SHROUTRESEARCH

Intel® Optane™ Storage Performance and Implications on Testing Methodology

October 27, 2017

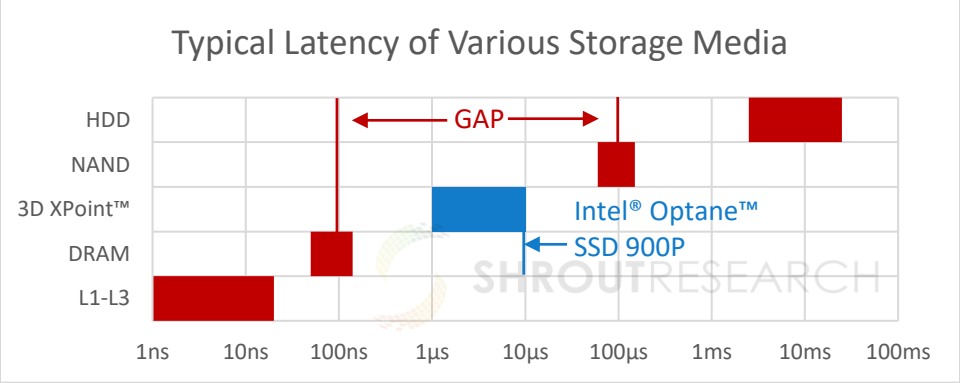
Version 1.0

Executive Summary

Since the unveiling in 2015, 3D XPoint™ Memory has shown itself to be the disruptive storage technology of the decade. Branded as Intel® Optane™ when packaged together with Intel's storage controller and software, this new transistor-less solid-state 'storage class memory' technology promises lower latencies and increased system responsiveness previously unattainable from a non-volatile memory product. When coupled with NVMe and ever faster interfaces, Intel® Optane™ seeks to bridge the gap between slower storage and faster system RAM.

Intel® Optane™ and 3D XPoint™ Technology

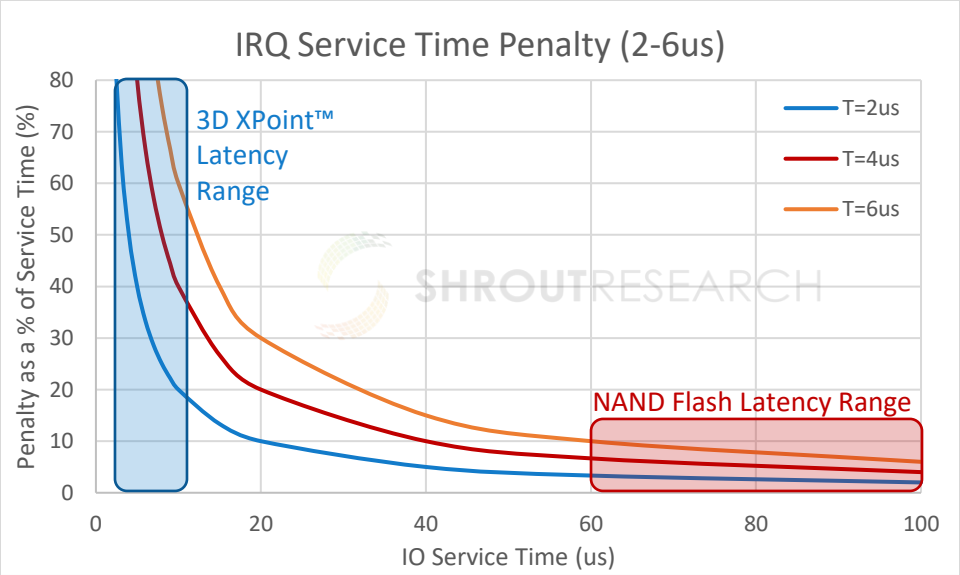
3D XPoint™ represents a radical departure from conventional non-volatile memory technologies. NAND flash memory stores bits by trapping an electrical charge within an insulated cell. Efficient use of die space mandates that programming be done by page and erasures by block. These limitations lead to a phenomenon called write amplification, where an SSD must manipulate relatively large chunks of data to achieve a given small random write operation, negatively impacting both performance and endurance. 3D XPoint™ is free of the block erase, page, and write amplification limitations inherent with NAND flash and can be in-place overwritten at the bit/byte/word level with no need for over-provisioning to maintain high random performance and consistency. 3D XPoint™ data access is more akin to that of RAM, and thanks to the significant reduction in write-related overhead compared to NAND, read responsiveness can be maintained even in the face of increased system write pressure. A deeper dive of how 3D XPoint™ Memory works is beyond the scope of this paper but can be found elsewhere on the web.



3D XPoint™ Technology bridges the ~100x performance gap between NAND flash and DRAM.

Low Latency Storage Impacted by System Configuration

Introducing such a low latency storage device into an insufficiently optimized software/hardware system can potentially run into diminishing return effects as the performance bottlenecks shift further into, and are more greatly amplified by, other portions of the system. The OS kernel's handling of Direct Memory Access (DMA) interrupts – a process integral to the completion of an input/output request, can add between two and six microseconds to each request, varying by OS and hardware platform. While six microseconds might only constitute a minor fraction of typical storage device latency, it is over 50% of the 10-microsecond latencies possible with Intel® Optane™.



