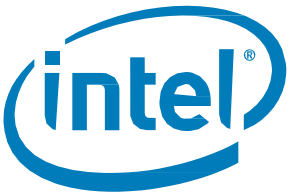


Intel® Optane™ Solid State Drive DC P4800X Series (Windows)

Performance Evaluation Guide

June 2020

Document Number: 336586-002US

Ordering Information

Contact your local Intel sales representative for ordering information.

Revision History

Revision Number	Description	Revision Date
001	• Initial release	October 2017
002	• Updated branding and legal disclaimers, changed disclosure to public	June 2020

Contents

Tables

Term and Acronyms

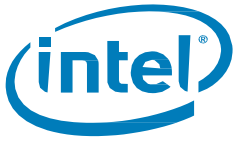
Table 1: Terms and Acronyms

Term	Definition
AHCI	Advanced Host Controller Interface
API	Application Programming Interface
ATA	Advanced Technology Attachment
DIPM	Device Initiated Power Management
GB	Gigabyte
HDD	Hard Disk Drive
KB	Kilobytes
I/O	Input/Output (the typical block used in specifications is 4kB)
IOPS	Input/Output Operations Per Second
MB	Megabytes
NCQ	Native Command Queuing
PCH	Platform Controller Hub
RAID	Redundant Array of Independent Disks
SATA	Serial Advanced Technology Attachment
SSD	Solid State Drive

Sources

Please see noted sources for additional information on topics discussed in this guide.

- https://sqlperformance.com/2015/08/io-subsystem/diskspd-test-storage
- https://gallery.technet.microsoft.com/DiskSpd-a-robust-storage-6cd2f223
- https://docs.microsoft.com/en-us/windows-hardware/drivers/kernel/avoid-polling-devices
- https://docs.microsoft.com/en-us/windows-hardware/drivers/kernel/general-i-o-programming-techniques



1 Introduction

This guide outlines the best know practices and configuration for evaluating Intel® Optane™ SSD DC P4800X performance in the Windows environment.

1.1 Overview

Intel® Optane™ SSD DC P4800X Series features Intel® Optane™ memory media. Intel Optane memory media is a class of memory technology that does not store data by trapping electrons in the memory cell, as NAND does; instead it utilizes the property-change of the memory material itself, to store the data. Intel Optane memory media, coupled with Intel-developed controller and firmware, takes SSD performance to the next level. For example, just two direct-attached Intel® Optane™ SSDs, deliver over 1 million I/O operations per second (IOPS). In addition, it achieves this performance at very low latencies. For example, at queue depth of one, 99 out of 100 read operations, of a 4KB-sized-aligned-workload, complete in single digit microseconds. With the reduction in latencies, Intel Optane SSD performance becomes sensitive to some of the variables that NAND SSDs are indifferent to. These variables include CPU speed, number of sockets, and relevant hardware and OS settings. This guide explores those variables in detail, enabling users to evaluate optimum SSD performance.

1.2 Typical I/O Execution

Let us first take a high-level look at a typical I/O access cycle. When an application requests data for processing, the kernel -- the core of the operating system, which controls system hardware -- checks for that data in main memory. If the required data does not exist in main memory, physical I/O is initiated through the device driver to the given device. The device processes the command and sends an interrupt at the completion of the command. The device driver acknowledges this interrupt command and the cycle ends.

Figure 1 shows the ways that a PCIe SSD can be installed in a system. Essentially, it is key to understand that end-to-end latency depends on many variables. For example, a faster CPU would allow the kernel to process and direct I/O requests faster. Similarly, less context-switching and better interrupt handling improves overall latency. Data transfer through various buses will contribute towards total latency. That is, if a PCIe SSD is attached through PCH, then the DMI bus would also incur latency. Finally, in multi-socket Intel server systems (Figure 2), if a process transitions from one socket to another, it uses QPI bus for data transfer which also adds to latency in the hundreds of nanoseconds. This evaluation guide will show you how to tune for latency improvement.

