Open-E JovianDSS

Intel® Optane™ SSD DC P5800X 400 GB / 1.6 TB

Certification report

Release date: 2022.04.12

Table of Contents

1. Introduction

The following certification report aims to present the results of various compatibility and performance tests run on the Intel Optane DC P5800X NVMe Solid State Drive when used in combination with Open-E JovianDSS. A description of the testing methods used during those procedures will also be provided. The results, obtained from the aforementioned testing, along with technical specifications were used as the basis for recommendations on specific roles in which to use the device under test (DUT) in Open-E JovianDSS-based systems.
The following applications were considered during the certification process:

- writelog device (ZFS SLOG device),
- read cache device (ZFS L2ARC device).

Usage as a boot medium device was not considered (due to the lack of cost efficiency and large unused storage space left over post installation) nor was usage as a data storage medium (as it would not be cost efficient).
For all the cases mentioned above, functional testing was performed for both the Single node and High Availability non-shared storage cluster configurations. In addition, performance tests were also conducted for the Single node configuration variant. Detailed descriptions for each of the aforementioned are included in the appropriate report chapters.

2. Device Under Test description

The following table includes the Intel Optane DC P5800X NVMe drive hardware specifications.

Table 1. Intel Optane P5800X hardware specifications.

Product name	Intel Optane SSD DC P5800X
Model name	SSDPF21Q400GB SSDPF21Q016TB
Storage capacity	400 GB 1.6 TB
Form factor	U.2 2.5”
Interface	PCIe 4.0 x4, NVMe
Technology	3D XPoint
Enhanced Power Loss Data Protection	Yes
Security	AES 256 bit encryption

3. Test environment description

Hardware specifications for environments used during certification testing are included in the following tables. The configuration described in Table 2a was used for every Single node test. Table 2b shows the configuration used when testing High Availability non-shared storage cluster nodes. More information can be found on HA cluster compatibility testing in chapter 6.

Table 2a. Hardware specifications for single node tests.

System name	Supermicro SYS-620U-TNR
Motherboard	Supermicro X12DPU-6
CPU	2x Intel Xeon Gold 6330
RAM	4x SK Hynix HMA82GR7DJR8N-XN 3200 MHz DDR 4 ECC 16 GB
Storage controller	Broadcom / LSI 9300 SAS HBA (used for HDD)
Storage devices	• 1x Intel Optane DC P5800X 1.6 TB • 1x Intel Optane DC P4800X 1.5 TB • 4x Intel D5-P4320 • 20x Toshiba AL13SEB900 SAS HDD 10K 900 GB
System	Open-E JovianDSS up29r1 b44475

Table 2b. Hardware specifications for HA non-shared storage cluster tests.

System name	Tarox ParX R2242i G6 Server
Motherboard	Intel S2600WFT
CPU	2x Intel Xeon Gold 5222
RAM	192 GB - 16x Micron MTA18ASF2G72PDZ-3G2E1 3200 MHz DDR416GB DDR4 ECC 16 GB
Storage devices	• 1x Intel Optane DC P5800X (400 GB / 1.6 TB) • 4x Intel D5-P4320 SSD 2 TB
System	Open-E JovianDSS up29r1 b44475

Tool used for performance benchmarking: Fio for Linux, v3.28.

4. Raw device tests

In order to be able to properly interpret the results acquired from the tests of the Intel Optane P5800X drive while it was used as a data storage device for Open-E JovianDSS, it's necessary to measure the raw performance of the device. The 1.6 TB model was chosen for this test. This benchmark then allows for a comparison to be made between the actual performance obtained in the test environment with the information provided by the vendor and guarantees that there are no bottlenecks on a hardware level allowing for effective storage performance tests to be done at the ZFS level.

4.1. Test description

Included test cases are described in Table 3. In all instances every combination of thread numbers (1, 4, 8, 16) and queue depths (1, 16, 64, 128) was applied to the fio test tool. For additional information, results of the same tests performed on the Intel Optane P4800X 1.5 TB NVMe disk are included.

Table 3. Test cases description for raw disk tests

Test case	IO pattern	Read to write %	Block size

Random read	random	100/0	4 kB
Random write	random	0/100	4 kB
Sequential read	sequential	100/0	1 MB
Sequential write	sequential	0/100	1 MB

4.2. Performance results

Figures 1 to 4 present the number of IOPS acquired in every test case.

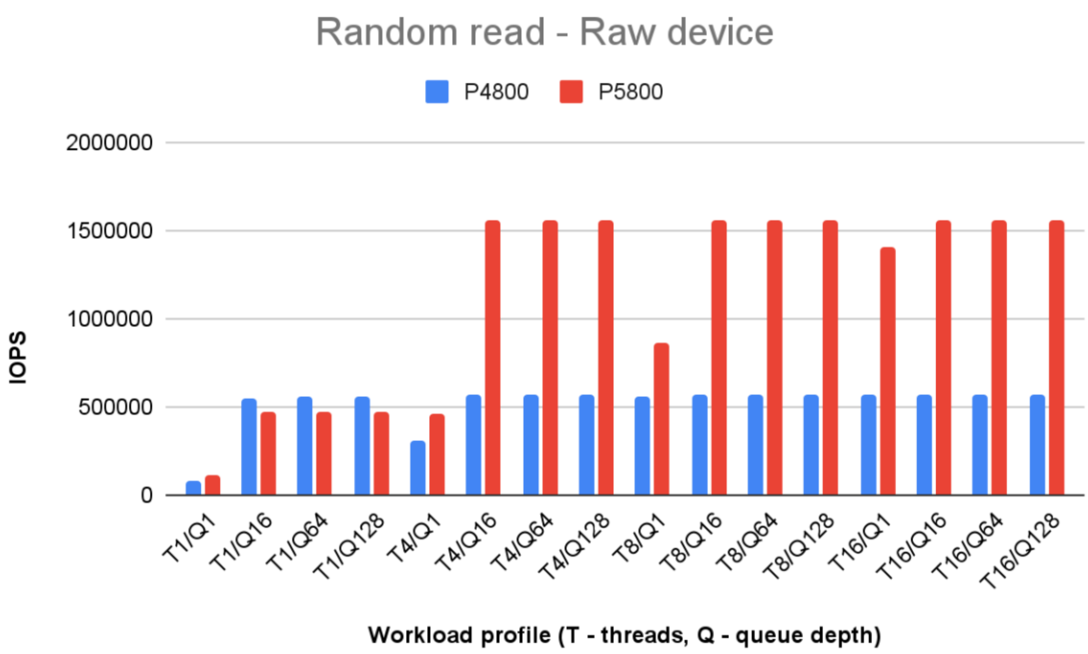

Fig. 1. Random read performance on raw device compared to P4800.

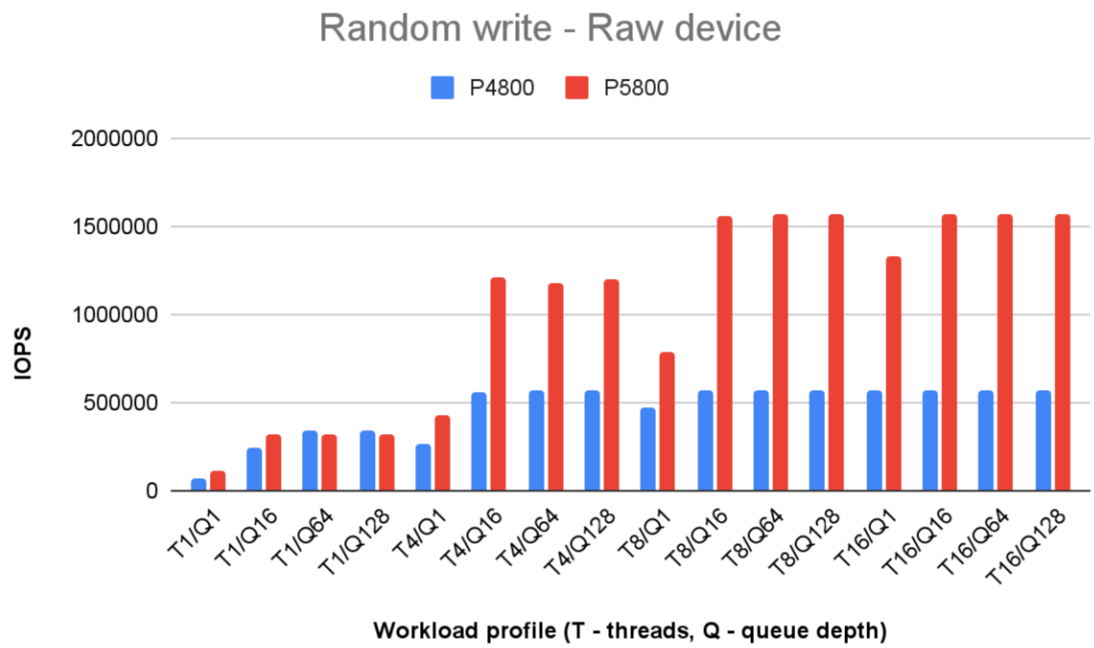

Fig. 2. Random write performance on raw device compared to P4800.