In addition to DMA handling, there exist other platform optimizations that may be necessary to realize the full performance benefits of Intel® Optane™. Overly aggressive processor power management resulting from an improperly tuned motherboard BIOS or setting may further impact responsiveness. Such tuning issues were observed in early generation BIOS revisions across several platforms, the worst offenders of which nullifying >80% of the potential responsiveness gains. It has been made clear from these observations that when operating at such low device latencies, unoptimized platforms can lead to significant negative impacts on storage performance and responsiveness.

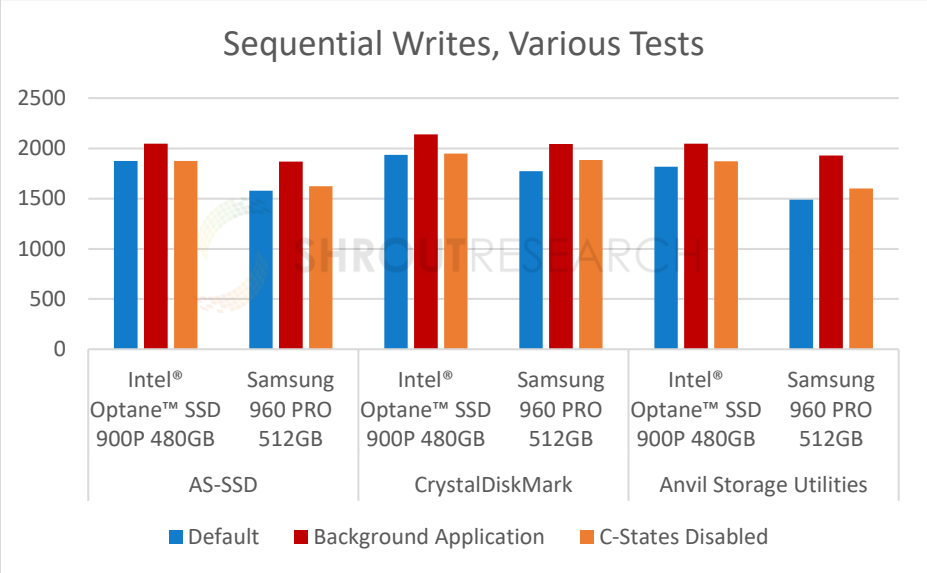
This added delay in low queue depth IO completion is shown as lower than expected results seen in legacy benchmarks run on systems with no other activity taking place (a condition common to storage benchmarking activities). The delay period is relatively constant, but it has a larger negative impact on storage devices capable of lower latencies, as the delay comprises a larger percentage of the response time and ultimately lowers the QD=1 performance reported by the benchmark.

Legacy Testing and Performance Issues

The added complexities of modern storage devices have caused device testing to evolve into an increasingly misunderstood topic, further complicated by a landscape of simplistic legacy benchmarks built upon an outdated understanding of how current generation devices function. Many of the tools available today are not suitable for obtaining accurate real-world performance figures from NAND-based SSDs. Short-run/short-throw tests (Anvil, AS-SSD, ATTO, CrystalDiskMark, PCMark, etc.) place a small test file on the SSD and mix the applied workload within that file, preventing any possibility of a steady-state condition from ever being reached, regardless of the number of times the test has been executed. While other tests (Iometer, PCMark Extended) may address some of the issues noted above, these take things too far in the opposite direction. Specifically, workloads are applied for longer durations and at saturation, overflowing SLC caches and forcing background garbage collection tasks to occur during the test, resulting in measurably reduced performance compared to what would have been seen in real-world usage.

Another issue common across the majority of benchmark applications is that while they do their best to focus their activity only on the device under test, there are cases where such a mentality is too efficient, in that they only issue IO requests and perform no other actions. After a single IO request has been issued (QD=1), the benchmark thread sleeps while waiting for the DMA interrupt signifying IO completion. While this is typical for an application requesting a piece of data, the application and even the overall system load is significantly lower while benchmarking than it would be running real software that would otherwise be performing some level of processing on that incoming data. In the legacy benchmarking scenario, the most active system task is the benchmark itself, and that primary thread entering a sleep state after issuing the request causes the OS scheduler to clock down the CPU. When the IO completes, the interrupt must not only trigger a context switch of the benchmark thread back to the appropriate processor core, it may also have to wake that core, as the CPU was otherwise idle at that time. Further compounding the idle sleep/wake cycle issue is that a typical storage performance testbed will be configured with a minimal amount of background tasks as to not interfere with the execution of the benchmarking application. This sterile environment, combined with applications cleanly issuing IO requests and doing nothing else with the data, results in CPU cores operating at a clock and power state significantly lower than they would be during actual usage of a fully configured system.

