SHROUTRESEARCH

Intel® Optane™ SSD 800P Low Queue Depth Performance and Implications on Testing Methodology

March 8, 2018

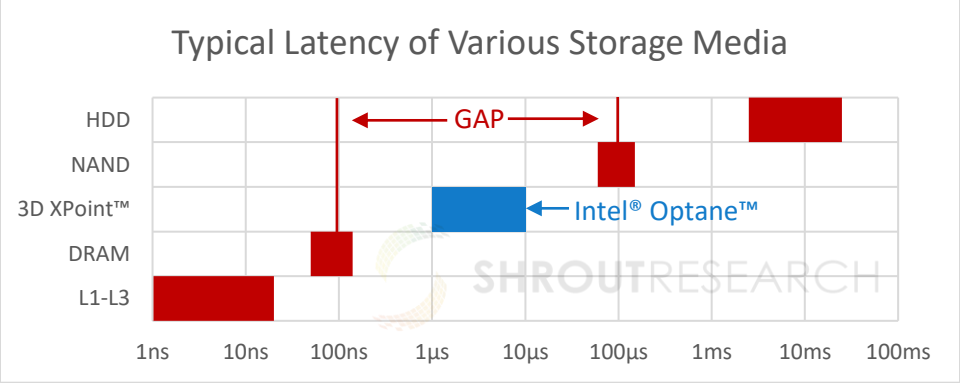
Version 1.0

Executive Summary

Since the unveiling in 2015, 3D XPoint™ Memory has shown itself to be one of the most disruptive storage technologies of the decade. Branded as Intel® Optane™ when packaged and productized, 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. The Intel® Optane™ SSD 800P seeks to bridge the gap between slower storage and faster system RAM to become the most responsive consumer system drive available.

Intel® Optane™ and 3D XPoint™ Technology

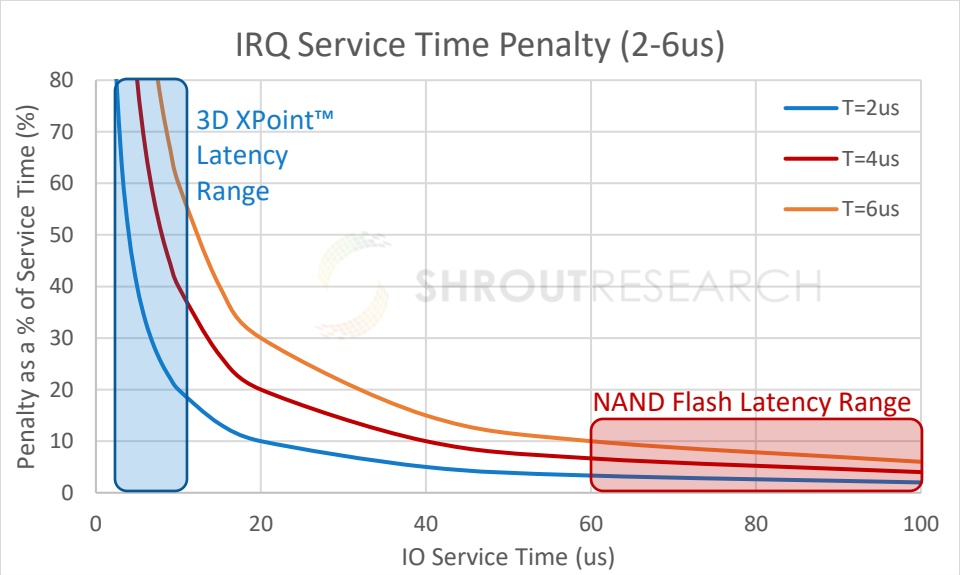
3D XPoint™ represents a radical departure from conventional non-volatile memory technologies. NAND flash memory stores bits by trapping an electrical charge within an insulated cell. Efficient use of die space mandates that programming be done by page and erasures by block. These limitations lead to 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™.



When operating at such low device latencies, unoptimized platforms can artificially limit the ultimate potential of the storage subsystem, hindering system performance and responsiveness. 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 nullified most of the potential responsiveness gains afforded when upgrading to lower latency storage.