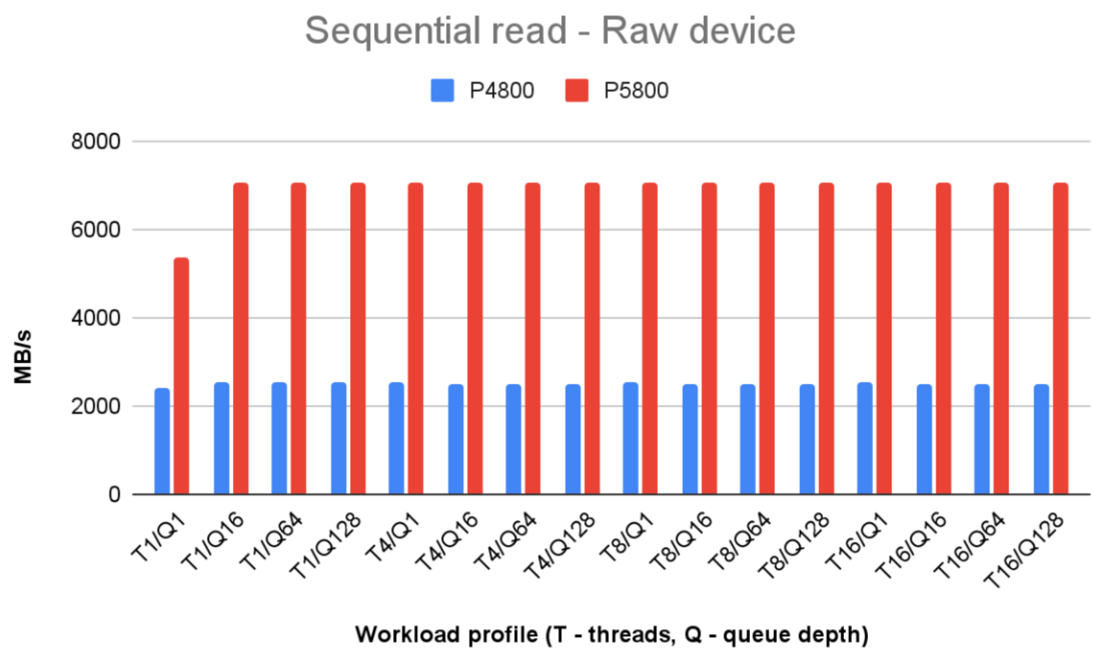

Fig. 3. Sequential read performance on raw device compared to P4800.

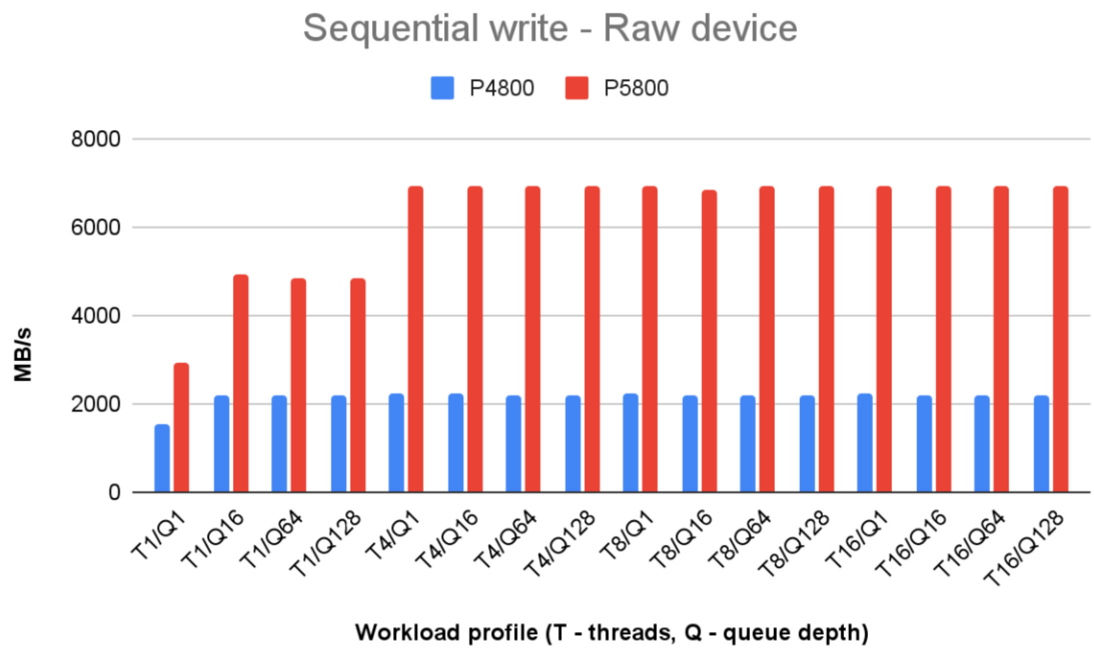

Fig. 4. Sequential write performance on raw device compared to P4800.

4.3. Test conclusions

All performance tests carried out on a raw device showed proper operation at the hardware level in the test environment. The results obtained are at a level similar to those declared by the manufacturer. Compared to the P4800 NVMe drive, the DUT outperformed it in sequential read and write tests. When using a random IO load, P5800X showed better results for cases in which 4 or more threads were used. Especially high results were observed for queue depths equal to or larger than 16.

5. Writelog device tests

In order to evaluate how the Intel Optane P5800X operated as a writelog for Open-E JovianDSS (ZFS SLOG device), several zpool configurations were assembled using HDD or SSD drives. Measurements included the number of IOPS as well as the latency distribution under a fio generated load for zpools with and without a writelog device, which was then added for each zpool in a single drive configuration. All tests were performed on zvols with the 'sync' parameter set to 'always', as well as enabled compression to simulate a recommended production environment configuration. A model with a capacity of 400 GB was chosen for this.
The following characteristics qualify the tested device to be considered as a writelog:

- form factor: U.2 NVMe interface - standard backplane connection for NVMe storage,
- suitable write endurance,
- exceptional declared sequential data transfer rates.

5.1. Zpool with 10x 2-way mirror HDD data groups

A zpool consisting of 20 Toshiba AL13SEB900 SAS HDD 10K disks, with a capacity of 900 GB, was assembled in a 2-way mirror configuration for each data group created. The test parameters are shown in Table 4. Figure 5 presents the IOPS results of a random write test case, with latency distribution for T1/Q1 and T16/Q1 combinations presented in figures 6 and 7. Figure 8 shows results for sequential write tests with latency presented in Fig. 9 and 10.

Table 4. Writelog device test parameters.

Test case	IO pattern	Read to write %	Zvol block size	Data block size
Random write	random	0/100	4 kB	4 kB
Sequential write	sequential	0/100	64 kB	1 MB

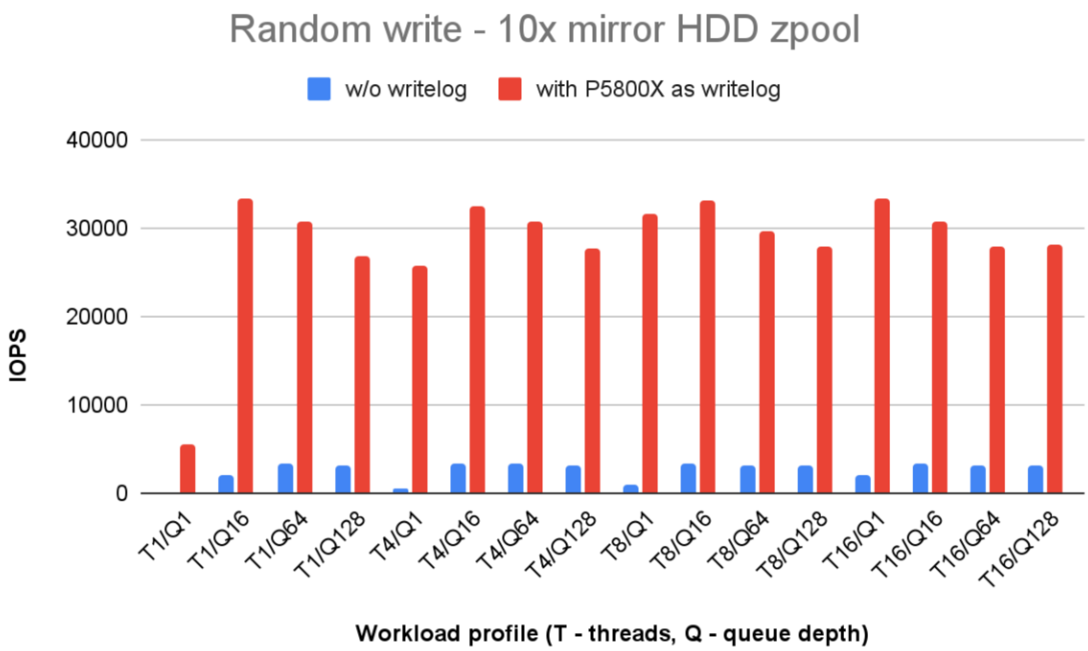

Fig. 5. Random write performance of a 10x mirror HDD zpool, with and w/o writelog.