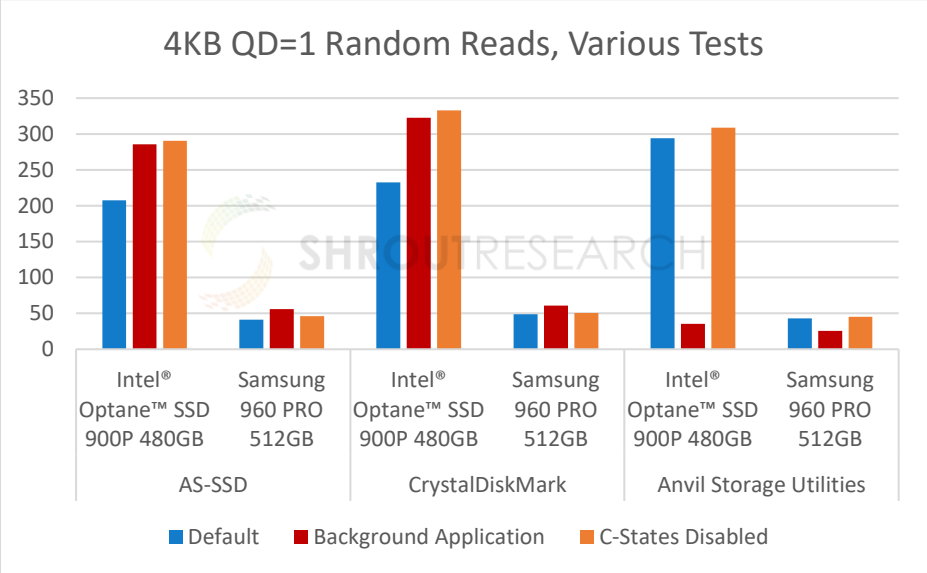
It is theoretically possible to simulate real-world application loads while executing legacy benchmarks, keeping the CPU at the higher clock rate expected during typical system activity without the need to disable processor C-States. Since we do not know which core the OS scheduler may shift the benchmark thread to, all cores would need to be loaded simultaneously to force all CPU cores to their highest clock rate. The instructions being executed must be simple as to minimize thermals and maximize the consistency of CPU boost clock rates. Care must also be taken to ensure the added threads are of the lowest priority as to minimally interfere with the execution of the legacy benchmark.



To demonstrate these points, we have sampled results from a few of the simple benchmark applications. Devices were sequentially filled to 90% capacity prior to test runs to minimize fresh-out-of-the-box conditions as much as practicable within the limitations of these tests. The following condition variants existed across three separate runs:

1. System at default state, running only the benchmark application.
2. Condition 1, plus a background application designed to increase CPU load.
3. Condition 1, but with hardware C-States disabled in system BIOS configuration.

In the above chart, we note that for sequential access, disabling C-States resulted in a slight uptick in measured performance while running the background application resulted in a more noticeable improvement. This is an apparent result of the sequential workload being light enough for the CPU to enter a lower power state between request completions. The additional background activity provided by our application kept the CPU more active and therefore able to respond to IO completions more quickly – even when compared to a system with C-States disabled.

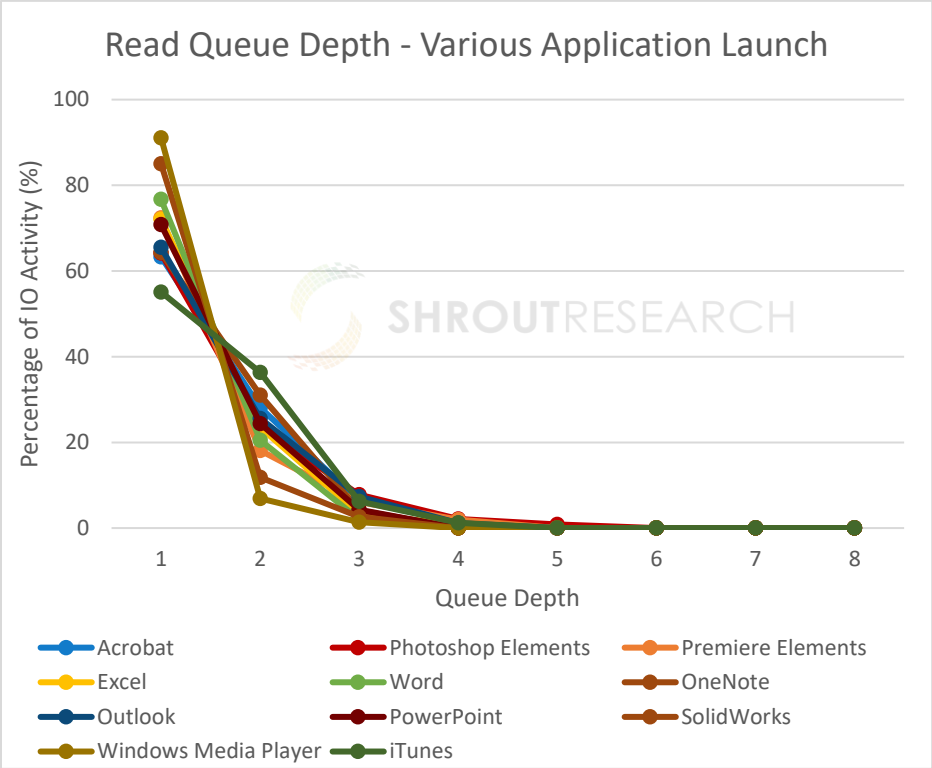


Looking at the same sets of conditions under a random workload, we have a few more interesting points. First is that the background activity, when added by an application separate from the benchmark itself, does not mesh well with all tests, as seen with the apparent interference with the Anvil Storage Utilities results. Second, and perhaps most noteworthy here as a testament to the large gains to be had with the 900P, is that keeping the CPU in a more active state during the test resulted in a **net** gain greater than the **total** performance offered by the 960 PRO. Had these benchmark applications better represented real-world CPU loads, there would be no need for background applications or other workarounds.

