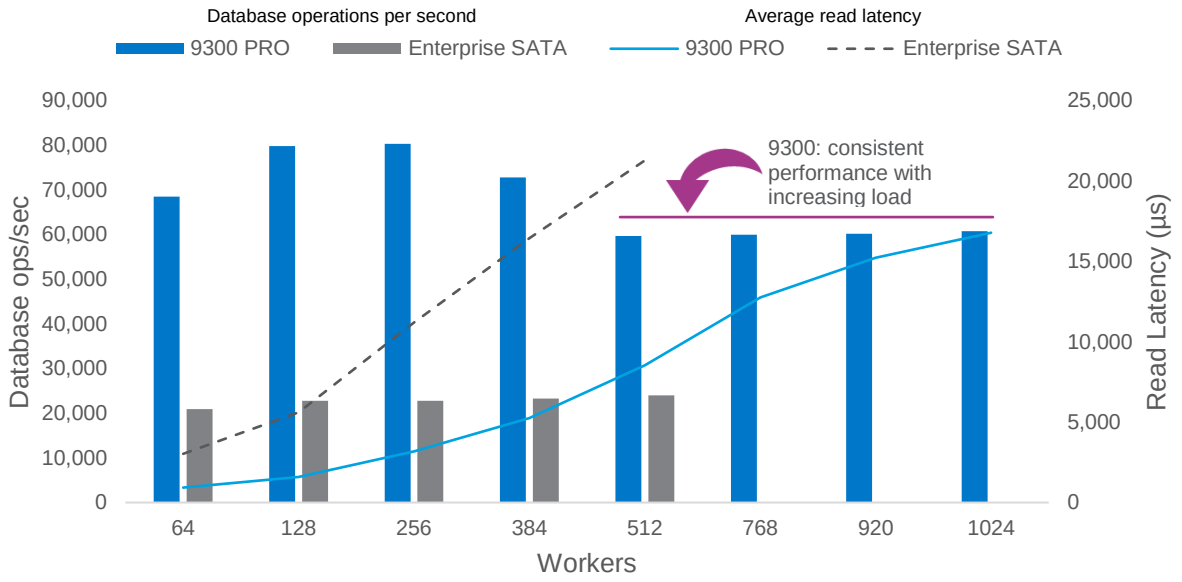


Figure 3: Workload C

The Micron 9300 cluster shows much higher performance with its global maximum of 80,265 OPS at 256 workers (128 worker performance is slightly lower). Its read latency again increases smoothly to 512 workers, after which it shows an increased latency at 768 workers. After 768 workers, its average read latency continues to increase, but at a lower rate. Performance from 512 to 1024 workers is consistent (as latency increases).

The SATA cluster performance is again consistently low, showing a very steep increase in average read latency beyond 128 workers. ("—" in the table below indicates that a comparison can't be made due SATA limitations.)

Workload C	Cluster Type		Micron 9300 Advantage
#workers	9300 NVMe	Enterprise SATA	
64	68,502	20,915	3.3X
128	79,806	22,749	3.5X
256	80,285	22,759	3.5X
384	72,776	23,288	3.1X
512	59,681	23,983	2.5X
768	59,963	(not tested)	---
920	60,147	(not tested)	---
1024	60,721	(not tested)	---

Table 3: OPS/sec, Workload C

Workload D

This workload inserts new records then queries the entire record set – giving preference to records most recently inserted. It models workloads where the latest updates are most popular, i.e., social media status updates.