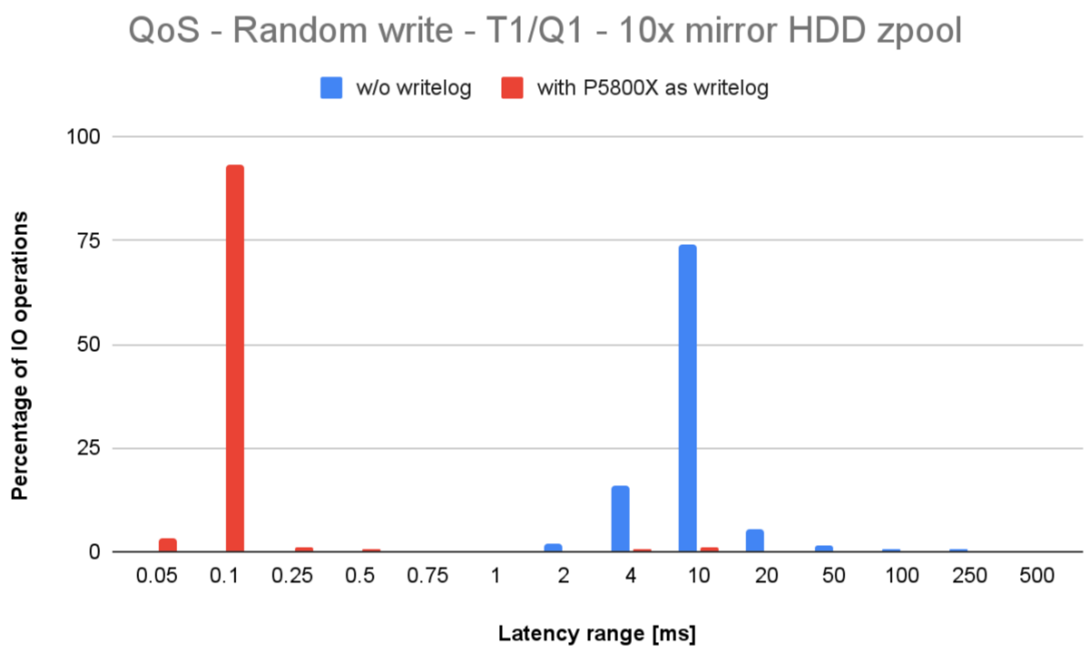

Fig. 6. Latency distribution for a random write test, done on a 10x mirror HDD zpool (T1/Q1), with and without writelog.

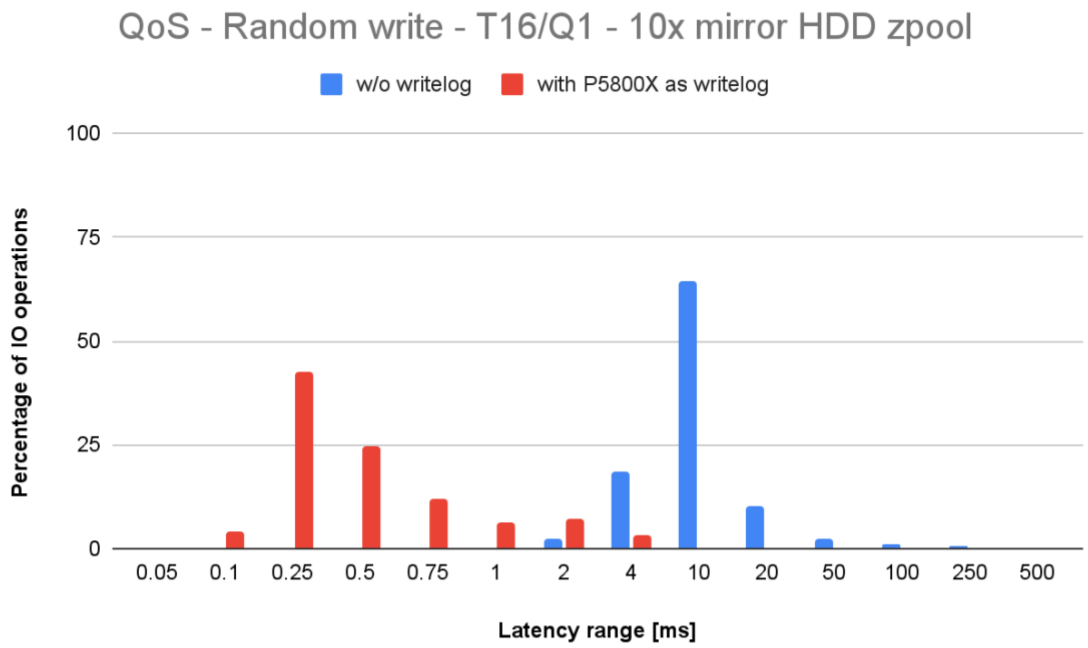

Fig. 7. Latency distribution for a random write test, done on a 10x mirror HDD zpool (T16/Q1), with and without writelog.

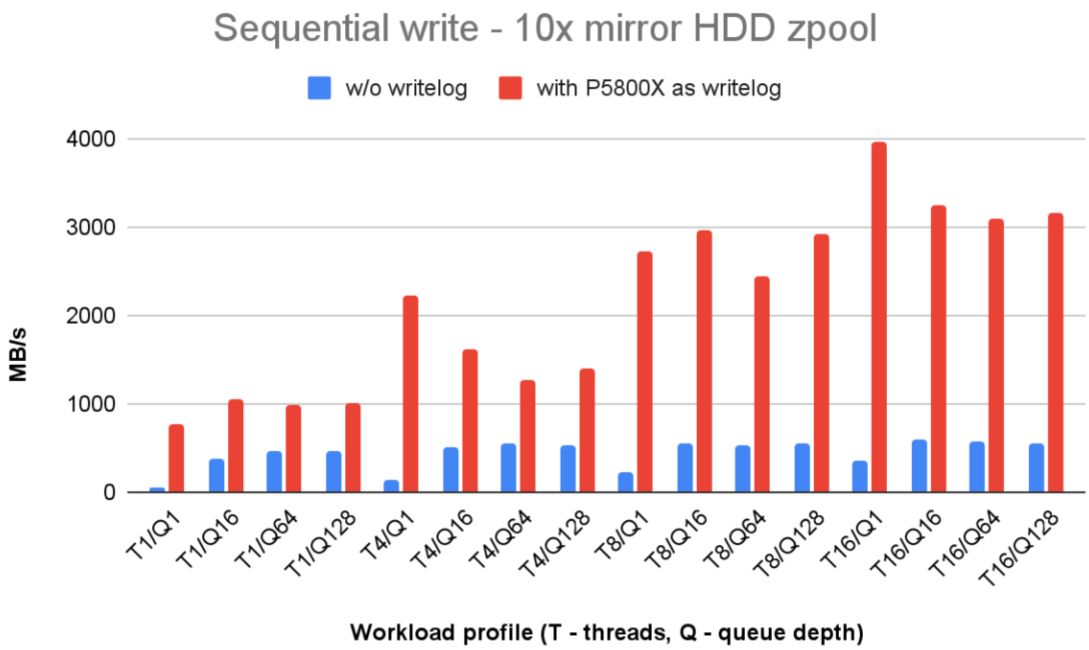

Fig. 8. Sequential write performance of a 10x mirror HDD zpool, with and w/o writelog.

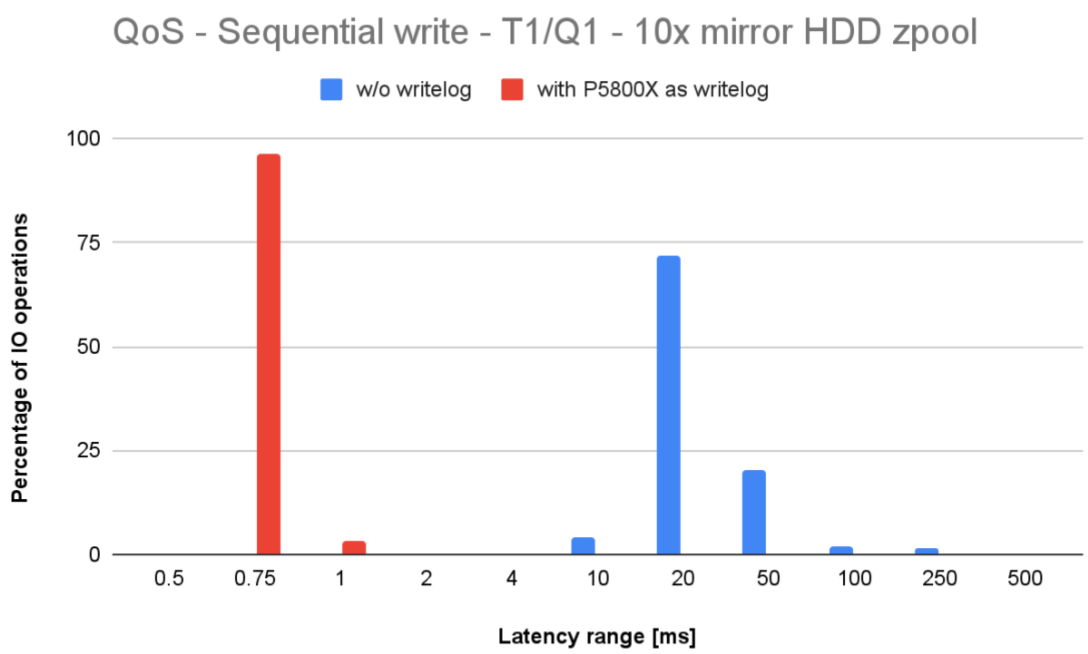

Fig. 9. Latency distribution for a sequential write test, done on a 10x mirror HDD zpool (T1/Q1), with and without writelog.

