TECHNOLOGY BRIEF

Data Center Memory and Storage
Intel® Optane™ Technology

Performance Where It Matters

Unlike NAND SSDs, Intel® Optane™ SSDs offer peak performance at queue depths relevant to real-world apps, not synthetic benchmarks.

Frank T. Hady, Ph.D.
Intel Fellow
Chief Optane Systems Architect
Intel Non-Volatile Memory Solutions Group

Memory and Storage Technical Series

The Direct Connection to Intel Fellows and Principal Engineers

This paper is part of a series designed to help system architects, engineers, and IT administrators understand the technological limitations of traditional memory and storage, how those limitations have led to performance and capacity gaps in the data center, and how Intel® Optane™ technology helps fill those gaps with an industry-disrupting architecture.

The series examines several topics that affect storage performance and capacity, including bandwidth, latency, queue depth, quality of service (QoS), and reliability.

You want a solid state drive (SSD) that will work the fastest for you and for your workload. Because you are reading this article, it's a good bet that you study SSD performance specifications when selecting an SSD for your system. When you read the specifications, you see throughput (also known as bandwidth) specified for both reads and writes. You also see the specified maximum accesses per second (commonly called input/output operations per second [IOPS]). It might surprise you to learn that these specifications assume highly idealized test scenarios. These scenarios might not—in fact likely don't—match the applications that you want to run quickly.

In this article, we explore the role that the number of outstanding accesses (commonly referred to as the queue depth [QD] of a workload) plays in SSD performance. We also examine the types of QDs commonly seen with real applications.

Simply put, most applications have relatively low QDs, and NAND SSDs need high QDs to deliver full performance. With their low latency, Intel® Optane™ SSDs deliver high performance at low QDs. So Intel Optane SSDs deliver high performance for a much wider set of applications.

The Prevalence of Low-QD Applications

QD is not something most people think about every day. An analogy can be used to illustrate QD, show its relationship to latency and throughput, and help explain why lower QDs matter most.

Imagine that your shed is on fire. You don't have a hose, but you have a bucket and a water faucet at the other side of a small field. So you turn on the faucet, fill the bucket, turn off the faucet, run across the field, and dump the water on the flames. Then you run back to the faucet and repeat the sequence.

In this example (Figure 1), the QD is one (QD=1) because there is only one person and one bucket. The throughput is equal to the average rate at which water is pulled from the faucet and applied to the fire (for example, 12 times per hour). Latency in this example is the time from the completion of the emptying of one bucket on the fire to the arrival of the next bucket to dump on the fire (for example, five minutes).

As you can see, there is a relationship between the latency and the throughput of water onto the fire. If the field is bigger, it takes longer to transit, so the latency for each bucket of water fetched will increase, and water throughput will drop.

QD = 1 (1 bucket)
Latency = 5 minutes
Throughput = 12 dumps/hour

Latency: Roundtrip time to dump water on the fire
Throughput: Rate water is applied to the fire over time

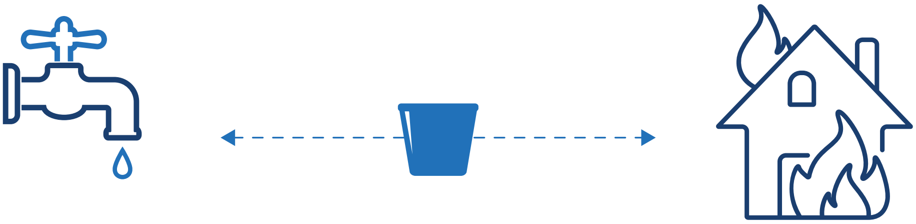


Figure 1. Throughput is determined by latency (roundtrip time) and QD (number of buckets)

If we could reduce the size of the field (Figure 2), moving the faucet closer to the shed, then we can get across the field faster and get more water to the fire more quickly. In this case, we reduce the latency, and, even with QD=1, we still increase the throughput and firefighting effectiveness.

