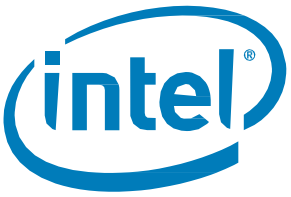


Intel® Optane™ Solid State Drive DC P4800X (VMware)

Performance Evaluation Guide

June 2020

Order Number: 336883-002US

Ordering Information

Contact your local Intel sales representative for ordering information.

Revision History

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

Contents

Tables

Figures

Terms and Definitions

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
OS	Operating System
PCH	Platform Controller Hub
RAID	Redundant Array of Independent Disks
SATA	Serial Advanced Technology Attachment
SSD	Solid State Drive
VMDK	Virtual Machine Disk
VMFS	Virtual Memory File System

Sources

Additional information for topics discussed in this guide:

- https://www.vmware.com/techpapers/2017/Perf_Best_Practices_vSphere65.html (Tips on maximizing performance)
- https://kb.vmware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=2001003 (Troubleshooting ESXi virtual machine performance issues)
- https://blogs.vmware.com/vsphere/launch (Many resources exist at this landing page)
- http://iometer.org/ (Download , installation and usage of IOMeter)

1 Introduction

This guide outlines the best known practices and configurations for evaluating Intel® Optane™ Solid State Drive Data Center P4800X performance in a VMware ESXi environment.

1.1 Overview

Intel® Optane™ SSD DC P4800X 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 more than 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 toward 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 SSDs

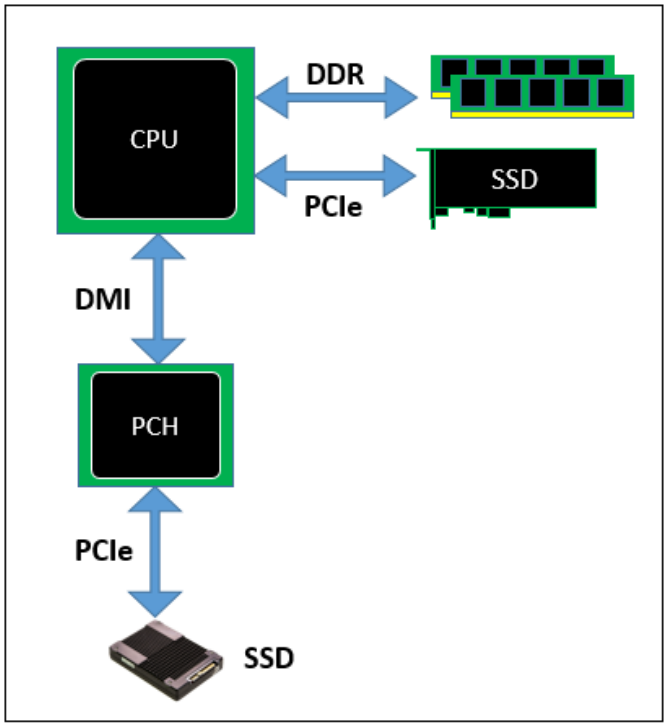
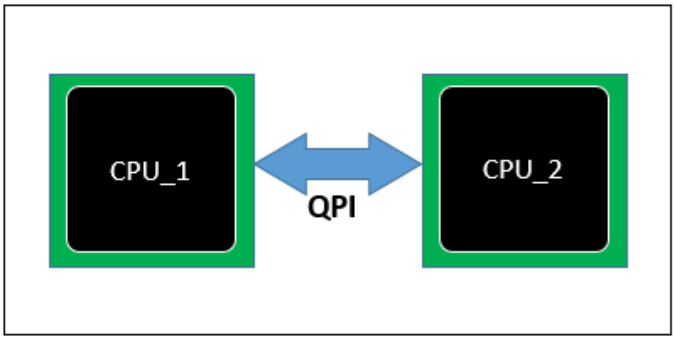


Figure 2: Quick Path Interconnect



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). 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

Intel Optane SSD P4800X is designed for a PCIe Gen 3.0 x4 interface and NVMe 1.0c protocol. Tuning method below focuses on VMware ESXi, 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 committed to more OS-specific guides.

Most of the recommendations in the table below would be enabled by default, but it's always good to confirm if the settings below are in an enable or disable state.

Table 1: Best Known System Configuration Summary

Hyper Threading	Enabled
EIST (Enhanced Intel® SpeedStep® Technology)	Enabled
Intel® Turbo Mode	Enabled
C-States[1]	Enabled
P-States[1]	Enabled
CPU Power and Performance Policy (through the BIOS)	Performance mode
Workload Configuration	I/O sensitive
Intel® Virtualization Technology	Enabled
ESXi Host – iSCSI.Max.IoSizeKB[2]	512

NOTES:

1. Based on Intel BIOS. C-States, and P-States are enabled through the BIOS, Power Management settings on Intel servers (2600WT).
2. Based on ESXi 6.5, ISCSI.Max.IoSizeKB is changed from default value 128 to maximum value 512 through the web client (Host -> Manage -> System tab -> Advanced Settings –> ISCSI.Max.IoSizeKB)

Best Known Configuration Details

Place the NVMe-based Intel Optane SSD in a PCIe slot that is directly connected to the CPU, and not the platform controller hub (PCH), as PCH would incur additional latency.

2.1 Prerequisites

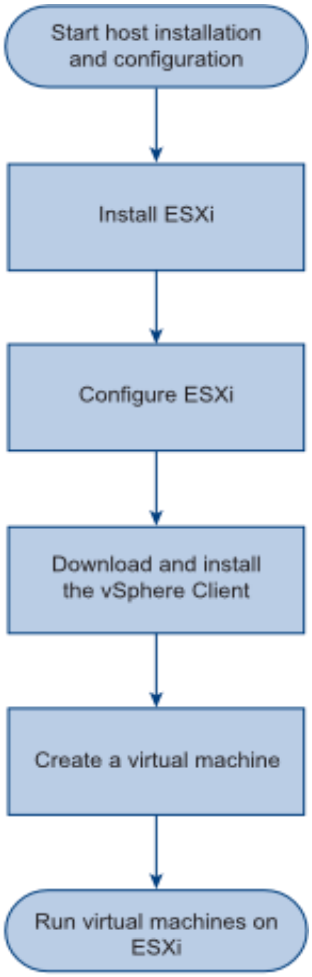
Version: ESXi 6.5 G.A. and later

Check for hardware configuration of the system on which ESXi is installed. The following are criteria to be checked before installation.

Supported server platform. For a list of supported platforms, see the VMware Compatibility Guide at http://www.vmware.com/resources/compatibility.

2.2 Installation Steps for ESXi Host

Figure 3: Workflow for ESXi Installation (Source: VMware)



2.2.1 Install ESXi

- Load ESXi installer media and set BIOS to boot from the installer media.
- Select ESXi installer media in boot menu and press enter.
- To accept the end user license agreement, press F11.
- Select the drive on which ESXi is to be installed and press Enter. Please choose the option of clean installation if you are installing ESXi on a disk that contains a previous installation of ESXi.
- To select the default keyboard, press enter.
- Type username, the host password, and retype the password to confirm it.
- Press F11 to begin the installation.
- After installation is completed, remove installation media and press enter to restart the host.

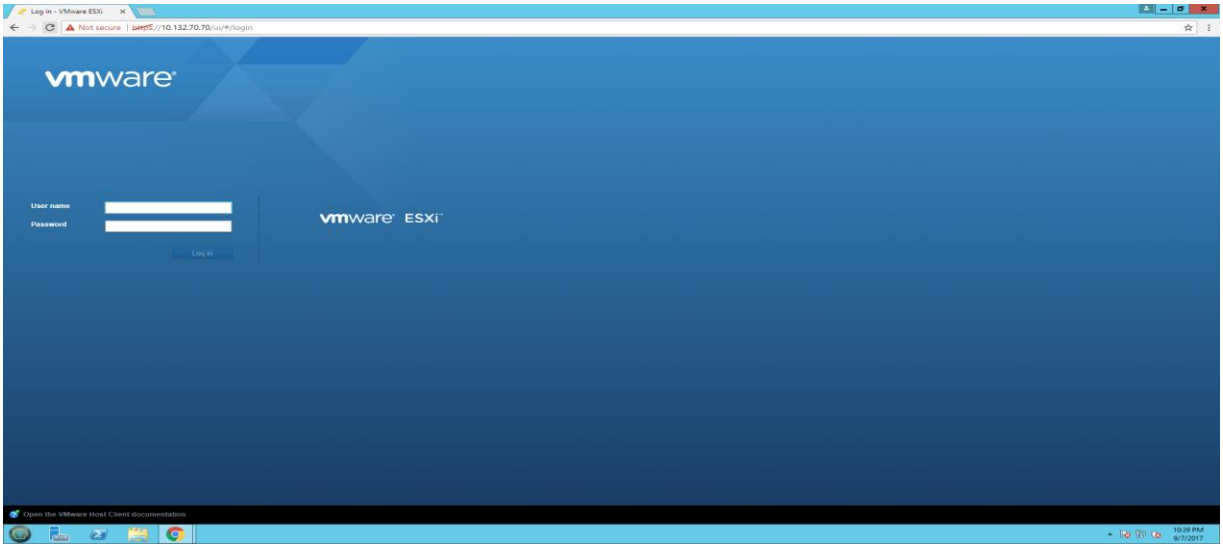
2.2.2 Configuring the ESXi Host

- Start the direct console on host.
- Press F2 to access configuration screen.
- Type the login credentials.
- Use keyboard and select Configure Management Network and press enter.
- Select network adapter option and press enter. Within the menu, choose the connected network adapter. Make changes if necessary. Press enter to accept network adapter changes and to go to previous menu.
- Select IPv4 configuration and press enter. Choose required IPv4 configuration (Static or DHCP). Press enter to accept IP configuration changes and to go to previous menu.
- Select DNS configuration and press enter. Type the primary server's DNS address, fully qualified domain name for host name field. Press enter to accept the changes and to go to previous menu.
- Press Esc to save changes and move to previous menu.

2.2.3 Logging to vSphere Web Client

- Login to the system that has access to the host network.
- Open a web browser and type in the IP address of the ESXi host.
- Enter the login credentials for the host and login to the vSphere Web Client.

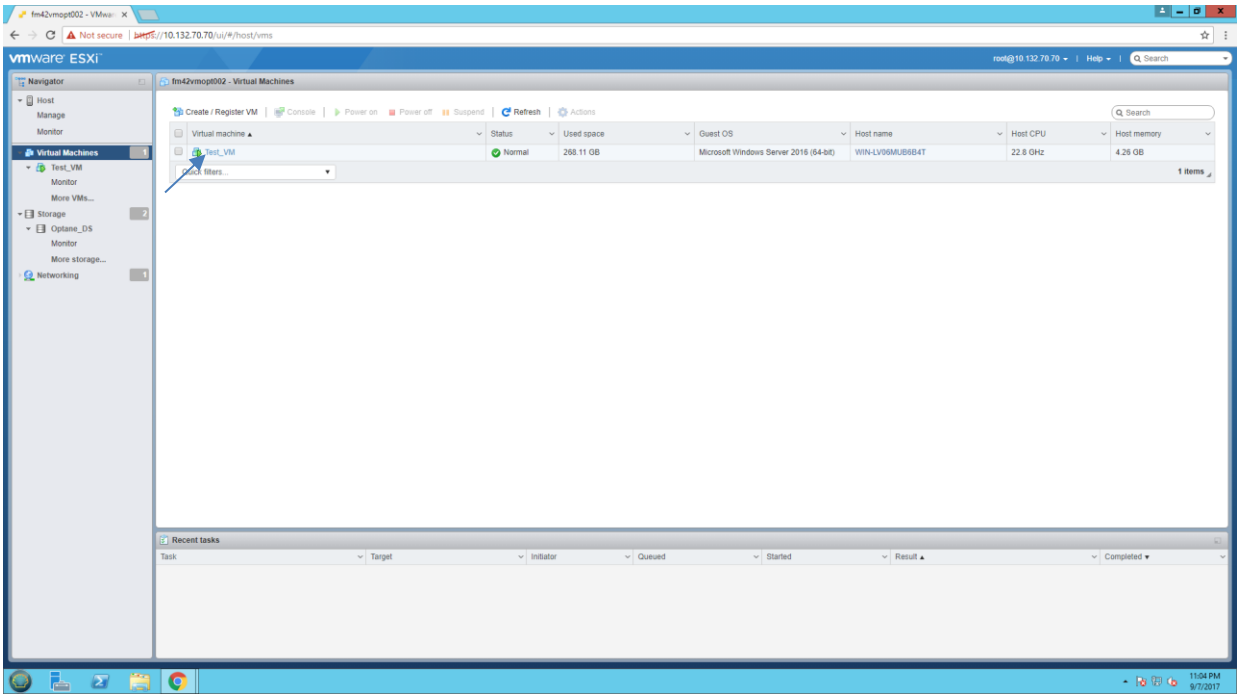
Figure 4: Login Screen of vSphere Web Client



2.2.4 Create a Virtual Machine

- After logging in to the vSphere Web Client, choose the Storage option from the left panel.
- If there is no default local data store listed, add a local data store by clicking New Datastore option.
- In the window that follows, choose the required drive from the data store and name it. Choose the partition method and finish. Now a local data store would have been created.
- Upload ISO file of Guest OS to the local data store.
- Click Virtual Machines in the left panel and select Create/Register VM option.
- Name the virtual machine and choose the operating system family and version from the list.
- Choose the local data store in the Storage option wizard.
- Customize the hardware settings as follows:
  1. Number of vCPUs = 8 (CPU cores =8, sockets=2, cores per socket=4)
  2. Memory capacity = 8GB
  3. Boot disk capacity = 100 GB (Location = Choose local data store, Thin provisioned = no, Controller = SATA controller, Mode= Dependent) Note: Boot drive capacity can be changed based on capacity of data store.
  4. CD/DVD drive = Upload the ISO image file of Guest OS from the local data store.
- Check the configurations and choose finish. A new virtual machine would be created with hard disk: 0 as boot drive backed by local data store.

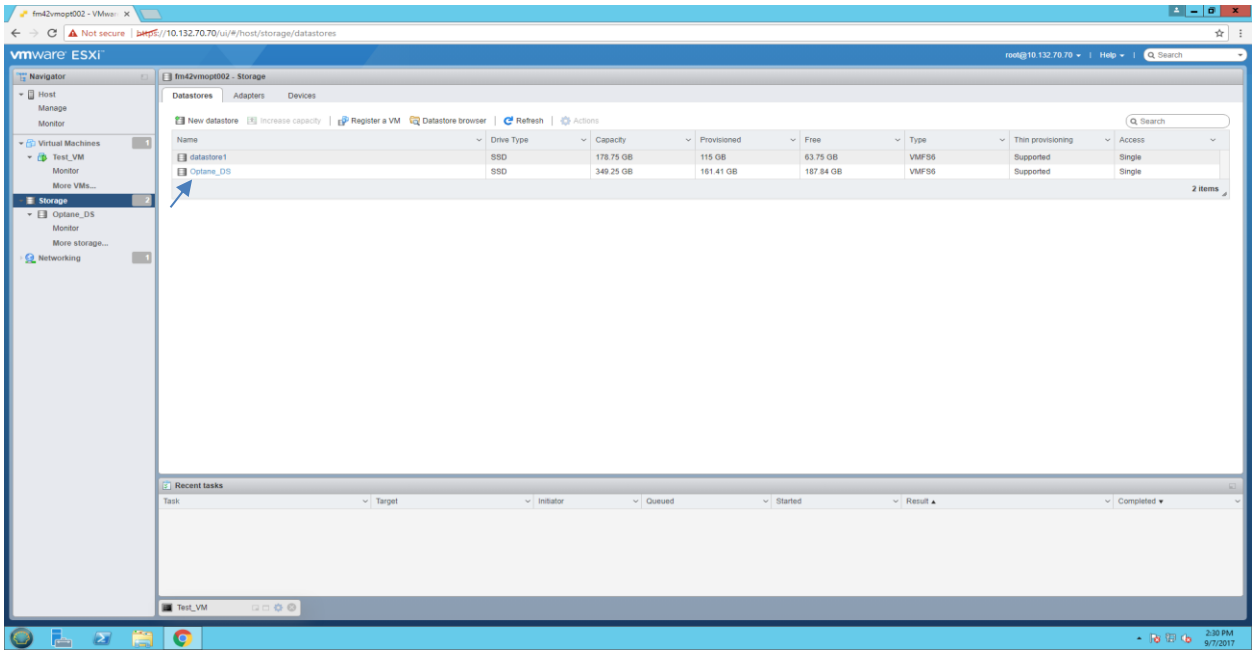
Figure 5: Virtual Machine in ESXi Host



2.3 Setting up Intel® Optane™ SSD P4800X to Test the Virtual Machine

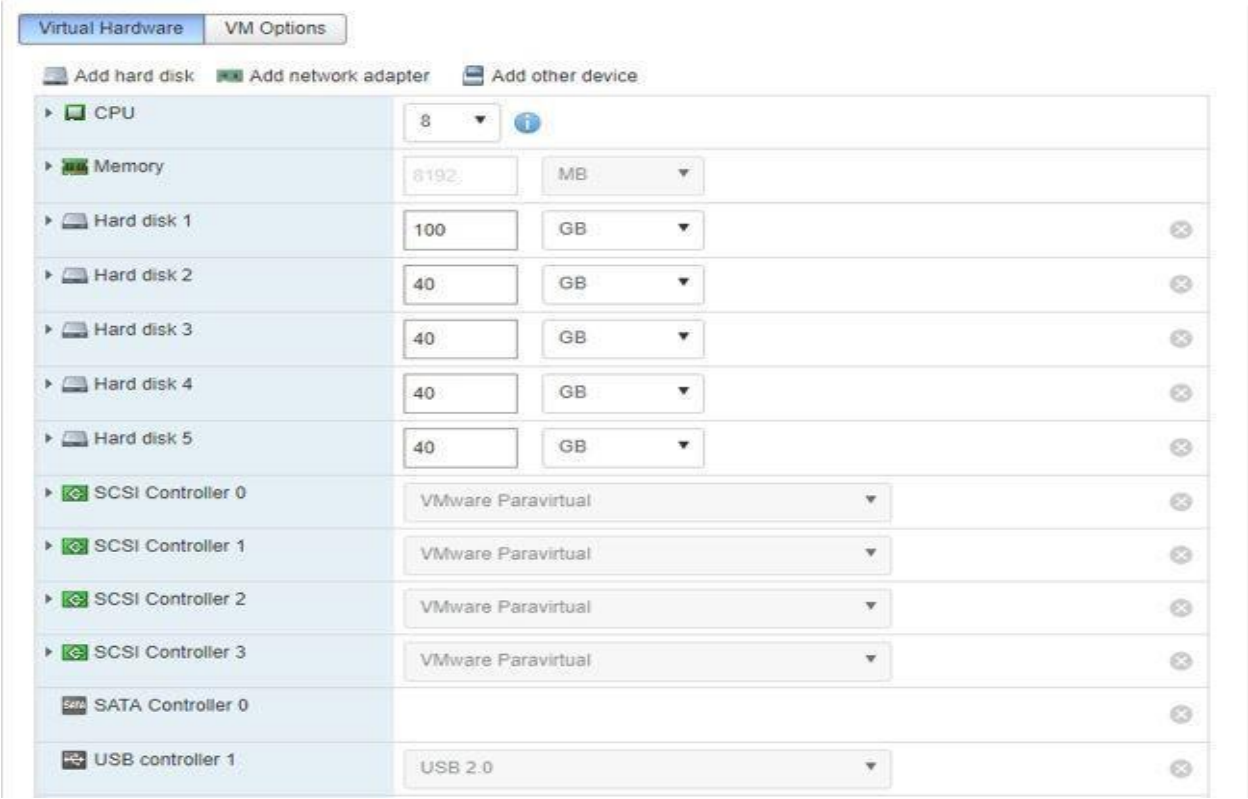
- Create a new Datastore by selecting the Intel Optane SSD P4800X under the devices tab.
- Upon creation, you must be able to see the new data store in the Datastore tab using the Intel Optane SSD P4800X as shown in Figure 6.

Figure 6: Intel® Optane™ SSD P4800X Drive Added to Datastore



- Right click on the Virtual Machine planned to use for this test and choose edit settings. On the edit settings tab follow the steps listed below:
  1. Click on Add other device option and select SCSI controller from the list.
  2. On the newly added SCSI controller change the type to VMware Paravirtual from the drop-down menu.
  3. Now click Add hard disk option and select new hard disk.
  4. Click on the new hard disk created and change the following settings:
     - Maximum size = 40 GB (Change the size as needed for the test case)
     - Location = Click on browse button and choose the Datastore created using Intel Optane SSD P4800X.
     - Type = Thick provisioned, lazily zeroed
     - Limit IOPS = Unlimited
     - Virtual Device Node = Choose the new Paravirtual SCSI controller created.
     - Disk Mode = Dependent
     - Click Save button.
  5. Add three more new hard disks using the Datastore with Intel Optane SSD P4800X, with each hard disk assigned to an individual Paravirtual SCSI controller by repeating steps 1 to 4, listed above, by opening the edit setting tab.
- Upon completion of the listed steps above, the Edit setting tab will look similar to Figure 7.

Figure 7: Screenshot of Edit Settings Tab after Setup



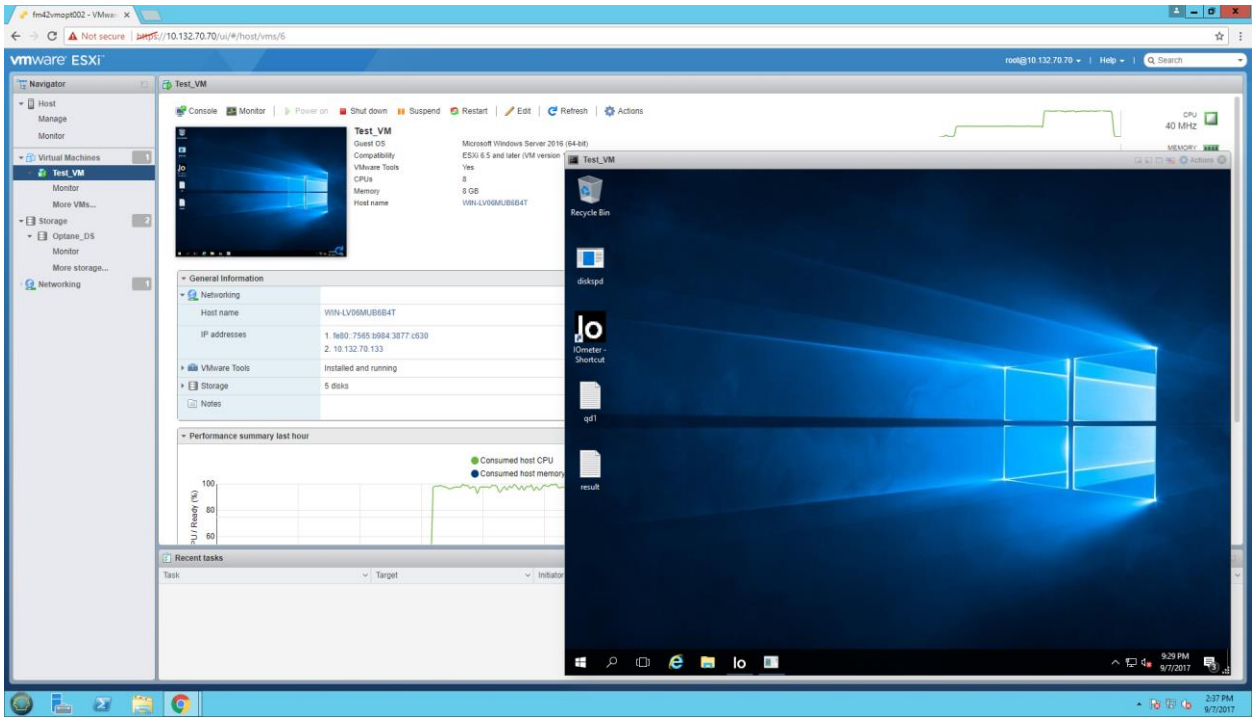
2.4 Setting up Windows Guest OS

Windows Guest OS version used: Microsoft Windows Server 2016 – Version 1607

- In the ESXi Web Client window, choose Actions -> Guest OS-> Install VMWare tools option.
- Open This PC window and choose VMware tools Installation CD to begin installation. Follow the wizard to complete the installation process and restart the guest OS.
- After restart, open Server Manager. In the Tools tab menu, select Computer Management. In the left panel of the window that appears, select Disk Management under Storage.
- In the Disk Management Window all the virtual disks are listed. All disks initially will be offline. Right click on each disk and select Online option to make the disk online.
- Right click on each disk and select Initialize Disk.
- Right click on the partition representation of each disk and select the New Simple Volume option. Follow the wizard to create a New Simple Volume of each disk.
- From the command prompt of the Windows Guest run the following command and restart the virtual machine for maximizing performance:

```
REG ADD HKLM\SYSTEM\CurrentControlSet\services\pvscsi\Parameters\Device /v
DriverParameter /t REG_SZ /d "RequestRingPages=32, MaxQueueDepth=1024"
```

Figure 8: Windows Guest OS on ESXi

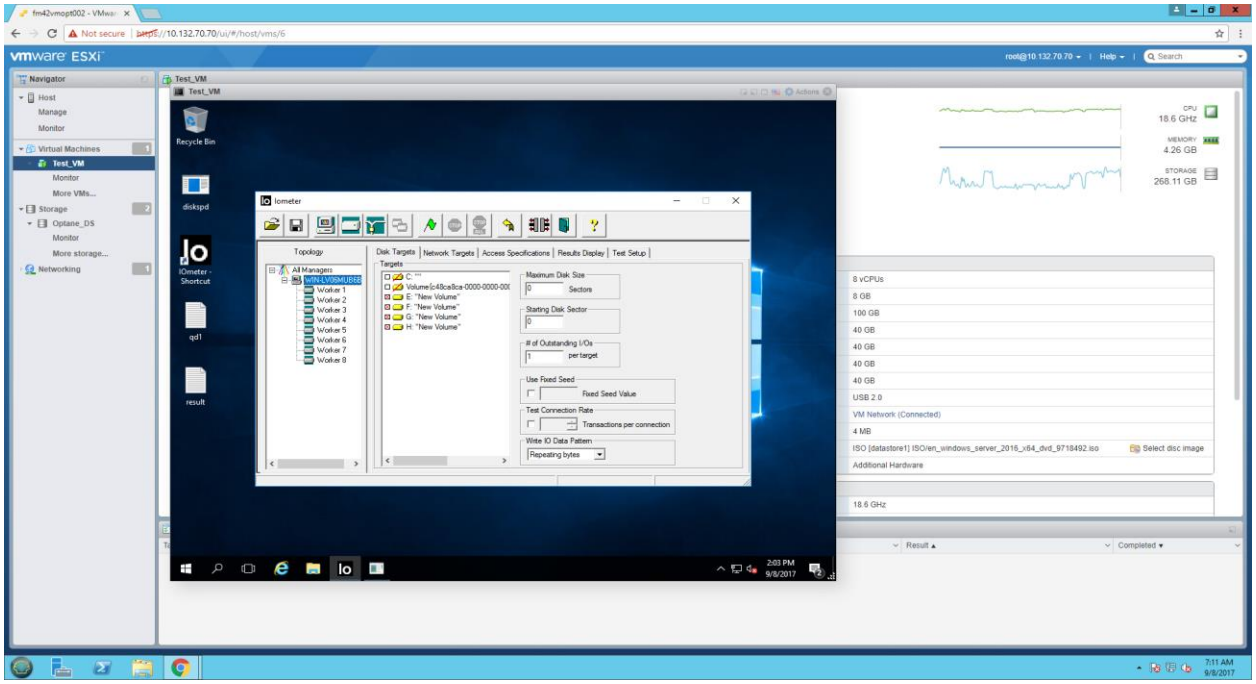


2.5 Setting up IOMeter

IOMeter Version used: 1.1.0

- Download IOMeter from http://www.iometer.org/doc/downloads.html
- To install IOMeter, run the executable file and accept the terms and conditions.
- After finishing the installation, click on the IOMeter icon to start GUI.
- Select the manager from the topology panel.
- The drives of the selected manager are displayed under the Disk Targets tab.
  Note: Blue folder icons represent physical drives; yellow folder icons represent logical drives; yellow folder with red slash icons represent that the drive needs to be prepared before the test begins.
- Select the drive(s) to be tested by adding one drive to each worker, and choose the number of outstanding I/O for the drive.
- Change the options such as maximum disk size, starting disk sector, speed, connection rate, and I/O pattern, if required.
- Move to the Access Specifications tab. Add the tests under Global Access Specifications or create a new test using the New button.
- In case you want to create a new test, enter a test name and change the access specification, percentage of reads and writes, percentage of sequential and random reads, and bursts and click ok. Add the newly created test to the assigned access specification.
- Move to Results Display tab. Change the update frequency from infinity to the time when the updated result is required.
- Under the Test Setup tab, ramp up time can be specified to give time for the disk to refresh.
- Click the green flag in the menu bar to start the test.
- A prompt would be displayed to save the result file in csv format.
- Whenever needed, select the stop icon to stop the test.
- All the test results would be shown in the results display tab and would be saved optionally to the csv file.

Figure 9: IOMeter Snapshot



§

3 Results

Table 2: System Configuration

Type	Configuration
CPU	44 CPUs Intel® Xeon® CPU E5-2699 v4 @2.20 GHz
Memory	32GB DDR4 @ 2133Mhz (32G X 1 DIMM)
BIOS version	SE5C610.86B.01.01.0019
BIOS configuration	Hyper-threading enabled, CPU C-state enabled, P-states enabled, Turbo boost mode-enabled, Performance mode –enabled, Workload configuration- I/O Sensitive
ESXi version	6.5

Table 3: Performance Results for a 4K Random Read Test

VM Queue Depth	VM IOPS	VM Average I/O Response Time (ms)	VM CPU Utilization (%)
1	126067	0.0314	14.41
2	216926	0.0366	21.66
4	362078	0.0438	39.30
8	430923	0.0739	43.53
16	476549	0.1340	47.33
32	494952	0.2583	47.50

Note: Lowest results are based on four test samples.

Figure 10: IOMeter Result Snapshot – 4k Random Read Queue Depth=1

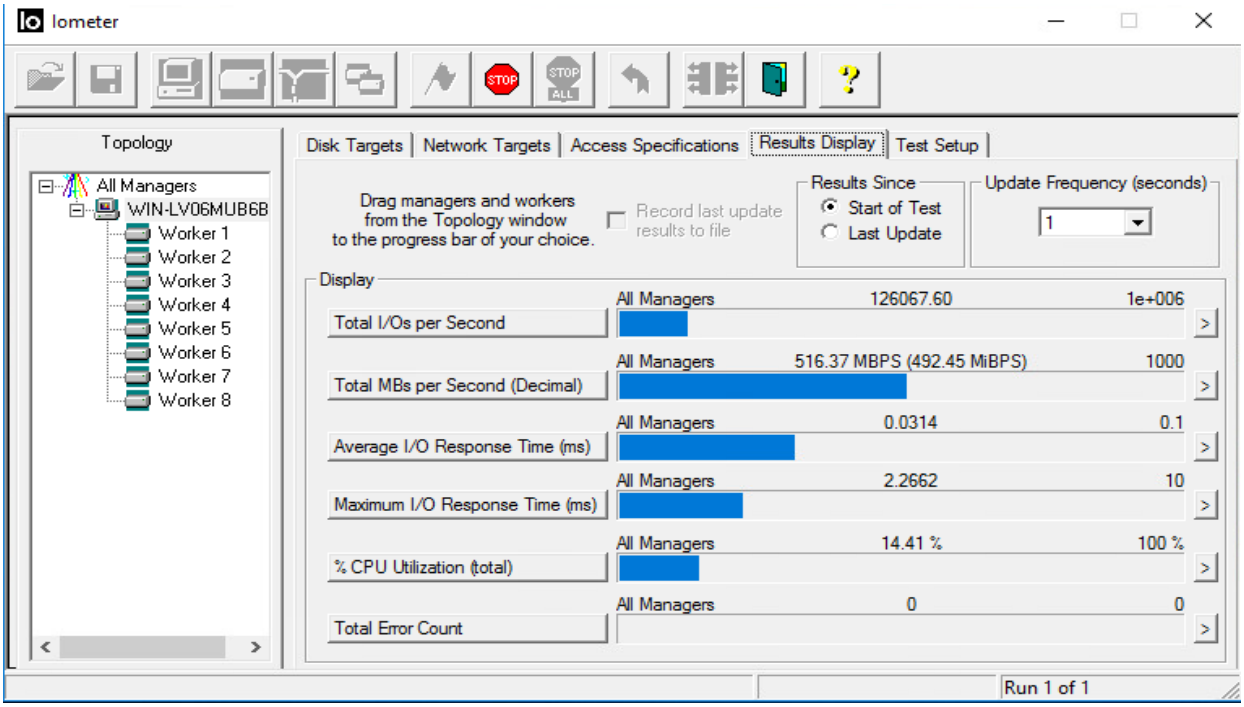
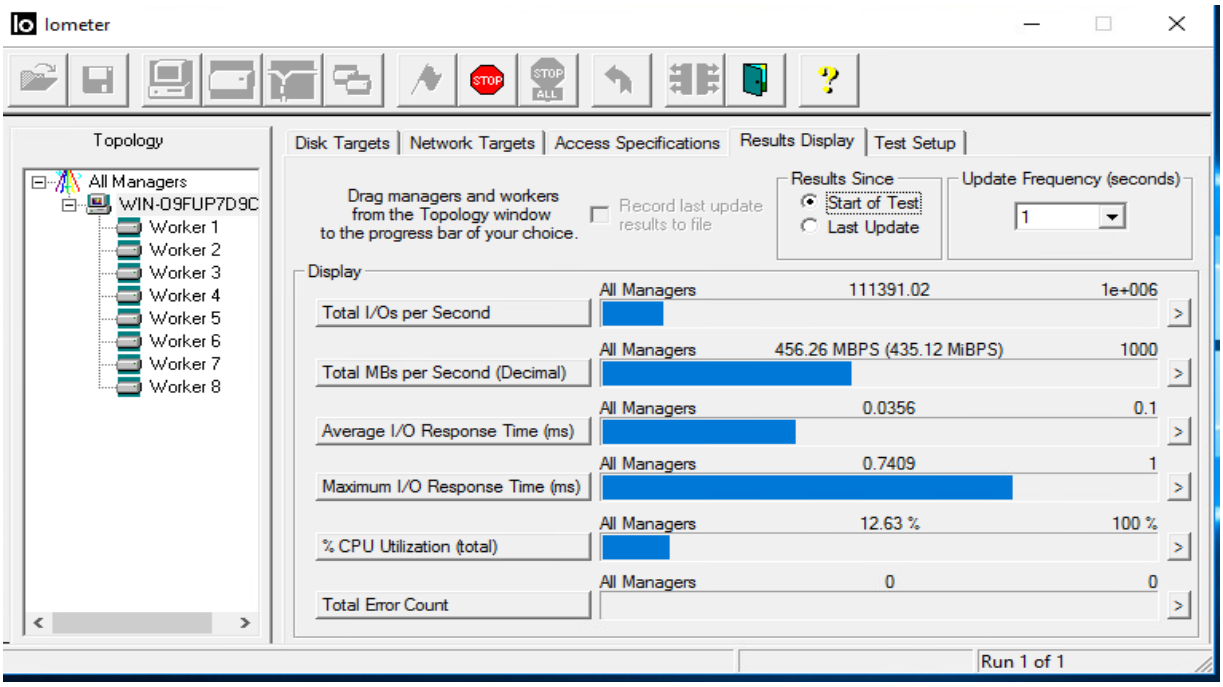


Table 4: Performance Results for a 4K Random Write Test

VM Queue Depth	VM IOPS	VM Average I/O Response Time (ms)	VM CPU Utilization (%)
1	111391	0.0356	12.63
2	181521	0.0439	23.58
4	310215	0.0514	33.42
8	370098	0.0863	42.93
16	415319	0.1539	46.24
32	444255	0.5760	63.22

Note: Lowest results are based on four test samples.

Figure 11: IOMeter Result Snapshot – 4k Random Write Queue Depth=1



3.1 Exhibit 1: Preconditioning

NAND devices are required to be written fully before users can 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 drive 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 the current length of faster background refresh to be set at ~3 hours.

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

§