Figure 1: Available slots for NVMe SSD

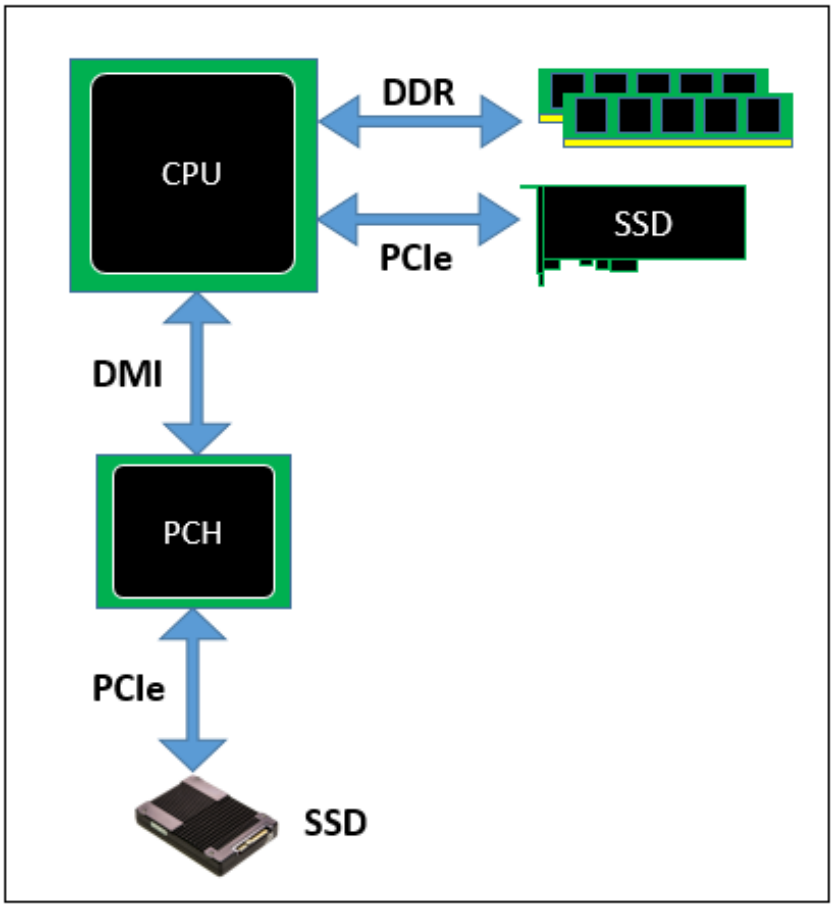


Figure 2: Intel Quick Path Interconnect

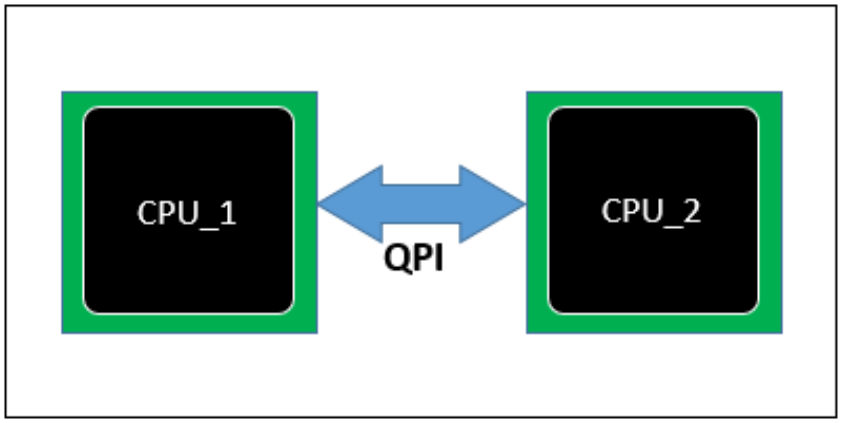
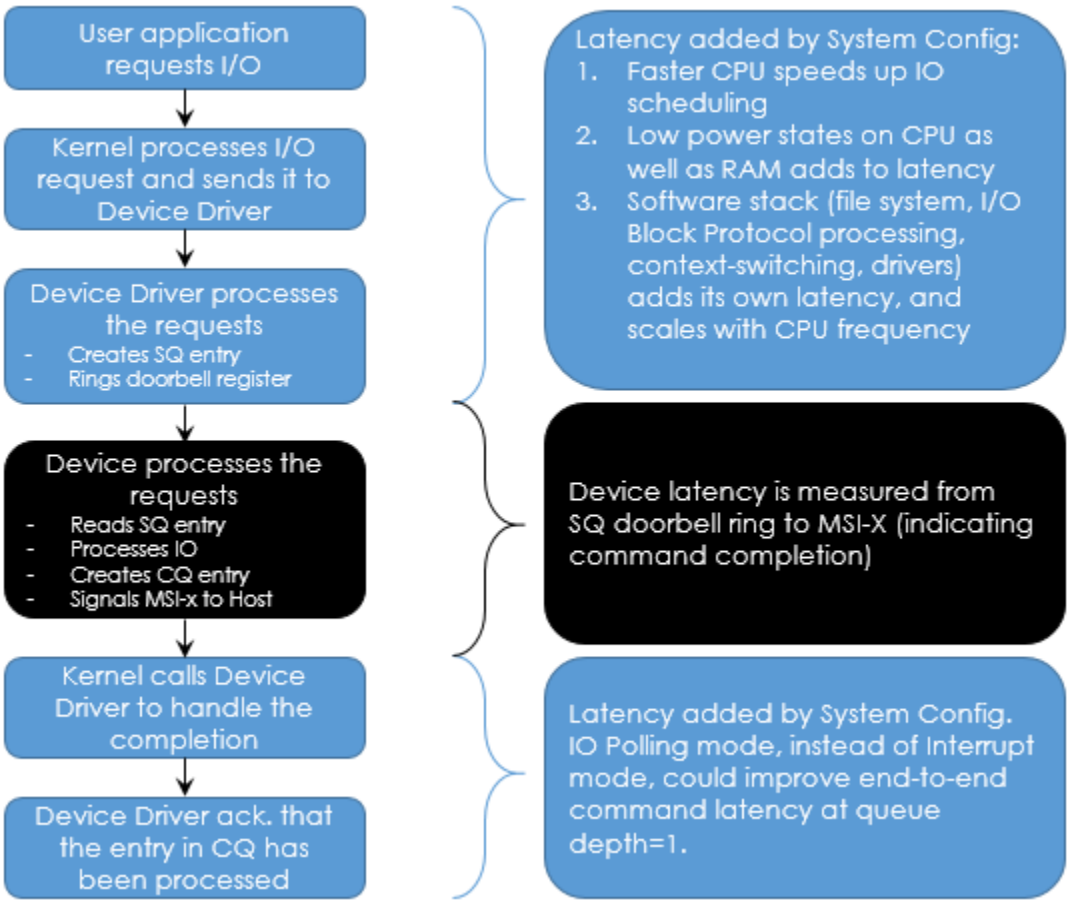


Figure 3 shows a typical command execution flow of a PCIe/NVMe SSD. Each stage incurs latency.
End-to-end command latency indicates the total latency from I/O request to command completion acknowledgement. For this evaluation guide, we will divide latencies into two broad categories: SSD Latency and System Latency.

Figure 3: Typical Flow of Command Execution on NVMe SSDs



SSD Latency: SSD latency is measured from the Submission Queue Tail Doorbell pointer update (submission command) to the MSI-X completion interrupt generated by the SSD (update of the Completion Queue Tail Doorbell pointer, indicating command completion). The Intel SSD controller has optimized this latency for 4KB aligned workloads through hardware acceleration. Note that sub-4KB or unaligned workloads, as well as any type of error correction, or any demands on firmware interception will increase SSD latency, when the SSD ramps the IO from the PCIe bus and to and from the submission and completions queues.

System Latency: This broad category includes end-to-end latency (from the originating user space application) for command completion, less SSD device latency, as captured above. This document is intended to guide you in tuning your system configuration to achieve low latency and properly evaluate SSD performance.

§

2 System Tuning For Optimum Performance

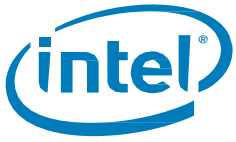
The Intel Optane SSD DC P4800X is designed for a PCIe Gen 3.0 x4 interface and NVMe 1.0 protocol. The tuning method shown here focuses on Windows Server 2016 OS, but the BIOS or Intel components can be generalized to any general-purpose OS. That said, implementation of actual settings will most likely vary with the OS and Intel is working on producing more OS-specific guides. Appendix-A includes performance results on Linux OS.