QD = 1 (1 bucket)
Latency = 2.5 minutes
Throughput = 24 dumps/hour

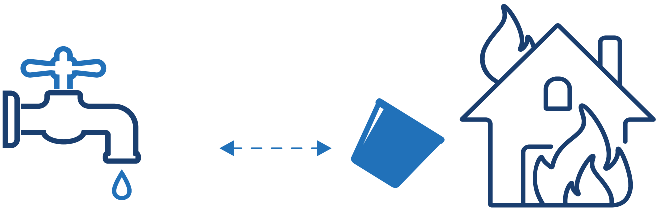


Figure 2. If you shorten the distance, latency is reduced and throughput is increased

Reducing latency sounds like magic. Is there another way? Let's take this example to QD=2. We need another bucket and a friend to help us. The two firefighters now pass each other in the field, one headed to the fire and one headed to the faucet. The latency hasn't changed because the field is the same size, but with QD=2, we now have twice the throughput: water is being applied to the fire faster (Figure 3).

QD = 2 (2 buckets)
Latency = 5 minutes
Throughput = 24 dumps/hour

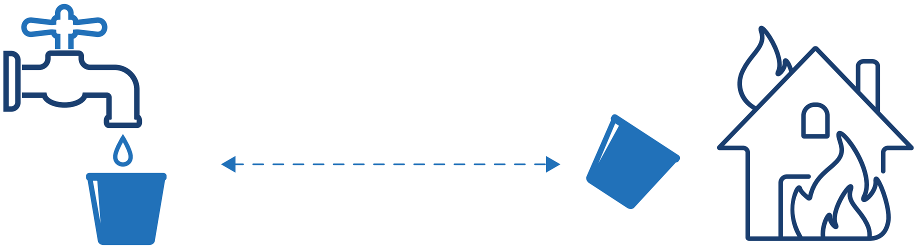


Figure 3. Another way to increase throughput is to increase QD

Until we run out of buckets and friends, we could continue to increase the throughput of water onto the fire by increasing the QD. As we increase the number of firefighters running across the field, we will start to run into each other (Figure 4). We've introduced inefficiency. Now, each added helper won't help as much as the first additional helper did. At some point, we will find that the faucet is never turned off, and someone is always filling a bucket. At this point, we will have reached the point of saturation (maximum throughput for the faucet), and adding more buckets (a higher QD) won't help.

QD = 4 (4 buckets)
Latency = 5 minutes
Theoretical Throughput = 48 dumps/hour
Actual Throughput: Much lower due to congestion

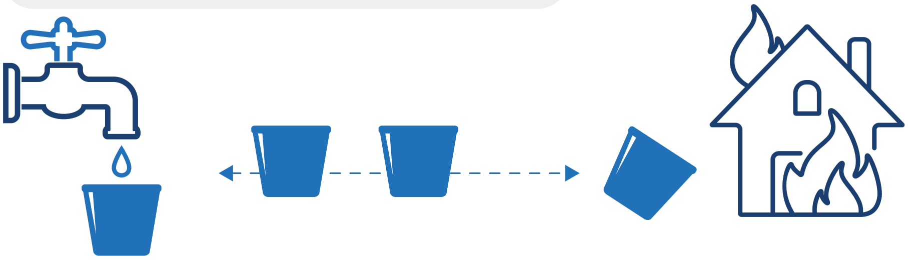


Figure 4. Eventually, increasing QD reaches a point of diminishing returns as saturation causes congestion

Storage systems work like the example above. The application running on the processor is the shed on fire—it needs buckets of data to move the computation forward. The application or operating system running on the processor makes individual requests of data from an SSD, and the returned data is used to move the computation forward. The number of data items that the application can request simultaneously (the QD, or the number of buckets) depends on the data parallelism of the computation, and on the capabilities of the application. The latency for each access depends on the latency of the SSD and of the system path to that SSD. Therefore, the throughput depends on both the application and the SSD used.

Application and Benchmark QD

SSD performance is usually measured with benchmarks like FIO (Linux) or CrystalDiskMark (Windows). These benchmarks are capable of high QDs. FIO is completely configurable in terms of QD—just specify the QD you want. FIO tests with QD equal to 128 or 256 are common when reporting SSD performance. CrystalDiskMark includes a test with 16 threads, each with a QD of 32, for a total QD of 512. Such high QDs make sense for fully exercising an SSD and for showing off the biggest possible performance in terms of IOPS and throughput.

However, those high-performance numbers—and their dependence upon high QDs—simply do not reflect the reality experienced daily in most data centers and on users' PCs. In real-world scenarios, a high QD is rarely achieved and maintained. Intel internal testing of real data center workloads has revealed that most applications are in the 1 to 9 QD range (Figure 5).[1] In fact, only an implementation of a transactional benchmark (such as TPC-H) reaches really large QDs.

