SPDK NVMe-oF TCP (Target & Initiator) Performance Report Release 23.09

Intel E810-CQDA2 version

Testing Date: November 2023

Performed by:

Jaroslaw Chachulski (jaroslawx.chachulski@intel.com)

Karol Latecki (karol.latecki@intel.com)

Acknowledgments:

Krzysztof Karas (krzysztof.karas@intel.com)

Tomasz Zawadzki (tomasz.zawadzki@intel.com)

Konrad Sztyber (konrad.sztyber@intel.com)

Contents

Audience and Purpose

This report is intended for people who are interested in evaluating SPDK NVMe-oF (Target & Initiator) performance. This report contains SPDK NVMe-oF Target and Initiator performance characteristics and provides comparison data between SPDK and its Kernel NVMe-oF Target and Initiator counterparts. This report covers the TCP transport only.

The purpose of reporting these tests is not to imply a single “correct” approach, but rather to provide a baseline of well-tested configurations and procedures that produce repeatable results. This report can also be viewed as information regarding best known method/practice when performance testing SPDK NVMe-oF (Target & Initiator).

Test Setup

Target Configuration

Table 1: Hardware setup configuration – Target system

Item	Description
Server Platform	SuperMicro® Ultra SuperServer SYS-220U-TNR 
Motherboard	Server board **X12DPU-6**
CPU	2 CPU sockets, Intel(R) Xeon(R) Gold 6348 CPU @ 2.60GHz Number of cores 28 per socket, number of threads 56 per socket Both sockets populated Microcode: 0xd000389
Memory	16 x 32GB SK Hynix HMA84GR7DJR4N-XN, DDR4, 3200MHz Total of 512GB
Operating System	Fedora 37
BIOS	1.4b
Linux kernel version	6.0.18-300.fc37.x86_64 Spectre-meltdown mitigations enabled
Storage	**OS:** 1x 250GB Crucial CT250MX500SSD1 **Storage Target**: 14x Kioxia® KCM61VUL3T20 3.2TBs (FW: 0105) (6 on CPU NUMA Node 0, 8 on CPU NUMA Node 1)
NIC	4x 100GbE Intel(R) Ethernet Network Adapter E810-CQDA2. Single port connected on each NIC. ice driver version: 1.11.14 irdma driver version: 1.11.58 NVM FW version: v4.20 2 NICs per CPU socket.

Initiator 1 Configuration

Table 2: Hardware setup configuration – Initiator system 1

Item	Description
Server Platform	Intel® Server System M50CYP2UR208
CPU	Intel® Xeon® Gold 6348 Processor @ 2.60GHz (42MB Cache) Number of cores 28 per socket, number of threads 56 per socket (Both sockets populated) Microcode: 0xd000389
Memory	16 x 32GB Micron 36ASF4G72PZ-3G2J3, DDR4, 3200MHz Total 512GBs
Operating System	Fedora 37
BIOS	SE5C620.86B.01.01.0007.2210270543
Linux kernel version	6.0.18-300.fc37.x86_64 Spectre-meltdown mitigations enabled
Storage	**OS:** 1x 250GB Crucial CT250MX500SSD1
NIC	2x 100GbE Intel(R) Ethernet Network Adapter E810-CQDA2. Single port connected on each NIC. ice driver version: 1.11.14 Irdma driver version: 1.11.58 NVM FW version: v4.20

Initiator 2 Configuration

Table 3: Hardware setup configuration – Initiator system 2

Item	Description
Server Platform	Intel® Server System M50CYP2UR208
CPU	Intel® Xeon® Gold 6348 Processor @ 2.60GHz (42MB Cache) Number of cores 28 per socket, number of threads 56 per socket (Both sockets populated) Microcode: 0xd000389
Memory	16 x 32GB Micron 36ASF4G72PZ-3G2J3, DDR4, 3200MHz Total 512GBs
Operating System	Fedora 37
BIOS	SE5C620.86B.01.01.0007.2210270543
Linux kernel version	6.0.18-300.fc37.x86_64 Spectre-meltdown mitigations enabled
Storage	**OS:** 1x 250GB Crucial CT250MX500SSD1
NIC	2x 100GbE Intel(R) Ethernet Network Adapter E810-CQDA2. Single port connected on each NIC. ice driver version: 1.11.14 Irdma driver version: 1.11.58 NVM FW version: v4.20

BIOS settings

Table 4: Test systems BIOS settings

Item	Description
BIOS ***(Applied to all 3 systems)***	Hyper threading Enabled CPU Power and Performance Policy: • "Extreme Performance" for Target • "Performance" for Initiators CPU C-state No Limit CPU P-state Enabled Enhanced Intel® SpeedStep® Tech Enabled Turbo Boost Enabled

Software

The tests and evaluations conducted in this report are based on the Storage Performance Development Kit (SPDK) version 23.09.

For I/O benchmarking tests, fio version 3.28 was utilized.

SPDK Build Options

All measurements included in this report document were done with SPDK build with "--enable-lto" option enabled. Link time optimization allows better SPDK performance thanks to code optimization done by inlining functions across compilation units, which in turn results in reduced function call overhead.

TCP configuration

Note that the SPDK NVMe-oF target and initiator use the Linux Kernel TCP stack. We tuned the Linux Kernel TCP stack for storage workloads over 100 Gbps NIC by settings the following parameters using sysctl:

```
net.core.busy_poll = 0
net.core.busy_read = 0
net.core.somaxconn = 4096
net.core.netdev_max_backlog = 8192
net.ipv4.tcp_max_syn_backlog = 16384
net.core.rmem_max = 268435456
net.core.wmem_max= 268435456
net.ipv4.tcp_mem = "268435456 268435456 268435456"
net.ipv4.tcp_rmem = "8192 1048576 33554432"
net.ipv4.tcp_wmem = "8192 1048576 33554432"
net.ipv4.route.flush = 1
vm.overcommit_memory = 1
```

SPDK iobufs

SPDK introduced a common pool of buffers to be used across libraries in SPDK called "iobuf". Over time more components are being converted to share the "iobufs". The default counts of elements are defined

at values common for some use cases. Depending on the test scenario those values might need to be increased via "iobuf_set_options" RPC. Please see "spdk/scripts/calc-iobuf.py" for guidance on minimum “iobuf” pool sizes.

Introduction to SPDK NVMe-oF (Target & Initiator)

The NVMe over Fabrics (NVMe-oF) protocol extends the parallelism and efficiencies of the NVM Express* (NVMe) block protocol over network fabrics such as RDMA (iWARP, RoCE, InfiniBand™), Fibre Channel and TCP. SPDK provides both a user-space NVMe-oF target and initiator that extends the software efficiencies of the rest of the SPDK stack over the network. The SPDK NVMe-oF target uses the SPDK user-space, polled-mode NVMe driver to submit and complete I/O requests to NVMe devices which reduces the software processing overhead. Likewise, it pins connections to CPU cores to avoid synchronization and cache thrashing so that the data for those connections is kept close to the CPU.