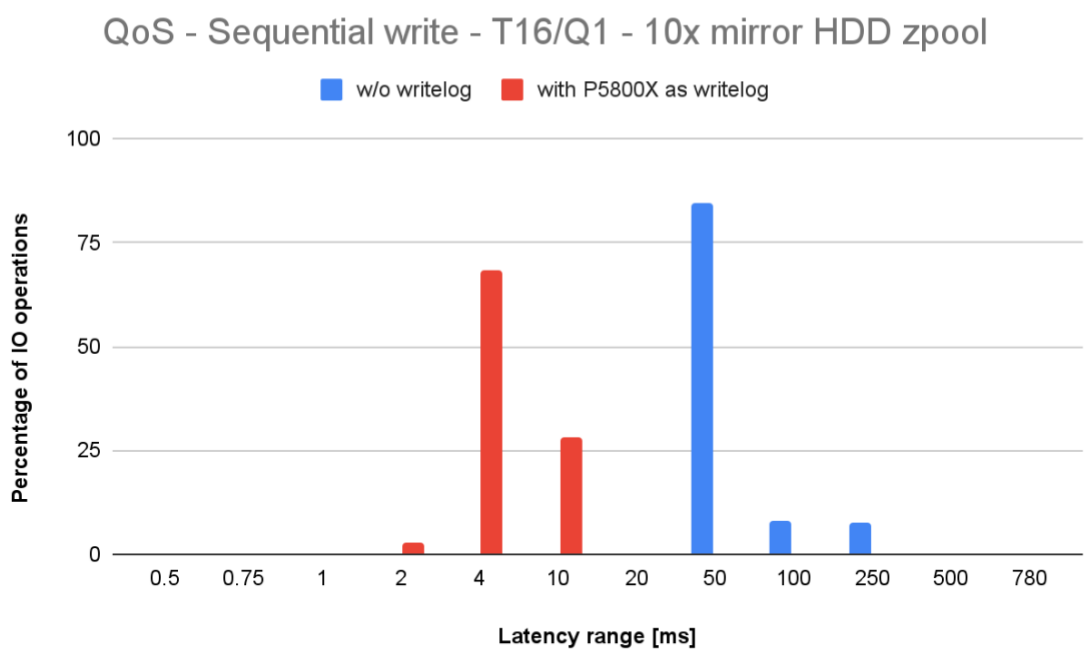

Fig. 10. Latency distribution for a sequential write test, done on a 10x mirror HDD zpool (T16/Q1), with and without writelog.

5.2. Zpool with 20x single HDD data groups

In this test, the zpool consisted of 20 single-disk data groups assembled from Toshiba AL13SEB900 SAS HDD 10K drives. Test parameters are shown in Table 4. Figure 11 presents the results of a random write test case with latency distributions for the T1/Q1 and T16/Q1 combinations presented in figures 12 and 13 respectively. Figure 14 shows the results for the sequential write test with the latency results being presented on Fig. 15 and 16.

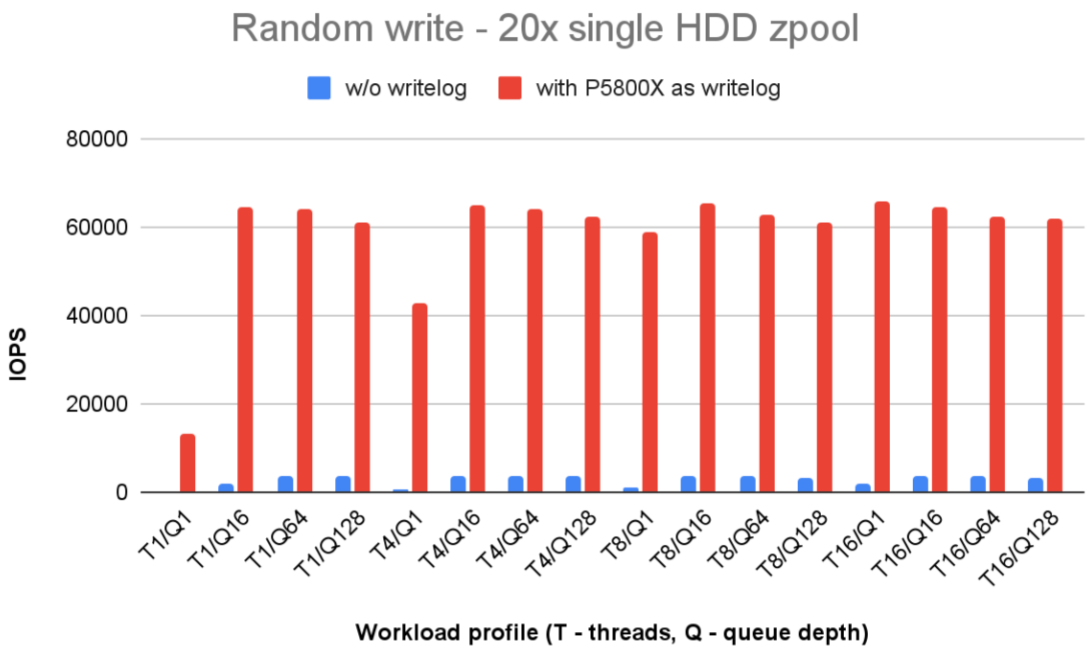

Fig. 11. Random write performance of a 20x single HDD zpool, with and w/o writelog.

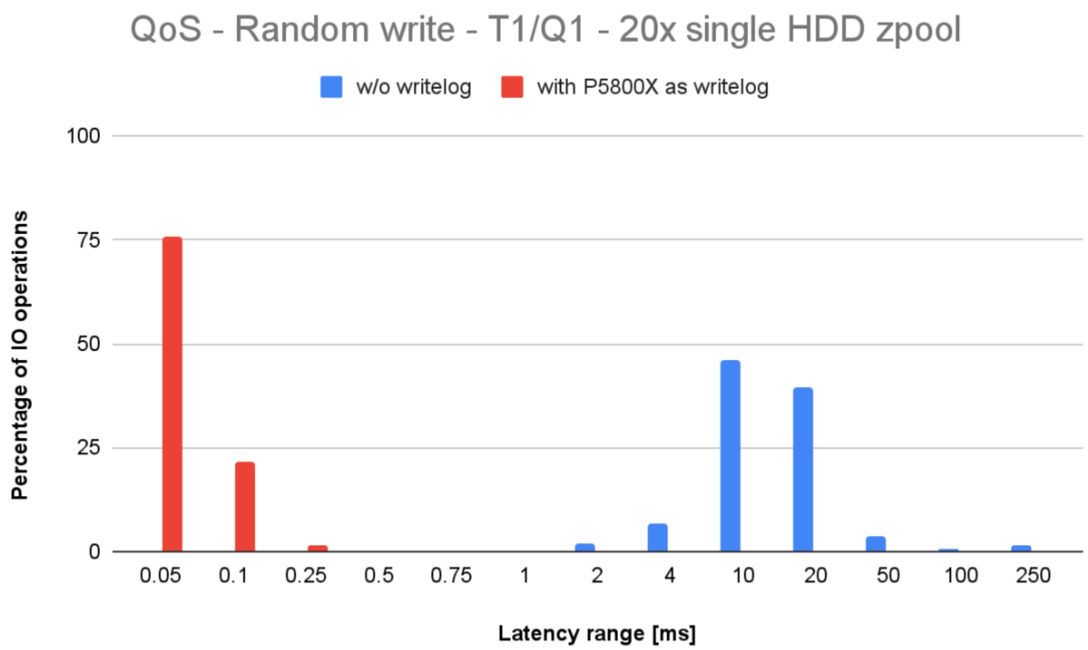

Fig. 12. Latency distribution for a random write test, done on a 20x single HDD zpool (T1/Q1), with and without writelog.

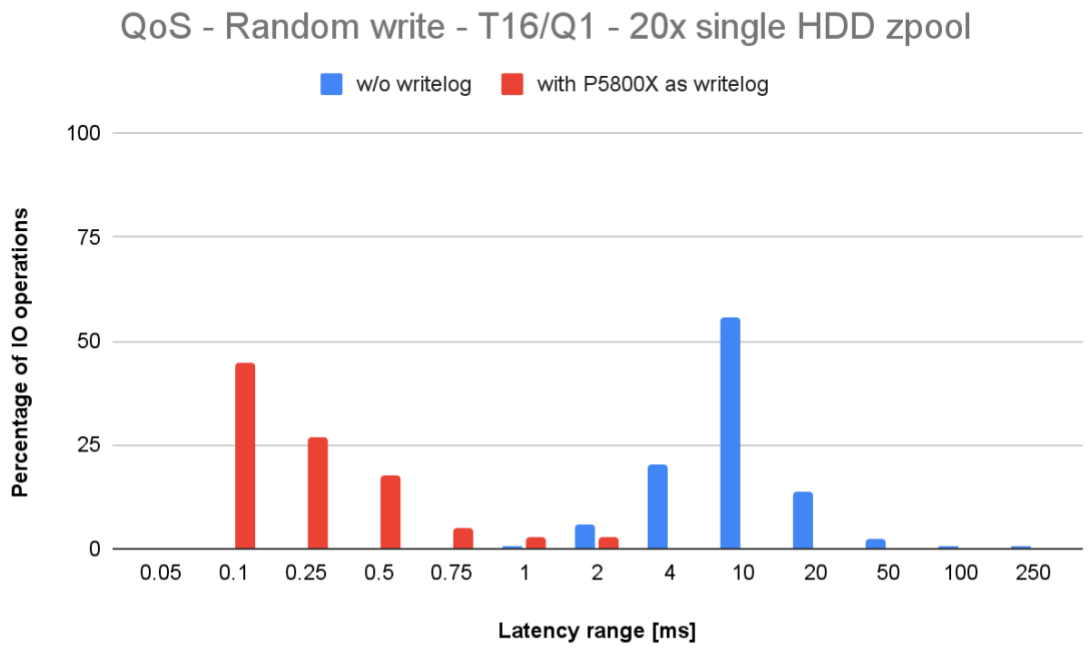

Fig. 13. Latency distribution for a random write test, done on a 20x single HDD zpool (T16/Q1), with and without writelog.