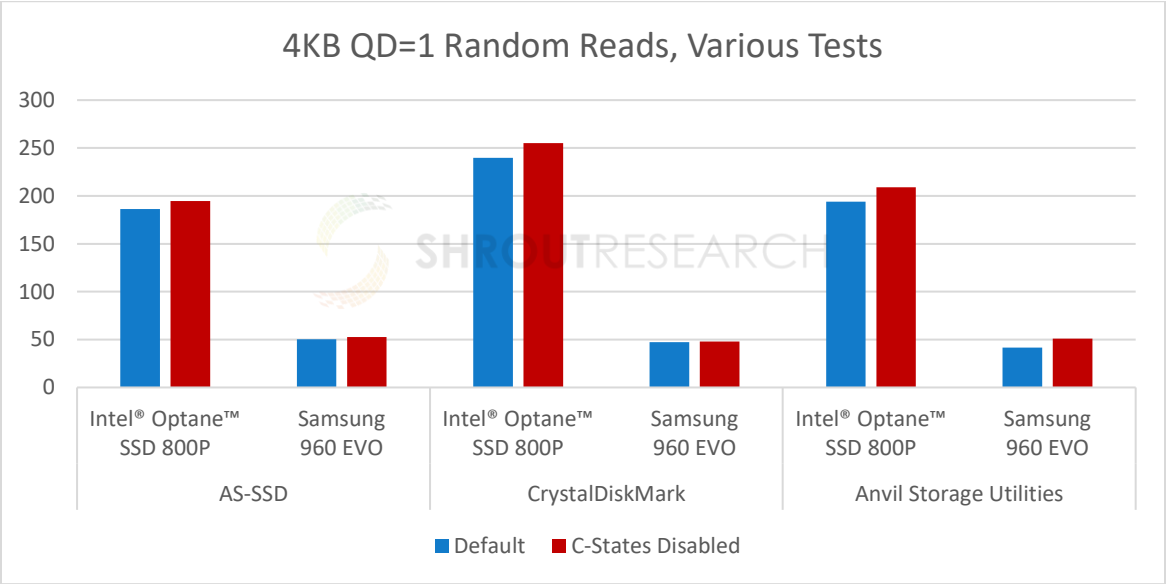
This added delay in low queue depth IO completion is shown as lower than expected results seen in legacy benchmarks run on systems with no other activity taking place (a condition common to storage benchmarking activities). The delay period is relatively constant, but it has a proportionally greater negative impact on the system results observed when using lower latency storage devices. While the storage device remains at its rated latency, the static IRQ delay comprises a larger percentage of the total system response time, ultimately lowering the QD=1 score reported by the benchmark.

Legacy Testing and Performance Issues

The added complexities of modern storage devices have led to storage evaluation becoming an increasingly misunderstood topic, further complicated by a landscape of simplistic legacy benchmarks built upon an outdated understanding of how modern storage devices function. Many of the tools available today are based on outdated test methodologies based on early SSD or even HDD testing, making them unsuitable for obtaining accurate real-world performance figures from NAND-based SSDs. Short-run/short-throw tests (Anvil, AS-SSD, ATTO, CrystalDiskMark, etc.) place a small test file on the device and mix the applied workload within that file, preventing any possibility of a steady-state condition from ever being reached on a tested SSD, 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, they 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.

We have also witnessed cases where the specific timing of the sleep/wake sequence may cause the Windows scheduler to assign the thread to a new core after each IO completes. 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.