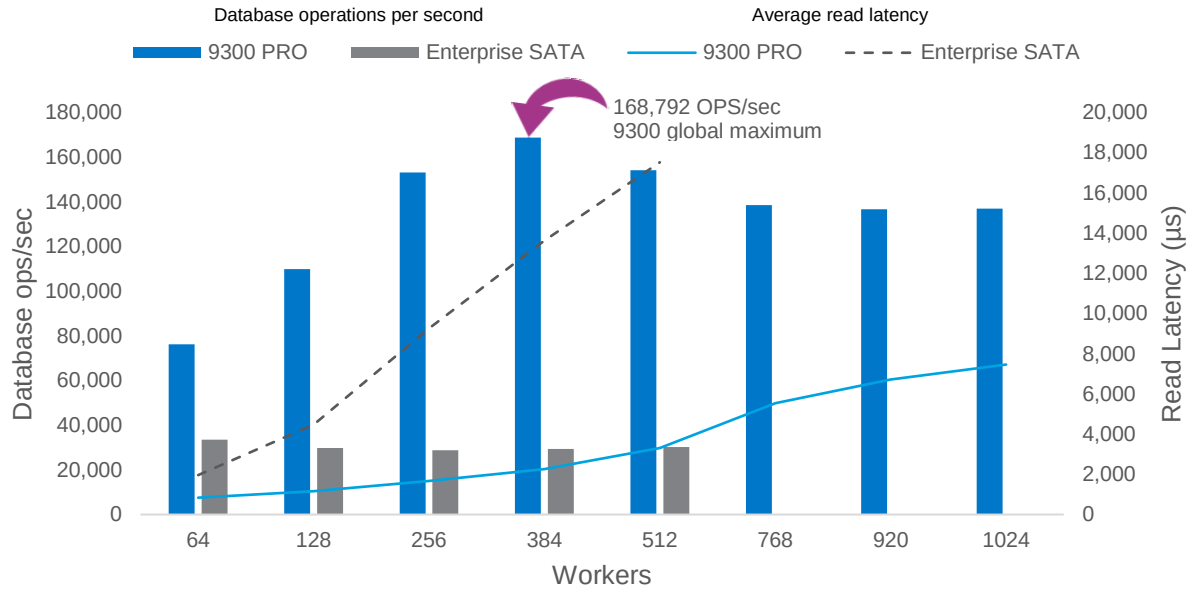


Figure 4: Workload D

The Micron 9300 cluster shows a performance maximum at 384 workers with 168,792 OPS. Performance increases steadily from 64 workers to 384, decreases slightly at 512 workers, then is steady through 1024 workers. Its average read latency gradually increases to 512 workers, increasing more rapidly with additional loading.

The enterprise SATA cluster performance is again consistently low, showing its global maximum at the smallest tested loading (64 workers). Its average read latency increases from 64 to 128 workers, increasing sharply thereafter.

Workload D	Cluster Type		Micron 9300 Advantage
#workers	9300 NVMe	Enterprise SATA	
64	76,285	33,553	2.3X
128	109,899	29,843	3.7X
256	153,185	28,845	5.3X
384	168,792	29,311	5.8X
512	154,165	30,186	5.1X
768	138,503	(not tested)	---
920	136,681	(not tested)	---
1024	136,998	(not tested)	---

Table 4: OPS/sec, Workload D

Workload F

Workload F reads a record, modifies it, then writes it back (read-modify-write). This workload models common database and user activity.

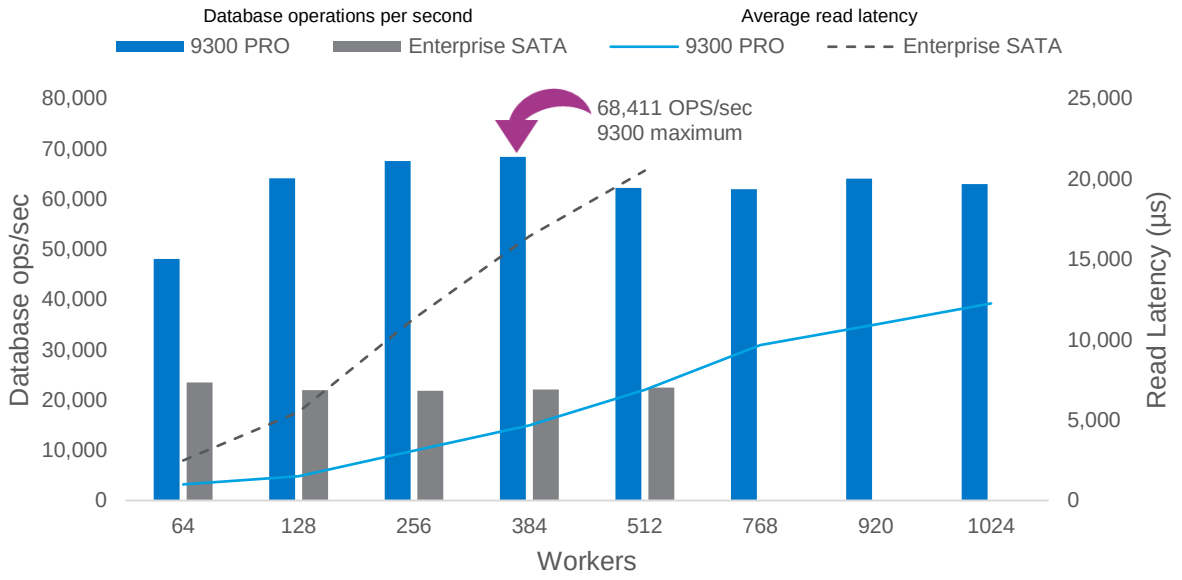


Figure 5: Workload F

The Micron 9300 cluster again shows a performance maximum at 384 workers with 68,411 OPS/sec. Its performance increases steadily from 64 workers to 384, then flattens from 512 through 1024 workers.