The SPDK NVMe-oF target and initiator use the underlying transport layer API which in case of TCP are POSIX sockets. Similar to the SPDK NVMe driver, SPDK provides a user-space, lockless, polled-mode NVMe-oF initiator. The host system uses the initiator to establish a connection and submit I/O requests to an NVMe subsystem within an NVMe-oF target. NVMe subsystems contain namespaces, each of which maps to a single block device exposed via SPDK's bdev layer. SPDK's bdev layer is a block device abstraction layer and general-purpose block storage stack akin to what is found in many operating systems. Using the bdev interface completely decouples the storage media from the front-end protocol used to access storage. Users can build their own virtual bdevs that provide complex storage services and integrate them with the SPDK NVMe-oF target with no additional code changes. There can be many subsystems within an NVMe-oF target and each subsystem may hold many namespaces. Subsystems and namespaces can be configured dynamically via a JSON-RPC interface.

Figure 1 shows a high-level schematic of the systems used for testing in the rest of this report. The set up consists of three systems (two used as initiators and one used as the target). The NVMe-oF target is connected to both initiator systems point-to-point using QSFP28 cables without any switches. The target system has fourteen Kioxia® KCM61VUL3T20 SSDs which were used as block devices for NVMe-oF subsystems and four 100GbE Intel® E810-CQDA2 NICs connected to provide up to 400GbE of network bandwidth. Each Initiator system has two Intel® E810-CQDA2100GbE NICs connected directly to the target without any switch.

One goal of this report was to make clear the advantages and disadvantages inherent to the design of the SPDK NVMe-oF components. These components are written using techniques such as run-to completion, polling and asynchronous I/O. The report covers four real-world use cases.

For performance benchmarking the fio tool is used with two storage engines:

1) Linux Kernel io_uring engine
2) SPDK bdev engine

Performance numbers reported are aggregate I/O per second, average latency, and CPU utilization as a percentage for various scenarios. Aggregate I/O per second and average latency data is reported from fio and CPU utilization was collected using sar (systat).

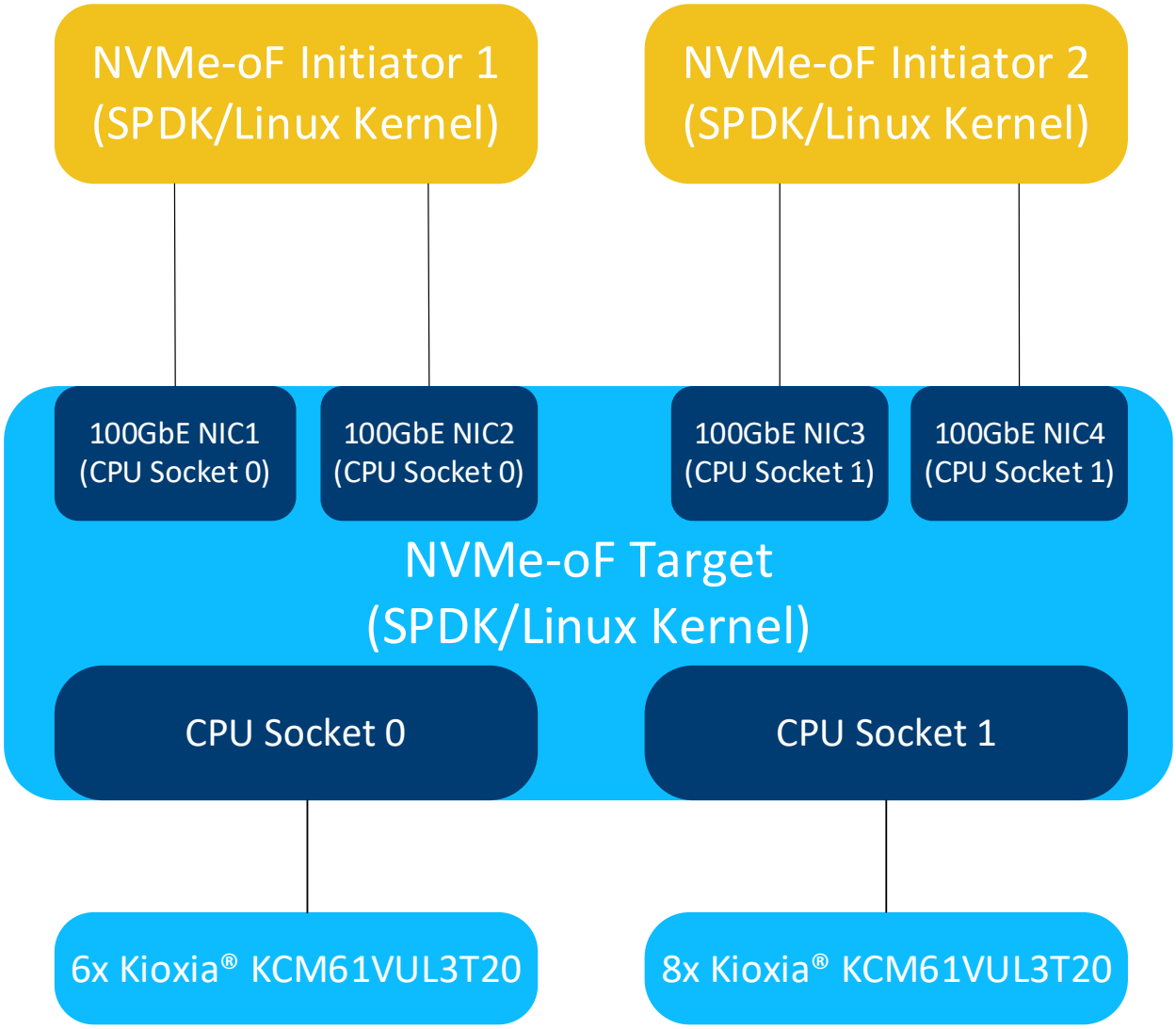

Figure 1: High-Level NVMe-oF TCP performance testing setup

Test Case 1: SPDK NVMe-oF TCP Target I/O core scaling

This test case was performed in order to understand the performance of SPDK TCP NVMe-oF target with I/O core scaling.

The SPDK NVMe-oF TCP target was configured to run with 14 NVMe-oF subsystems. Each NVMe-oF subsystem ran on top of an individual NVMe bdev backed by a single Kioxia KCM61VUL3T20 NVMe drive. Each of the 2 host systems was connected to 7 NVMe-oF subsystems which were exported by the SPDK NVMe-oF Target over 2 x 100GbE NIC. The SPDK bdev fio plugin was used to target 7 NVMe-oF bdevs on each of the host. The SPDK Target was configured to use 1, 4, 8, 12, 16, 24, 32, 40 and 48 CPU cores. We ran the following workloads on each initiator:

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

We scaled the fio jobs using fio parameter numjobs=4 in order to generate more I/O requests. When using the SPDK fio plugin it is important to note the difference between the fio I/O depth parameter and the NVMe device I/O depth because we can configure an fio job to send I/Os to more than one NVMe device and we can also scale the number of fio jobs using the numjobs parameter. The parameter values presented in the table below are actual queue depths used for each of the NVMe devices specified by the filename. These values were calculated in test based on number of fio job sections, numjobs parameter and the number of "filename" targets grouped in each of the fio job sections.