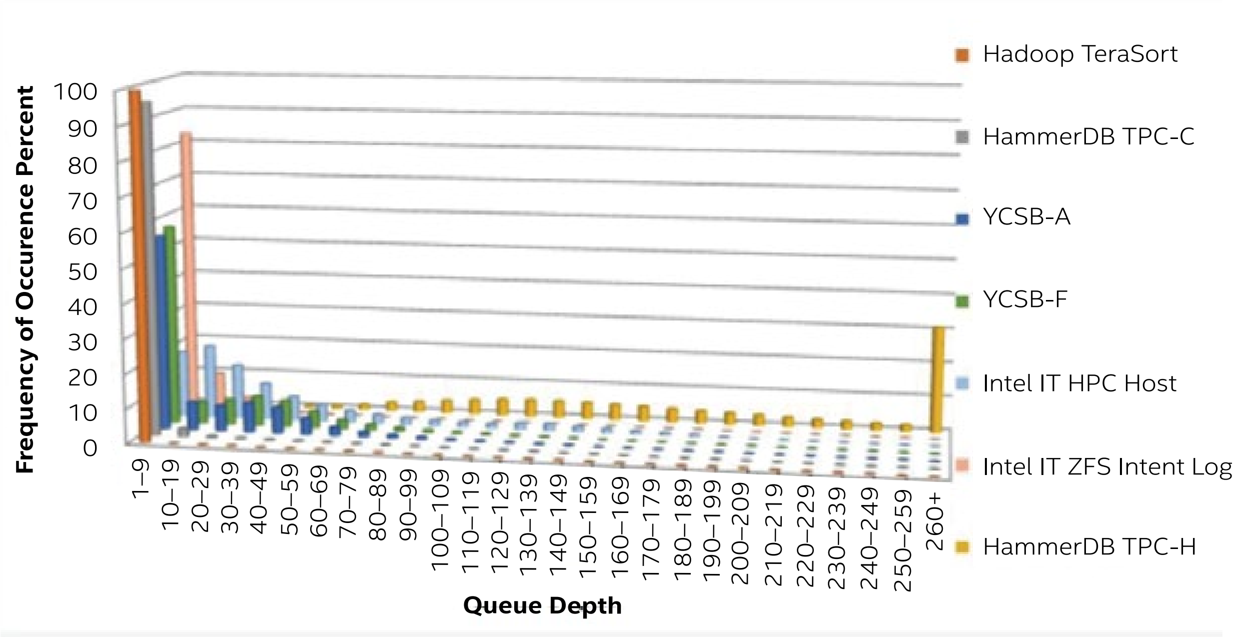


Figure 5. Many enterprise workloads occur at low QD ranges[1]

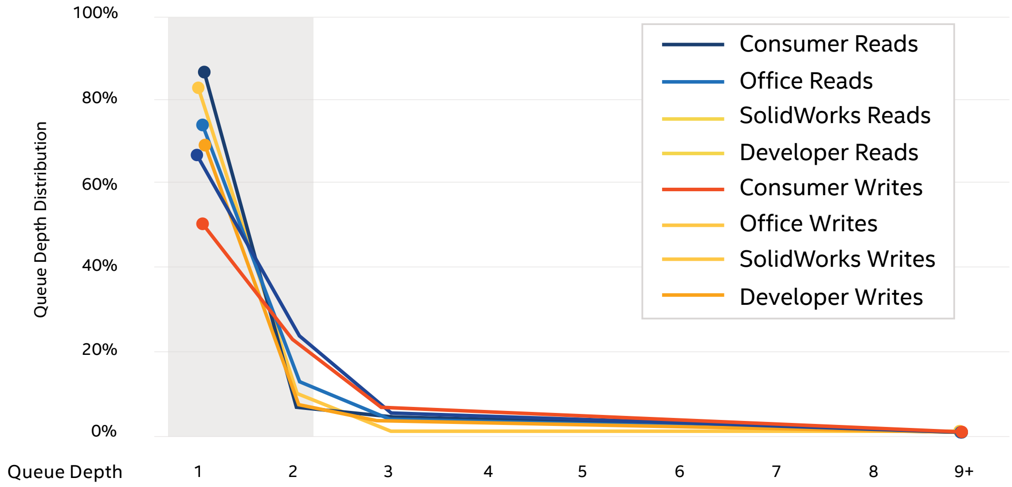


Figure 6. Various client workloads and their associated QDs; all of the measured workloads operate primarily at low QDs[2]

The situation is even more acute for PC applications. With our own measurements, we find that many desktop applications support a QD of just one, two, or four. As Figure 6 illustrates, real-world workloads for many of the most popular applications occur at less than QD=3.