Professional product reviewers and power users more familiar with benchmarking storage devices are wise to this deficiency and typically disable processor C-States to keep the CPU clock rate at the maximum, minimizing the delays noted above. This attempted workaround has an adverse impact on power consumption and should not be employed as a default system configuration, as the CPU must constantly remain at a higher than normal power state to support the higher clock rate, even while idle. This significantly reduces power efficiency, particularly during prolonged idle conditions. Further, as the two preceding charts indicate, the disabling of processor C-States is not a perfect substitute for true real-world CPU loading, as it apparently undercompensates for sequential workloads while overcompensating for those that are more random in nature.

Since adding background activity to the system resulted in significant positive improvements to the low queue results of many legacy benchmarks, we can conclude that these tests are not coded in a way that can accurately report the low queue depth performance of modern storage devices being tested on modern high core count systems. Ironically, this very metric is the most important in demonstrating the strengths of next-generation low-latency storage devices.

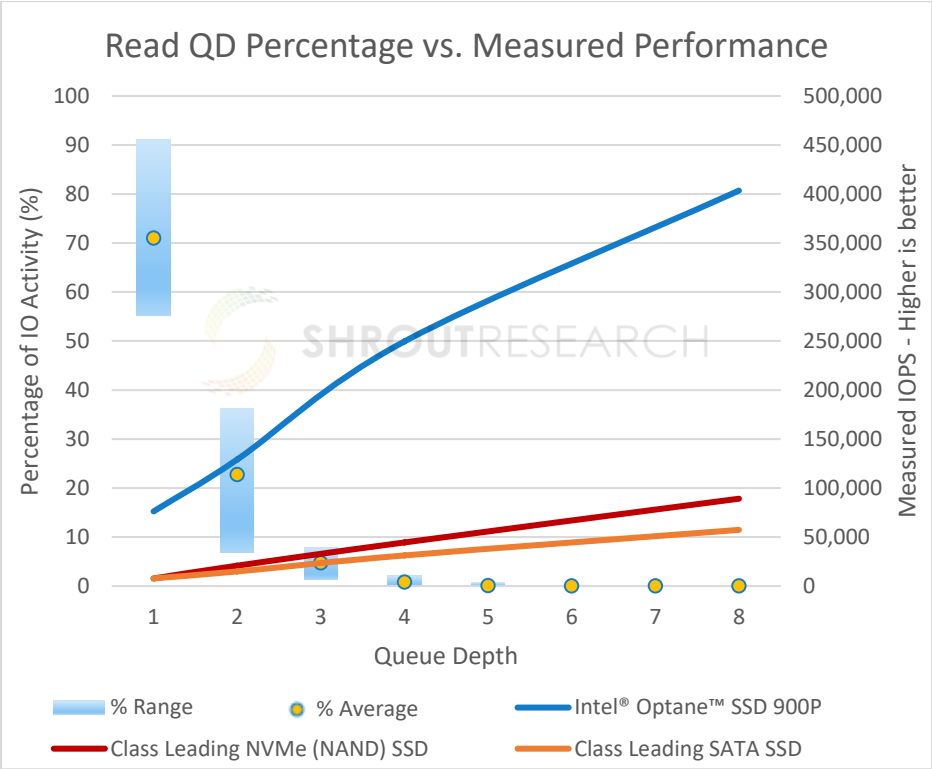
Focusing on Real-World Queue Depths



Queue depths recorded during typical application launches.
Data provided by Intel and validated by Shrout Research.

The significance of lower Queue Depth (QD) workloads cannot be understated. While SSD specifications lend themselves to a 'megapixel race' where manufacturers strive for ever higher 'maximum IOPS' ratings, it is vital to consider that SSDs are typically rated at queue depths of 32, 128, or even 256. The above table shows that real-world workloads are nowhere near those very high values.