For detailed configuration please refer to the table below. The actual SPDK NVMe-oF configuration was done using JSON-RPC and the table contains the sequence of commands used by spdk/scripts/rpc.py script rather than a configuration file. The SPDK NVMe-oF Initiator (bdev fio_plugin) still uses plain configuration files.

Each workload was run three times at each CPU count and the reported results are the average of the 3 runs. For workloads which need preconditioning, 4KiB Random Read and 4KiB Random 70%/30% Read /Write we ran preconditioning once before running all of the workload to force the NVMe devices into a steady state so that we get consistent results.

Table 5: SPDK NVMe-oF TCP Target Core Scaling test configuration

Item	Description
Test Case	Test SPDK NVMe-oF Target I/O core scaling
SPDK NVMe-oF Target configuration	All the commands below were executed with spdk/scripts/rpc.py script. **Set iobuf buffer pool options:** iobuf_set_options --small-pool-count 32767 --large-pool-count 16383 **Enable zero-copy send on Target side before initializing all other subsystems.** sock_impl_set_options –impl=posix –enable-zerocopy-send-server (note: zerocopy for Client side is disabled by default)

	Construct NVMe bdevs:
	bdev_nvme_attach_controller -t PCIe -b Nvme0 -a 0000:17:00.0
	bdev_nvme_attach_controller -t PCIe -b Nvme1 -a 0000:18:00.0
	bdev_nvme_attach_controller -t PCIe -b Nvme2 -a 0000:65:00.0
	bdev_nvme_attach_controller -t PCIe -b Nvme3 -a 0000:66:00.0
	bdev_nvme_attach_controller -t PCIe -b Nvme4 -a 0000:67:00.0
	bdev_nvme_attach_controller -t PCIe -b Nvme5 -a 0000:68:00.0
	bdev_nvme_attach_controller -t PCIe -b Nvme6 -a 0000:98:00.0
	bdev_nvme_attach_controller -t PCIe -b Nvme7 -a 0000:99:00.0
	bdev_nvme_attach_controller -t PCIe -b Nvme8 -a 0000:9a:00.0
	bdev_nvme_attach_controller -t PCIe -b Nvme9 -a 0000:9b:00.0
	bdev_nvme_attach_controller -t PCIe -b Nvme10 -a 0000:e3:00.0
	bdev_nvme_attach_controller -t PCIe -b Nvme11 -a 0000:e4:00.0
	bdev_nvme_attach_controller -t PCIe -b Nvme12 -a 0000:e5:00.0
	bdev_nvme_attach_controller -t PCIe -b Nvme13 -a 0000:e6:00.0
	
	Create a TCP transport:
	nvmf_create_transport -t TCP
	{
	"trtype": "TCP",
	"max_queue_depth": 128,
	"max_io_qpairs_per_ctrlr": 127,
	"in_capsule_data_size": 4096,
	"max_io_size": 131072,
	"io_unit_size": 131072,
	"max_aq_depth": 128,
	"num_shared_buffers": 8192,
	"buf_cache_size": 32,
	"dif_insert_or_strip": false,
	"c2h_success": true,
	"sock_priority": 0,
	"abort_timeout_sec": 1
	}
	
	Create NVMe-oF subsystems and add NVMe bdevs as namespaces:
	for i in $(seq 1 14); do
	nvmf_subsystem_create nqn.2018-09.io.spdk:cnode${i} -s SPDK00${i} -a -m 8
	nvmf_subsystem_add_ns nqn.2018-09.io.spdk:cnode${i} Nvme$((i-1))n1
	done
	
	Add listeners to NVMe-oF Subsystems:
	i=1
	ips=(20.0.0.1 20.0.1.1 10.0.0.1 10.0.1.1)
	for ip in ${ips[@]}; do
	for j in $(seq 1 4); do
	nvmf_subsystem_add_listener nqn.2018-09.io.spdk:cnode${i} -t tcp \
	-f ipv4 -s 4420 -a ${ip}
	((i++))
	done
	done

SPDK NVMe-oF Initiator - fio plugin configuration	**BDEV.conf:** See appendix A. **fio.conf** [global] ioengine=/tmp/spdk/examples/bdev/fio_plugin/fio_plugin spdk_conf=/tmp/spdk/bdev.conf thread=1 group_reporting=1 direct=1

	norandommap=1 rw=randrw rwmixread={100, 70, 0} bs=4k iodepth={128, 384} time_based=1 numjobs=4 ramp_time=60 runtime=300 [filename0] filename=Nvme0n1 [filename1] filename=Nvme1n1 [filename2] filename=Nvme2n1 [filename3] filename=Nvme3n1 [filename4] filename=Nvme4n1 [filename5] filename=Nvme5n1 [filename6] filename=Nvme6n1

4KiB Random Read Results

Table 6: SPDK NVMe-oF TCP Target Core Scaling results, Random Read IOPS, QD=384

# of Cores	Throughput (IOPS k)	Bandwidth (Gbps)	Avg. Latency (usec)
1 core	673.1	22.06	7979.2
4 cores	2503.0	82.02	2165.2
8 cores	5910.8	193.69	902.1
12 cores	9028.9	295.86	584.6
16 cores	9811.2	321.49	536.9
24 cores	10523.6	344.84	499.9
32 cores	10652.5	349.06	492.7
40 cores	10728.2	351.54	489.6
48 cores	10680.5	349.98	491.9