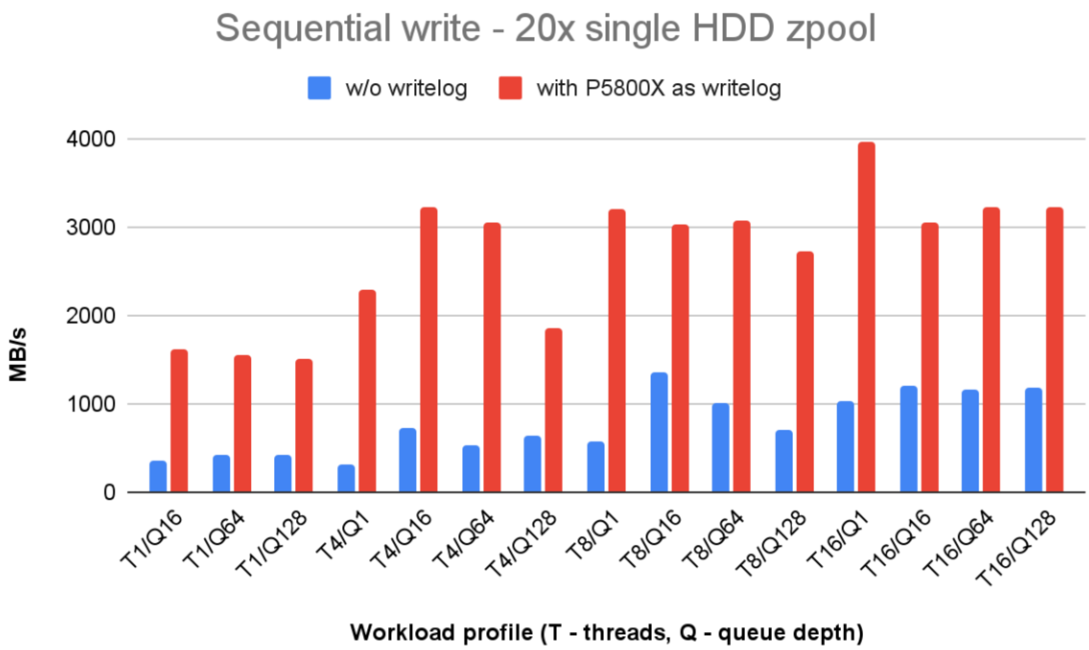

Fig. 14. Sequential write performance of a 20x single HDD zpool, with and w/o writelog.

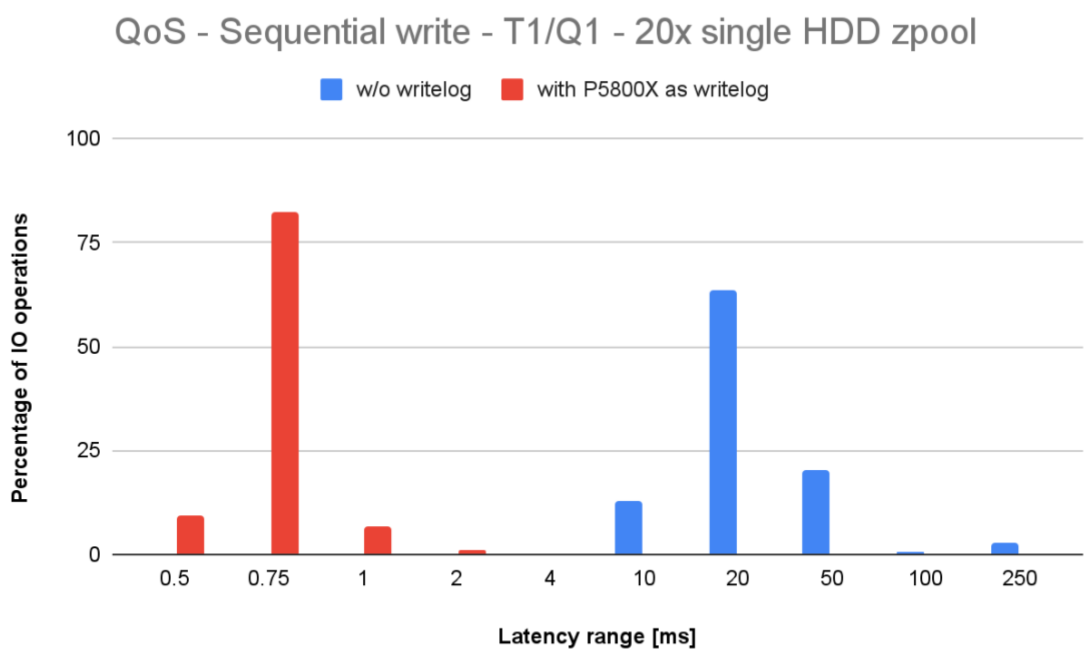

Fig. 15. Latency distribution for a sequential write test, done on a 20x single HDD zpool (T1/Q1), with and without writelog.

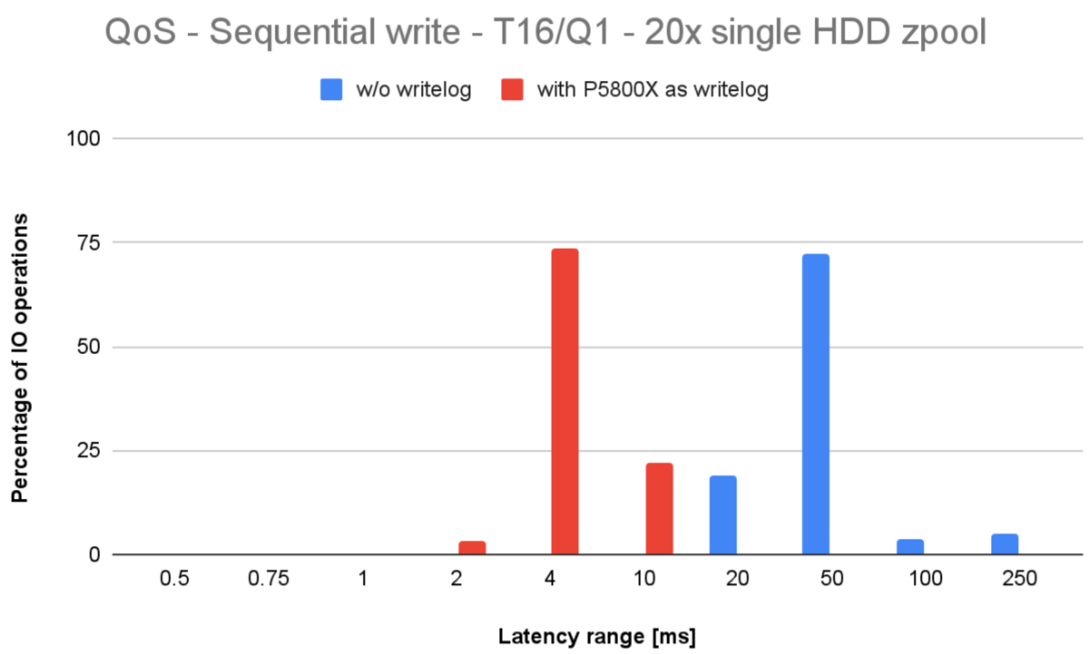

Fig. 16. Latency distribution for a sequential write test, done on a 20x single HDD zpool (T16/Q1), with and without writelog.

5.3. Zpool with 2x 2-way mirror QLC NAND SSD data groups

In addition to HDD disks, tests with QLC NAND SSD drives were also carried out to check the influence of the P5800X used as a writelog device for all-flash storage. In this case 4 Intel P4320 2 TB drives were used to create a zpool in a 2x 2-way mirror configuration. Figure 17 presents the results of the performance for random writes on the zpool both with and without a writelog device. Latency distribution for both the T1/Q1 and T16/Q1 combinations are included in Figures 18 and 19. Sequential write results are shown in Figure 20 along with the latencies, in Fig. 21-22 respectively. Test parameters are described in Table 4.