The enterprise SATA cluster performance is again consistently low, showing its global maximum at the smallest tested loading (64 workers). Its average read latency increases from 64 to 128 workers and rises sharply thereafter.

Workload F	Cluster Type		Micron 9300 Advantage
#workers	9300 NVMe	Enterprise SATA	
64	48,061	23,494	2.0X
128	64,152	21,987	2.9X
256	67,586	21,865	3.1X
384	68,412	22,119	3.1X
512	62,185	22,459	2.8X
768	61,949	(not tested)	---
920	64,044	(not tested)	---
1024	62,980	(not tested)	---

Table 5: OPS/sec, Workload F

Micron 9300 PRO Clusters Provide More Consistent Read Response

Since many Cassandra deployments rely heavily on fast, consistent read responses, we compared the 99th percentile read response times for each test cluster, workload and worker count. Figure 6 shows the results for each configuration. (In Figure 6, a lower value indicates more consistent latency.)

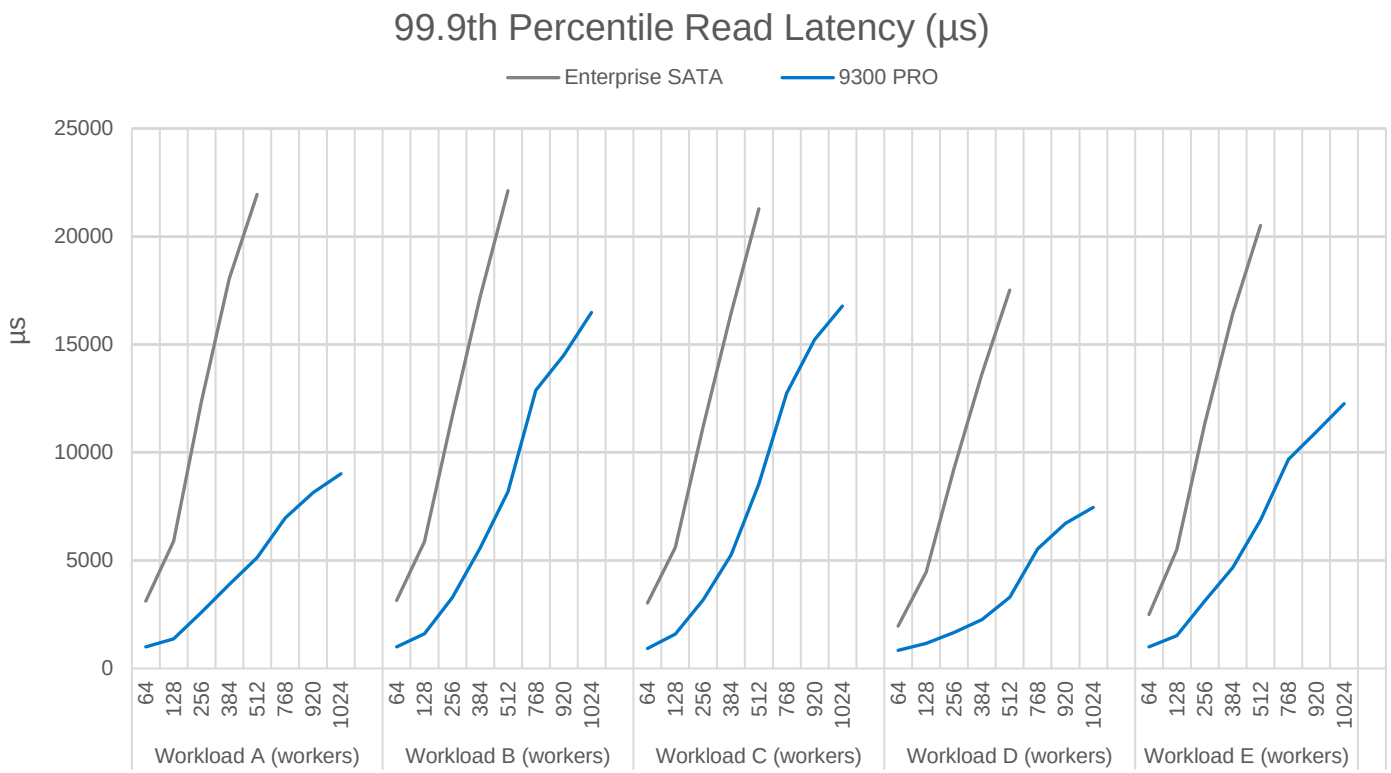


Figure 6: Relative Read Response Consistency

For each workload and worker count in Figure 6, the Micron 9300 cluster shows much lower 99.9th percentile read latency, indicating a more consistent read response time. The enterprise SATA configuration read latency was not measured for worker counts greater than 512, hence its 99.9th percentile latency does not appear in Figure 6 for tests using more than 512 workers.