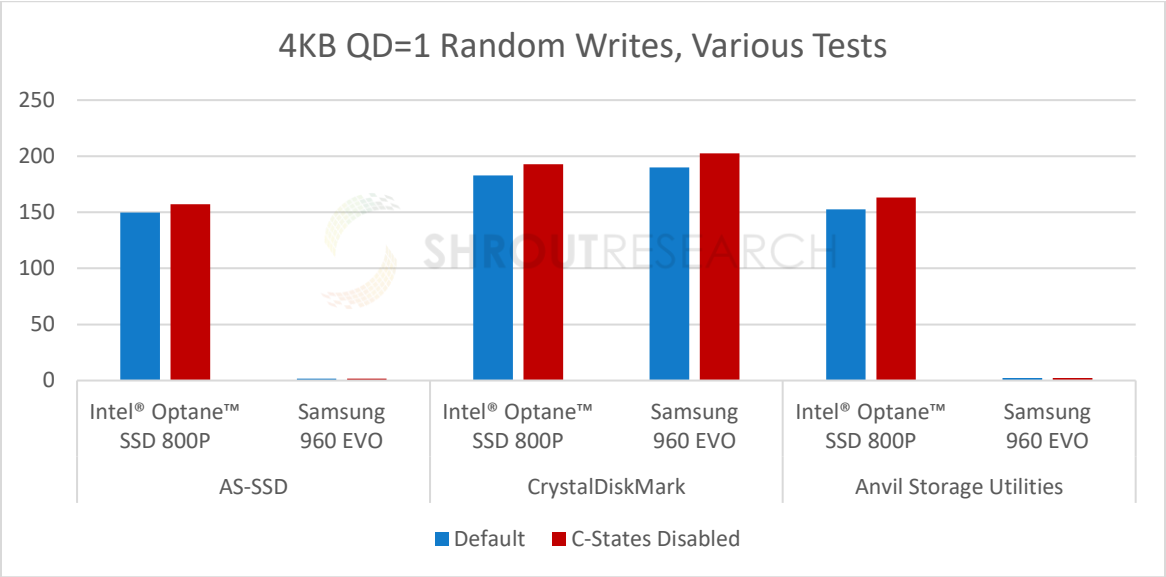


To demonstrate these points, we have sampled results from a few of the simple benchmark applications. Devices were sequentially filled to 90% capacity prior to test runs to minimize fresh-out-of-the-box conditions as much as practicable within the limitations of these tests. The following condition variants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

1. System at default state, running only the benchmark application.
2. Condition 1, but with hardware C-States disabled in system BIOS configuration.

In the above chart we note that for random reads, disabling C-States resulted in a slight uptick in measured performance. This is an apparent result of the QD1 workload being light enough for the CPU to enter a lower power state between request completions.

Also noted was a relative difference in performance recorded across the three benchmarks used for testing. CrystalDiskMark showed 800P performance significantly higher than the same metric recorded with AS-SSD and Anvil. Upon further investigation, this was due to differences in workload application by those other tests. While testing the 800P, CrystalDiskMark correctly maintained its workload on a single processor thread/core, while the IO call / waiting method of AS-SSD and Anvil allowed the Windows scheduler to park the core as the system was fully idle while the SSD serviced the IO. IO completion initiated an interrupt, which caused the scheduler to resume a different core, effectively forcing a context switch upon the completion of each IO and resulting in lower scores for those two tests. The process described does not match real-world usage since real applications would be actively processing the data received, keeping the core active on the current thread, and therefore suffering from significantly fewer performance robbing context switches.



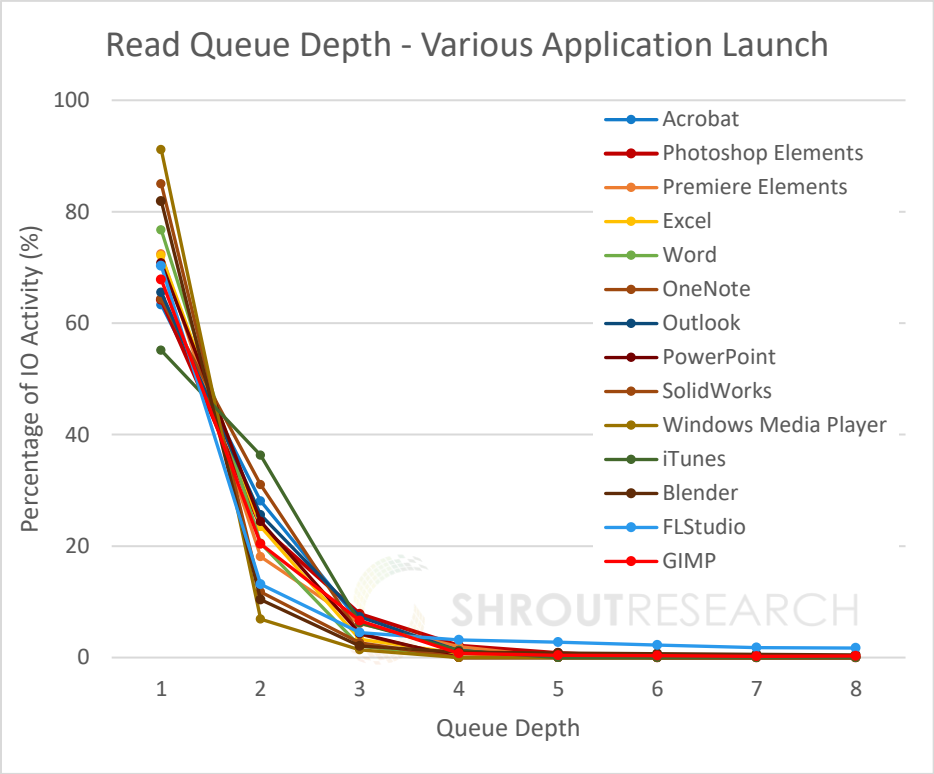
Looking at a random write workload, we have a few more interesting points. We can observe the same uptick in performance with C-States disabled. We also observe the same higher 800P performance seen with CrystalDiskMark due to AS-SSD and Anvil allowing the scheduler to bounce the application thread across all of the host system's cores.