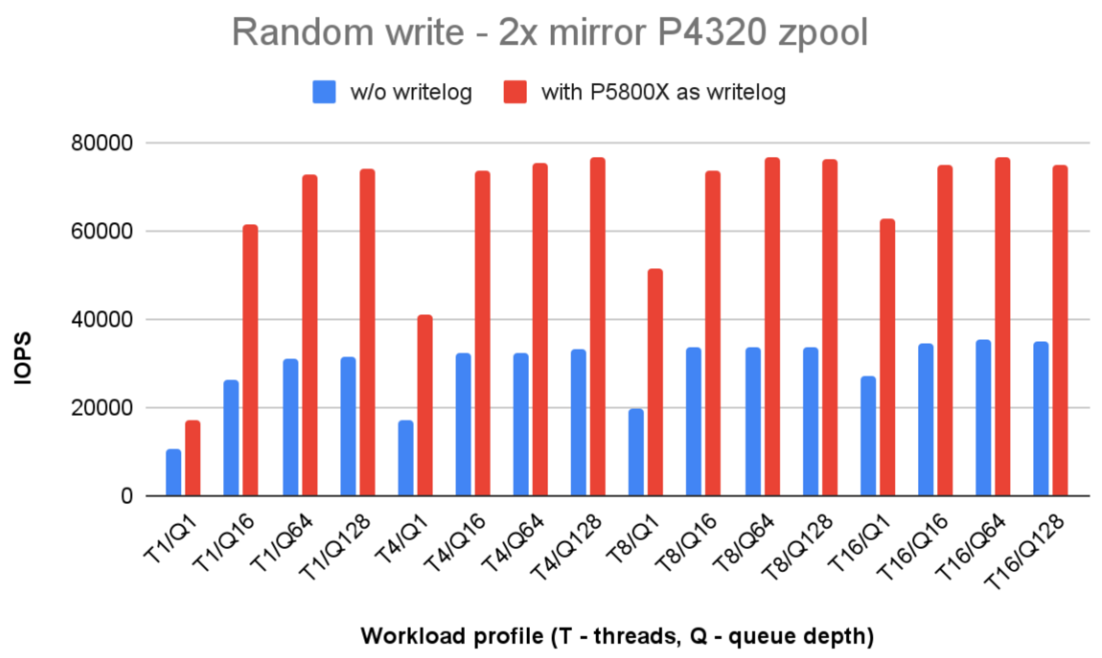

Fig. 17. Random write performance of a 2x 2-way mirror SSD zpool, with and w/o writelog.

Fig. 18. Latency distribution for a random write test, done on a 2x 2-way mirror SSD zpool (T1/Q1), with and without writelog.

Fig. 19. Latency distribution for a random write test, done on a 2x 2-way mirror SSD zpool (T16/Q1), with and without writelog.

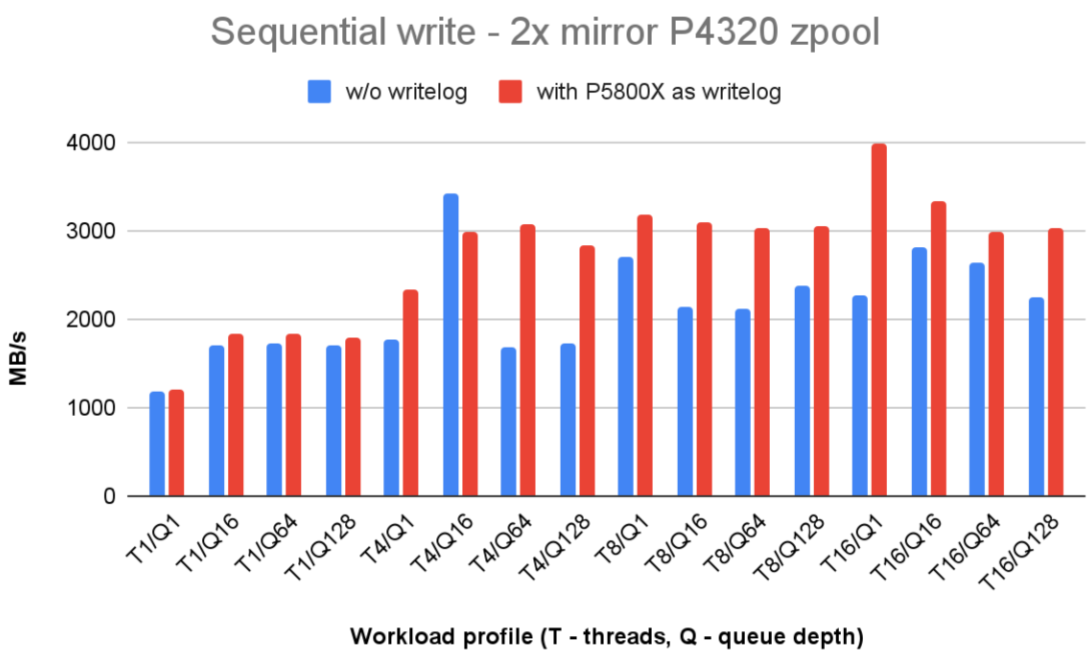

Fig. 20. Sequential write performance of a 2x 2-way mirror SSD zpool, with and w/o writelog.

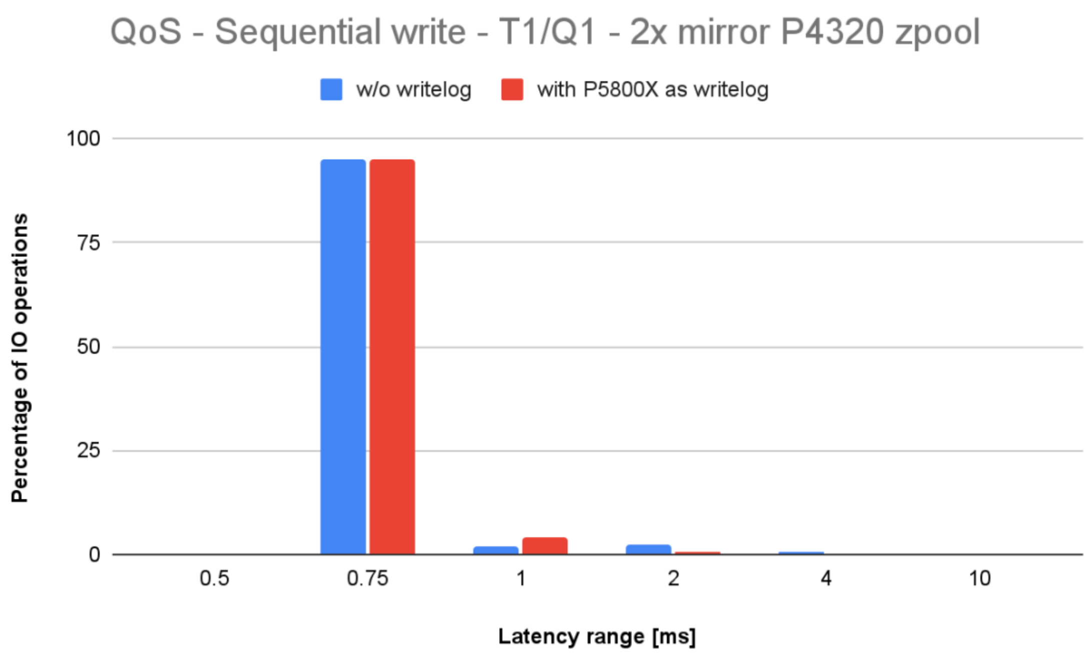

Fig. 21. Latency distribution for a sequential write test, done on a 2x 2-way mirror SSD zpool (T1/Q1), with and without writelog.

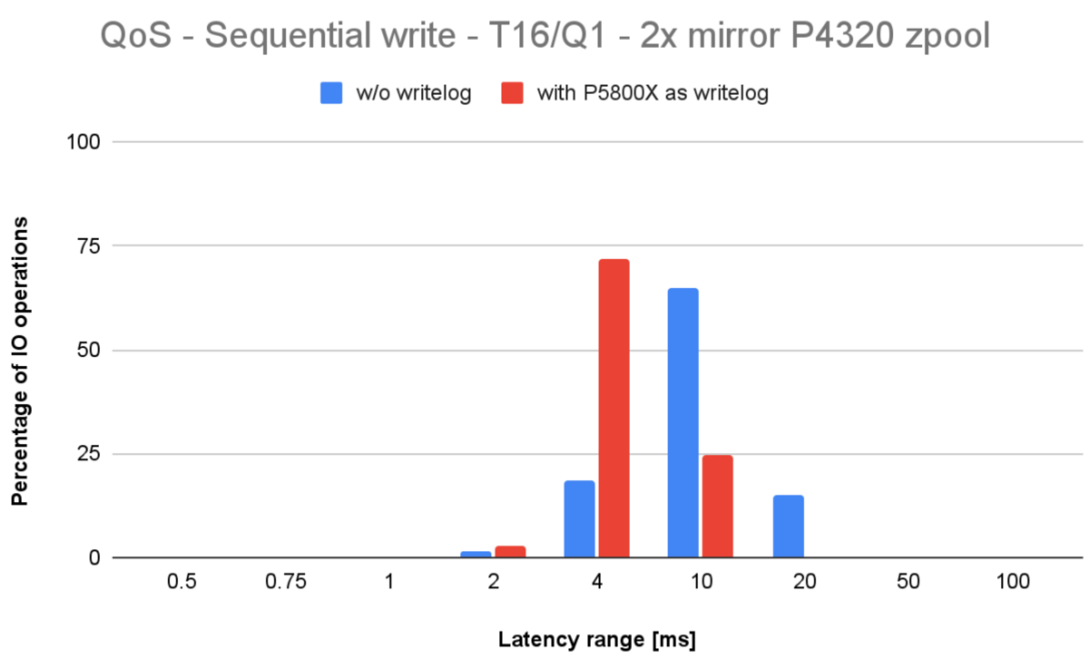

Fig. 22. Latency distribution for a sequential write test, done on a 2x 2-way mirror SSD zpool (T16/Q1), with and without writelog.

1.1 5.4. Zpool with 4x single QLC NAND SSD data groups

Another test including 4 Intel P4320 NAND SSD drives was carried out on zpool created with 4 single-disk data groups. Similarly to the previous test, the IOPS and latency were measured under a load generated by fio on a zpool with and without the Intel P5800X as a writelog device. Test parameters are shown in Table 4. Figures 23 and 26 present the IOPS results for random and sequential write both with and without writelog respectively.