Figures 5 and 6 vividly illustrate the disconnect between high QD measurements employed for SSD specification sheets, and the needs of real-world applications. SSD benchmarks provide lots of buckets to move data, while applications provide only a few. With this background, let's look at NAND and Intel Optane SSD performance versus QD.

NAND SSD Performance

It's no surprise that NAND SSDs are built from NAND memory. A single NAND SSD contains many NAND integrated circuits. The latency for a read of data from a NAND integrated circuit itself dominates SSD latency for all but less-frequent tail latencies.[3] Due to this NAND read latency, modern NAND SSDs typically have an idle average of about 80 microseconds (µs).[4] For a single 3 GHz CPU, that translates to 240,000 processor instructions—a big field to run across with a bucket.

Because of this relatively high latency, low QD performance is a challenge for a NAND SSD. A little math—4,096 bytes x (1/80 µs) = 50 MB/sec—shows us how slow the throughput would be. Of course, larger transfers (a bigger bucket) will increase this throughput. That is why you see SSD benchmarks use large transfers for throughput measurements. Note that only some applications can use large transfers.

A little more math—(1/80 µs) = 12K IOPS—shows how low the IOPS would be for QD=1. A higher QD number will increase this rate. That is why you see larger QD measurements for these values. Larger transfers will also increase the throughput number, which is why you will see high QD levels for IOPS measurements for SSDs.

There are lots of secondary impacts on NAND SSD performance that also drive the need for a higher QD to reach maximum NAND SSD performance. Only one is worth mentioning here: the Yahtzee effect, named by an Intel colleague, Knut Grimsrud. Each NAND integrated circuit (IC) can sustain only one read through its entire latency. Therefore, to get higher performance, the NAND SSD must have many ICs, and each read must exercise a different IC. But data is held on specific ICs, so incoming accesses may collide with a previous access for a specific IC and have to wait, even though other ICs are idle. It's as if we have multiple faucets, but each is slow, and each bucket can only be filled by a specific faucet. As the QD increases, the likelihood of collisions of reads for a single IC increases, causing performance to increase more slowly than QD. This is why SSD specification sheets include such large QDs to show high IOPS. Intel Optane SSDs do not suffer from the Yahtzee effect because of their more capable memory and SSD architecture.

How Intel Optane SSDs Outperform NAND SSDs in Real-World Data Center Operations

Unlike NAND SSDs, Intel Optane SSDs are designed to provide peak performance at real-world QDs, by using a revolutionary memory and SSD architecture that provides consistent low latency. The low latency of the Intel Optane memory media allows the SSD to achieve extremely low latencies (for an SSD) of around ~8 µs (a much smaller field to run across). Additionally, unlike NAND SSDs, the latency of Intel Optane SSDs is not dominated by memory latency and does not suffer from a Yahtzee effect. An Intel Optane SSD assembles even a single 4 KB read from multiple Intel Optane memory media ICs and those ICs are ready for another read very quickly. Intel Optane SSDs avoid the location and address-based collisions NAND SSDs exhibit. It is like Intel Optane SSDs use multiple faucets at once to fill a single bucket making them ready to fill the next bucket very quickly. This means the Intel Optane memory media is ready for another read in much less time than the NAND SSD, so it doesn't need input/output (I/O) parallelism to achieve high IOPS.

Stated simply, Intel Optane SSDs deliver peak performance at QDs that are consistent with the lower QDs at which most applications work. NAND SSDs typically require QD ranges of 128 or more to deliver peak performance while Intel Optane SSDs can reach full performance for much smaller QDs often seen with real applications (see Figure 7).[5] The chart also highlights the performance difference between a NAND SSD (Intel® SSD DC P4610) and an Intel Optane SSD (Intel® Optane™ SSD DC P4800X). The results show a real-world speed advantage for Intel Optane SSDs of four to five times the real-world relevant performance of the tested Intel NAND SSDs.