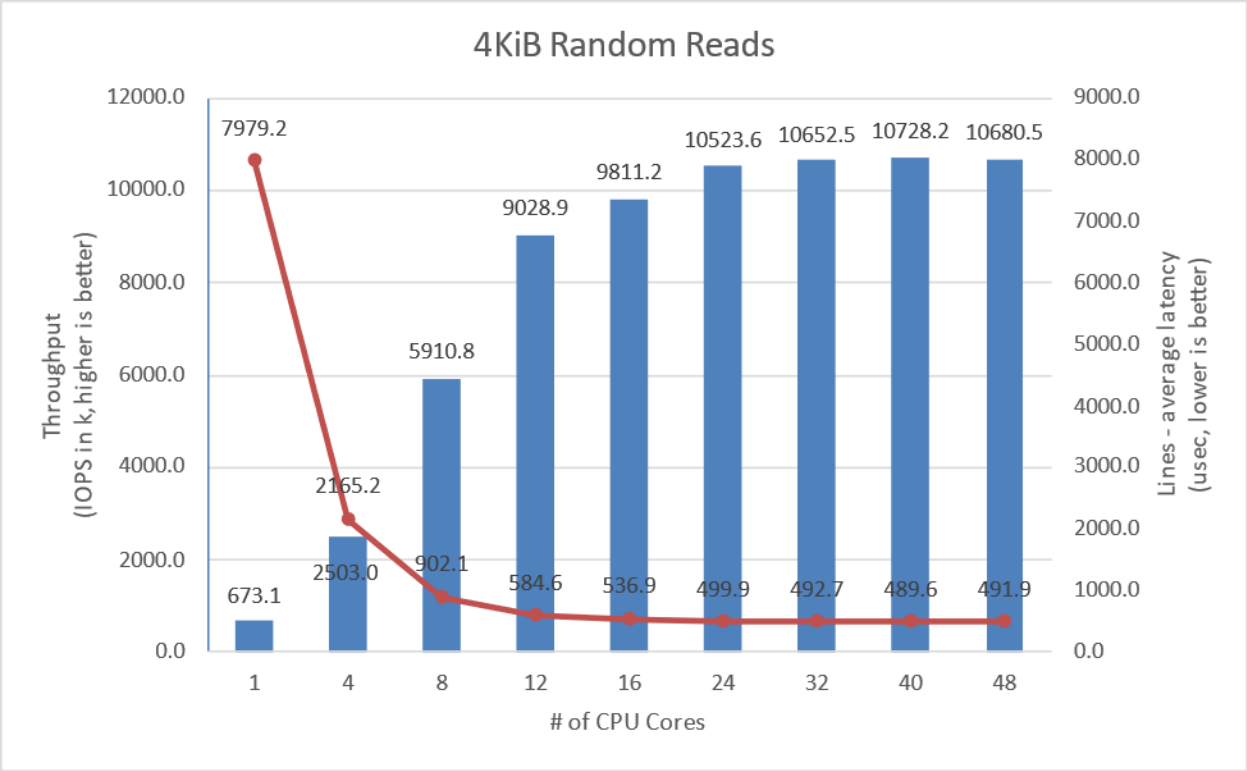

Figure 2: SPDK NVMe-oF TCP Target I/O core scaling: IOPS vs. Latency while running 4KiB 100% Random Read workload at QD = 384

4KiB Random Write Results

Note that the SSDs were not preconditioned for the 4K random write workload because that would limit the workload performance to the SSDs steady state performance.

Table 7: SPDK NVMe-oF TCP Target Core Scaling results, Random Write IOPS, QD=128

# of Cores	Throughput (IOPS k)	Bandwidth (Gbps)	Avg. Latency (usec)
1 core	372.0	12.19	4830.6
4 cores	1528.8	50.09	1170.4
8 cores	2825.0	92.57	631.3
12 cores	3815.7	125.03	466.4
16 cores	4521.4	148.16	392.6
24 cores	5257.1	172.27	337.7
32 cores	5457.5	178.83	324.8
40 cores	5530.9	181.24	320.9
48 cores	5546.3	181.74	320.0

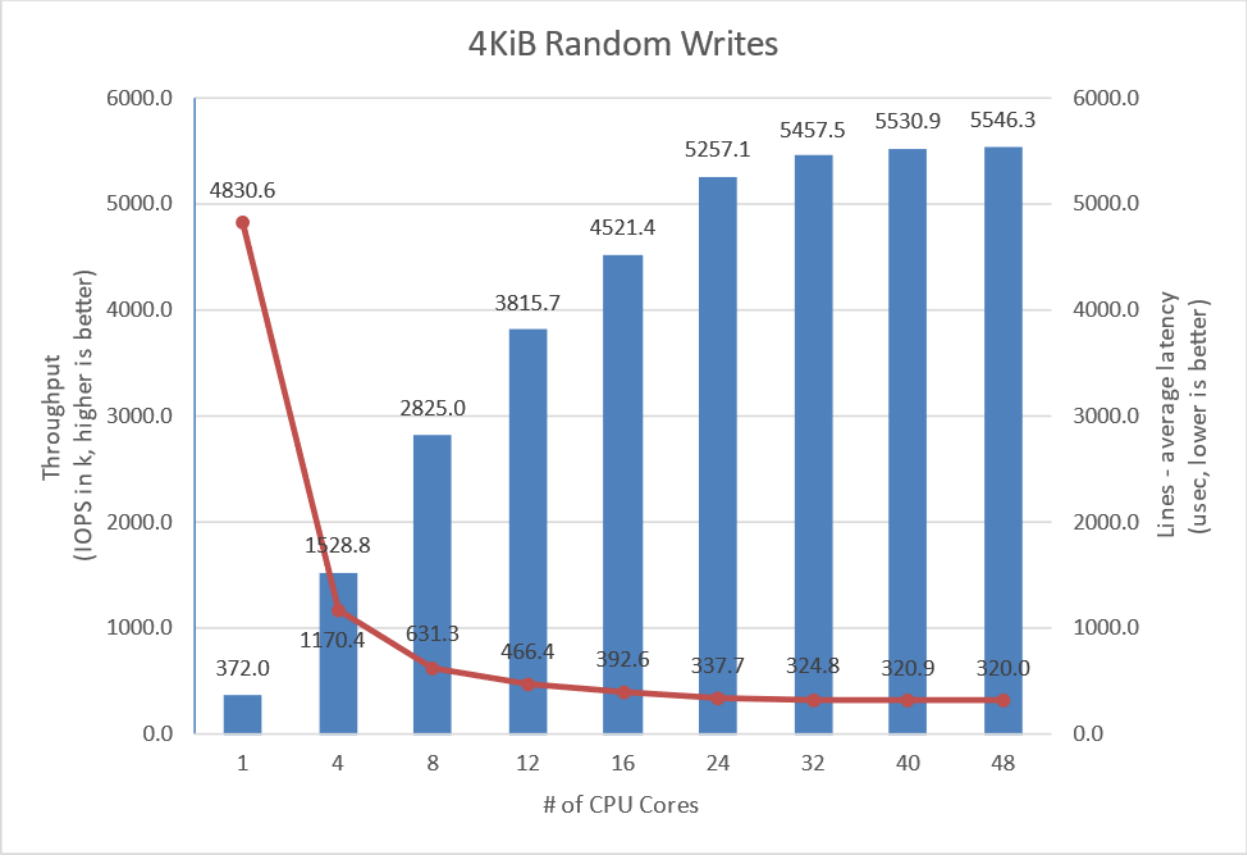

Figure 3: SPDK NVMe-oF TCP Target I/O core scaling: IOPS vs. Latency while running 4KiB 100% Random Write Workload at QD=128

4KiB Random Read-Write Results

Table 8: SPDK NVMe-oF TCP Target Core Scaling results, Random Read/Write 70%/30% IOPS, QD=384

# of Cores	Throughput (IOPS k)	Bandwidth (Gbps)	Avg. Latency (usec)
1 core	459.0	15.04	11707.7
4 cores	1939.7	63.56	2768.2
8 cores	4065.7	133.23	1318.6
12 cores	5819.8	190.70	920.6
16 cores	7474.7	244.93	715.5
24 cores	8783.2	287.81	608.5
32 cores	9728.3	318.78	548.7
40 cores	9629.3	315.53	554.5
48 cores	9839.3	322.41	542.4