To benchmark Intel Optane SSD DC P4800X against its specification, use a system with a CPU frequency of 3.0GHz or higher. We actually lock the CPU frequency to maintain the best specification matching. This does not mean Turbo is not good for Intel Optane SSDs, but to get the expected best quality of service, it is recommended to set up your configuration as follows:

The platform settings in Table 1 ensure stable, consistent, and repeatable results during benchmarking:

Table 2: Platform Settings Summary

Type	Configuration	
CPU	Intel® Xeon® E5-2687W v4 @ 3.0 GHz (12 cores x 2)	
Motherboard	Intel® S2600WT	
Memory	32GB DDR4 @ 2133Mhz (32G X 1 DIMM)	
BIOS version	SE5C610.86B.01.01.0019	
BIOS configuration	Hyper threading:	Disabled
	EIST (Enhanced Intel Speed Step Technology):	Disabled
	Intel Turbo Mode:	Disabled
	PCIe ASPM (Active State Power Management):	Disabled
	C-States:	Disabled
	P-States:	Disabled
	Power Scheme:	Performance mode
Windows version	Windows Server 2016 R2 64-bit Server OS	
Intel NVMe driver version	Intel Driver	2.0.0.1015
	Inbox Driver	Windows Server 2016 1607 (Build 14393.0)
Diskspd version	v2.0.17 (run on raw disk)	



The OS settings in Table 2 ensure stable, consistent, and repeatable results during benchmarking.

Table 3: Best Known Operating System Configuration Summary

Setting	Status
Power Scheme:	Performance mode
Power Options > Turn off hard disk after	0 Minutes (Never)
Power Options > Internet Explorer > JavaScript Timer Frequency	Maximum Performance
Power Options > Desktop background settings > Slide show	Paused
Power Options > Wireless Adapter Settings > Power Saving Mode	Maximum Performance
Power Options > Sleep > Sleep after	0 Minutes (Never)
Power Options > Sleep > Allow wake timers	Enabled
Power Options > USB settings > USB selective suspend setting	Disabled
Power Options > Intel Graphics Power Plan	Maximum Performance
Power Options > Power buttons and lid > Power button action	Do nothing
Power Options > Power buttons and lid > Sleep button action	Do nothing
Power Options > PCI Express > Link Power Management	Off
Power Options > Processor power management > Minimum processor state	100%
Power Options > Processor power management > System cooling policy	Active
Power Options > Processor power management > Maximum processor state	100%
Power Options > Display > Turn off display after	0 Minutes (Never)
Power Options > Display > Display Brightness	100%
Power Options > Display > Dimmed display brightness	100%
Power Options > Display > Enable adaptive brightness	Off
Indexing Service	Disabled
Scheduled Defragmentation	Disabled
System Protection (System Restore)	Disabled
Paging File (Swap File)	Disabled
Prefetch and Superfetch	Disabled
Hibernate and Sleep Mode	Disabled

§

3 Results

The Intel Optane SSD DC P4800X is designed for a PCIe Gen 3.0 x4 interface and NVMe 1.0 protocol.

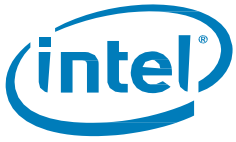
3.1 Diskspd Syntax for Performance Evaluation