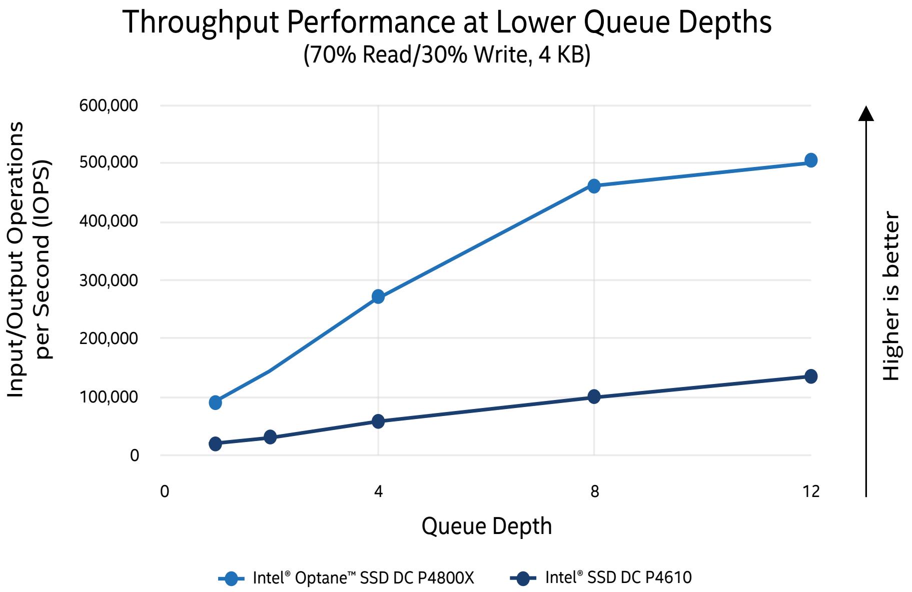


Figure 7. Intel® Optane™ SSDs deliver peak performance at lower QDs, where most applications work; NAND SSDs typically require QD ranges of 128 or more to deliver peak performance[6]

While it is an important chart, it only tells part of the story. Figure 8 shows the same workload, but it is plotted to show the operating point of the system in terms of both the throughput delivered (x-axis) and the resulting per-I/O read latency (y-axis). QD is included as the number on the NAND and Intel Optane SSD lines. Suppose we have an application capable of QD=4 operation. The Intel Optane SSD allows that application to operate at greater than 1.2 GB/s throughput with a latency per read-I/O of only about 10 µs. The NAND SSD, on the other hand, provides the application with an operating point of less than 0.3 GB/s and a latency per read-I/O of about 100 µs. Those are very different operating points that will, in turn, result in very different application performance.

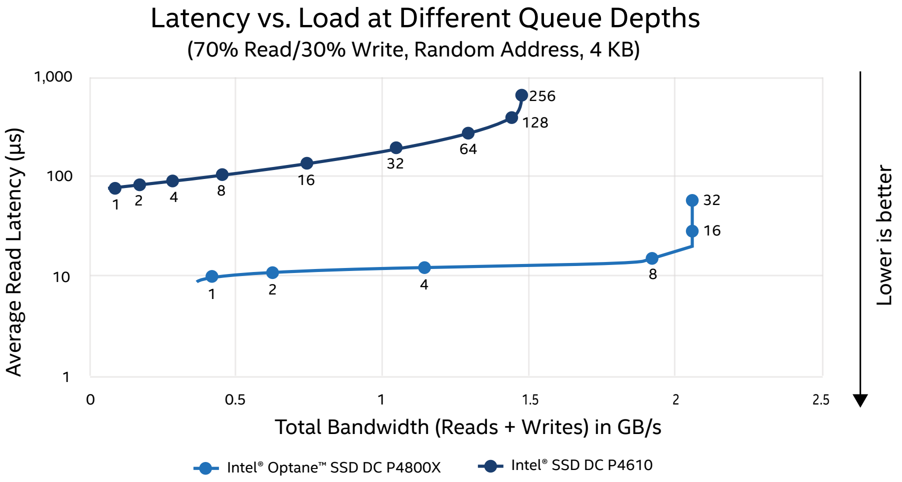


Figure 8. At lower QDs, Intel® Optane™ SSDs provide higher bandwidth and lower latency than NAND SSDs[6]

Also note in Figure 8 that the NAND SSD requires QDs of 128 or even 256 to reach full performance. Even if your application could get to that operating point, it would come at the cost of higher latency for reads. Now you can see why NAND SSD maximum performance is specified for such high QDs, and why you should ask about the latency for a read at that operating point. For this reason, several benchmarks, such as CrystalDiskMark, include QD=1 measurements as a part of their test suites. Intel Optane SSDs reach full performance for a QD of just over 8, and they maintain low read latency at that operating point. For realistic application QDs, an Intel Optane SSD delivers high throughput and simultaneously low latency. When it's time to put out the fire, I want an Intel Optane SSD in my system.