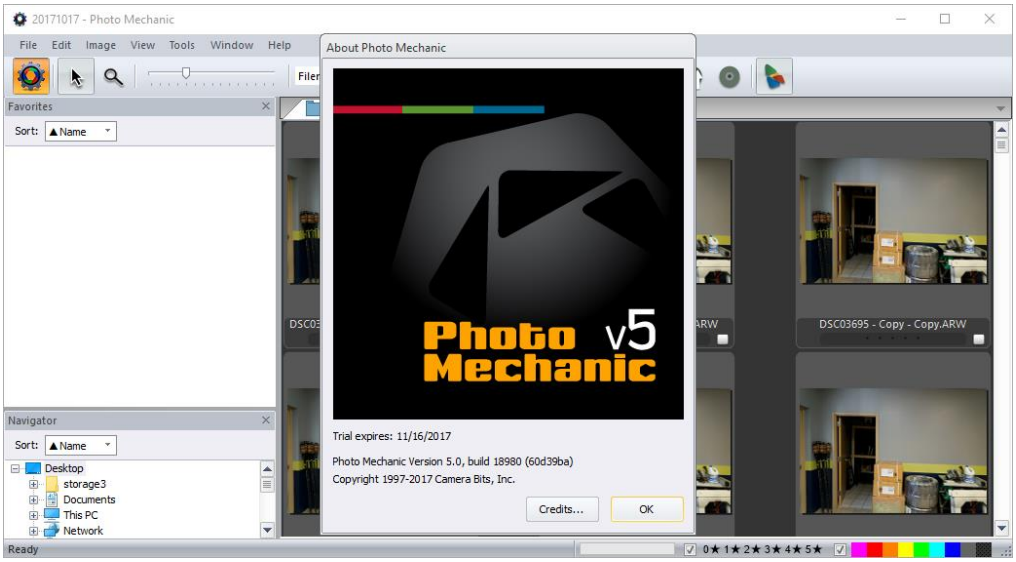
The above data was obtained from a system running a NAND-based SSD. Had an Intel® Optane™ SSD been used, the reduced read latency would have naturally driven the recorded queue depths even lower. This is due to faster storage response times' natural effect of 'shallowing the queue', as the lower latencies reduce the likelihood of an IO request waiting on the completion of its predecessor. Even as it stands, the data clearly shows a current state in which low QD performance is more telling of real-world consumer performance.



Here we have taken the percent activity ranges from the previous data set and overlaid the measured performance of the Intel® Optane™ SSD 900P, as well as class-leading NVMe and SATA products. The data shows the SSD 900P retains a strong performance lead at the queue depths most commonly seen during real-world usage scenarios. High performance at these low queue depths leads to a more responsive system, significantly reducing the wait times for most user initiated actions.

The NAND-based products above do indeed have respectable maximum IOPS ratings, but the higher media latency means those figures can only be realized at queue depths never reached in actual use. Meanwhile, the Intel® Optane™ SSD 900P climbs quickly, reaching its maximum possible performance sooner than the competing products.

Real-world Testing Examples



Camera Bits Photo Mechanic is a media tool used to manage and organize digital photos.

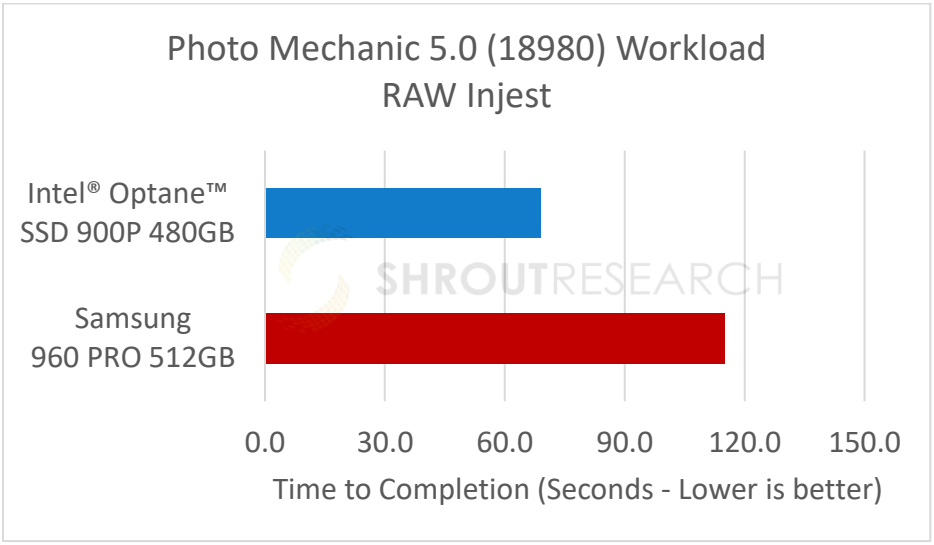
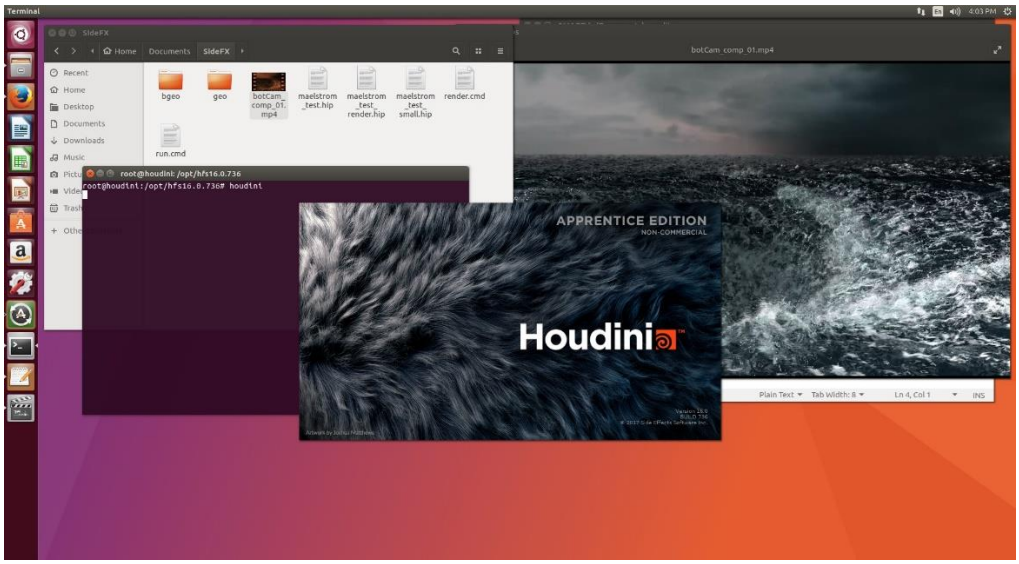
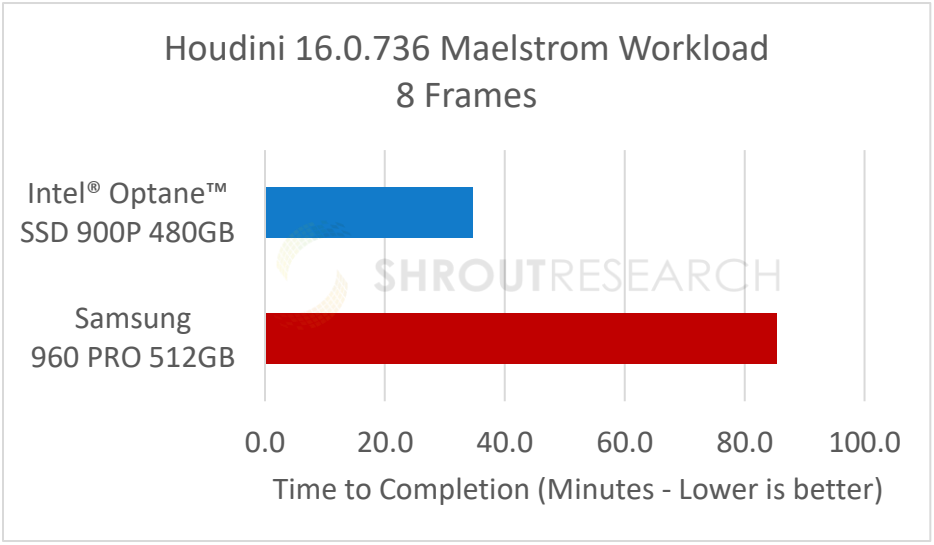