That is where the similarities end. We expected the NAND SSD random write performance to meet or exceed that of the 800P in these tests, and while that was the case for CrystalDiskMark, the other two benchmarks employ IO API calls that do not mesh nicely with the default write cache policy set by the Microsoft Windows Inbox NVMe driver. The poor performance reported by those tests was not seen in actual system usage, but it severely impacted these benchmarks, demonstrating that their results do not match reality. Installing the Samsung NVMe driver would alleviate this issue, however the testing in this paper revolves around typical OEM configurations, which heavily rely on default in-box drivers for NVMe SSDs.

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.

The observed inconsistencies in the reported QD1 random write performance typifies the fact that legacy benchmark applications are not coded in a way that accurately reports the low queue depth performance of modern storage devices being tested on modern high core count systems.

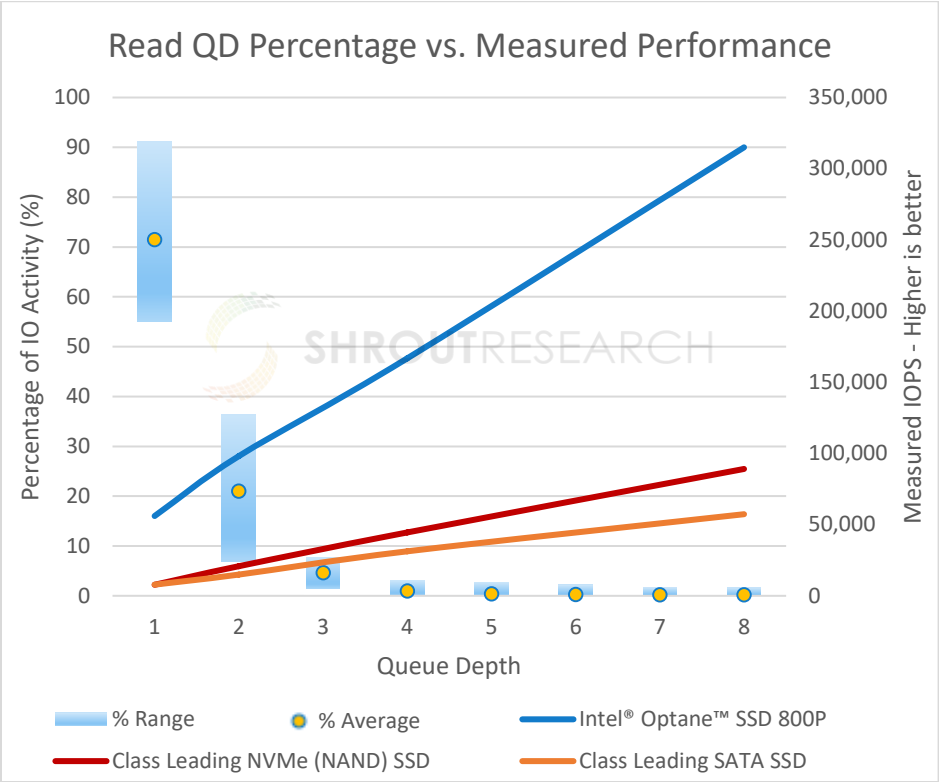
Focusing on Real-World Queue Depths



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 typically cite 'maximum IOPS' figures 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.