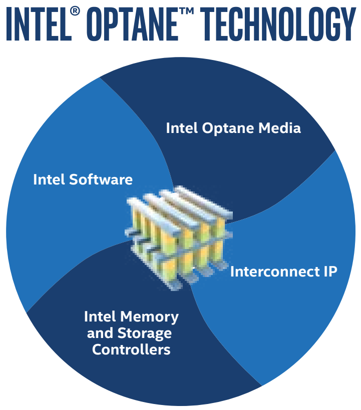
The Bonus Benefit of Intel Optane SSDs' Low-Latency Performance: Easier Code

As David Clark at MIT once put it, "Bandwidth problems can be cured with money. Latency problems are harder because the speed of light is fixed; you can't bribe God."[7] Clark was talking about networking, but the same is true for storage; low latency is powerful and has far ranging impact. We've noticed a recurring theme as we have worked with operating system and application developers to integrate low-latency Intel Optane SSDs into systems. These developers have incurred costs in the form of developer time, extra code, and extra compute cycles to overcome the high latency of storage. Over the years, developers of operating systems and key data center applications have expended great effort to increase application throughput in spite of the high latencies of NAND SSDs (and even hard disk drives [HDDs]). Significant code and complex heuristics have been developed to try to shorten the long wait times incurred when transferring data to and from storage. With Intel Optane SSDs, this extra code and extra developer time are no longer needed. The low latency provided by Intel Optane SSDs solves the root of the problem: quick access to data.

To illustrate this concept, let's look at a commercially important database benchmark, TPC-C. Another colleague at Intel, Jeff Smits, conducted extensive experiments comparing NAND SSD performance to Intel Optane SSD performance. TPC-C is all about throughput—transactions per second (TPS). Database implementations of TPC-C are heavily optimized at the code and system level. Jeff discovered that simply inserting Intel Optane SSDs into the system didn't deliver the full benefit. He had to reduce the number of outstanding transactions this heavily optimized system generated. When he did this, he saw a strong application-level performance gain. The system assumed high-latency storage, so it included complex code capable of generating lots of simultaneous transactions. Interestingly, dialing back the number of outstanding transactions even allowed CPU caches to function more effectively, because the working set size of the application was reduced. We've seen similar simplification-for-performance opportunities with operating system virtual memory paging.

Intel® Optane™ Technology Breaks through the NAND Barrier

Intel Optane technology is built on a revolutionary memory media that is byte addressable like DRAM, non-volatile like NAND, and has a read/write latency between the two. Intel Optane technology combines Intel Optane memory media with Intel controllers, software, and system interconnects that can be deployed as memory or storage.


INTEL® OPTANE™ TECHNOLOGY
Intel Optane Media
Intel Software
Interconnect IP
Intel Memory and Storage Controllers

Intel Fellow Frank Hady

Frank Hady is an Intel Fellow and the Chief Optane Systems Architect in Intel's Non-Volatile Memory Solutions Group (NSG). Frank leads research and definition of Intel® Optane™ technology products and their integration into the computing system. Frank has served as Intel's lead platform I/O architect, delivered research foundational to Intel® QuickAssist Technology, and driven significant platform performance advances. He has authored or co-authored more than 30 published papers on topics related to networking, storage, and I/O innovation and presents often on memory and storage. He holds more than 30 U.S. patents. Frank received his bachelor's and master's degrees in electrical engineering from the University of Virginia, and his Ph.D. in electrical engineering from the University of Maryland.

So the bonus benefit of Intel Optane SSDs is a reduction in code complexity and smaller working sets. From that reduced complexity, we see even more increases in system performance. If you are a developer, think about your application and how you could simplify it to achieve higher performance and productivity by using Intel Optane SSDs.