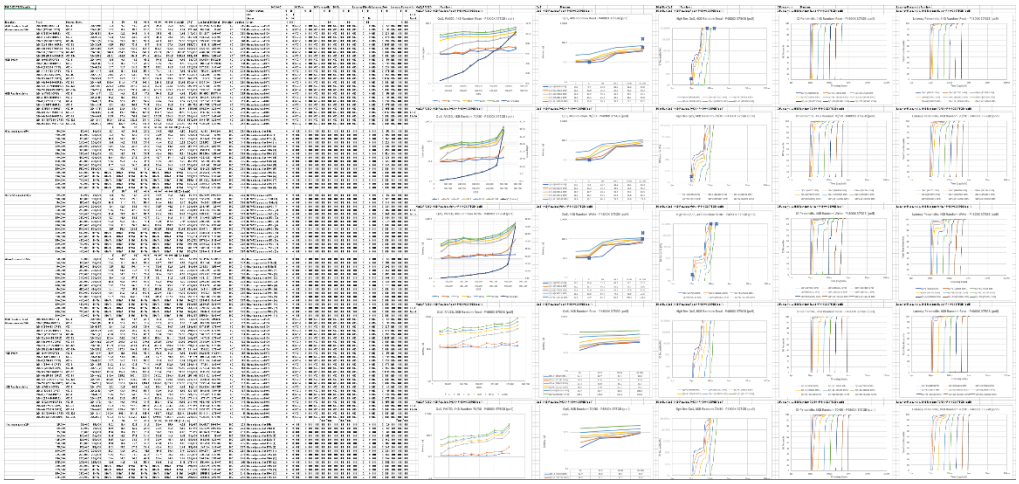
Photo ingest/import is one of the first steps of a professional photographer's workflow. Using the products as a working drive for photo processing, the 900P saw a 67% improvement over a NAND SSD in the time it took to ingest 3,075 photos totaling 37.4GB. This time reduction is further amplified throughout the course of a project since the mixed workload performance gains of Intel® Optane™ equally apply to other operations which simultaneously read and write (photo duplication, export, etc.)