Latency distributions for T1/Q1 and T16/Q1 thread/queue depth combinations are presented in Figures 24-25 and 27-28.

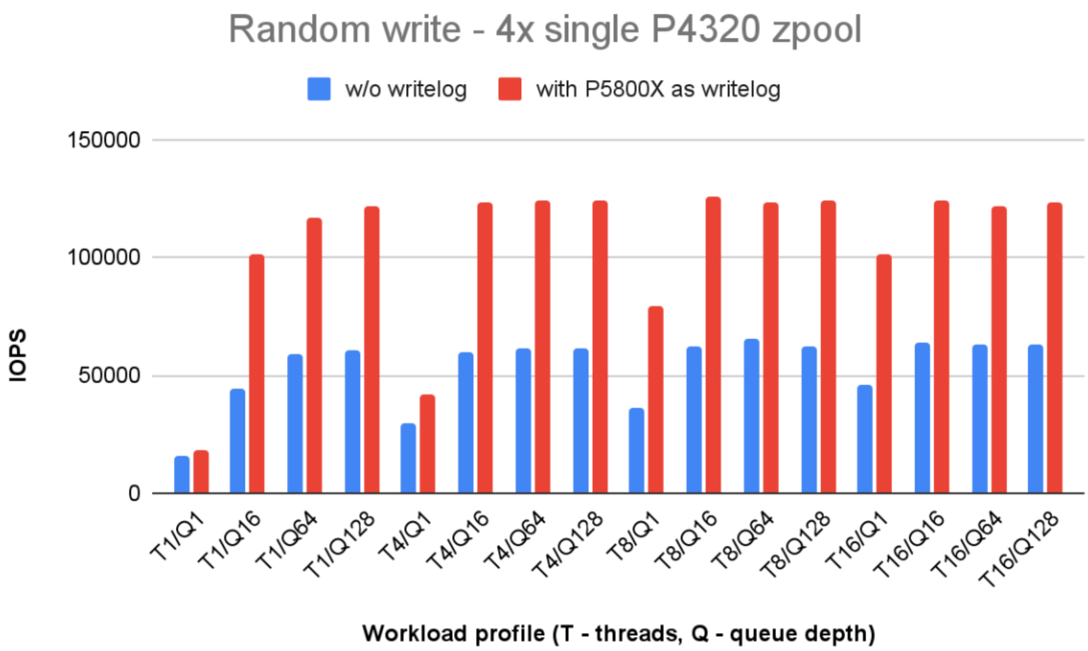

Fig. 23. Random write performance of a 4x single SSD zpool, with and w/o writelog.

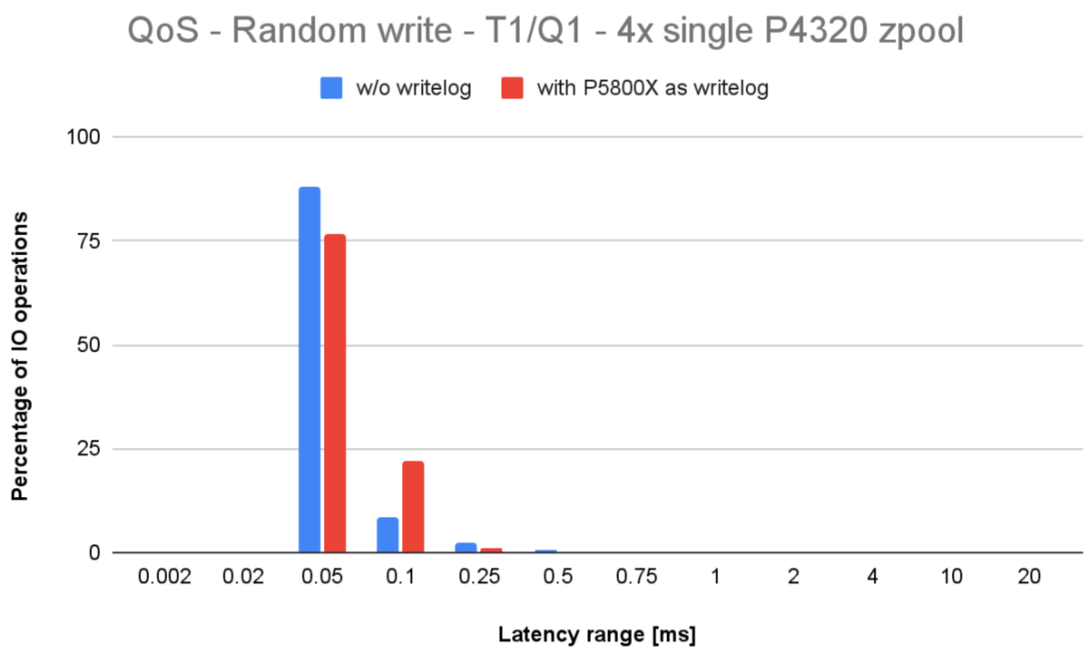

Fig. 24. Latency distribution for a random write test, done on a 4x single SSD zpool (T1/Q1), with and without writelog.

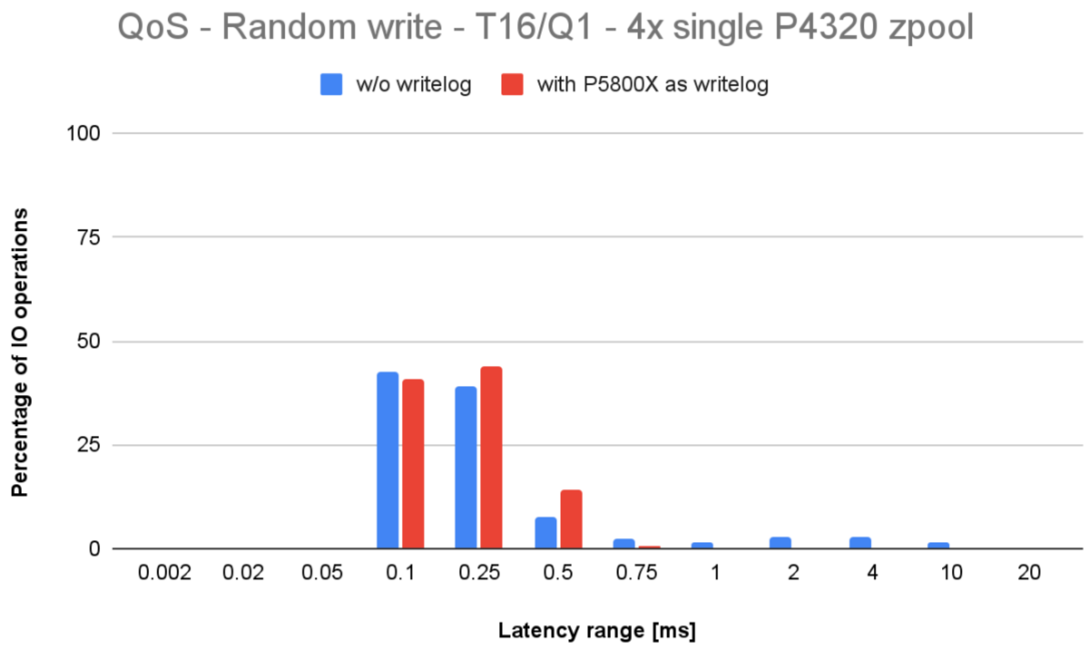

Fig. 25. Latency distribution for a random write test, done on a 4x single SSD zpool (T16/Q1), with and without writelog.

Fig. 26. Sequential write performance of a 4x single SSD zpool, with and w/o writelog.

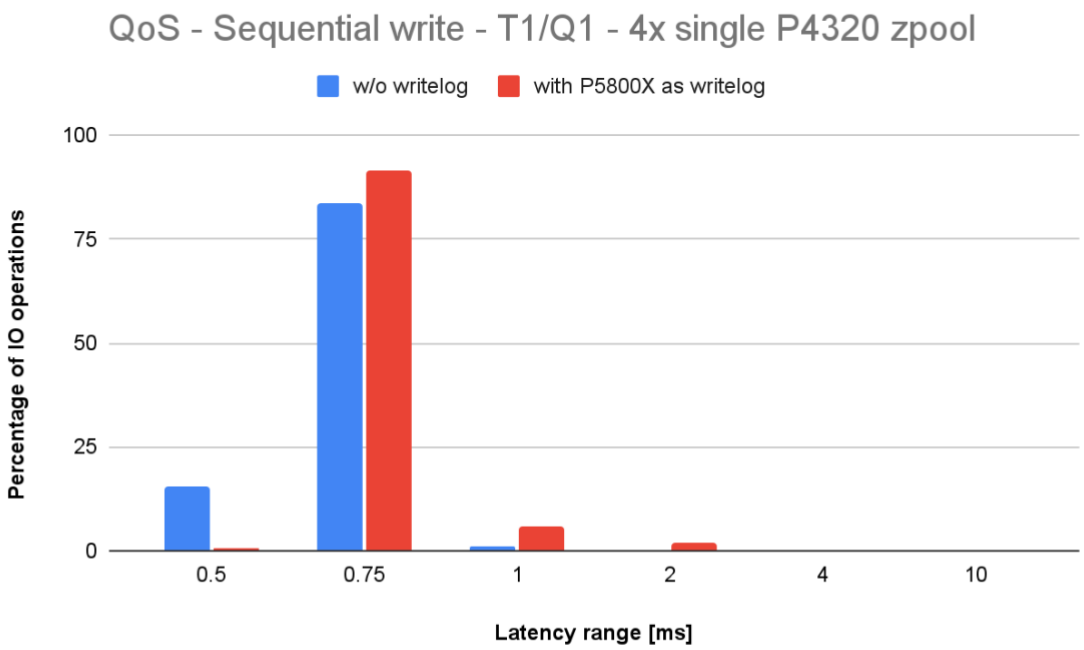

Fig. 27. Latency distribution for a sequential write test, done on a 4x single SSD zpool (T1/Q1), with and without writelog.