“Real-World” Performance Is Really All That Matters

The term “real-world” is sprinkled liberally throughout this paper. That’s as it should be. After all, published performance stats, no matter how breathtakingly impressive, are of little consequence if the same results cannot be achieved in actual practice. While NAND SSD performance stats might impress when browsing sales brochures, Intel Optane SSD performance will impress day-in and day-out in real-world data center operations and PC applications.

Learn More

Learn more about how Intel® Optane™ technology is disrupting the memory and storage hierarchy in the data center by exploring other papers in the Memory and Storage Technical Series.

To learn more about Intel® Optane™ SSDs, visit: **intel.com/content/www/us/en/products/memory-storage/solid-state-drives/data-center-ssds/optane-dc-ssd-series.html**

[1] Intel. “Performance Benchmarking for PCIe* and NVMe* Enterprise Solid-State Drives.” February 2015. intel.com/content/dam/www/public/us/en/documents/white-papers/performance-pcie-nvme-enterprise-ssds-white-paper.pdf.

[2] Source: Intel testing as of July 2018. System configuration: CPU: Intel® Core™ i7-8086K processor; BIOS version 9008 (x64) build date: 5/16/2018, EC version MBEC-Z370-0203, Intel Management Engine (Intel ME) firmware Ver11.8.50.3399; motherboard: ASUS Z370-A; operating system: Windows 10 RS4 1803; driver: Microsoft Inbox Driver; DRAM: 8 GB x 2 Corsair Vengeance LPX DDR4 (Model: CMK16GX4M2A2666C16R); 1 TB WD Blue 2.5" hard-disk drive (HDD) (model: WD10JPVX); 32 GB Intel Optane memory, 118 GB Intel Optane SSD 800P; 900P; SATA SSD: 512 GB Intel SSD 545s; NVM Express (NVMe) SSD: 512 GB Intel SSD 760p PCIe, M.2, NVMe SSD; all testing done internally by Intel.

[3] Intel. “Achieve Consistent Low Latency for Your Storage-Intensive Workloads.” December 2019. intel.com/content/www/us/en/architecture-and-technology/optane-technology/low-latency-for-storage-intensive-workloads-tech-brief.html.

[4] Based on Intel testing as of July 24, 2018. Average read latency measured at queue depth 1 during 4K random write workload. Measured using FIO 3.1 comparing Intel Reference Platform with 375 GB Intel Optane DC SSD P4800X and 1.6 TB Intel SSD DC P4600 compared to SSDs commercially available as of July 1, 2018.

[5] Intel-tested: 4K 70/30 read/write performance at low queue depth. Test and system configuration: CPU: Intel® Xeon® Gold 6140 processor FC-LGA14B (2.3 GHz, 24.75 MB, 140 W, 18 cores), CD8067303405200, CPU sockets: 2, RAM capacity: 32 GB, RAM model: DDR4, RAM stuffing: NA, DIMM slots populated: 2 slots, PCIe attach: CPU (not PCH lane attach), chipset: Intel C620 Series Chipset BIOS: SE5C620.86B.00.01.0013.030920180427, switch/retimer model/vendor: cable OCuLink 800 mm straight SFF-8611 to right angle SFF-8611 Intel AXXCBL800CVCR, OS: CentOS 7.5, kernel: 4.14.50 (LTS), FIO version: 3.5; NVMe driver: inbox, C-states: disabled, Intel Hyper-Threading Technology (Intel HT Technology): disabled, CPU governor (through OS): performance mode. Enhanced Intel SpeedStep® Technology (EIST), Intel Turbo Boost Technology: disabled, and P-states: enabled.

[6] Based on Intel testing as of November 15, 2018: Measured using FIO 3.1. Common configuration: Intel 2U server system, CentOS 7.5, kernel 4.17.6-1.el7.x86_64, 2 x Intel Xeon 6154 Gold processors at 3.0 GHz (18 cores), 256 GB DDR4 RAM at 2,666 MHz. Configuration: 375 GB Intel Optane SSD DC P4800X and 3.2 TB Intel SSD DC P4610. Intel microcode: 0x2000043; system BIOS: 00.01.0013; Intel Management Engine (Intel ME) firmware: 04.00.04.294; baseboard management controller (BMC) firmware: 1.43.91f76955; FRUSDR: 1.43.

[7] Attributed to David Clark, Massachusetts Institute of Technology (MIT).

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors.

Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit intel.com/benchmarks.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. **No product or component can be absolutely secure**.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

Printed in USA 0820/JR/PRW/PDF Please Recycle 343308-002US