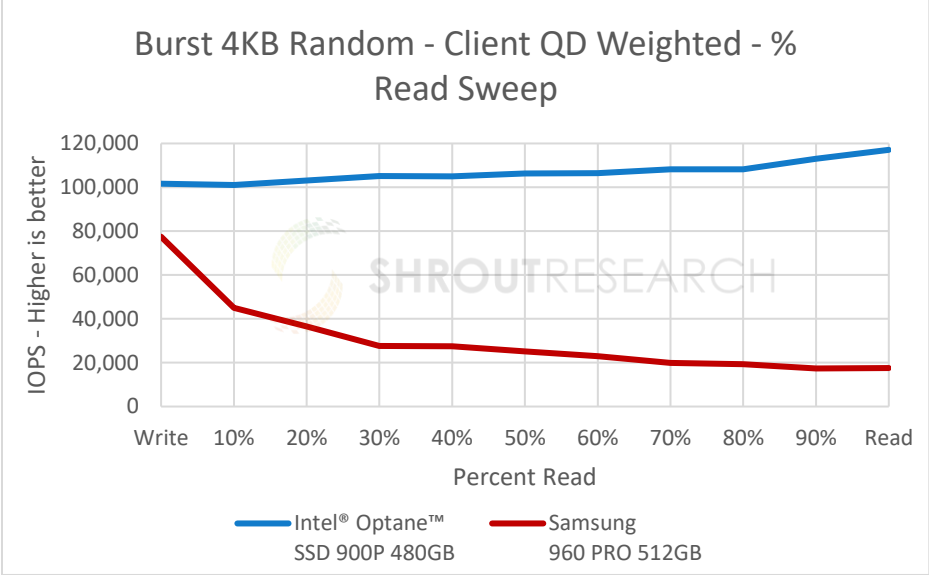
SideFX Houdini FX is a tool used for 3D animation and visual effects (VFX).



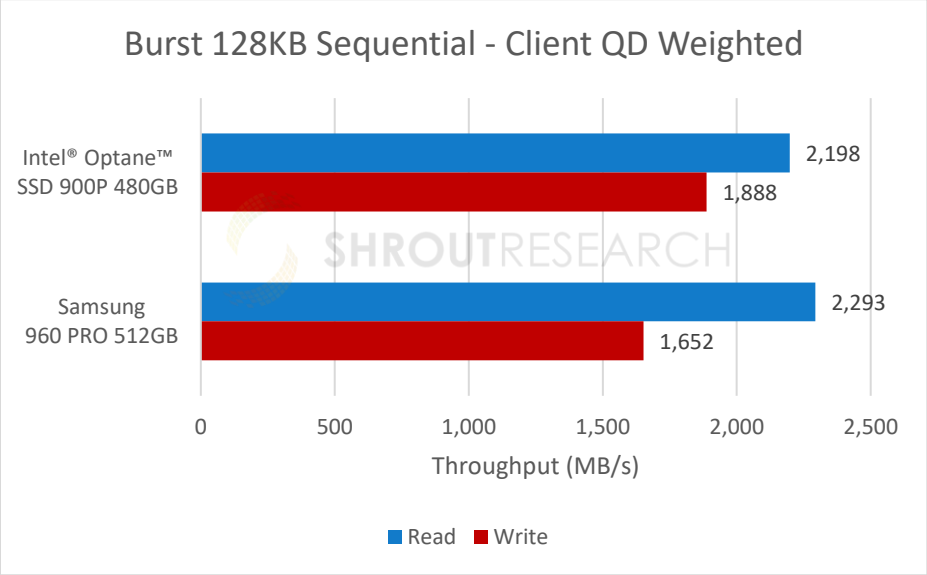
Utilizing the 900P as a swap space augmenting 32GB of DRAM under Linux resulted in a 2.5x improvement to render times. The superior mixed workload performance of Optane™ cut time to complete by more than half. A render shop considering building multiple workstations to achieve their desired timeline could potentially halve the number of machines needed, significantly reducing costs.



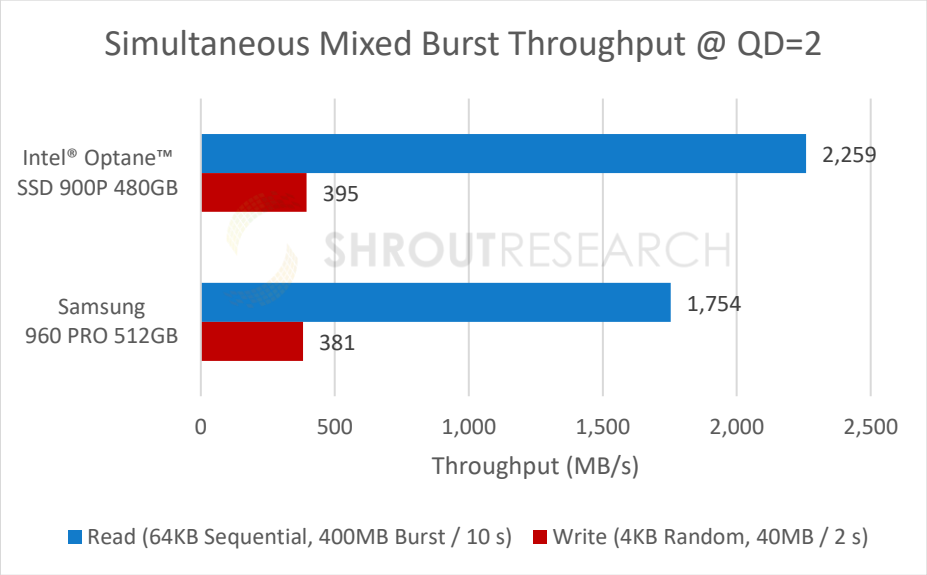
The Shrout Research Storage Performance Analysis Software Suite (S-PASS) is an in-house developed toolset that ensures realistic conditioning of the storage device under test. Workload application granularity is superior to that of any off the shelf benchmark tool. Precise IO-level latency telemetry enables tracking of instantaneous throughput and IOPS of even the shortest of workload bursts.