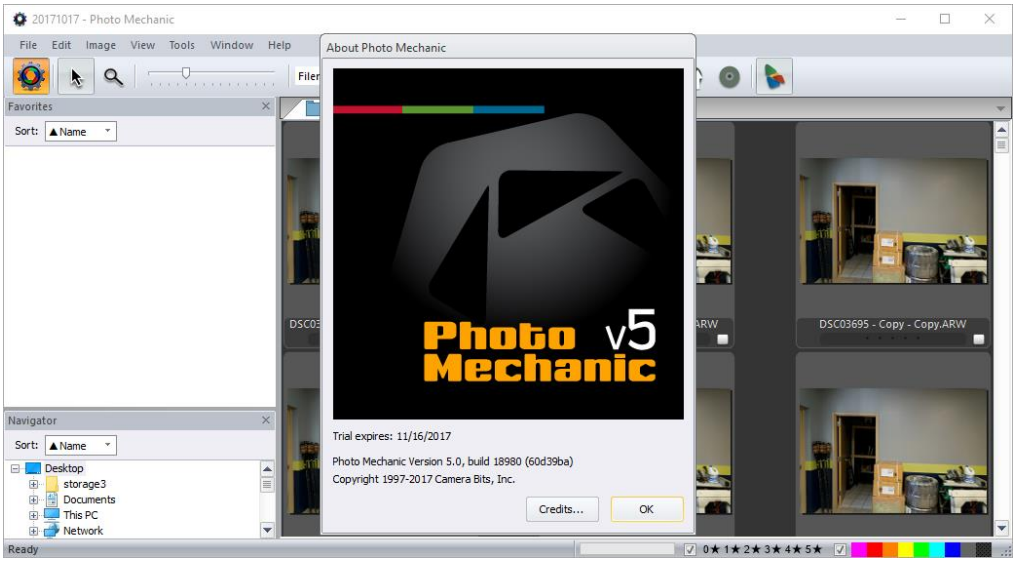
While simultaneous parallel tasks running on a single host may run at higher queue depths than seen above and may therefore see an additional benefit from lower latency storage in that it can more aggressively act to 'shallow the queue', modern client applications are not coded to take advantage of higher queue depths. Typical client workloads, focusing on singular applications in a given session, will rely heavily on the very low queue depth performance of the storage device. IO trace recordings taken across over a dozen commonly used application show that more than half of data requests occur individually and without any parallelization (QD=1), with the vast majority of IO servicing taking place with no greater than four simultaneous requests. Given that modern applications have yet to optimize for higher queue depth operation, an alternative optimization would be to introduce storage devices capable of lower latencies while operating at the lowest possible queue depths.



Here we have taken the percent activity ranges from the previous data set and overlaid the measured performance of the Intel® Optane™ SSD 800P, as well as class leading NVMe and SATA products. The SSD 800P outperforms all other options 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, 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 during actual use. Meanwhile, the Intel® Optane™ SSD 800P not only starts higher, but also climbs faster, reaching its maximum performance sooner than competing products.