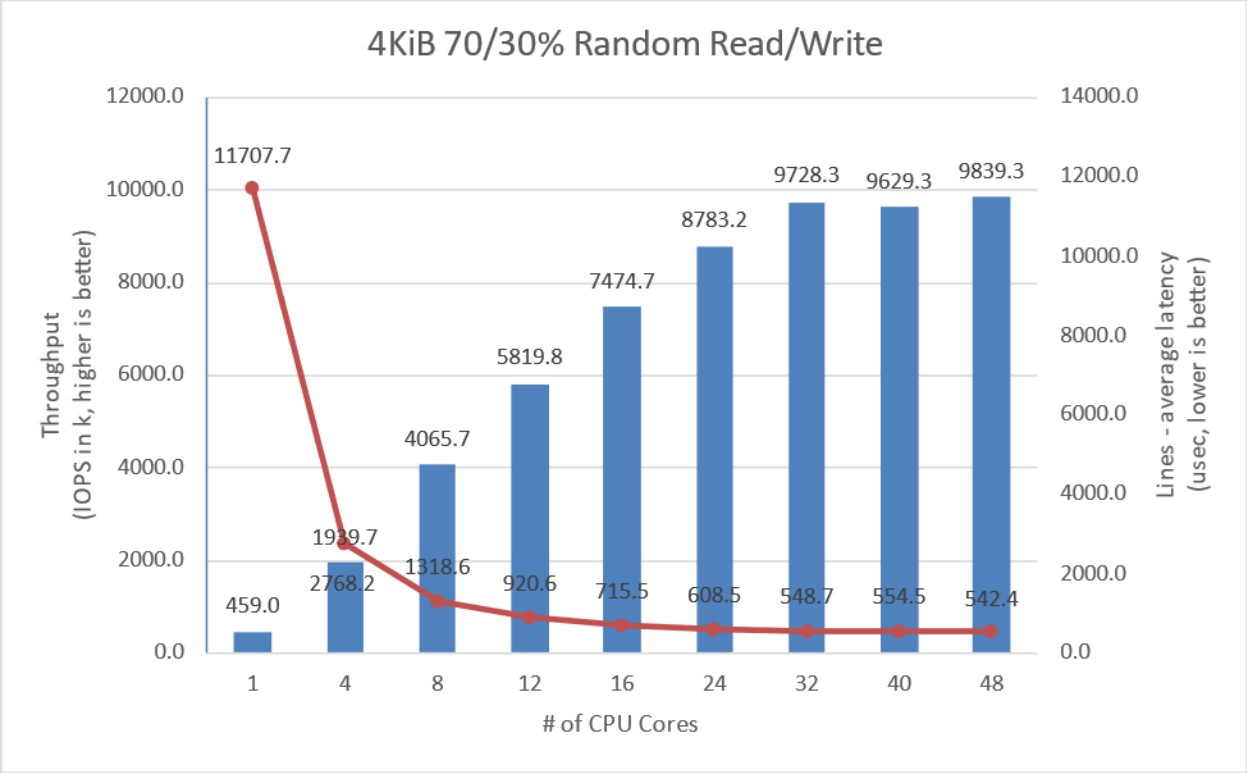

Figure 4: SPDK NVMe-oF TCP Target I/O core scaling: IOPS vs. Latency while running 4KiB Random 70/30 Read/Write workload at QD=384

Large Sequential I/O Performance

We measured the performance of large block I/O workloads by performing sequential I/Os of size 128KiBs at queue depth 16 We used iodepth=16 because higher queue depth resulted in negligible bandwidth gain and a significant increase in the latency. The rest of the fio configuration is similar to the 4KiB test case in the previous part of this document.

Table 9: SPDK NVMe-oF TCP Target Core Scaling results, 128KiB Sequential Read IOPS, QD=16

# of Cores	Throughput (IOPS k)	Bandwidth (Gbps)	Avg. Latency (usec)
1 core	149.6	156.87	1502.5
4 cores	357.2	374.50	626.4
8 cores	358.9	376.36	623.2
12 cores	358.9	376.37	623.2

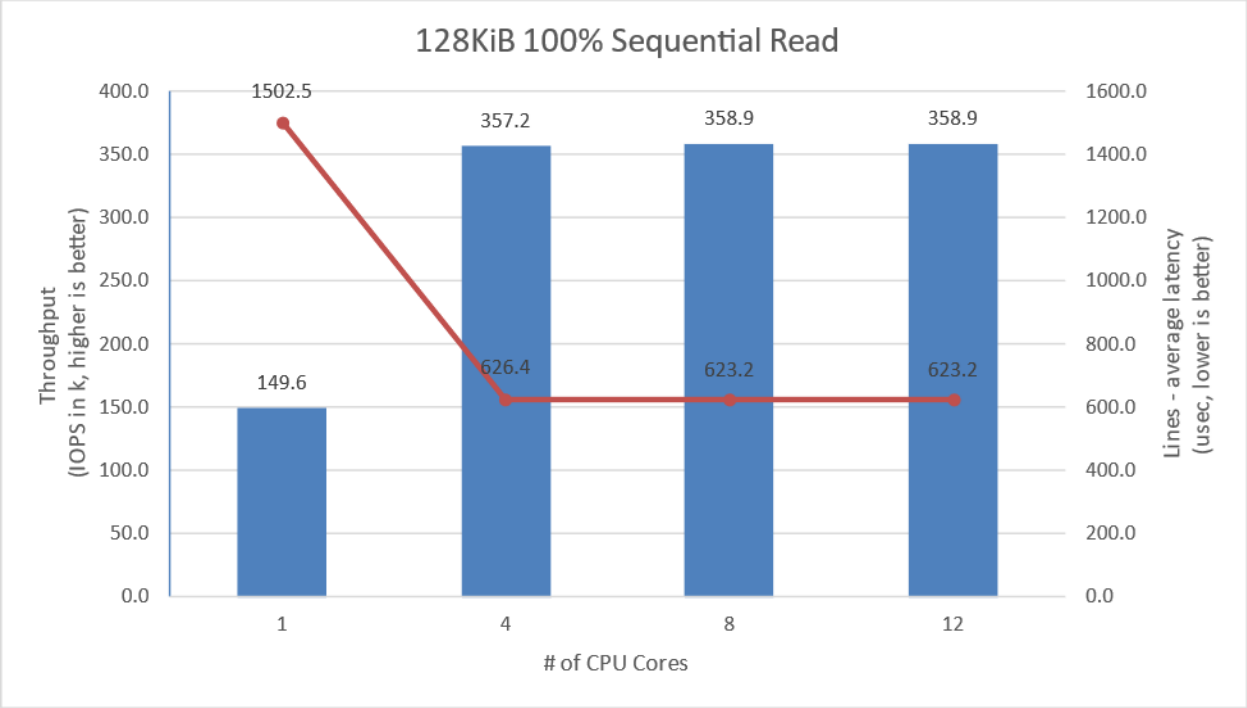

Figure 5: SPDK NVMe-oF TCP Target I/O core scaling: IOPS vs. Latency while running 128KiB 100% Sequential Read Workload at QD=16 and initiator fio numjobs=4