The above result is based on a synthesized workload applied to the storage devices at varying percentages of fill and Queue Depth. Client workloads typically fall in the center-right section of the percent read spread, where the 900P enjoys up to a 6.7x lead. Higher performing NAND-based SSDs can approach Optane™ performance during low QD write workloads as their controllers effectively 'hide latency' by acknowledging the incoming IO, temporarily storing data in registers located on the flash memory dies, where it is later written in the background. This opens the door to possible in-flight data loss on power failure events, as flash memory Page Program Time (tPROG) and Block Erase Time (tBERS) far exceed the Read Time (tREAD) of the media, evidenced by NAND SSD read ratings far exceeding those of writes.

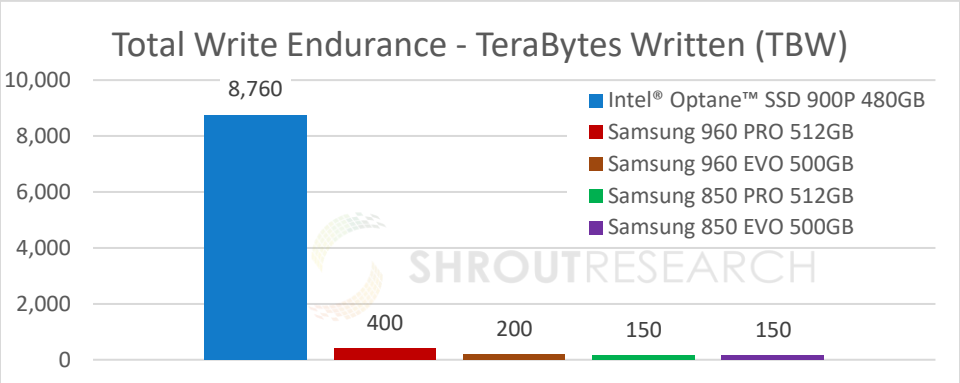


In purely sequential burst workloads, the 900P offers faster write performance while falling slightly behind on reads, but it should be noted that in actual real-world usage, sequential reads do not occur in a vacuum:



This mixed throughput test introduces a relatively light (20 MB/s) write load while requesting 400MB bursts of sequential data from the product under test. The background write load taxed the NAND flash to such a degree that sequential read performance dropped by nearly 25%, while the 900P saw no such performance deficit and realized a 28% gain, shaving a full second off the total time required to read 4GB of data during this test.

Endurance Implications



A final point worth consideration is that of endurance. While NAND-based products are typically sufficient for lighter general client usage, workstation and power users frequently execute repeated heavier workloads that can prematurely wear out consumer-grade NAND flash media. Additionally, 3D XPoint™ media does not suffer from the same Write Amplification effects as compared with NAND flash, meaning an even greater effective endurance than the minimum 20x improvement shown above.

The Changing Storage Landscape

The Intel® Optane™ SSD 900P has shown itself to offer impressive low-latency performance, enabling great time savings for demanding tasks such as those sampled in this paper. But the potential improvements do not stop with what can be tested today. Improving storage device responsiveness by an order of magnitude above high-performance NAND-based NVMe SSDs opens the door for a host of possible software enhancements driven by this new hardware.

Had 3D XPoint™ and Intel® Optane™ been available a decade ago, the modern hardware and software landscape would likely look very different. Architectures might have integrated small amounts as a Level-4 cache. Operating Systems might have fully transitioned to ‘instant-on’, where all necessary code would be executed directly from the 3D XPoint™ media. Games might have been better optimized to rely less on DRAM, instead streaming richer content and larger textures on-the-fly. While we do not have a time machine to affect the above changes, they remain exciting innovations worth looking forward to, and there are relatively minor things that can be done to help make low-latency storage more effective today. At the lower levels, Operating Systems can be updated to more optimally handle IO-related thread scheduling, and improved drivers that implement hybrid polling techniques can remain resource efficient while also mitigating IRQ-related latencies. Intel® Optane™ will spark the momentum for software to adapt, providing the platform on which this new class of storage will evolve.

Author: Allyn Malventano, Technology Analyst at Shrout Research

Editor: Ryan Shrout, President and Analyst at Shrout Research

Please direct questions about this paper to allyn@shroutresearch.com.

Disclosure: This paper was commissioned by Intel. All testing, evaluation, and analysis was performed in-house by Shrout Research and its contractors. Shrout Research provides consulting and research services for many companies in the technology field, other of which are mentioned in this work.

The information and data presented in this document are for informational purposes only and Shrout Research is not responsible for any inaccuracies, typographical errors, or omissions. Any and all warranties are disclaimed in regard to the accuracy, adequacy or completeness of data and information contained within. The document includes opinions of Shrout Research.

Appendix

The following test system configuration was used in the preparation of this paper:

Component	Product / Version
CPU	Intel® Core™ i9-7900X
Motherboard	ASUS Prime X299-A (BIOS 0802)
RAM	Corsair 32GB (8GB x4) DDR 3000 C15
GPU	NVIDIA GeForce Titan X Pascal 12GB
OS	Windows 10 Pro RS2 / Ubuntu 17.10