Real-world Testing Examples



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

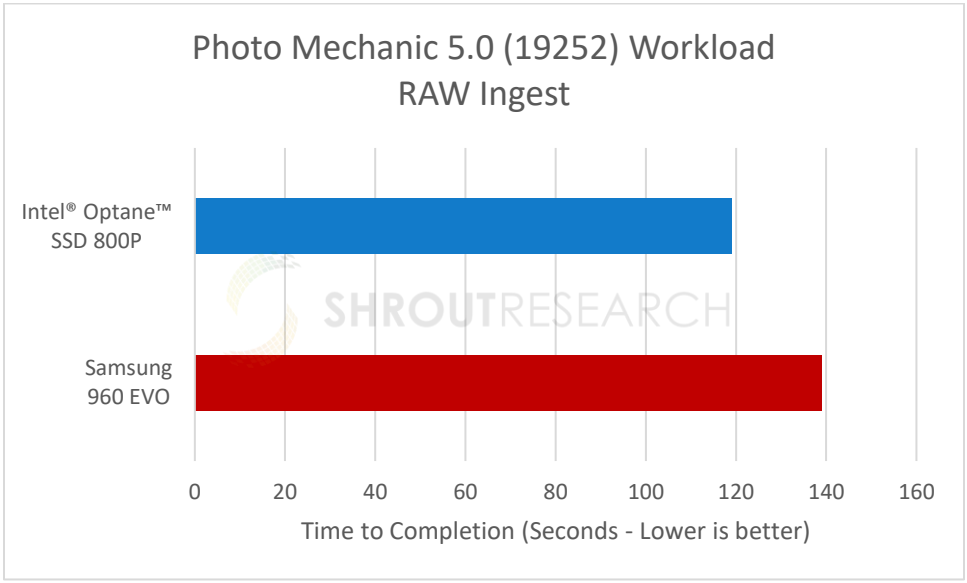
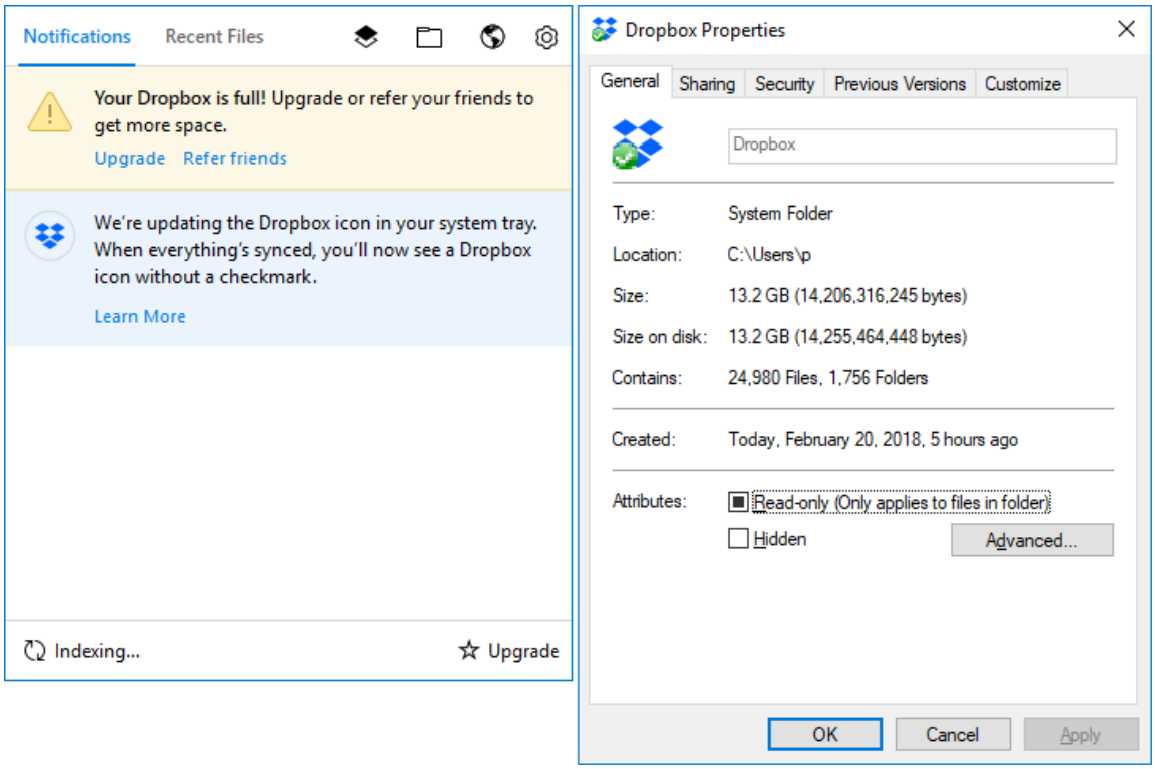
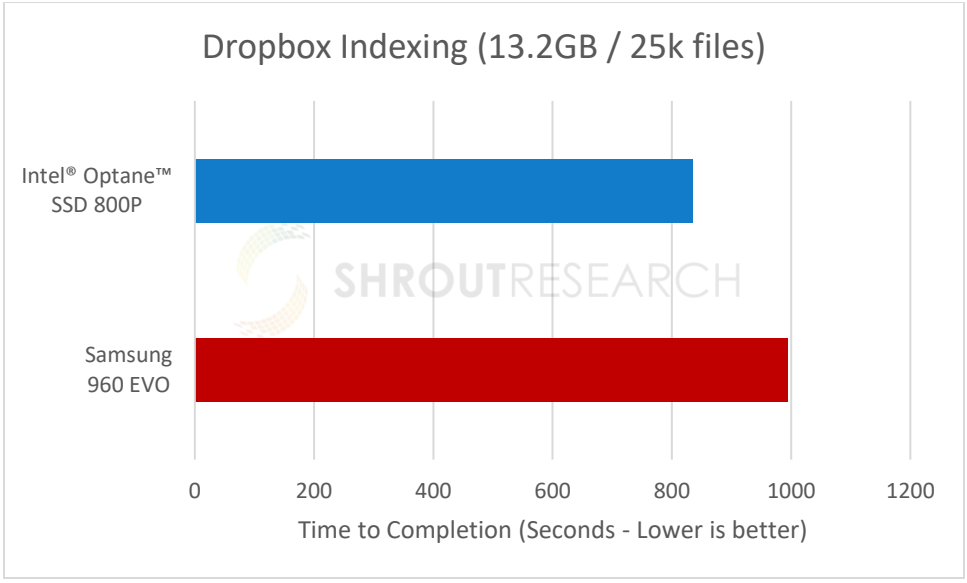