The Bottom Line

High-capacity, high-performance NVMe SSDs like the Micron 9300 PRO can produce amazing results with Cassandra. Whether you are scaling your local or cloud-based Cassandra deployment for higher performance or faster, more consistent read responses, the 9300 PRO is a great option.

We tested two clusters for database performance and read responsiveness across multiple workloads and numbers of workers. We built an NVMe cluster using a pair of Micron 9300 PRO 3.84TB NVMe SSDs in each node and a second cluster using a pair of enterprise SATA 3.84TB SSDs in each node. Both clusters used four nodes of otherwise identical hardware.

The results were more than interesting — they were amazing!

The Micron 9300 test cluster showed a tremendous increase in performance over all the workloads and numbers of workers tested, with much lower and more consistent read responses.

We expect good performance when the data set fits into memory. When it is too large to fit into memory, SATA SSDs have shown good results. For the next level of results, Micron 9300 PRO NVMe SSDs are the clear choice for faster and more consistent responses.

We are at a crossroads. Our demands drive us toward higher performance, and data growth drives us toward forward-thinking designs. When we combine these, the answer is clear: The Micron 9300 PRO SSDs deliver Cassandra performance and capacity that's lightning-fast and easily extensible.

Learn more about the Micron 9300 PRO and the complete line of Micron NVMe SSDs at www.micron.com.

How We Tested

All SSDs were restored to fresh out of box (FoB) state and preconditioned prior to measurement. Table 6 shows the tested configurations, types of storage devices used, the number and capacity of each as well as the number of nodes in each Cassandra test cluster. Table 7 shows the hardware and software configuration parameters used.

Configuration	Drive Type	Drives per Node	Capacity per Node	Nodes per Cluster
Micron 9300 NVMe	3.84TB PRO	2	7.68TB	4
Enterprise SATA	3.84TB mixed-use	2	7.68TB	4

Table 6: Tested Configuration Capacities

Component	Test Cluster	Configuration/Description
Controller	NVMe SSD	Not used for database storage
	Legacy	Broadcom SAS 3008 (HBA)
Database Storage	NVMe SSD	2X Micron 9300 PRO 3.84TB U.2 SSD with NVMe
	SATA SSD	2X enterprise SATA 3.84TB mixed-use SSD
OS Settings for Cassandra	Both clusters	/etc/security/limits.d/cassandra.conf: cassandra - memlock unlimited cassandra - nofile 100000 cassandra - nproc 32768 cassandra - as unlimited /etc/sysctl.conf: vm.max_map_count = 131072

Table 7: Configuration Parameters

Our test methodology approximates real-world deployments and uses for a Cassandra database. Although the test configuration is relatively small (four nodes in each cluster), Cassandra's scaling technology means these results are also relevant to larger deployments.

- Four nodes host the database.
- The replication factor for the database was set to 3 (there are three copies of the data and the cluster can sustain the loss of two data nodes and continue to function).

The database is initially created by utilizing YCSB workload A's load parameter, which generated a dataset of approximately 1TB, far exceeding available DRAM (ensuring we measure storage system IO). The database is then backed up to a separate location on the server for quick reload of data between test runs. For each configuration under test, the database was restored from this backup, starting every test from a consistent state.

Table 8 shows the percentage of data owned by each of the four nodes.

Node	Capacity	Tokens	Percent Owned
Node01	763.56 GB	256	75.2%
Node02	798.76 GB	256	78.6%
Node03	720.66 GB	256	71.0%
Node04	763.84 GB	256	75.2%

Table 8: Data Distribution Across Nodes

Table 9 shows the testing parameters used in the tested workloads.

Parameter	Value	Description
Threads	64, 128, 256, 384, 512, 768, 920, 1024	Database load
Field Count	10	Standard 1KB record size
Recordcount	1 billion	Number of database records
Operationcount	1 billion	Dataset size within database
ExecutionTime	60 minutes	Test duration

Table 9: Test Parameters

Dim_stat was used to capture statistics on the server running Apache Cassandra. It captures IOStat, VMStat, mpstat, network load, processor load, and several other statistics. Dim_stat was configured to capture statistics on a 10-second interval.

Table 5 shows the IO profiles for tested YCSB workloads (additional details are available at YCSB Core Workloads).

Name	Type	IO Profile
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 (R/M/W)	50% read, 50% R/M/W

Table 10: Workloads