Table 4: Diskspd Syntax for Performance Evaluation

Workloads	Diskspd v2.0.17 syntax (raw disk)
4KB Random Read QD32[1]	diskspd.exe -d600 -b4k -w0 -r -o4 -t8 -h -L #1
4KB Random Write QD32	diskspd.exe -d600 -b4k -w100 -r -o4 -t8 -h -L #1
4K Random70/30 Read/Write QD32	diskspd.exe -d600 -b4k -w30 -r -o4 -t8 -h -L #1
64KB Sequential Read QD32	diskspd.exe -d600 -b64k -w0 -r -o4 -t8 -h -L #1
64KB Sequential Write QD32	diskspd.exe -d600 -b64k -w100 -r -o4 -t8 -h -L #1
4KB Random Read QD1 Avg Latency	diskspd.exe -d600 -b4k -w0 -r -o1 -t1 -h -L #1
4KB Random Write QD1 Avg Latency	diskspd.exe -d600 -b4k -w100 -r -o1 -t1 -h -L #1
4K Random Reads , QD1	diskspd.exe -d600 -b4k -w0 -r -o1 -t1 -h -L #1
4K Random Reads , QD32	diskspd.exe -d600 -b4k -w0 -r -o4 -t8 -h -L #1
4K Random Writes, QD1	diskspd.exe -d600 -b4k -w100 -r -o1 -t1 -h -L #1
4K Random Writes,QD32	diskspd.exe -d600 -b4k -w100 -r -o4 -t8 -h -L #1

NOTE:

1. QD32 = Queue Depth 32 @ 8 threads x 4 outstanding I/O requests per thread



3.2 Performance Results[1]

Table 5: Performance Table

Workloads	Measurement	Windows Inbox NVMe Driver[2]	Intel NVMe Driver[3]
4KB Random Read QD32	IOPS	583558	584461
4KB Random Write QD32	IOPS	536782	554095
4K Random70/30 Read/Write QD32	IOPS	506194	506466
64KB Sequential Read QD32	GBs	2.5	2.54
64KB Sequential Write QD32	GBs	2.17	2.14
4KB Random Read QD1 Avg Latency	µs <=	12	12
4KB Random Write QD1 Avg Latency	µs <=	12	12
4K Random Reads , QD1	µs <=99%	15	13
	µs <=99.999%	58	47
4K Random Reads , QD32	µs <=99%	81	81
	µs <=99.999%	120	122
4K Random Writes, QD1	µs <=99%	14	14
	µs <=99.999%	73	55
4K Random Writes,QD32	µs <=99%	118	114
	µs <=99.999%	190	186

NOTES:

1. Performance measured by Intel using Diskspd with 8 workers and total Queue Depth of 32 in most case, except where specified. Diskspd is run on raw disk. System configuration: Intel® Xeon® CPU E5-2687W v4 @ 3.00GHz, Intel® R2208WT2YS-IDD Server Board, 32GB DRAM, Hyper-threading and CPU C-states disabled, P-states disabled, Windows Server 2016 R2 64bit Server, Diskspd v2.0.17. QD1= 1 thread x 1 I/O per thread; QD32= 8 threads x 4 I/O per thread
2. Windows Server 2016 1607 (Build 14393.0)
3. Intel Driver 2.0.0.1015. Intel® NVMe Windows NVMe driver can be downloaded at the link below. https://downloadcenter.intel.com/download/26906/NVMe-Drivers-for-Intel-SSDs

3.3 Run to Run Variation

We noticed large run to run variation in Quality of Service at higher Queue Depth for a Windows system. This observation seems Windows specific as we didn't observe such behavior on a Linux system, see Appendix A for details. We are in process investigating the root cause of the issue. Any new learnings will be captured in the future revision of this guide.

§

4 Exhibits

4.1 Exhibit-1: Pre-conditioning

NAND devices are required to be written fully before you will achieve stable performance. Background Data Refresh is a "periodic refresh" of the data stored on the drive to prevent the cell voltage levels from shifting outside an acceptable range. When the drive detects that data stored on it has reached a certain age, it will rewrite that data in the background.

Intel Optane memory media is a completely new technology. Intel Optane memory media also need periodic refreshing. Intel Optane drives have a background refresh policy that ensures that the data in the media remains refreshed even when it is not accessed.

If an Intel Optane SSD DC P4800X remains de-energized for an extended period of time, faster background data refresh will be invoked by firmware design, impacting performance.

The length and temperature of the de-energized state will impact the likelihood of triggering faster background data refresh. We designed current length of faster background refresh is currently set ~3 hours.

With this, Intel recommends minimum power-on time of 3 hours before running performance evaluations.