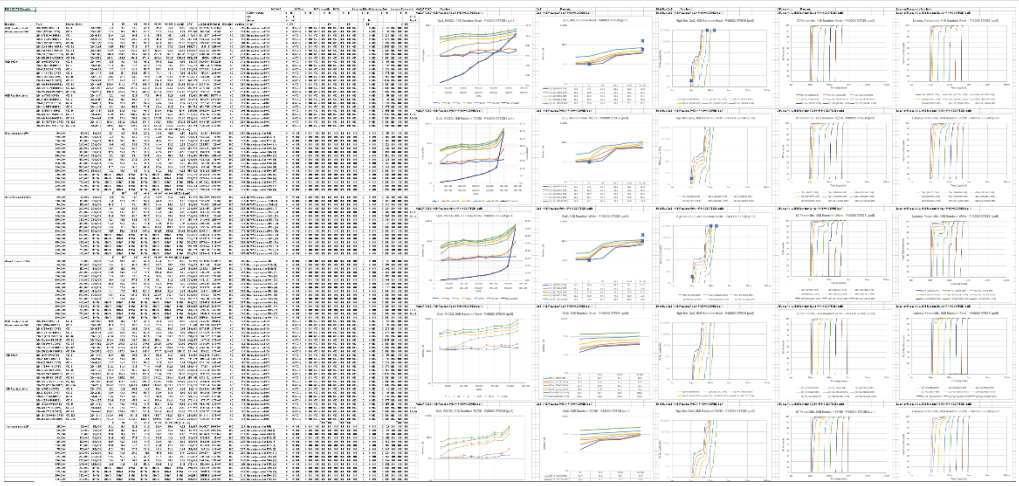
Photo ingest/import is one of the first steps of a professional photographer's workflow. Using the products as a working drive for photo processing, the 800P completed the photo import operation over 16% faster than the NAND SSD in the time it took to ingest 3,075 photos totaling 37.4GB. This time reduction 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.)



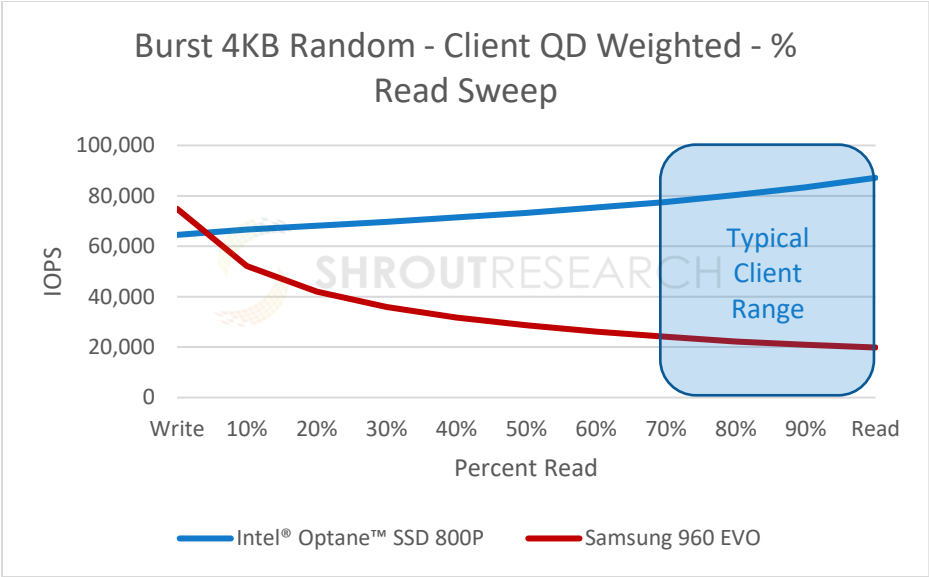
Dropbox is used in many desktop and mobile PC installations.



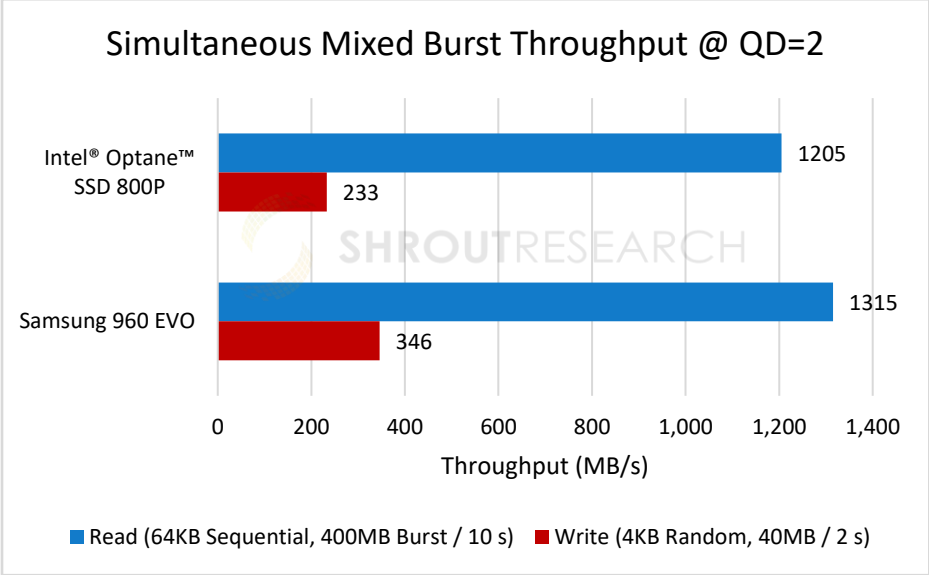
Dropbox occasionally re-indexes its synchronized files. This operation takes a relatively long time and is taxing on the host system drive as the indexed files must be read while simultaneously updating the file index database. Across multiple test runs, the 800P saw a 19% advantage. This time savings is significant especially on mobile platforms, as the system remains in a high-power state during the indexing operation.