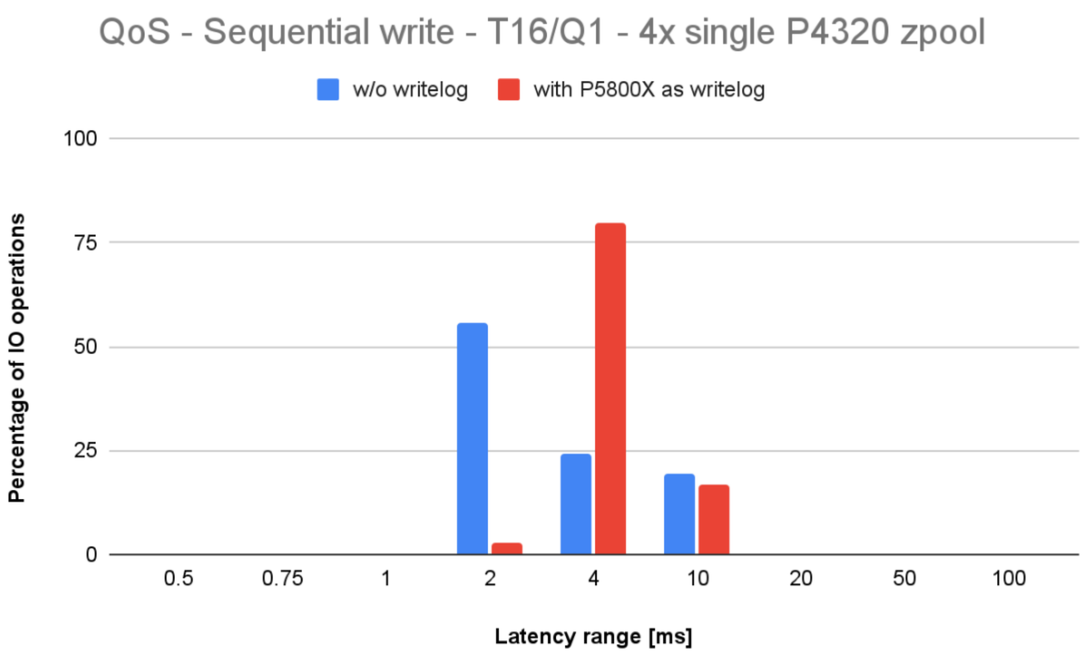

Fig. 28. Latency distribution for a sequential write test, done on a 4x single SSD zpool (T16/Q1), with and without writelog.

5.5. Test conclusions

Random write IO results for both performance and latency improved significantly for zpools with an added writelog over those without one, which can be seen in Figures 5-7 and 11-13. On average IOPS increased 15 times for the 10x mirror HDD configuration and 37-fold for the 20x single HDD zpool. Similarly in both sequential write tests, performance rose while latency decreased. In the first case throughput averaged a 6.3-fold improvement and in the second case a 4.5-fold rise was observed.
Similarly to the tests with HDD devices, the 4x Intel P4320 drives were used as storage devices for a zpool consisting of SSD drives. In both the 2x 2-way mirror and the 4x single drive configurations more than doubled

IOPS results were observed for all random writes (except in the 1T/1Q instance) in test cases with an added writelog as opposed to those without one (Figures 22 and 28). When using sequential IO as a test load, the 2x 2-way mirror zpool showed little improvement when a writelog was added (x1.3 on average). In the case of a 4x single SSD zpool however, a decrease in performance was observed for all threads and queue depths combinations (average 20% decrease) when a writelog was added.

6. Read cache device tests

To evaluate the Intel Optane P5800X drive as a read cache device for Open-E JovianDSS, a model with 1.6 TB capacity was chosen and a zpool consisting of 10 mirrored HDD drives was assembled. In order to lessen the influence of ZFS' ARC cache on the results, its size was decreased to 10% of the nominal value. This ensured that most of the cache hits would come from the read cache device (ZFS L2ARC). The amount of data read by fio was set to be equal to 5 times that of the configured ZFS ARC size. The generated load was 100% random read. Before actual measurements were conducted, fio was run until all reads went from cache.

6.1. Performance results

Figure 29 presents the IOPS acquired for every thread/queue depth combination. Latency distributions for the T1/Q1 and T16/Q1 cases are shown in figures 30 and 31.

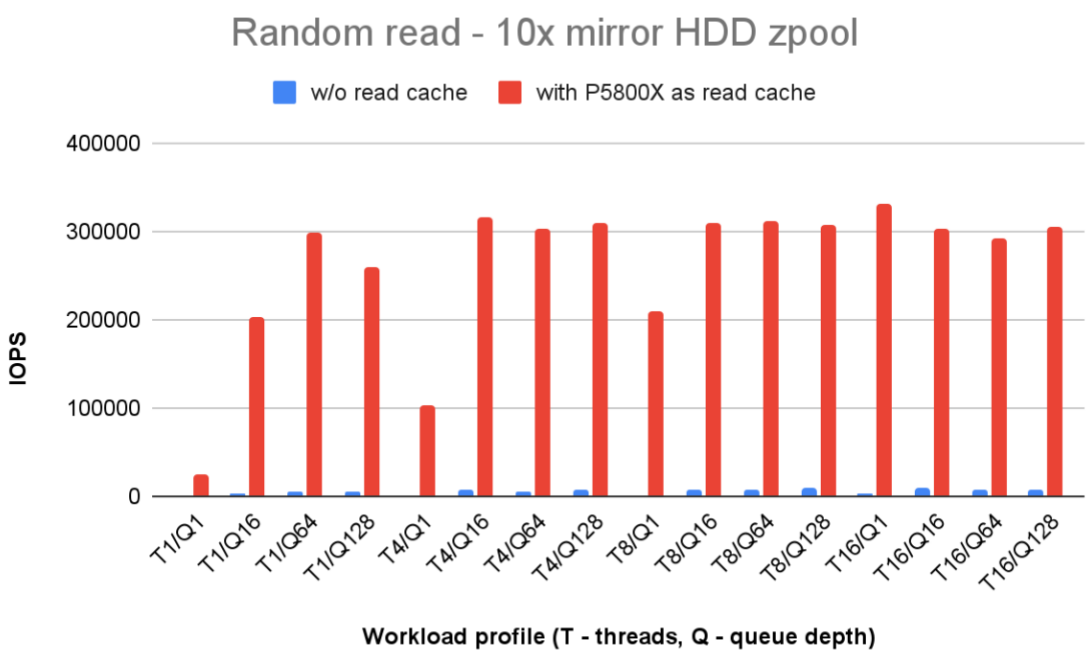

Fig. 29. Random read performance of a 10x mirror HDD zpool, with and w/o read cache.

Fig. 30. Latency distribution for a random read test, done on a 4x single SSD zpool (T1/Q1), with and without read cache.

Fig. 31. Latency distribution for a random read test, done on a 4x single SSD zpool (T16/Q1), with and without read cache.

6.2. Test conclusions

A zpool with an added Intel Optane P5800X 1.6 TB drive operating as a read cache device showed much higher IOPS values for every test case, with an average 70-fold improvement over a zpool without a read cache. Most of the IO requests (above 70%) were handled with a latency between 20 and 50 μs, compared to 4-10 ms in cases of tests without a read cache.

7. HA Storage Cluster compatibility test

In order to ensure proper operation of the Intel D7-P5510 drive in Open-E JovianDSS High Availability Storage cluster environments, various compatibility tests were performed.

7.1. Functional tests

All essential and critical cluster mechanisms were examined for correct operation with the tested drives used as read cache (1.6 TB drive model) and writelog device (400 GB model). They are summarized in Table 5.

Table 5. Results for the HA Storage Cluster compatibility test.

Tested functionality	Result
Manual Failover	passed
Automatic Failover triggered after network failure	passed
Automatic Failover triggered after system shutdown	passed
Automatic Failover triggered after system reboot	passed
Automatic Failover triggered after system power-off	passed
Automatic Failover triggered after I/O failure	passed

7.2. Test conclusions

Compatibility of the tested device with essential HA cluster operations was extensively checked. None of the test cases described in Table 5 showed any undesirable behaviour, indicating full compatibility with Open-E JovianDSS in cluster configurations.

8. Summary

The Intel Optane P5800X drive was comprehensively tested for full functional compatibility with Open-E JovianDSS. Performance characteristics were also examined in several use cases. Both the Single node and HA cluster environments were taken into consideration. Tests were designed to find any abnormalities in the device, whether used as a read cache or a writelog for Open-E JovianDSS. Its usage showed significant improvement in random IO handling in most cases, especially for loads with large thread number and queue depth. It was also beneficial for throughput in sequential write tests with zpools consisting of HDD drives. Given the results, the tested device has now been added to the Hardware Certification List and granted "Certified by Open-E" status