4.2 Exhibit-2: Why use the Intel NVMe driver for Windows OS

There are three distinct areas that allow Intel to provide a better experience with our Intel device drivers for the Windows operating system. Intel's device drivers are architected, designed and implemented to integrate seamlessly with other Intel SSD software products such as Intel® SSD Toolbox and Intel® Data Center Tool which are unique products for management and monitoring of the Intel SSD's. Second, Intel device driver debug efforts will be more responsive and customer or OEM specific support requests will be handled in time critical manners since Intel's device drivers are not tied to any operating system release cadence. Finally, any new Intel SSD features are developed jointly across the I/O stack from hardware, firmware, and host software (device driver and tools) which allows Intel to release device drivers with focused development and validation efforts and on a cadence that is specific to Intel's SSD's, instead of infrequent operating system releases.

4.3 Exhibit-3: I/O Latency in the Host Operating System

Linux has a very fast block mode storage stack and if you read the Linux version of Intel Optane SSD DC P4800X Series Performance Evaluation Guide you will see slightly faster numbers than what is exhibited as default behavior for "off the shelf" Windows Operating system, for either the Windows in-box driver or the Windows NVMe Driver for Intel SSDs (v2.0.0.1015) from the Intel Download Center. Data transfer speed -- to and from storage to main memory -- depends on many variables. These include device driver, I/O scheduler, and file system to name a few. When an application requests data for processing, the kernel -- core of the operating system, which controls system hardware -- checks for that data in main memory. If the required data does not exist in main memory, physical I/O is initiated through the device driver to the given device. The device processes the command and sends an interrupt at the completion of the command. The device driver acknowledges this interrupt command and the cycle ends. To summarize, end-to-end latency depends not only on the device, but also on system configuration and how the kernel is designed for use with the hardware. A faster CPU enables kernel to process and direct I/O requests faster. Similarly, less context-switching and better interrupt handling improves overall latency. With the reduction in device latencies, Intel Optane SSD performance becomes sensitive to some of the variables that NAND SSDs are indifferent to, including the architecture of the Operating System and the device driver models that are available for those Operating Systems. In Linux there are a number of driver types to choose from including polling mode drivers that focus on different needs and ways of processing I/O.

§

Appendix A Performance Measurement on Intel Server System with Linux OS

Table 6: Linux System Configuration

Type	Configuration
CPU	Intel® Xeon® E5-2687W v4 @ 3.0 GHz (12 cores x 2)
Motherboard	Intel® S2600WT
Memory	32GB DDR4 @ 2133Mhz (32G X 1 DIMM)
BIOS version	SE5C610.86B.01.01.0019
BIOS configuration	Hyper-threading disabled, CPU C-state disabled, P states disabled
Linux version	CentOS- 7.3
Kernel version	4.10.8
FIO version	2.18

Table 7: Linux Performance Table

Settings	4KB Random Read, Queue Depth =1[1]					4KB Random Write, Queue Depth =16				
	Latency (µs)				IOPS	Latency (µs)				IOPS
	Avg	99%	99.999%	Max		Avg	99%	99.999%	99.99999%	
Stock OS Results	8.43	17	45	139	115k	25.37	65	334	4576	550k
Governor set to Performance	8.16	16	45	77	119k	25.61	64	119	1192	555k
IRQ Balancing Service Turned off	8.01	16	39	69	120k	25.69	65	237	2416	554k
SMP Affinity Set	7.96	15	44	141	122k	25.81	64	123	221	553k
Turbo off 2.3 GHz	9.2	11	36	87	107K	24.44	64	255	2576	556K
Turbo off 3 GHz	7.57	9	53	135	126k	24.82	63	129	1416	556k

Note:

Max values can be affected by system services (example: udevd system service) and are not deterministic or statistically significant based on configuration variances. Therefore 99,99999% data is shown.

[1]QD1 performance uses ioengine=pvsync2 w/ hipri, kernel 4.8.6

Performance measured by Intel using FIO rev 2.18, with 4 workers and total Queue Depth of 16 in most case, except where specified. System configuration: Intel® Xeon® CPU E5-2687W v4 @ 3.00GHz, Intel® R2208WT2YS-IDD Server Board, 32GB DRAM, Hyper-threading and CPU C-states disabled, P-states disabled, CentOS 64bit Server, Linux Kernel 4.8 with polling enabled.

§