The Shrout Research Storage Performance Analysis Software Suite (S-PASS) is an in-house developed tool set 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.

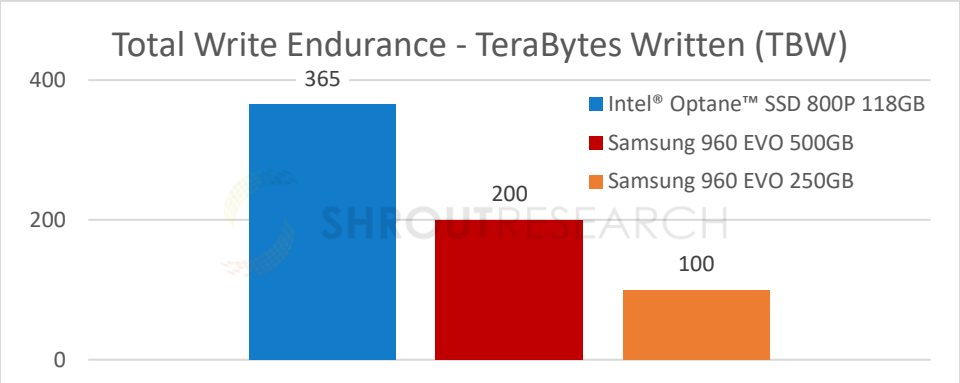


The above result is based on a synthesized workload applied to the storage devices at varying percentages of fill and Queue Depth. Client workloads typically fall in the center-right section of the percent read spread, where the 800P leads by up to 4.4x over the competing NAND product. Higher performing NAND-based SSDs can match 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.



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 800P did not win here, but it is highly competitive. This is surprising given the 800P's more limited PCIe 3.0 x2 interface combined with the sequential nature of the reads in this test. The 960 EVO shown above would normally read at greater than 2GB/s but is forced to slow considerably while under a relatively small background write pressure.

Endurance Implications



Despite coming in at smaller capacities than competing units, the 800P retains a superior endurance rating. When scaling this endurance rating proportionally to SSD capacity, the 118GB 800P comes in at 7.7x the 960 EVO 250GB. That number extends to 15.7x for the smaller 58GB capacity, which retains the same 200GB/day / 365TBW (5-year) endurance rating.

It is also worth noting that Intel® Optane™ SSDs do not suffer from the effects of write amplification. When receiving a write from the host, NAND-based SSDs must manipulate pages and erase the flash media in relatively large blocks which can be several megabytes in size, potentially wearing the flash faster than would have been caused by the host writes. Since 3D XPoint™ is a write-in-place media that does not require such coarse page/block operations, host writes are directly equal to media writes, which translates to even greater effective endurance with no negative write amplification effects.

The Changing Storage Landscape

The Intel® Optane™ SSD 800P shows itself to be a competitive product, offering impressive low-latency performance which gives it the edge in some of the more storage centric client workloads. Care must be taken to select the correct test system configuration and software in order to measure these potential gains in responsiveness observed at the lower queue depths in real-world usage.

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, others of which may be mentioned in this work.

The information and data presented in this document is 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 were used in the preparation of this paper:

	Product / Version
Motherboard	ASUS STRIX Z370-E Gaming
CPU	Intel® Core™ i7-8700K
RAM	16GB DDR4-2400
GPU	NVIDIA GeForce GTX 1080
OS	Windows 10 Pro RS3
Storage	Intel® Optane™ SSD 800P 118GB Samsung 960 EVO 250GB

Benchmark results were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown". Implementation of these updates may make these results inapplicable to your device or system. Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase.