SPDK NVMe BDEV Performance Report Release 22.01

Testing Date: February 2022

Performed by:

Karol Latecki (karol.latecki@intel.com)

Maciej Wawryk (maciejx.wawryk@intel.com)

Acknowledgments:

James Harris (james.r.harris@intel.com)

John Kariuki (john.k.kariuki@intel.com)

Contents

Audience and Purpose

This report is intended for people who are interested in comparing the performance of the SPDK block device layer vs the Linux Kernel (5.11.20-200.fc33.x86_64) block device layer. It provides performance and efficiency information between the two block layers under various test workloads.

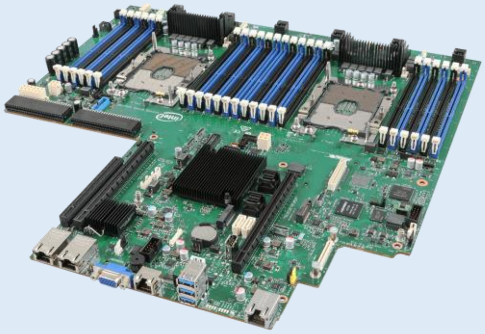
The purpose of the report is not to imply a single “correct” approach, but rather to provide a baseline of well-tested configurations and procedures with repeatable and reproducible results. This report can be viewed as information regarding best known method/practice when performance testing the SPDK NVMe block device.

Test setup

Hardware configuration

Table 1: Hardware setup configuration

<table>
<tr><th>Item</th><th>Description</th></tr>
<tr><td><b>Server Platform</b></td><td>Intel WolfPass <b>R2224WFTZS</b>

Server board <b>S2600WFT</b>
</td></tr>
<tr><td><b>Motherboard</b></td><td>S2600WFT</td></tr>
<tr><td><b>CPU</b></td><td>2 CPU sockets, Intel(R) Xeon(R) Gold 6230N CPU @ 2.30GHz
Number of cores 20 per socket, number of threads 40 per socket
Both sockets populated
Microcode: 0x4003003</td></tr>
<tr><td><b>Memory</b></td><td>12 x 32GB Micron DDR4 36ASF4G72PZ-2G9E2
Total 384 GBs
Memory channel population:
<table>
<tr><th>P1</th><th>P2</th></tr>
<tr><td>CPU1_DIMM_A1</td><td>CPU2_DIMM_A1</td></tr>
<tr><td>CPU1_DIMM_B1</td><td>CPU2_DIMM_B1</td></tr>
<tr><td>CPU1_DIMM_C1</td><td>CPU2_DIMM_C1</td></tr>
<tr><td>CPU1_DIMM_D1</td><td>CPU2_DIMM_D1</td></tr>
<tr><td>CPU1_DIMM_E1</td><td>CPU2_DIMM_E1</td></tr>
<tr><td>CPU1_DIMM_F1</td><td>CPU2_DIMM_F1</td></tr>
</table></td></tr>
<tr><td><b>Operating System</b></td><td>Fedora 33</td></tr>
<tr><td><b>BIOS</b></td><td>SE5C620.86B.02.01.0013.121520200651</td></tr>
</table>

Linux kernel version	5.11.20-200.fc33.x86_64
SPDK version	SPDK 22.01
Fio version	3.19
Storage	**OS:** 1x 120GB Intel SSDSC2BB120G4 **Storage**: 24x Intel® P4610™ 1.6TBs (FW: VDV10170) (6 on CPU NUMA Node 0, 18 on CPU NUMA Node 1)

BIOS Settings

Table 2: Test setup BIOS settings

Item	Description
BIOS	VT-d = Enabled CPU Power and Performance Policy = <Performance> CPU C-state = No Limit CPU P-state = Enabled Enhanced Intel® Speedstep® Tech = Enabled Turbo Boost = Enabled Hyper Threading = Enabled

Storage distribution across NUMA Nodes and PCIe Switches

Wolfpass server platforms PCIe Lanes are not symmetrically distributed between CPU NUMA nodes, which is an important factor in performance tests. Additionally, the total amount of PCIe Lanes available was not enough to accommodate 24 NVMe drives. Therefore, PCIe switches were used to fan out PCIe lanes to NVMe SSDs on the riser cards. For more information on PCIe capabilities of the platform please refer to its technical specification.

Table 3: Test platform NVMe storage setup

Item	Description	
PCIe Riser cards	Risers 1&2: 2x Intel A2UL16RISER2 (PCI gen 3 1x16 Riser) ○ Installed in Riser Slot #1 ○ Installed in Riser Slot #2 Riser 3: 1 x A2UX8X4RISER (PCI gen 3 1x8 Riser) ○ Installed in Riser Slot #3	
PCIe Switches	5 x Intel 4-Port PCIe Gen3 x8 Switch AIC AXXP3SWX08040 Installed in: ○ PCIe Switch 1: Riser Slot #1 port 1 (using CPU1 PCIe Lanes) ○ PCIe Switch 2: Riser Slot #1 port 2 (using CPU2 PCIe Lanes) ○ PCIe Switch 3: Riser Slot #2 port 1 (using CPU2 PCIe Lanes) ○ PCIe Switch 4: Riser Slot #2 port 2 (using CPU2 PCIe Lanes) ○ PCIe Switch 5: Riser Slot #3 port 1 (using CPU2 PCIe Lanes)	
NVMe Drives distribution across the system	Nvme0 – 1	Motherboard ports (CPU1 PCIe Lanes)
	Nvme2 – 3	Motherboard ports (CPU2 PCIe Lanes)
	Nvme4 – 7	PCIe Switch 1 (CPU1 PCIe Lanes)
	Nvme8 – 11	PCIe Switch 2 (CPU2 PCIe Lanes)

	Nvme12 – 15	PCIe Switch 3 (CPU2 PCIe Lanes)
	Nvme16 – 19	PCIe Switch 4 (CPU2 PCIe Lanes)
	Nvme20 – 23	PCIe Switch 5 (CPU2 PCIe Lanes)

SSD Preconditioning

An empty NAND SSD will often show read performance far beyond what the drive claims to be capable of because the NVMe controller knows that the device is empty and completes the read request successfully without performing any actual read operation on the device. Therefore, prior to running each performance test case we preconditioned the SSDs by writing 128K blocks sequentially across the namespace’s full LBA range twice to ensure the controller accesses the NAND media for each subsequent I/O. Additionally, the 4K 100% random writes performance decreases from one test to the next until the NAND management overhead reaches steady state because the wear-levelling activity increases dramatically until the SSD reaches steady state. Therefore, to obtain accurate and repeatable results for the 4K 100% random write workload, we ran the workload for 60 minutes before starting the benchmark test and collecting performance data. For a highly detailed description of exactly how to force an SSD into a known state for benchmarking see the SNIA Solid State Storage Performance Test Specification.

Kernel & BIOS Spectre-Meltdown information

Host server system uses 5.11.20 kernel version which is available from the DNF repository. The default Spectre-Meltdown mitigation patches for this kernel version have been left enabled.

Introduction to SPDK Block Device Layer

SPDK Polled Mode Driver

The NVMe PCIe driver is something that is usually expected to be part of the system kernel and your application would interact with the driver via the system call interface. SPDK takes a different approach. SPDK unbinds the NVMe devices from the kernel NVMe driver and binds them to a userspace NVMe driver instead. This allows a userspace application to directly access the device and its queues from userspace.

The SPDK NVME Driver is a C library that may be linked directly into an application that provides direct, zero-copy data transfer to and from NVMe SSDs. It is entirely passive, meaning that it spawns no threads and only performs actions in response to function calls from the application. The library controls NVMe devices by directly mapping the PCI BAR into the local process and performing MMIO. The SPDK NVMe driver is asynchronous, which means that the driver submits the I/O request as an NVMe submission queue entry on a queue pair and the function returns immediately, prior to the completion of the NVMe command. The application must poll for I/O completion on each queue pair with outstanding I/O to receive completion callbacks.

SPDK Block Device Layer

SPDK further provides a full block stack as a user space library that performs many of the same operations as a block stack in an operating system. The SPDK block device layer often simply called **bdev**, is a C library intended to be equivalent to the operating system block storage layer located above the device drivers in traditional kernel storage stack.

The bdev module provides an abstraction layer that provides common APIs for implementing block devices that interface with different types of block storage device. An application can use the APIs to enumerate and claim SPDK block devices, and then perform asynchronous I/O operations (such as read, write, unmap, etc) in a generic way without knowing if the device is an NVMe device or something else, for example Ceph RBD or malloc ramdisk block device. The SPDK NVMe bdev module can create block devices for both local PCIe-attached NVMe device and remote devices exported over NVMe-oF.

In this report, we benchmarked the performance and efficiency of the bdev for the local PCIe-attached NVMe devices use case. We also demonstrated the benefits of the SPDK approaches, like user-space polling, asynchronous I/O, no context switching etc. under different workloads.

FIO Integration

SPDK provides an FIO plugin for integration with Flexible I/O benchmarking tool. The quickest way to generate a configuration file with all the bdevs for locally PCIe-attached NVMe devices is to use the gen_nvme.sh script with “—json-with-subsystems” option as shown in Figure 1.

```
[user@localhost spdk]$ sudo scripts/gen_nvme.sh --json-with-
subsystems | jq
{
  "subsystems": [
    {
      "subsystem": "bdev",
      "config": [
        {
            "method": "bdev_set_options",
            "params": {
              "bdev_io_pool_size": 65535,
              "bdev_io_cache_size": 2048,
              "bdev_auto_examine": true
            }
        },
        {
          "method": "bdev_nvme_attach_controller",
          "params": {
            "trtype": "PCIe",
            "name": "Nvme0",
            "traddr": "0000:1a:00.0",
            "bdev_io_cache_size": 2048
          }
        },

        [...]

        {
          "method": "bdev_nvme_attach_controller",
          "params": {
            "trtype": "PCIe",
            "name": "Nvme23",
            "traddr": "0000:df:00.0",
            "bdev_io_cache_size": 2048
          }
        }
      ]
    }
  ]
```

Figure 1 : Example NVMe bdev configuration file

Add SPDK bdevs to the fio job file, by setting the *ioengine=spdk_bdev* and adding the *spdk_json_conf* parameter whose value points to the NVMe bdev configuration file.

The example fio configuration file in figure 2, shows how to define multiple fio jobs and assign NVMe bdevs to each job. Each job is also pinned to a CPU core on the same NUMA node as the NVMe SSDs that the job will access.

Finally, to use the bdev fio plugin specify the LD_PRELOAD when running fio.

LD_PRELOAD=<path to spdk repo>/examples/bdev/fio_plugin/fio_plugin fio <fio job file>

```
[global]
direct=1
thread=1
time_based=1
norandommap=1
group_reporting=1
ioengine=spdk_bdev
spdk_json_conf=/tmp/bdev.conf

rw=randread
rwmixread=70
bs=4096
numjobs=1
runtime=300
ramp_time=60

[filename0]
iodepth=192
cpus_allowed=0
filename=Nvme0n1
filename=Nvme1n1
filename=Nvme4n1
filename=Nvme5n1
filename=Nvme6n1
filename=Nvme7n1

[filename1]
iodepth=192
cpus_allowed=21
filename=Nvme2n1
filename=Nvme3n1
filename=Nvme8n1
filename=Nvme9n1
filename=Nvme10n1
filename=Nvme11n1

[filename2]
iodepth=192
cpus_allowed=22
filename=Nvme12n1
filename=Nvme13n1
filename=Nvme14n1
filename=Nvme15n1
filename=Nvme16n1
filename=Nvme17n1

[filename3]
iodepth=192
cpus_allowed=23
filename=Nvme18n1
filename=Nvme19n1
filename=Nvme20n1
filename=Nvme21n1
filename=Nvme22n1
filename=Nvme23n1
```

Figure 2: Example SPDK Fio BDEV configuration file

Test Case 1: SPDK NVMe BDEV IOPS/Core Test

Purpose: The purpose of this test case was to measure the maximum performance in IOPS/Core of the NVMe block layer on a single CPU core. We used different benchmarking tools (SPDK bdevperf vs. SPDK FIO BDEV plugin vs SPDK NVMe perf) to understand the overhead of benchmarking tools. Measuring IOPS was key in this test case, so latency measurements were either disabled or skipped.

The following Random Read/Write workloads were used:

- 4KiB 100% Random Read
- 4KiB 100% Random Write
- 4KiB Random 70% Read 30% Write

For each workload we followed the following steps:

1) Precondition SSDs as described in "Test Setup" chapter.
2) Run each test workload: Start with a configuration that has 24 Intel P4610x NVMe devices and decrease the number of SSDs by 2 on each subsequent run.
    - This shows us the IOPS scaling as we add SSDs till the maximum IOPS/Core is reached.
    - Starting with 24 SSDs and reducing the number of SSDs on subsequent eliminates having to precondition between runs because all SSDs used in the subsequent run were used in the previous run so they should still be in a steady state.
3) Repeat three times. The data reported is the average of the 3 runs.

Table 4: SPDK NVMe BDEV IOPS Test configuration

Item	Description
Test case	SPDK NVMe BDEV IOPS/Core Test
Test configuration	**FIO Version**: fio-3.19 **Number of NVMe SSDs**: {1, 2, 4, 6 ,8, 10, 12, 16, 20, 24} **SPDK_BDEV_IO_CACHE_SIZE** changed from 256 to 2048 (using bdev_set_options RPC call). **NUMA optimization:** The test platform has PCIe lanes unevenly distributed between NUMA nodes, most of the NVMe SSDs (18 out of total 24) are located on NUMA node 1. Therefore, a CPU Core from NUMA node 1 was selected as primary core for test, in order to reduce the overhead of cross-numa operations.
Bdevperf	`./bdevperf -c bdev.conf -q ${iodepth} -o ${block_size} -w ${rw} -M ${rwmixread} -t 300 -m 20 -p 20`

configuration	

FIO configuration	[global]
	ioengine=spdk_bdev
	spdk_json_conf=bdev.conf
	
	gtod_reduce=1
	direct=1
	thread=1
	norandommap=1
	time_based=1
	ramp_time=60s
	runtime=300s
	
	bs=4k
	rw={randread, randwrite, randrw}
	rwmixread={100,70,0}
	iodepth={32, 64, 128}
	numjobs=1

SPDK NVMe BDEV Single Core Throughput

The first test was performed using SPDK bdevperf, which is lightweight benchmarking tool that adds minimal latency to the I/O path. The charts below show the Single core IOPS results for the SPDK Block Layer with increasing number of NVMe SSDs.

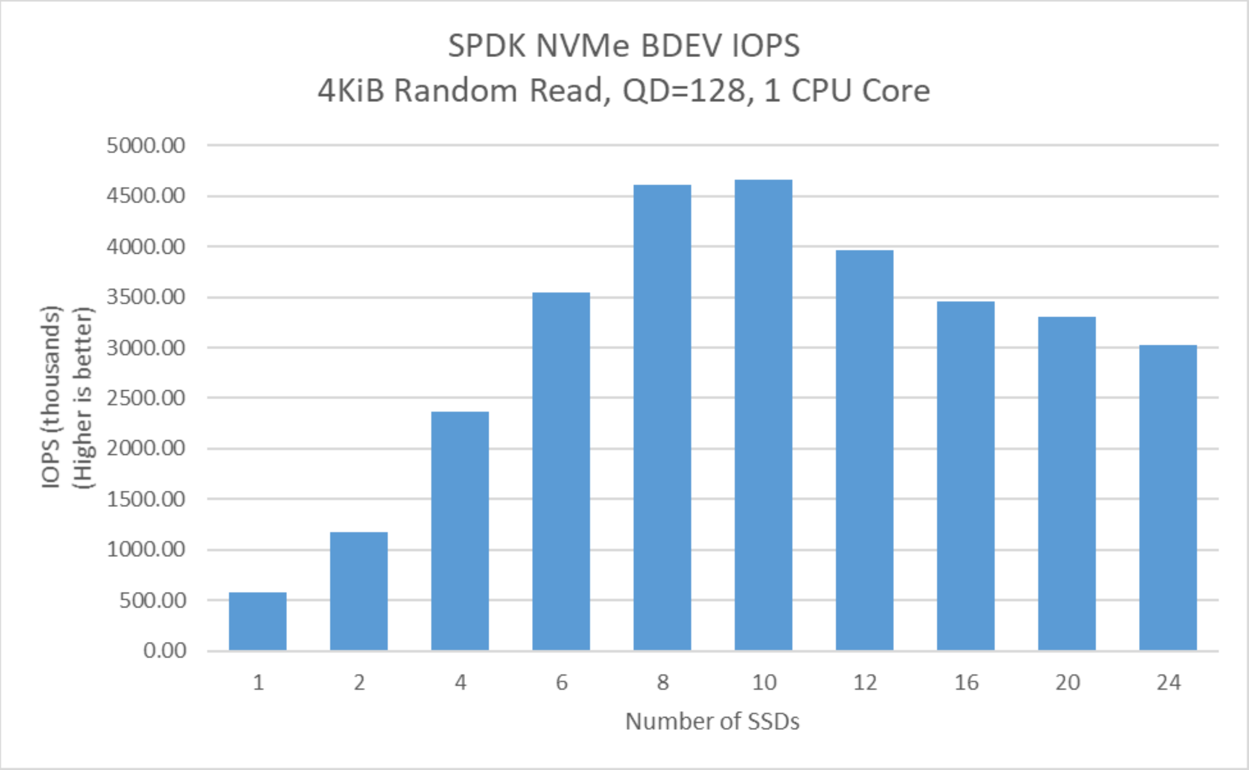


Figure 3: SPDK NVMe BDEV IOPS scalability with addition of SSDs (4KiB Random Read, 1CPU Core, QD=128, using bdevperf tool)

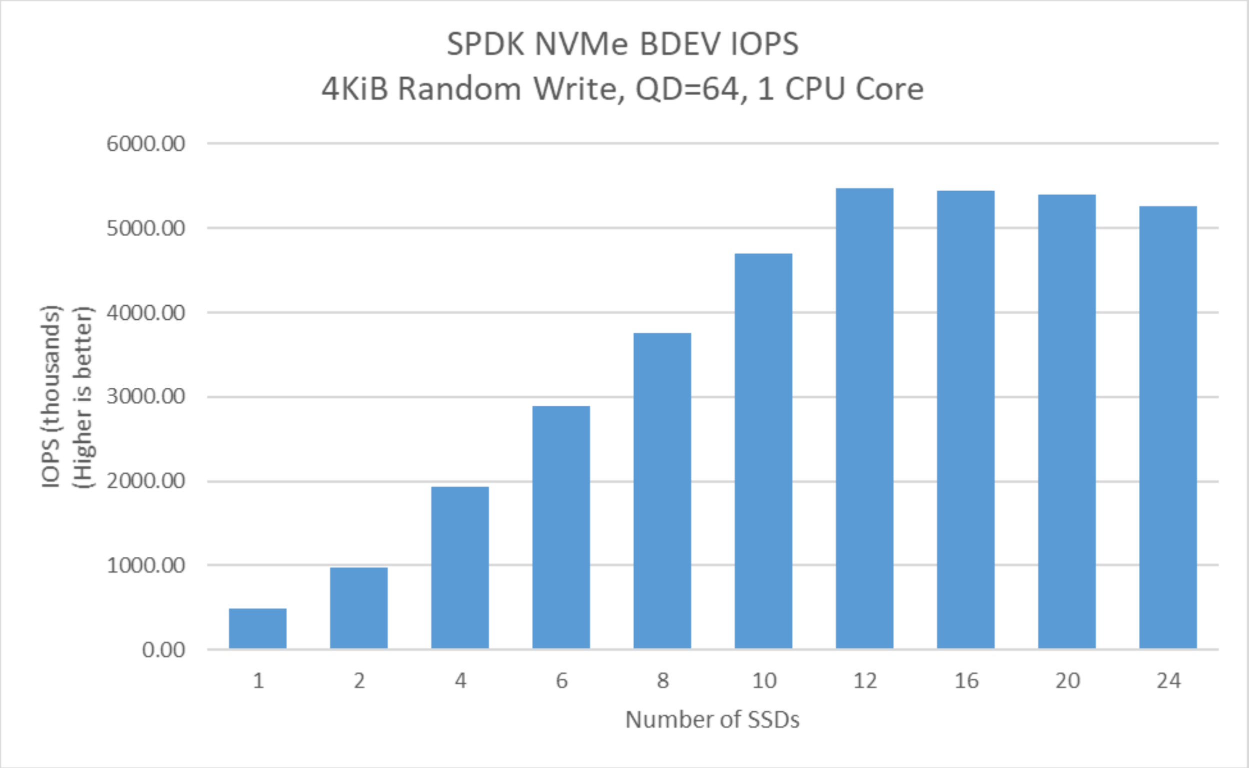


Figure 4: SPDK NVMe BDEV IOPS scalability with addition of SSDs (4KiB Random Write, 1CPU Core, QD=64, using bdevperf tool)

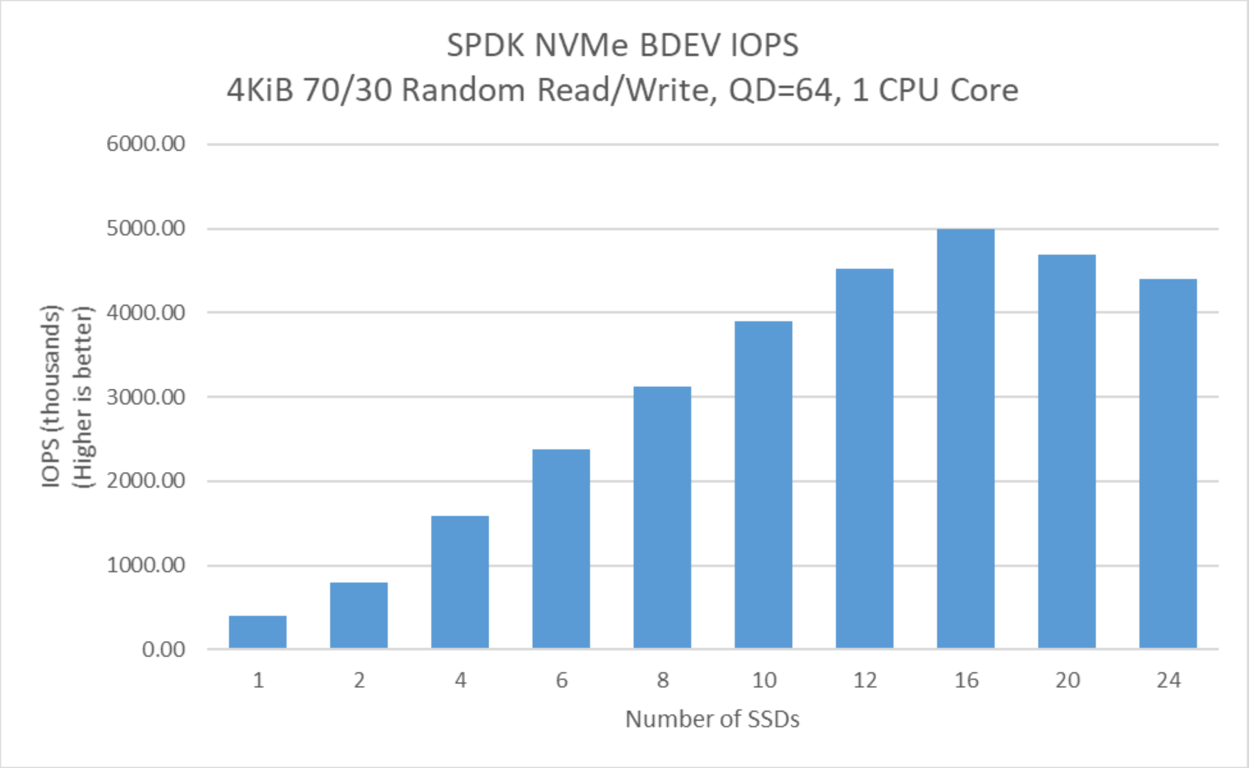


Figure 5: SPDK NVMe BDEV IOPS scalability with addition of SSDs (4KiB 70/30 Random Read/Write, 1CPU Core, QD=64, using bdevperf tool)

Bdevperf vs. FIO IOPS/Core results

SPDK provides the bdevperf benchmarking tool that provides minimal capabilities needed to define basic workloads and collects a limited amount of data. The FIO benchmarking tool provides a lot of great features to enable users to quickly define workloads, scale the workloads and collect many data points for detailed performance analysis, however, at cost of higher overhead. This test compares the performance in IOPS/core of the bdevperf and FIO benchmarking tools.

Table 5: IOPS/Core performance; SPDK FIO bdev plugin vs SPDK bdevperf (Blocksize=4KiB, 1 CPU Core)

Workload	SDPK Fio BDEV Plugin (IOPS, thousands)	SPDK Bdevperf (IOPS, thousands)	Performance gain
4KiB Random Read, QD=128, 10 SSDs	2483.60	4663.84	87.8%
4KiB Random Write, QD=64, 12 SSDs	2827.02	5478.23	93.8%
4KiB 70/30 Random Read/Write, QD=128, 12 SSDs	2218.73	4675.78	110.7%

The overhead of the benchmarking tools is important when you are testing a system that is capable of millions of IOPS/Core. Using a benchmarking tool that has minimal overhead like the SPDK bdevperf yields up to 110.7% more IOPS/Core than FIO.

NVMe BDEV vs. Polled-Mode Driver IOPS/Core

In this test case, we compared the throughput of the NVMe BDEV with that of the polled-mode driver. How to read this data? The SPDK block layer provides several key features at a cost of approximately 26.4% and 39.6% more CPU utilization for Random Read and Random Write workloads. If you are building a system with many SSDs that is capable of millions of IOPS, you can take advantage of the block layer features at the cost of approximately 1 additional CPU core for every 4 I/O cores for Random Read workload and 1 CPU core for every 2 I/O cores for Random Write workload. Comparison was done using SPDK Bdevperf and Nvmeperf test tools.

Table 6: SPDK NVMe Bdev vs SPDK NVMe PMD IOPS/Core (Blocksize=4KiB, 1 CPU Core)

Workload	SPDK Bdevperf (IOPS, thousands)	SPDK Nvmeperf (IOPS, thousands)	Performance gain
4KiB Random Read, QD=128, 10 SSDs	4663.84	5894.03	26.4%
4KiB Random Write, QD=64, 12 SSDs	5478.23	5533.87	39.6%

Conclusions

1. The SPDK NVMe block device module adds approximately 26.4 and 39.6% overhead compared to using only the SPDK NVMe Polled-Mode Driver without the block device module for Random Read and Random Write workloads respectively.

2. Performance scales linearly with addition of NVMe SSDs up to 8 and 12 SSDs for Random Read and Random Read/Write workloads, reaching around 4.60 and 5.47 million IOPS respectively.

3. Performance scaling is linear for Random Write workload up to 12 NVMe SSDs, reaching around 5.47M IOPS.

4. For all workloads there is a noticeable performance degradation with addition of more NVMe SSDs after peak performance point has been reached.

5. The IOPS for the 4 KiB Random Write workload exceeded the expected NVMe SSDs maximum throughput is about 4.8M IOPS. We suspect this is due to a not perfect preconditioning process, which wears off over time. The results, however, were repeatable for a number of test runs.

Test Case 2: SPDK NVMe BDEV I/O Cores Scaling

Purpose: The purpose of this test case is to demonstrate the I/O throughput scalability of the NVMe BDEV module with the addition of more CPU cores to perform I/O. The number of CPU cores used was scaled as 1, 2, 3, 4 and 5.

Test Workloads: We use the following Random Read/Write mixes

- 4KiB 100% Random Read
- 4KiB 100% Random Write
- 4KiB Random 70% Read 30% Write

Table 7: SPDK NVMe BDEV I/O Cores Scalability Test

Item	Description
Test case	Test SPDK NVMe BDEV I/O Cores Scalability Test
Test configuration	**Number of CPU Cores:** 1, 2, 3, 4, 5 **Number of NVMe SSDs:** 6 per each CPU Core used in test, up to maximum of 24 NVMe SSDs **NUMA optimization:** The test platform has PCIe lanes unevenly distributed between NUMA nodes, most of the NVMe SSDs (18 out of total 24) are located on NUMA node 1. Therefore, only CPU Cores from NUMA node 1 were selected for test, in order to reduce the overhead of cross-numa operations.

Bdev perf configuration	`spdk/test/bdev/bdevperf/bdevperf --json bdev.conf \`
	`-q 128 -o 4096 -w randrw -M ${MIXREAD} \`
	`-t 300 -m ${CORE_MASK} -p ${PRIMARY_CORE}`

Results

Table 8: SPDK NVMe BDEV I/O Cores Scalability Test (4KiB 100% Random Read IOPS at QD=128; 4KiB 100% Random Write IOPS at QD=64; 4KiB 70/30 Random Read/Write IOPS at QD=128)

		IOPS (thousands)		
CPU Cores	**NVMe SSDs**	**Random Read QD=128**	**Random Write QD=64**	**70/30 Random Read/Write QD=128**
1	**6**	3593.35	2670.59	2693.06
2	**12**	7164.58	5437.77	5400.39
3	**18**	10345.97	8268.57	8134.68
4	**24**	10928.00	10069.52	10851.79
5	**24**	10914.65	10078.57	10886.56

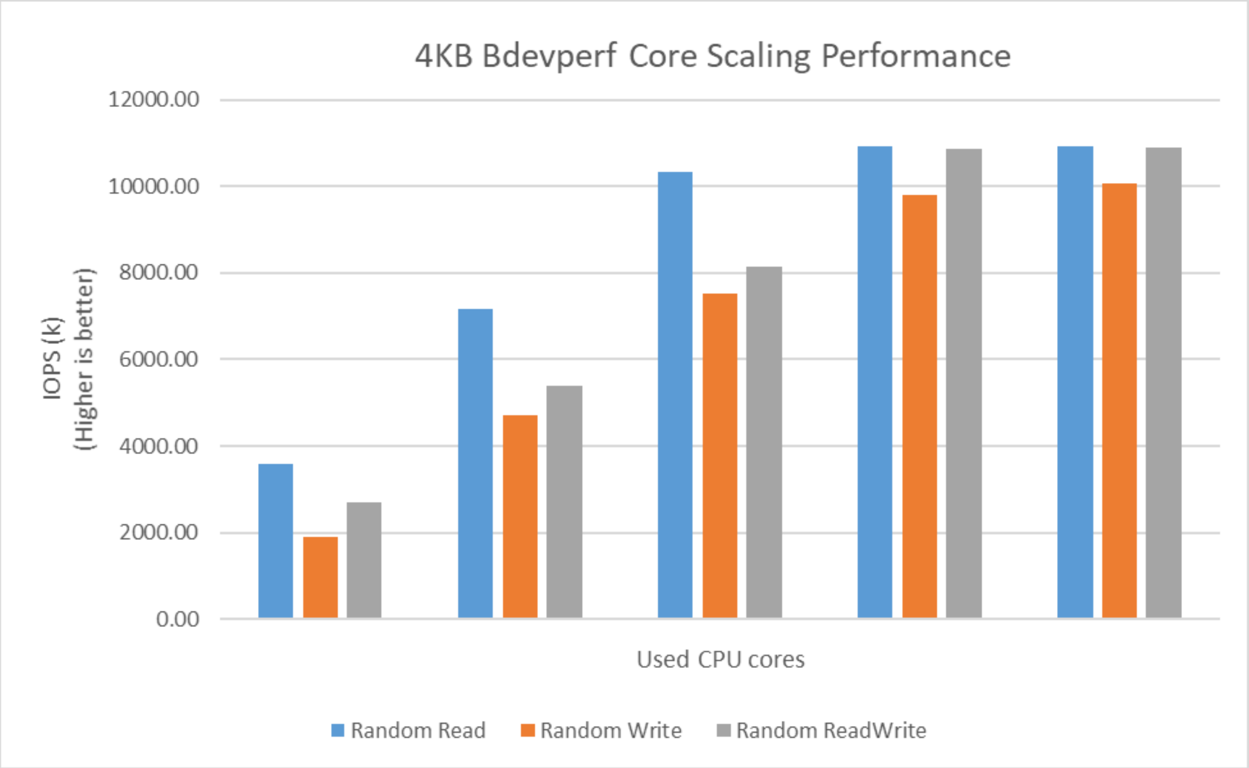


Figure 6: SPDK NVMe BDEV I/O Cores Scalability (4KiB 100% Random Read IOPS at QD=128; 4KiB 100% Random Write IOPS at QD=64; 4KiB 70/30 Random Read/Write IOPS at QD=128)

Conclusions

1. The IOPS for the 4 KiB Random Read and 4 KiB Random Read/Write workloads scale up linearly with the addition of I/O cores until the PCIe switches in platform are saturated (at about 10.9M IOPS; see "Test setup" chapter for more information).

2. The IOPS for the 4 KiB Random Write workload scale up linearly until platform PCIe switches are saturated. The IOPS exceeded the expected NVMe SSDs throughput for this workload which is about 4.8M IOPS. We suspect this is due to a not perfect preconditioning process, which wears off over time. However, the results were repeatable and showed SPDK's high scalability with addition of I/O cores.

Test Case 3: SPDK NVMe BDEV Latency

This test case was carried out to understand latency characteristics while running SPDK NVMe bdev and its comparison to Linux Kernel NVMe block device layer. We used SPDK FIO BDEV Plugin instead of the SPDK Bdevperf tool, as it allowed us to gather detailed latency metrics. FIO was ran for 15 minutes targeting a single block device over a single NVMe drive. This test compares consistency between latency of the SPDK and Linux Kernel block layers over time in a histogram. The Linux Kernel block layer provides I/O polling capabilities to eliminate overhead such as context switch, IRQ delivery delay and IRQ handler scheduling. This test case includes a comparison of the I/O latency for the Kernel vs. SPDK.

Test Workloads: We use the following workloads:

- 4KiB 100% Random Read
- 4KiB 100% Random Write

Important note: For 21.01 benchmark tests we have been unable to successfully run tests for Kernel io_uring engine with sqthread_pool option enabled when workload was using write path and Queue Depth was set to 1. Because of this, Random Write QD=1 workload was run without sqthread_poll option enabled. This is caused by some unidentified bug, probably in Fio or Kernel itself. For more information, please see the issue for this problem on Github.

Table 9: SPDK NVMe BDEV Latency Test

Item	Description
Test case	Test SPDK NVMe BDEV Latency Test
Test configuration	**FIO Version:** fio-3.19 **Number of CPU Cores:** 1 **Number of NVMe SSDs:** 1
SPDK NVMe Driver Configuration	ioengine=spdk_bdev
Linux Kernel Default (libaio) Configuration	ioengine=libaio

Linux Kernel io_uring	ioengine=io_uring
	System NVMe block device configuration:
	`echo 0 > /sys/block/nvme0n1/queue`
	`echo 0 > /sys/block/nvme0n1/rq_affinity`
	`echo 2 > /sys/block/nvme0n1/nomerges`
	`echo -1 > /sys/block/nvme0n1/io_poll_delay`

FIO configuration (common part)	`[global]`
	`direct=1`
	`thread=1`
	`time_based=1`

	norandommap=1
	group_reporting=1
	
	rw={randread | randwrite}
	bs=4096
	runtime=900
	ramp_time=120
	numjobs=1
	log_avg_msec=15
	write_lat_log=/tmp/tc3_lat.log

FIO configuration (SPDK specific)	[global]
	ioengine=spdk_bdev
	spdk_conf=/tmp/bdev.conf
	
	[filename0]
	iodepth=1
	cpus_allowed=20
	filename=Nvme0n1

FIO configuration (Linux Kernel common)	[global]
	ioengine={libaio | io_uring}
	
	[filename0]
	iodepth=1
	cpus_allowed=20
	filename=/dev/nvme18n1

FIO configuration (Linux Kernel io_uring specific)	[global]
	fixedbufs=1
	hipri=1
	registerfiles=1
	sqthread_poll=1

The Linux block layer implements I/O polling on the completion queue. Polling can remove context switch(cs) overhead, IRQ delivery and IRQ handler scheduling overhead[1].

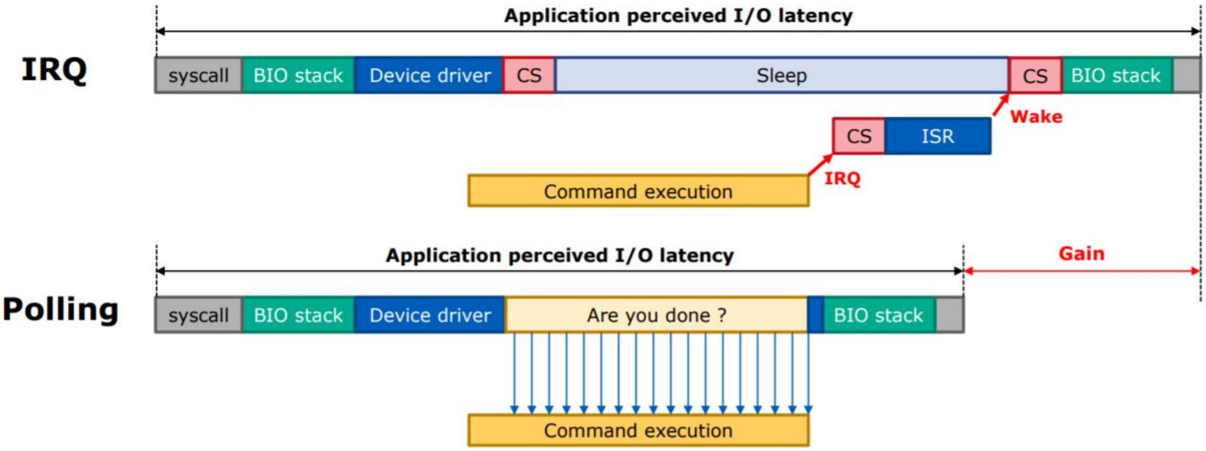


Figure 7: Linux Block Layer I/O Optimization with Polling. Source [1]

Furthermore, the Linux block I/O polling provides a mechanism to reduce the CPU load. In the *Classic Polling* model, the CPU spin-waits for the command completion and utilizes 100% of a CPU core [1]. There's also an adaptive hybrid polling which reduces the CPU load by putting the I/O polling thread to sleep for about half of the command execution time, but the polling thread must be woken up before the I/O completes with enough heads-up time for a context switch[1]. Hybrid polling mode was not used for testing in this document.

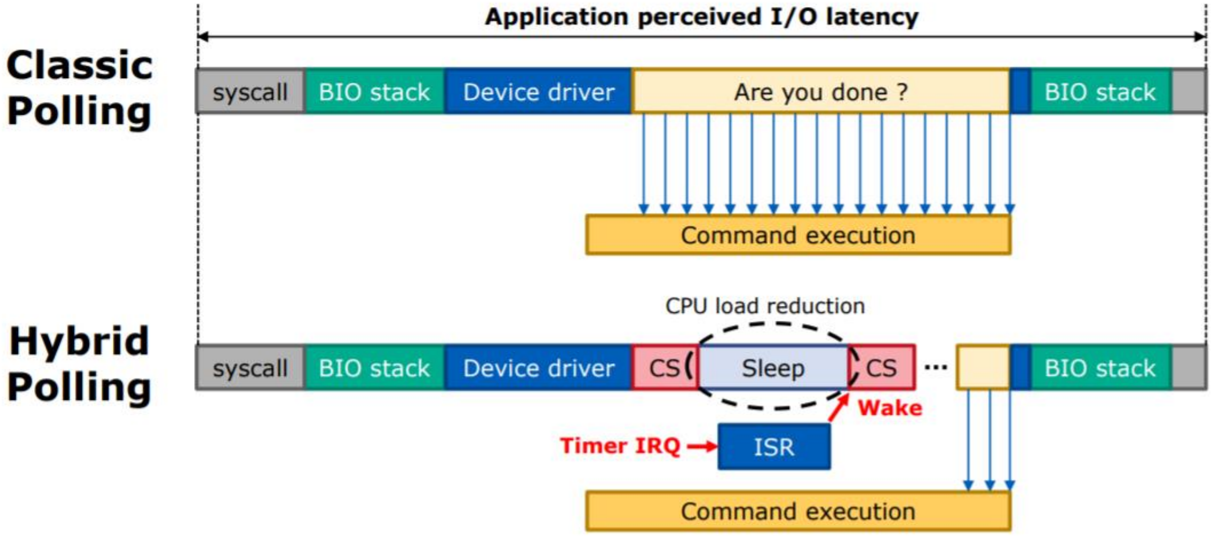


Figure 8: Linux Block I/O Classic and Hybrid Polling latency breakdown. Source [1]

The data in tables and charts compares the I/O latency for a various 4KiB workloads performed using the SPDK BDEV vs. Linux block layerI/O model libaio and io_uring with polling mode enabled.

Average and tail latency comparison

Table 10: SPDK bdev vs. Linux Kernel latency comparison (4KiB Random Read, QD=1, runtime=900s)

Latency metrics (usec)	SDPK Fio BDEV Plugin	Linux Kernel (libaio)	Linux Kernel (io_uring)
Average	70.99	85.69	73.21
P90	98.82	108.03	100.86
P99	99.84	112.13	102.91
P99.99	305.15	346.11	292.86
Stdev	21.54	22.18	20.62
Average submission latency	0.13	5.15	0.00
Average completion latency	70.86	80.23	73.17

Table 11: SPDK bdev vs. Linux Kernel latency comparison (4KiB Random Write, QD=1, runtime=900s)

Latency metrics (usec)	SDPK Fio BDEV Plugin	Linux Kernel (Default libaio)	Linux Kernel (io_uring)
Average	10.18	11.89	12.01
P90	19.33	18.05	19.58
P99	36.61	34.56	35.58
P99.99	82.43	75.26	78.34
Stdev	7.06	6.17	6.21
Average submission latency	0.18	1.23	0.00
Average completion latency	10.00	10.60	11.69

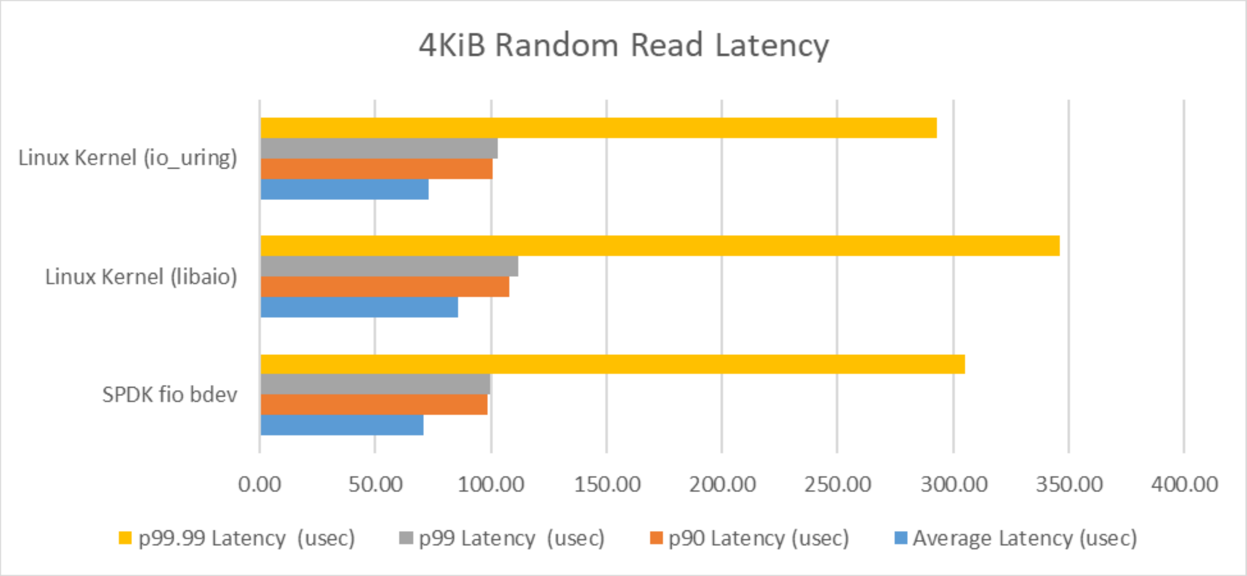


Figure 9: SPDK bdev vs Linux Kernel Latency comparison (4KiB Random Read)

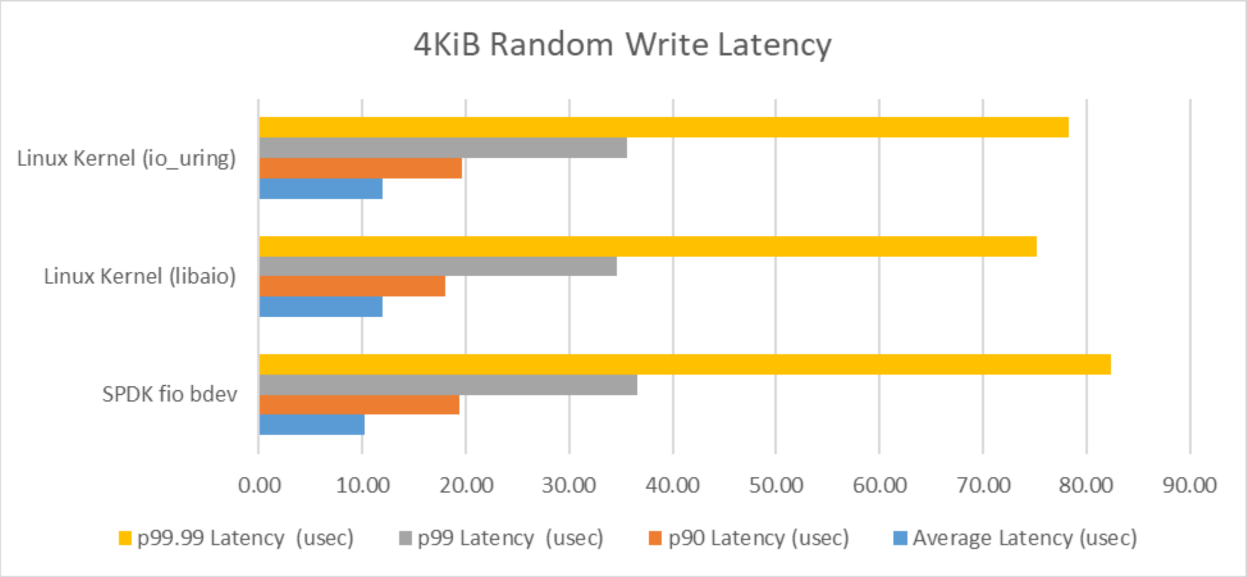


Figure 10: SPDK bdev vs Linux Kernel Latency comparison (4KiB Random Write)

Linux Kernel libaio Histograms

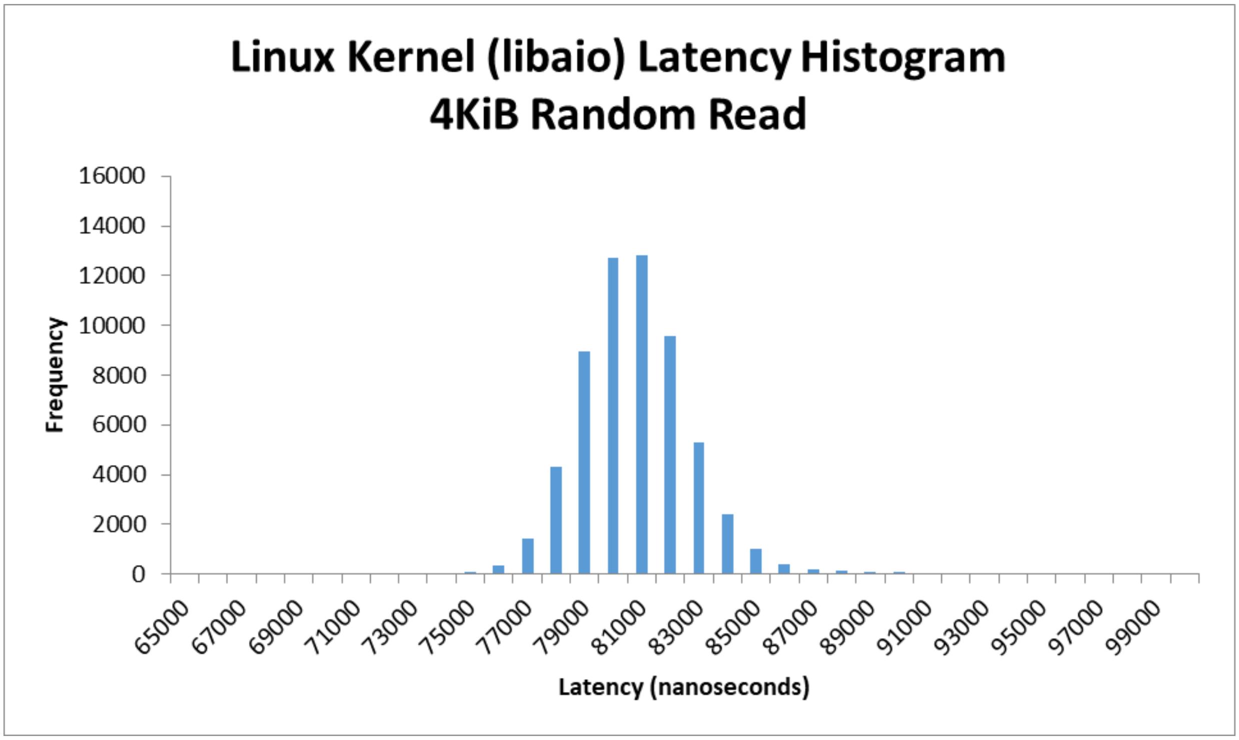


Figure 11: Linux Kernel (Default libaio) 4KiB Random Read Average Latency Histogram (QD=1, Runtime=900s, fio, sampling interval = 15msec)

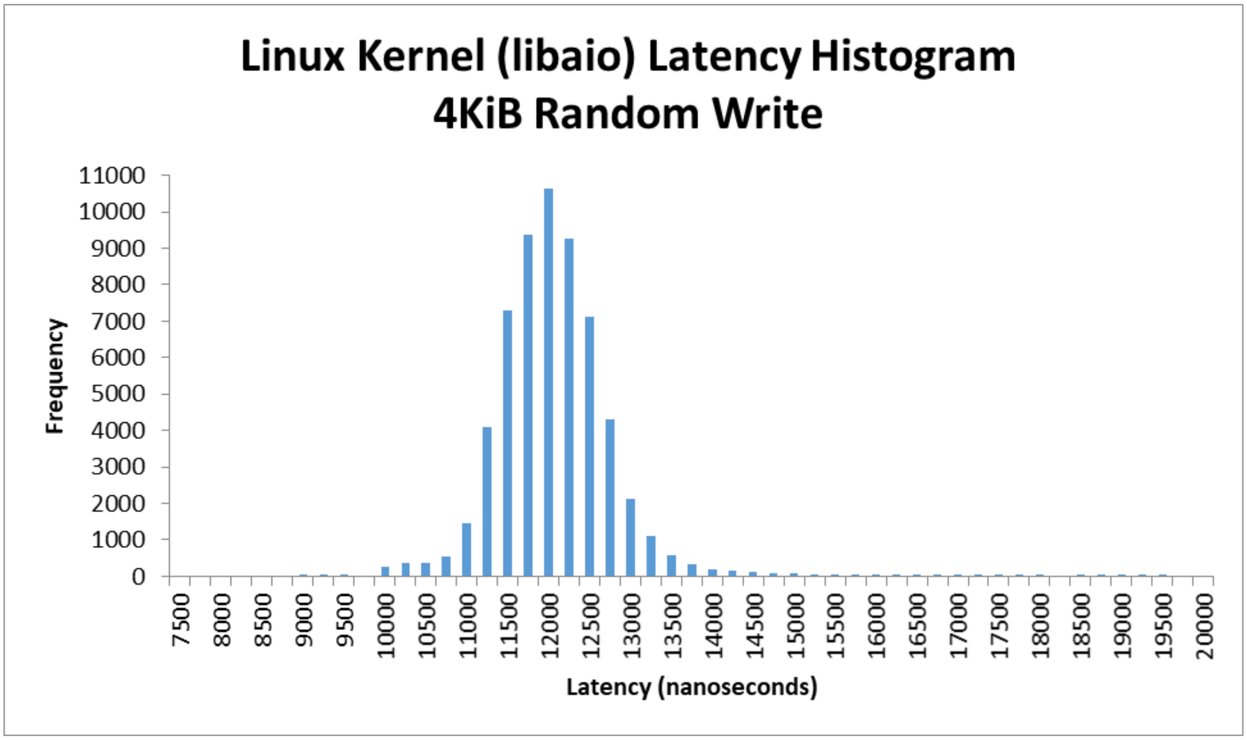


Figure 12: Linux Kernel (Default libaio) 4KiB Random Write Average Latency Histogram (QD=1, Runtime=900s, fio, sampling interval = 15msec)

Linux Kernel io_uring Histograms

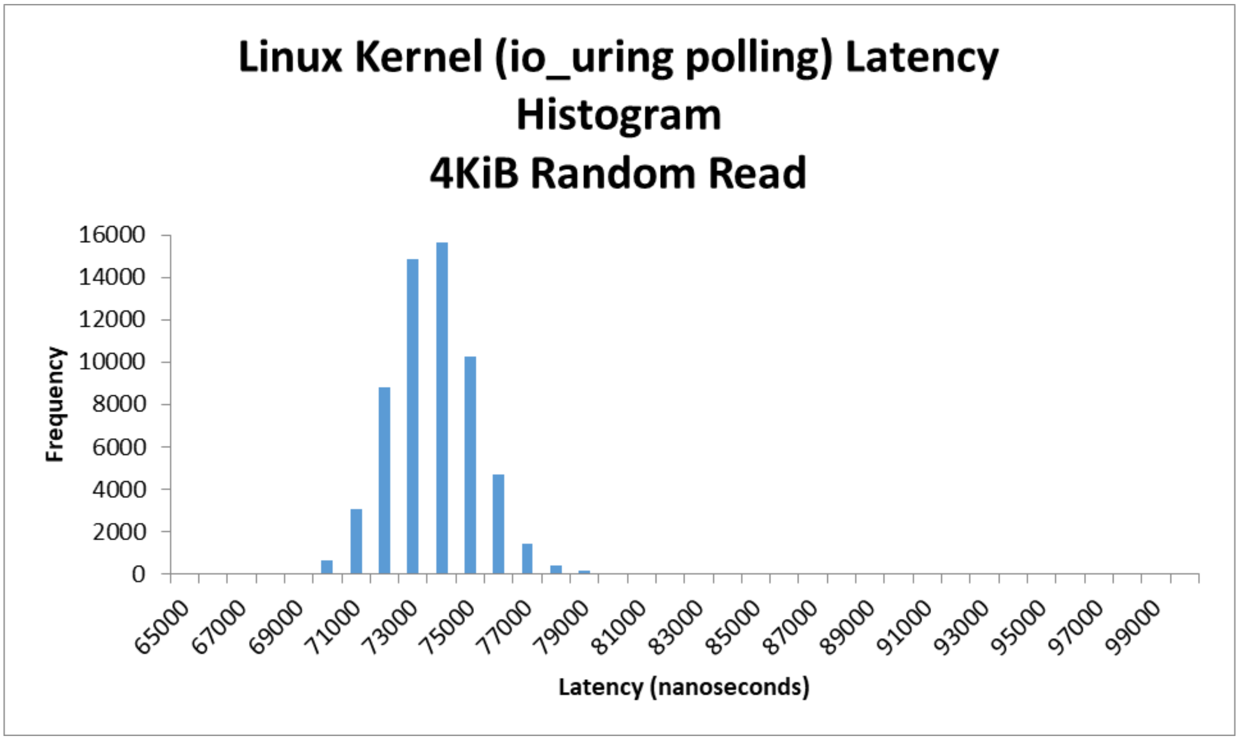


Figure 13: Linux Kernel (io_uring polling) 4KiB Random Read Average Latency Histogram (QD=1, Runtime=900s, fio, sampling interval = 15msec)

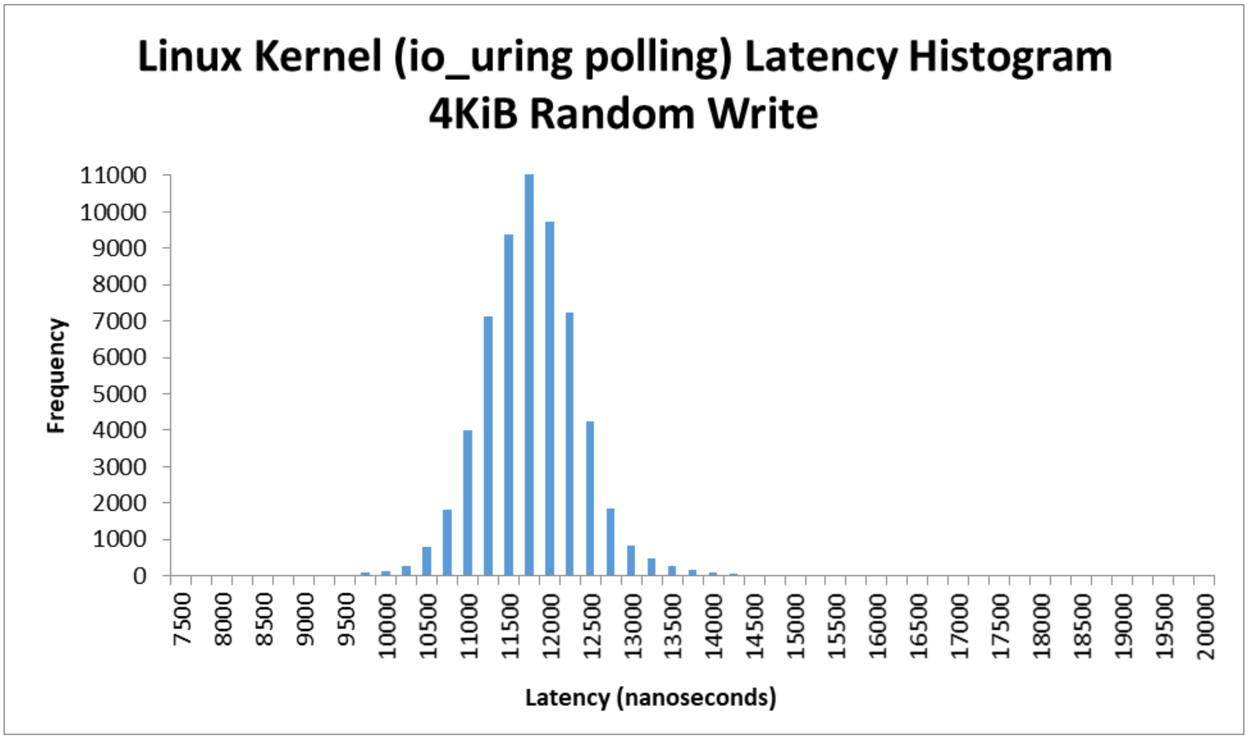


Figure 14: Linux Kernel (io_uring polling) 4KiB Random Write Average Latency Histogram (QD=1, Runtime=900s, fio, sampling interval = 15msec)

SPDK FIO Bdev Histograms

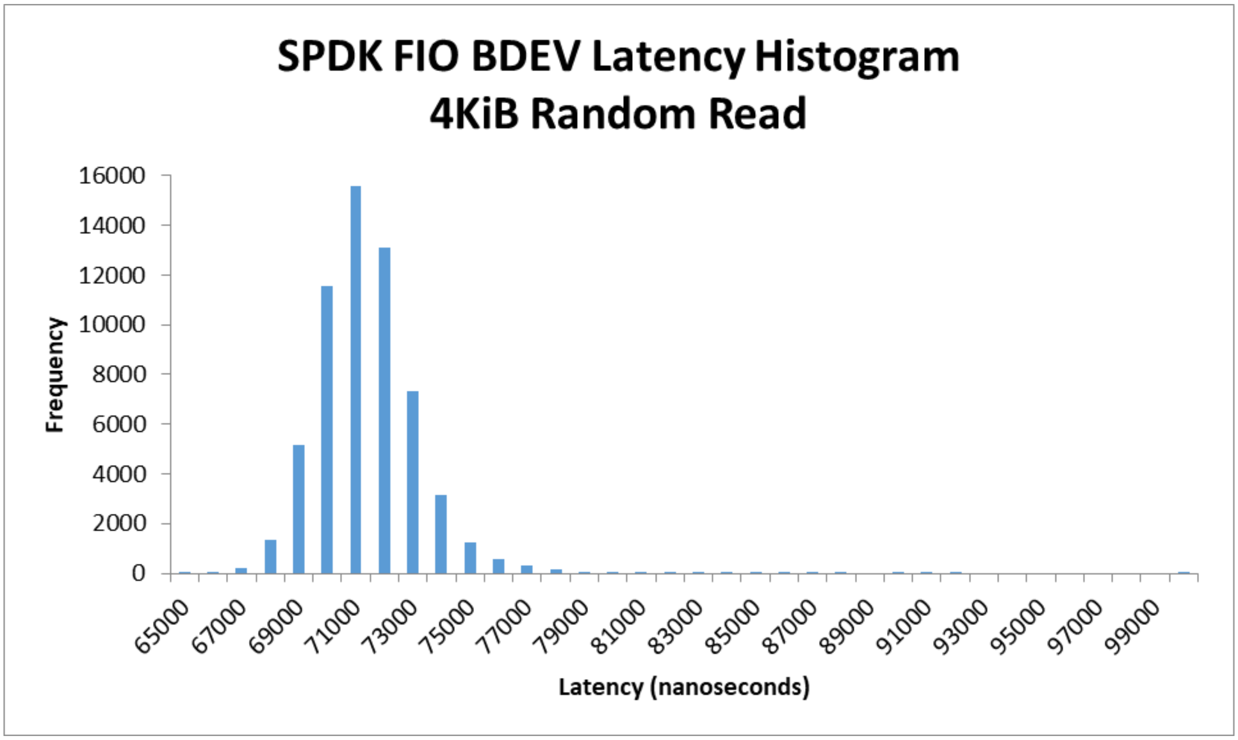


Figure 15: SPDK BDEV NVMe 4KiB Random Read Average Latency Histogram (QD=1, Runtime=900s, fio, sampling interval = 15msec)

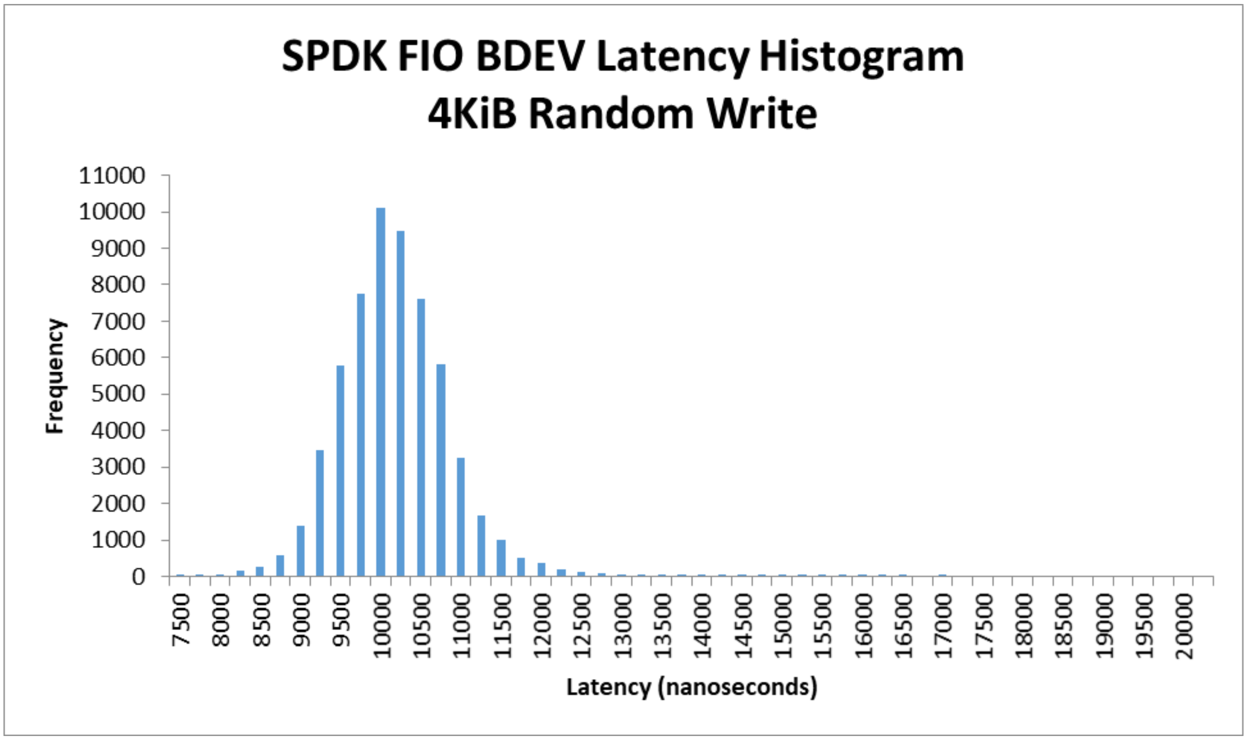


Figure 16: SPDK BDEV NVMe 4KiB Random Write Average Latency Histogram (QD=1, Runtime=900s, fio, sampling interval = 15msec)

Performance vs. increasing Queue Depth

Purpose: Understand the performance in IOPS and average latency of SPDK vs. the Linux io_uring polling and libaio block layer as the queue depth increases by powers of 2 from 1 to 512 for single NVMe SSD and single CPU Core.

Table 12: Performance at increasing Queue Depth; SPDK NVMe BDEV vs Linux Default libaio vs Linux io_uring polling (4KiB Random Read, 1 NVMe SSD, 1 CPU Core, Numjobs=1)

	SPDK		Linux Kernel (Default libaio)		Linux Kernel (io_uring polling)	
QD	**IOPS**	**Avg. Lat. (usec)**	**IOPS**	**Avg. Lat. (usec)**	**IOPS**	**Avg. Lat. (usec)**
1	13975	71	11383	86	13673	73
2	27716	72	23393	84	27178	73
4	54201	74	51276	78	53509	75
8	103576	77	99635	80	102366	78
16	189191	84	182079	88	187181	85
32	319195	100	307681	104	315736	101
64	476522	134	444757	144	469789	136
128	590937	216	464281	275	592706	216
256	621401	412	465092	550	626837	408
512	619421	826	465761	1099	629633	813

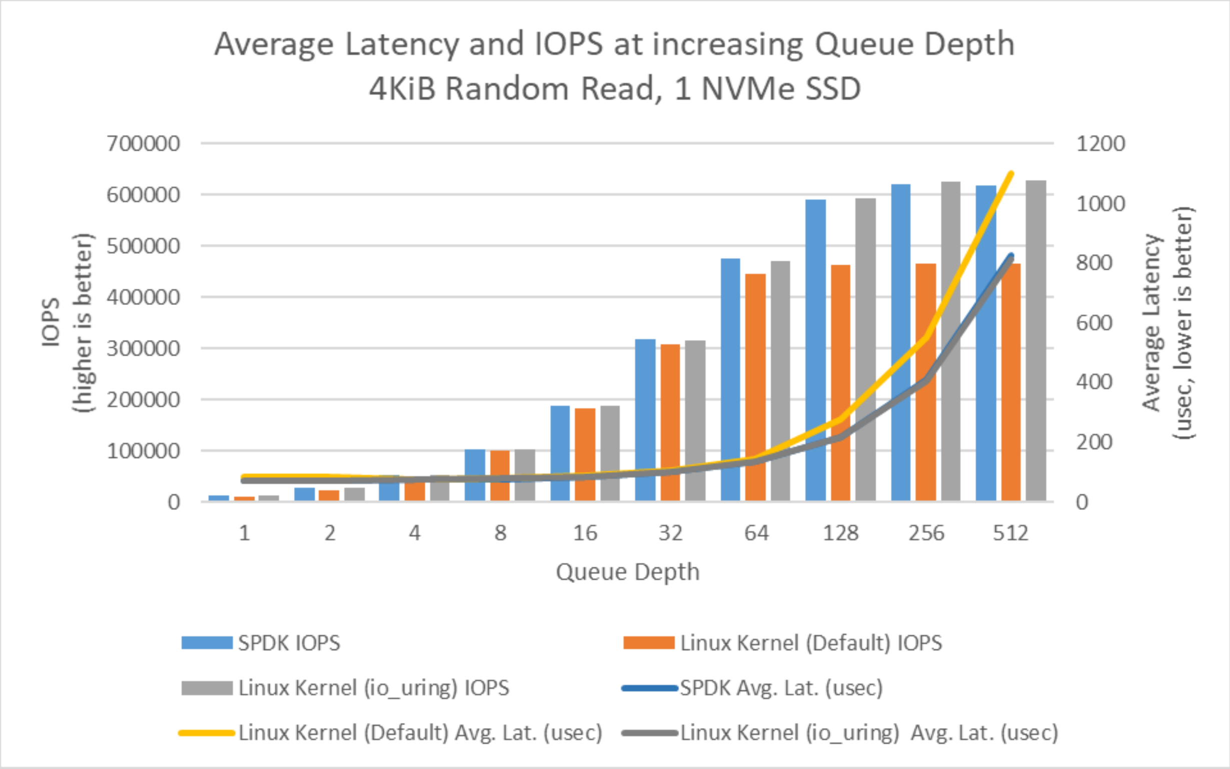


Figure 17: Performance at increasing Queue Depth; SPDK NVMe BDEV vs Linux Default libaio vs Linux io_uring polling (4KiB Random Read, 1 NVMe SSD, 1 CPU Core, Numjobs=1)

Table 13: Performance at increasing Queue Depth; SPDK NVMe BDEV vs Linux Default libaio vs Linux io_uring polling (4KiB Random Write, 1 NVMe SSD, 1 CPU Core, Numjobs=1)

	SPDK		Linux Kernel (Default libaio)		Linux Kernel (io_uring polling)	
QD	**IOPS**	**Avg. Lat. (usec)**	**IOPS**	**Avg. Lat. (usec)**	**IOPS**	**Avg. Lat. (usec)**
1	144739	7	107839	9	106618	9
2	257010	8	194733	10	189402	10
4	372443	11	306844	13	308352	13
8	452918	18	403138	20	440197	18
16	461116	35	434440	37	471167	34
32	458275	70	439656	73	467950	68
64	457938	140	441610	145	467575	137
128	454688	281	441453	290	462915	276
256	460050	557	440488	581	462688	554
512	445798	1149	426795	1200	454141	1128

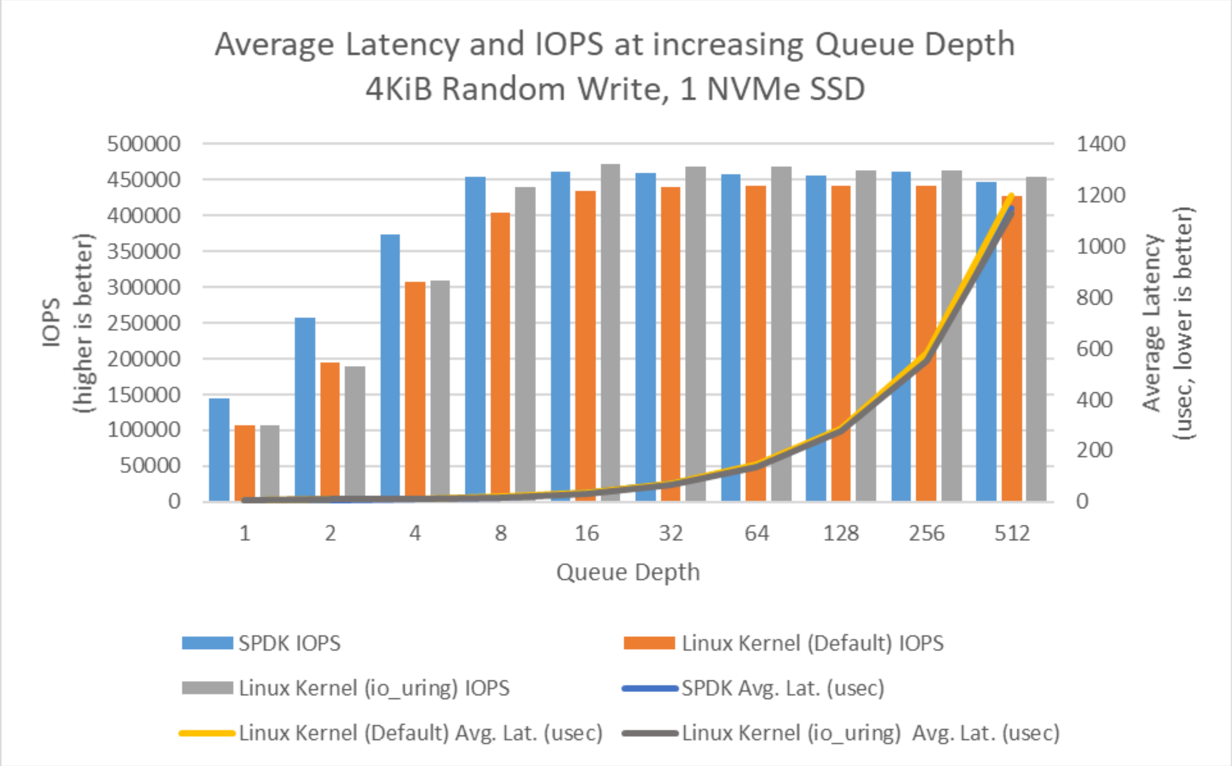


Figure 18: Performance at increasing Queue Depth; SPDK NVMe BDEV vs Linux Default libaio vs Linux io_uring polling (4KiB Random Write, 1 NVMe SSD, 1 CPU Core, Numjobs=1)

Conclusions

1. Polling hardware for completion instead of relying on interrupts lowers both total latency and its variance.

2. SPDK NVMe Bdev average latency was up to 17% and 14% lower than Linux Kernel Libaio, for Random Read and Random Write workloads respectively.

3. SPDK NVMe Bdev average latency was up to approximately 3% and 15% lower than Linux Kernel io_uring for Random Read and Random Write workloads respectively.

4. SPDK NVMe Bdev and Kernel io_uring IOPS throughput scaled almost linearly with increasing queue depth until QD=128 is reached for the 4KiB Random Read Workload. Beyond that point scaling became non-linear and increased until NVMe SSDs was saturated.

5. Kernel libaio IOPS throughput scaling was linear until QD=16 and the peak IOPS of approximately 465K IOPS was achieved at QD=256. We were unable to fully saturate NVMe SSD by increasing queue depth with just one CPU core. Both IOPS and latency results are worse than SPDK and io_uring.

Test Case 4: IOPS vs. Latency at different queue depths

Purpose: This test case was performed in order to understand throughput & latency trade-offs with varying queue depth while running SPDK vs. Kernel NVMe block layers.

Results in the table represent performance in IOPS and average latency for the SPDK and Linux Kernel NVMe block layers. We limited both the SPDK and Linux NVMe block layers to use the same number of CPU Cores.

Test Workloads: We use the following Random Read/Write mixes

- 4KiB 100% Random Read
- 4KiB 100% Random Write
- 4KiB Random 70% Read 30% Write

Table 14: SPDK NVMe BDEV Latency Test at different Queue Depths configuration

Item	Description
Test case	Test SPDK NVMe BDEV Latency Test at different Queue Depths
Test configuration	**FIO Version:** fio-3.19 **Number of CPU Cores:** 4 **Number of NVMe SSDs:** 24

Linux Kernel io_uring NVMe block device configuration	`echo 0 > /sys/block/nvme0n1/queue`
	`echo 0 > /sys/block/nvme0n1/rq_affinity`
	`echo 2 > /sys/block/nvme0n1/nomerges`
	`echo -1 > /sys/block/nvme0n1/io_poll_delay`

FIO configuration (common part)	`[global]`
	`direct=1`
	`thread=1`
	`time_based=1`
	`norandommap=1`
	`group_reporting=1`
	
	`rw={randread | randwrite | randrw}`
	`rwmixread={100 | 0 | 70}`
	`bs=4096`
	`runtime=240`
	`ramp_time=60`
	`numjobs=1`

FIO configuration (SPDK specific)	`[global]`
	`ioengine=spdk_bdev`
	`spdk_conf=/tmp/bdev.conf`

	[filename0]
	iodepth={6, 12, 24, 48, 96, 192, 512, 768}*
	cpus_allowed=0
	filename=Nvme0n1
	…
	filename=Nvme5n1
	
	[filename1]
	iodepth={6, 12, 24, 48, 96, 192, 512, 768}*
	cpus_allowed=21
	filename=Nvme6n1
	…
	filename=Nvme11n1
	
	[filename2]
	iodepth={6, 12, 24, 48, 96, 192, 512, 768}*
	cpus_allowed=22
	filename=Nvme12n1
	…
	filename=Nvme17n1
	
	[filename3]
	iodepth={6, 12, 24, 48, 96, 192, 512, 768}*
	cpus_allowed=23
	filename=Nvme18n1
	…
	filename=Nvme23n1
	
	* - - actual iodepth parameter value used in test; this was multiplied by the number of "filename" objects in job section to achieve desired queue depth value per NVMe SSD (e.g. QD=768 in this case is QD=128 per SSD)

FIO configuration (Linux Kernel common)	[global]
	ioengine={libaio | io_uring}
	
	[filename0]
	iodepth={6, 12, 24, 48, 96, 192, 512, 768}*
	cpus_allowed=21
	filename=/dev/nvme0n1
	…
	filename=/dev/nvme5n1
	
	[filename1]
	iodepth={6, 12, 24, 48, 96, 192, 512, 768}*
	cpus_allowed=22
	filename=/dev/nvme6n1
	…
	filename=/dev/nvme11n1
	
	[filename2]
	iodepth={6, 12, 24, 48, 96, 192, 512, 768}*
	cpus_allowed=0
	filename=/dev/nvme12n1
	…
	filename=/dev/nvme17n1
	
	[filename3]
	iodepth={6, 12, 24, 48, 96, 192, 512, 768}*

	cpus_allowed=23
	filename=/dev/nvme18n1
	…
	filename=/dev/nvme23n1
	
	* - actual iodepth parameter value used in test; this was multiplied by the number of "filename" objects in job section to achieve desired queue depth value **per SSD** (e.g. queue depth=768 in this case is queue depth=128 per SSD)

FIO configuration (Linux Kernel io_uring specific)	[global]
	fixedbufs=1
	hipri=1
	registerfiles=1
	sqthread_poll=1

4KiB Random Read Results

Table 15: Performance at increasing Queue Depth; SPDK NVMe BDEV vs Linux Default libaio vs Linux io_uring polling (4KiB Random Read, 24 NVMe SSDs, 4 CPU Cores)

	SPDK		Linux Kernel (Default libaio)		Linux Kernel (io_uring polling)	
QD	**IOPS (millions)**	**Avg. Lat. (usec)**	**IOPS (millions)**	**Avg. Lat. (usec)**	**IOPS (millions)**	**Avg. Lat. (usec)**
1	0.33	72	0.32	76	0.40	74
2	0.66	73	0.62	77	0.75	76
4	1.29	74	1.19	80	1.40	81
8	2.46	78	1.71	112	2.48	92
16	4.44	86	1.72	223	3.77	123
32	7.24	104	1.72	446	4.09	223
64	9.91	149	1.72	893	3.69	492
128	9.48	316	1.71	1800	3.43	1048

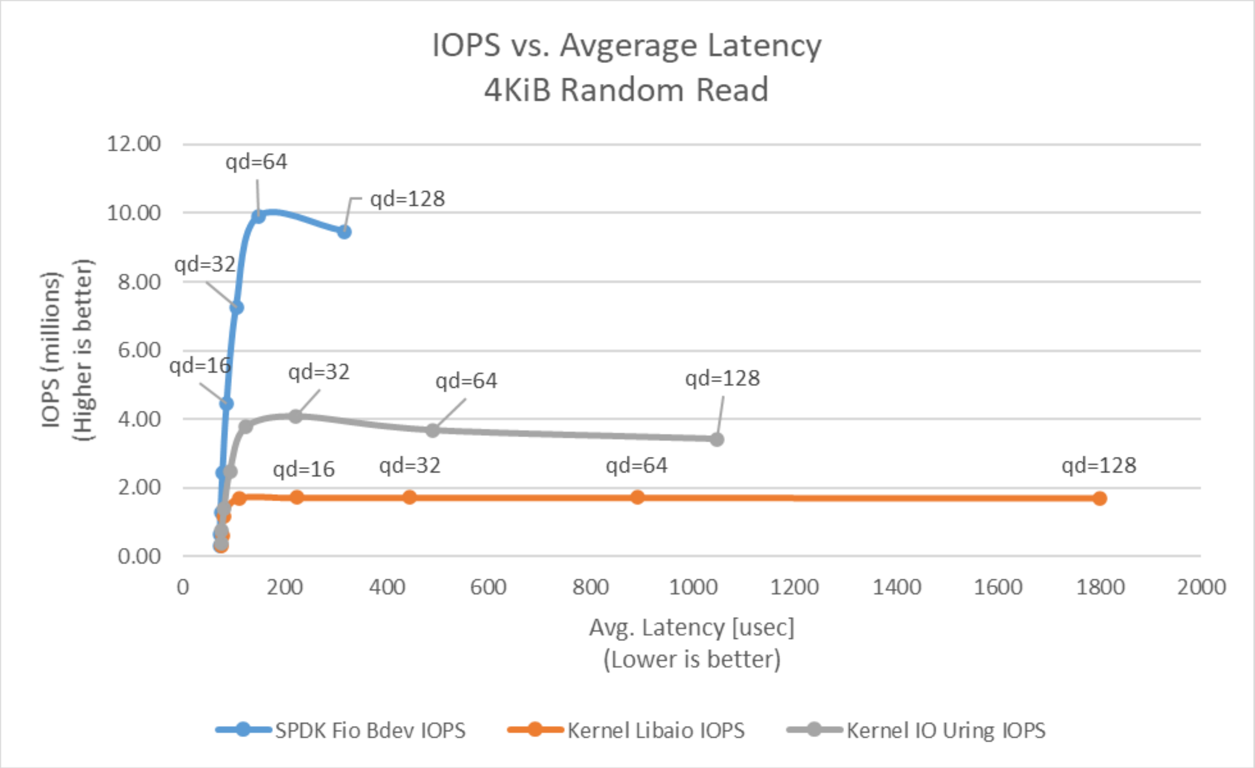


Figure 19: Performance at increasing Queue Depth; SPDK NVMe BDEV vs Linux Default libaio vs Linux io_uring polling (4KiB Random Read, 24 NVMe SSDs, 4 CPU Cores)

4KiB Random Write Results

Table 16: Performance at increasing Queue Depth; SPDK NVMe BDEV vs Linux Default libaio vs Linux io_uring polling (4KiB Random Write, 24 NVMe SSDs, 4 CPU Cores)

	SPDK		Linux Kernel (Default libaio)		Linux Kernel (io_uring polling)	
QD	IOPS (millions)	Avg. Lat. (usec)	IOPS (millions)	Avg. Lat. (usec)	IOPS (millions)	Avg. Lat. (usec)
1	2.92	8	1.64	14	2.21	4
2	4.53	10	1.67	28	2.79	7
4	5.79	15	1.66	57	3.35	11
8	7.20	23	1.65	116	4.28	18
16	8.09	37	1.64	234	4.67	33
32	7.16	82	1.61	476	4.39	72
64	7.28	162	1.55	989	3.62	177
128	5.60	500	1.48	2074	3.44	374

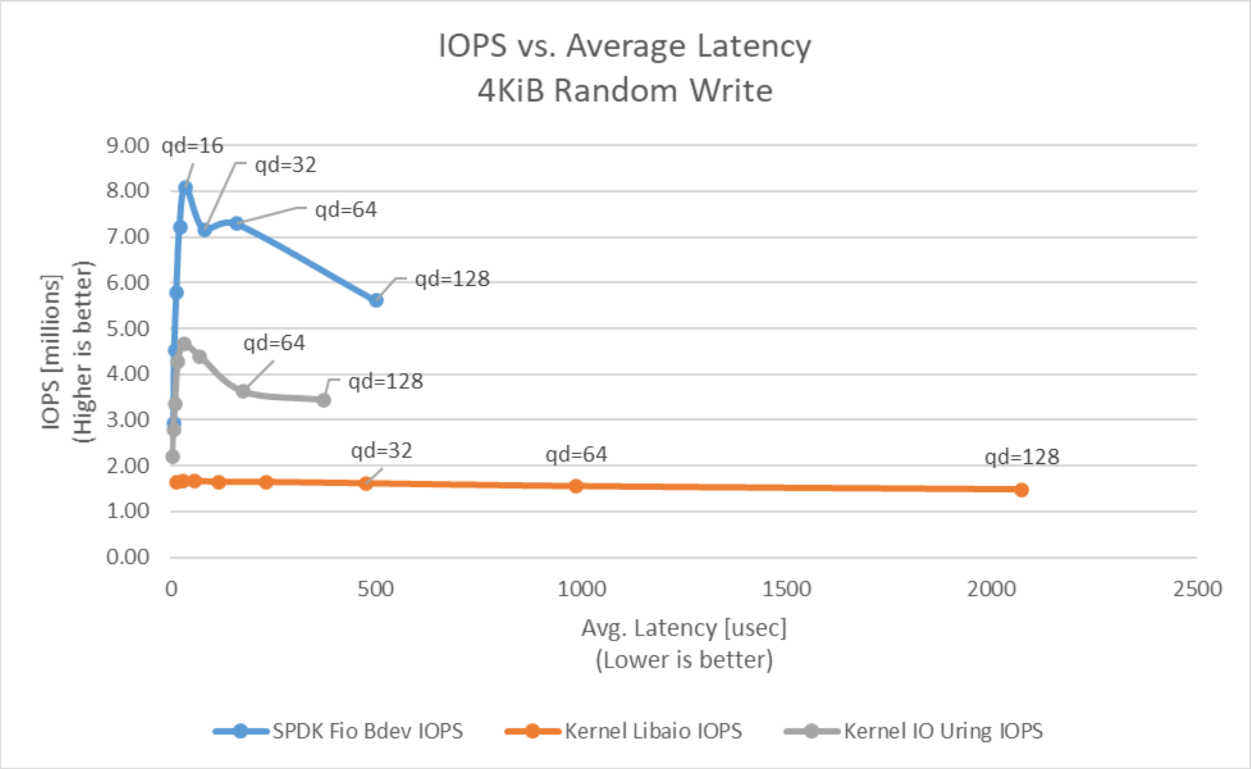


Figure 20: Performance at increasing Queue Depth; SPDK NVMe BDEV vs Linux Default libaio vs Linux io_uring polling (4KiB Random Write, 24 NVMe SSDs, 4 CPU Cores)

4KiB Random 70%/30% Read/Write Results

Table 17: Performance at increasing Queue Depth; SPDK NVMe BDEV vs Linux Default libaio vs Linux io_uring polling (4KiB 70/30 Random Read/Write, 24 NVMe SSDs, 4 CPU Cores)

	SPDK		Linux Kernel (Default libaio)		Linux Kernel (io_uring polling)	
QD	IOPS IOPS (millions)	Avg. Lat. (usec)	IOPS (millions)	Avg. Lat. (usec)	IOPS (millions)	Avg. Lat. (usec)
1	0.44	55	0.41	58	0.52	19
2	0.83	57	0.78	61	0.93	21
4	1.51	63	1.33	72	1.55	24
8	2.49	77	1.69	113	2.36	32
16	3.73	103	1.68	228	3.02	50
32	5.44	140	1.68	456	3.56	85
64	7.47	202	1.65	932	3.75	160
128	7.63	383	1.62	1898	3.27	364

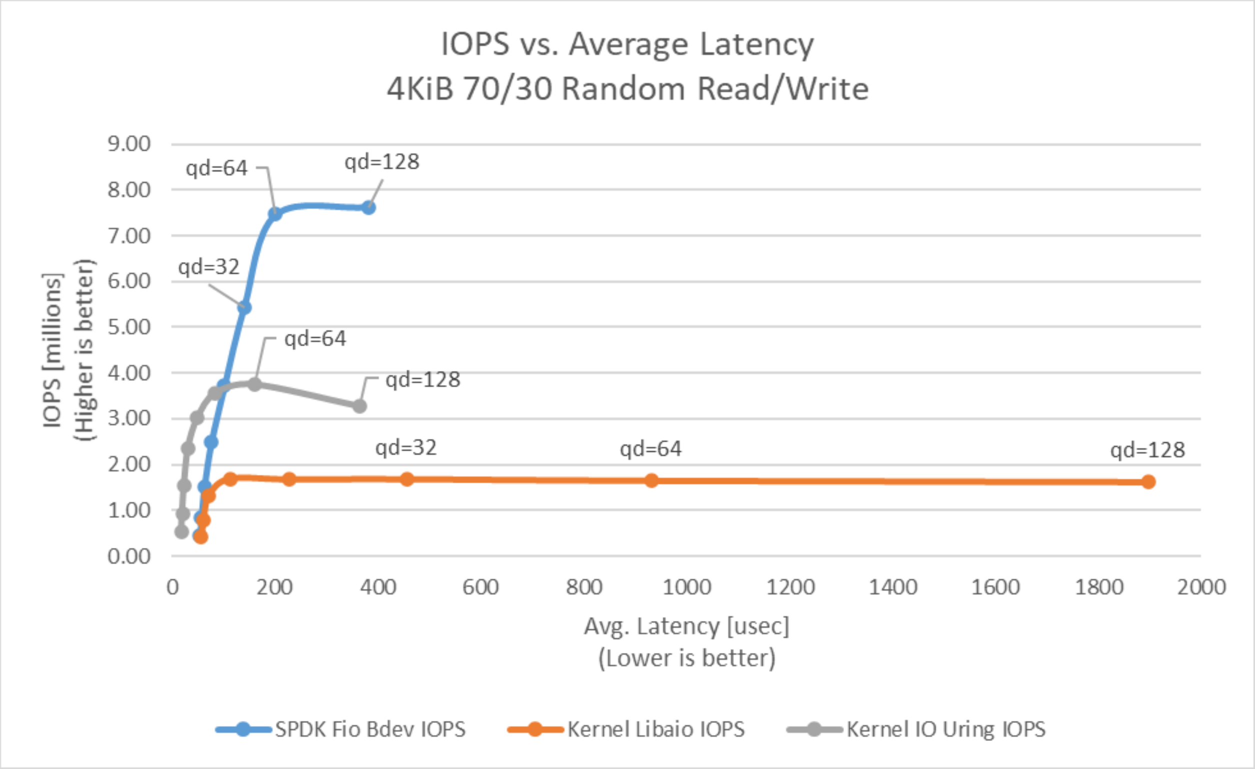


Figure 21: Performance at increasing Queue Depth; SPDK NVMe BDEV vs Linux Default libaio vs Linux io_uring polling (4KiB 70/30 Random Read/Write, 24 NVMe SSDs, 4 CPU Cores)

Conclusions

1. SPDK NVMe BDEV fio plugin reached up to around 9.9 million IOPS for Random Read workload at Queue Depth = 64. This is close to result measured in Test Case 2 - I/O Cores Scaling using Bdevperf.

2. SPDK NVMe BDEV fio plugin reached up to around 8.1 million IOPS for Random Read/Write workload at Queue Depth = 32, which is worse than what we measured in Test Case 2 - I/O Cores Scaling using Bdevperf.

3. The results for the Random Write workload exceeded what the platforms NVMe SSDs are capable of (around 4.8M IOPS). This is probably due to a not perfect preconditioning process, which wears off over time. However, these results were repeatable and still show SPDK's high scalability with increase in the I/O requests.

4. For Random Write workload (when running SPDK NVMe BDEV fio plugin) with increasing queue depth, after reaching peak performance there is a noticeable performance drop.

5. The Kernel libaio ioengine achieved maximum performance of up to 1.72M IOPS with 4 CPU cores and was unable to saturate platforms NVMe disks or PCIe switches throughput. Peak performance was reached at QD=8 for Random Read and Random Read/Write workloads and at QD=2 for Random Write workload. Beyond these queue depth values there was no IOPS improvement, but the latency increased.

6. The Kernel io_uring engine reached a performance peaks of 4.09 million IOPS at Queue Depth = 32 for Random Read workload, 4.67 million at QD = 16 for Random Write and 3.75 million at QD

= 128 for Random Read/Write workload. Similarly, to SPDK NVMe BDEV fio plugin, the performance starts to drop beyond these Queue Depth values. However, when we looked at htop we noticed that io_uring was using 8 CPU cores, because when we configured the sqthread_poll parameter to eliminate system calls io_uring starts a special kernel thread that polls the shared submission queue for I/O added by the fio thread. Therefore, in terms of CPU efficiency we measured up to 511K IOPS/Core for io_uring vs up to about 2.48M IOPS/Core for the SPDK NVMe bdev. The Submission Queue Polling blog provides more information about how to eliminate system calls with io_uring.

7. We were unable to run tests using Kernel io_uring ioengine with Queue Depth = 256 and 512 when we configured 6 NVMe SSDs per fio job. The reason for that was fio job configuration paired with limiting system settings. In this test 4 CPU cores were used, which in fio job configuration translates to 4 job sections, each with multiple “filename” arguments for target NVMe devices and an upscaled iodepth argument to match the number of devices. For example: a single job section was limited to a single cpu core using cpus_allowed argument; 6 NVMe devices were attached to this section using 6 “filename” arguments, and iodepth was set to iodepth=1536 (6*256). Queue depth of this value makes the test impossible to run because of fio “registerfiles” option (which is required to enable polling). When “registerfiles” is used the test fails because of the default UIO_MAXIOV limitation in sys/uio.h header file.

Summary

1. SPDK NVMe BDEV Block Layer using SPDK Bdevperf benchmarking tool can deliver up to 5.6 million IOPS on a single Intel® Xeon® Gold 6230N with Turbo Boost enabled.

2. The SPDK NVMe BDEV IOPS scale linearly with addition of CPU cores. We demonstrated up to 10.9 million IOPS on just 3 CPU cores (Intel® Xeon® Gold 6230N with Turbo Boost enabled).

3. The SPDK NVMe BDEV has lower QD=1 latency than the Linux Kernel NVMe block driver for small (4KiB) blocks.

    a. SPDK BDEV latency was 17% and 14% lower than Linux Kernel Libaio latency for Random Read and Random Write workloads.

    b. SPDK BDEV latency was about 3% lower than Linux Kernel io_uring latency for Random Read workload and 15% lower for Random Write workload.

4. SPDK NVMe Bdev Fio reaches up to 9.9 million IOPS with an average latency of around 150usec while using 4 CPU cores an queue depth of 64. With the same fio workloads Kernel io_uring and Kernel libaio reach up to 3.69 million (using 8 cores: 4 for fio and 4 for submission queue polling) and 1.72 million IOPS respectively.

List of tables

List of figures

References

[1] Damien Le Moal, “I/O Latency Optimization with Polling”, Vault – Linux Storage and Filesystem Conference – 2017, May 22nd, 2017.

Notices & Disclaimers

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates.

Your costs and results may vary.

No product or component can be absolutely secure.

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

§