Table 10: SPDK NVMe-oF TCP Target Core Scaling results, 128KiB Sequential Write IOPS, QD=16

# of Cores	Throughput (IOPS k)	Bandwidth (Gbps)	Avg. Latency (usec)
1 core	23.5	24.61	9561.7
4 cores	85.8	89.95	2610.7
8 cores	150.4	157.74	1490.2
12 cores	188.5	197.62	1189.7
16 cores	217.6	228.12	1029.7
24 cores	248.2	260.27	908.9
32 cores	250.0	262.19	895.6
40 cores	266.0	278.96	843.5
48 cores	246.5	258.47	914.4

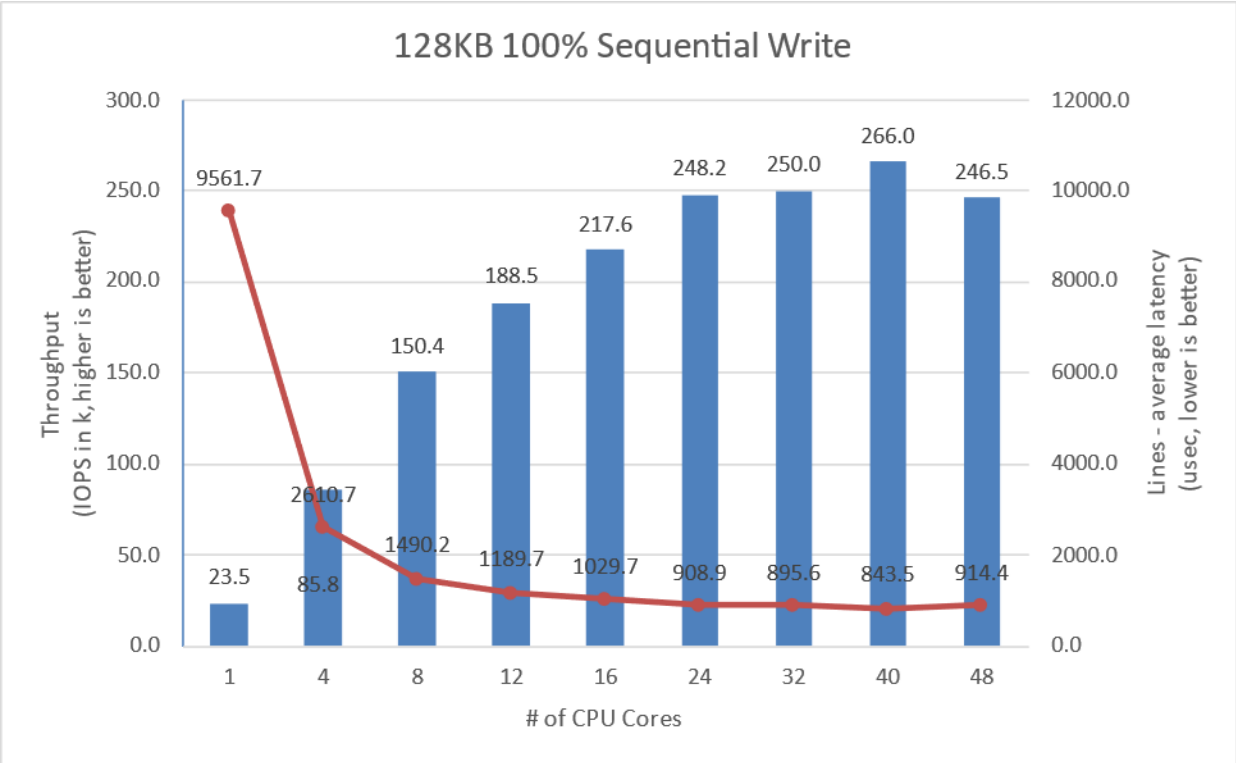

Figure 6: SPDK NVMe-oF TCP Target I/O core scaling: IOPS vs. Latency while running 128KiB 100% Sequential Write Workload at QD=16 and Initiator fio numjobs=4

Table 11: SPDK NVMe-oF TCP Target Core Scaling results, 128KiB Sequential 70% Read 30% Write IOPS, QD=16

# of Cores	Throughput (IOPS k)	Bandwidth (Gbps)	Avg. Latency (usec)
1 core	53.2	55.74	4223.5
4 cores	185.0	194.01	1210.2
8 cores	291.0	305.14	768.9
12 cores	339.3	355.75	660.4
16 cores	371.1	389.10	602.7
24 cores	389.5	408.47	574.2
32 cores	395.1	414.24	566.3

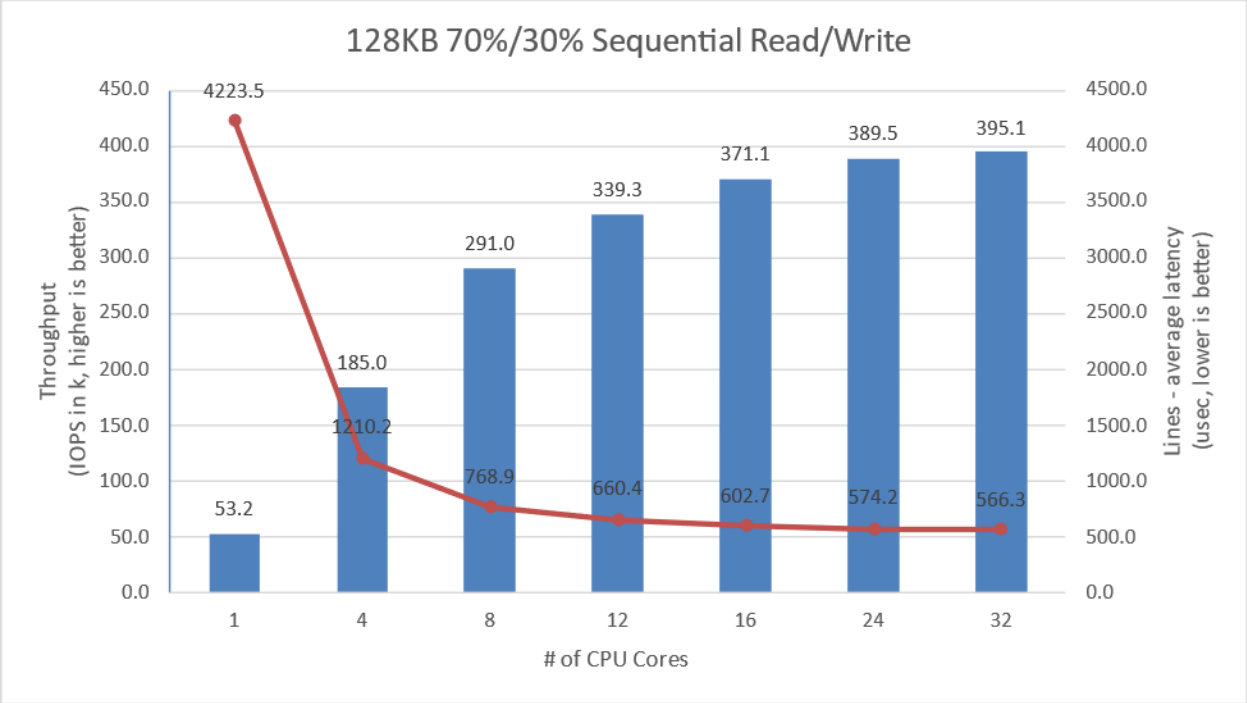

Figure 7: SPDK NVMe-oF TCP Target I/O core scaling: IOPS vs. Latency while running 128KiB Sequential 70% Read 30% Write Workload at QD=16 and Initiator fio numjobs=4

Conclusions

1. The SPDK NVMe-oF TCP Target IOPS throughput scales nonlinearly with addition of CPU cores for 4KiB Random Read workload up to 12 CPU cores, reaching 296 Gbps bandwidth and 9 million IOPS. Adding more CPUs to Target configuration results in non-linear performance gains peaking at about 10.7 million IOPS at 40 CPU cores, reaching 400GbE network link saturation.

2. For the 4KiB Random Write workload 100GbE link is saturated at 8 CPU cores. Performance scales up almost linearly to 12 CPU cores reaching 3.8 million IOPS. It further scales up to 40 CPU cores reaching 5.5 million IOPS.

3. 4KiB Random Read-Write workload throughput scaling is close to linear up to 24 CPU cores reaching 8.78 million IOPS. Increasing the number of CPU cores beyond 24 results in non-linear performance improvement that peaks at 9.8 million IOPS at 48 CPU cores.

4. The best trade-off between CPU efficiency and network saturation is when the Target is configured with 24CPU cores. The performance we achieved was near saturation of a 400Gbps link between Target and Initiators for all Random Read and Random Read/Write workloads, and 200Gbps link saturation for Random Write workload.

5. For the 4KiB Random Write workload, we did not precondition the NVMe drives because preconditioning, them would limit the IOPS throughput to a maximum of about 4.9 million IOPS. Not preconditioning the drives enabled us to serve more IO requests than usual.

6. The throughput of large block workloads scaled up with addition of CPU cores reaching peak performance at different CPU core counts. For the 128K Sequential Reads workload, the peak throughput of 357 Gbps was observed at 4 CPU cores. For the 128K Sequential Writes, the throughput scaled to 150 Gbps at 8 cores and reached peak performance of 266 Gbps at 40 CPU cores. For the 128K Sequential 70/30 Read/Write workload, the scaling was up to 8 CPU cores, reaching 305 Gbps. Beyond 12 cores the scaling was not linear, reaching beyond 400 Gbps due to bi-directional characteristic of the workload.

Test Case 2: SPDK NVMe-oF TCP Initiator I/O core scaling

This test case was performed in order to understand the performance of SPDK NVMe-oF TCP Initiator as the number of CPU cores is scaled up.

The test setup for this test case is slightly different than the set up described in introduction chapter, as we used just a single SPDK NVMe-oF TCP Initiator. The Initiator was connected to Target server with two 100 Gbps network links.

The SPDK NVMe-oF TCP Target was configured similarly as in test case 1, using 24 cores. We used 24 CPU cores based on results of the previous test case which show that the target can easily serve about 6 million IOPS for all workloads, which is enough IOPS to saturate 200 Gbps network connection.

The SPDK bdev fio plugin was used to target 14 individual NVMe-oF subsystems exported by the Target. The number of CPU threads used by the fio process was managed by setting the fio job sections and numjobs parameter and ranged from 1 to 48 CPUs. For detailed fio job configuration see table below. fio was run with following workloads:

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

It is important to note that fio IO depth parameter values presented in the table below are actual queue depths used for each of the connected subsystem. These values were calculated in test based on number of fio job sections, numjobs parameter and the number of "filename" targets grouped in each of the fio job sections.

Table 12: SPDK NVMe-oF TCP Initiator Core Scaling test configuration

Item	Description
Test Case	Test SPDK NVMe-oF TCP Initiator I/O core scaling
SPDK NVMe-oF Target configuration	Same as in Test Case #1, using 24 CPU cores.
SPDK NVMe-oF Initiator 1 - fio plugin configuration	**fio.conf** **For X*4 CPU (up to 48) initiator configuration:** [global] ioengine=/tmp/spdk/examples/bdev/fio_plugin/fio_plugin spdk_conf=/tmp/spdk/bdev.conf thread=1 group_reporting=1 direct=1 norandommap=1 rw=randrw rwmixread={100, 70, 0} bs=4k iodepth={128, 192, 256, 384}

	time_based=1 ramp_time=60 runtime=300 numjobs=X [filename0] filename=Nvme0n1 filename=Nvme1n1 filename=Nvme2n1 filename=Nvme3n1 [filename1] filename=Nvme4n1 filename=Nvme5n1 filename=Nvme6n1 filename=Nvme7n1 [filename2] filename=Nvme8n1 filename=Nvme9n1 filename=Nvme10n1 [filename3] filename=Nvme11n1 filename=Nvme12n1 filename=Nvme13n1

4KiB Random Read Results

Table 13: SPDK NVMe-oF TCP Initiator Core Scaling results, 4KiB Random Read IOPS, QD=256

# of Cores	Throughput (IOPS k)	Bandwidth (Gbps)	Avg. Latency (usec)
1 core	417.5	13.68	8563.1
4 cores	1542.2	50.53	2316.2
8 cores	2806.6	91.97	1268.2
12 cores	3966.4	129.97	890.7
16 cores	4681.8	153.41	751.0
24 cores	5595.2	183.34	627.3
32 cores	5694.0	186.58	618.2
40 cores	5680.0	186.12	620.5
48 cores	5689.2	186.42	621.0

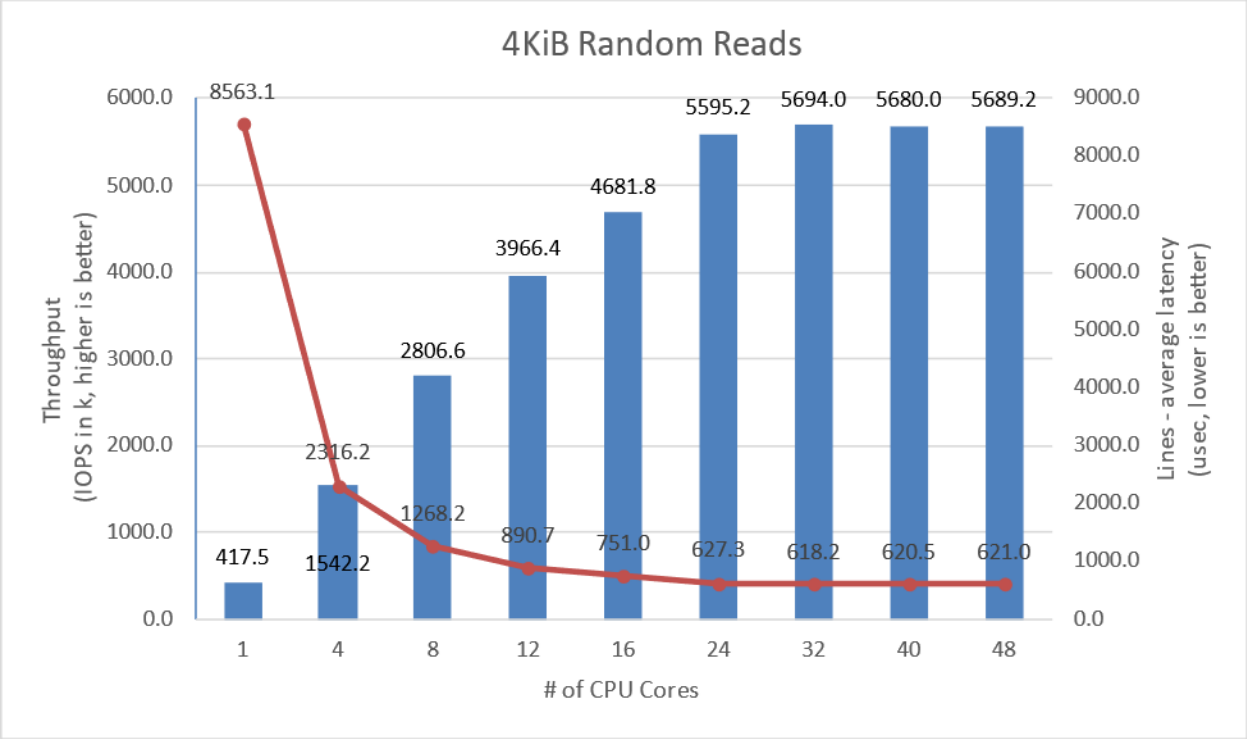

Figure 8: SPDK NVMe-oF TCP Initiator I/O core scaling: IOPS vs. Latency while running 4KiB 100% Random Read QD=256 workload

4KiB Random Write Results

Table 14: SPDK NVMe-oF TCP Initiator Core Scaling results, 4KiB Random Write IOPS, QD=128

# of Cores	Throughput (IOPS k)	Bandwidth (Gbps)	Avg. Latency (usec)
1 core	602.1	19.73	2088.4
4 cores	2863.6	93.83	553.9
8 cores	3660.6	119.95	473.6
12 cores	3869.7	126.80	457.1
16 cores	3693.5	121.03	478.9
24 cores	3627.2	118.86	489.5
32 cores	3573.2	117.09	499.3
40 cores	3428.9	112.36	517.2
48 cores	3400.2	111.42	527.7

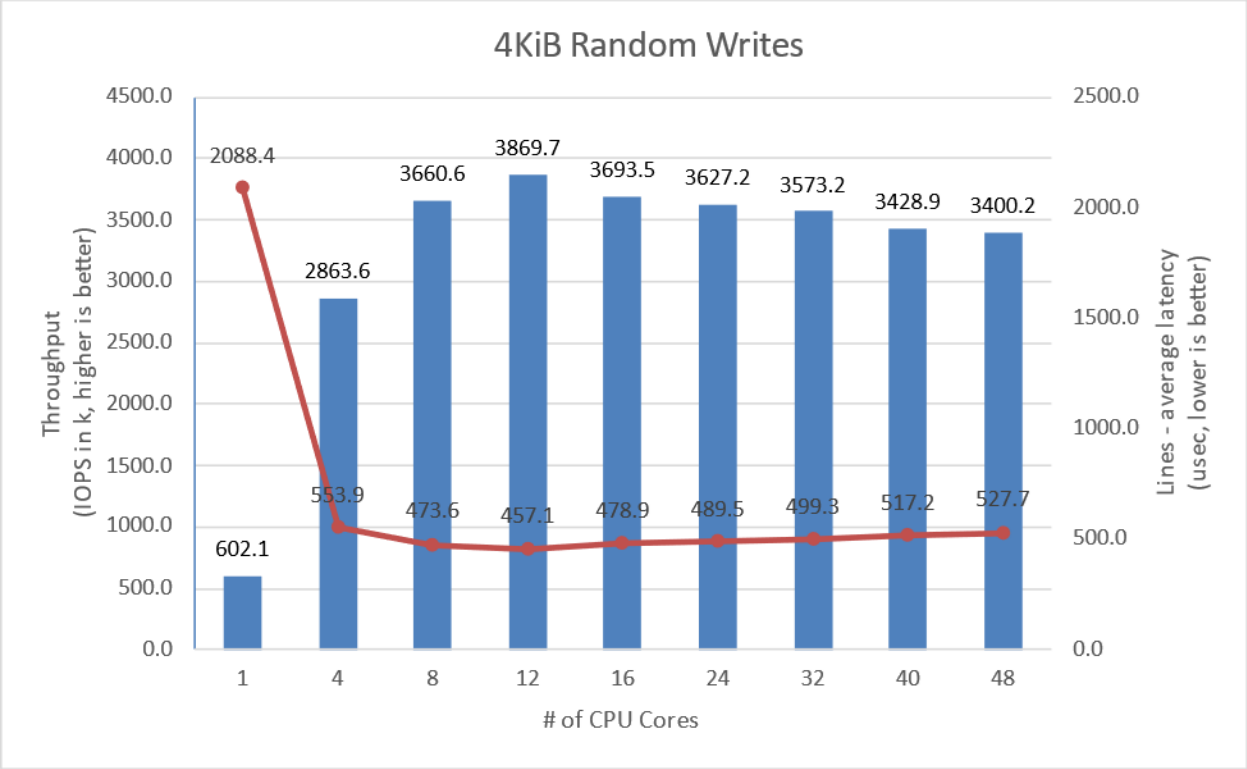

Figure 9: SPDK NVMe-oF TCP Initiator I/O core scaling: IOPS vs. Latency while running 4KiB 100% Random Write Workload at QD=128

4KiB Random Read-Write Results

Table 15: SPDK NVMe-oF TCP Initiator Core Scaling results, 4KiB Random 70%/30% Read/Write IOPS, QD=256

# of Cores	Throughput (IOPS k)	Bandwidth (Gbps)	Avg. Latency (usec)
1 core	427.6	14.01	8340.0
4 cores	1847.3	60.53	1929.3
8 cores	3352.6	109.86	1056.6
12 cores	4596.3	150.61	765.2
16 cores	5648.9	185.10	621.5
24 cores	6605.8	216.46	531.8
32 cores	6492.2	212.73	543.5
40 cores	6610.5	216.61	534.3
48 cores	6426.9	210.60	550.6

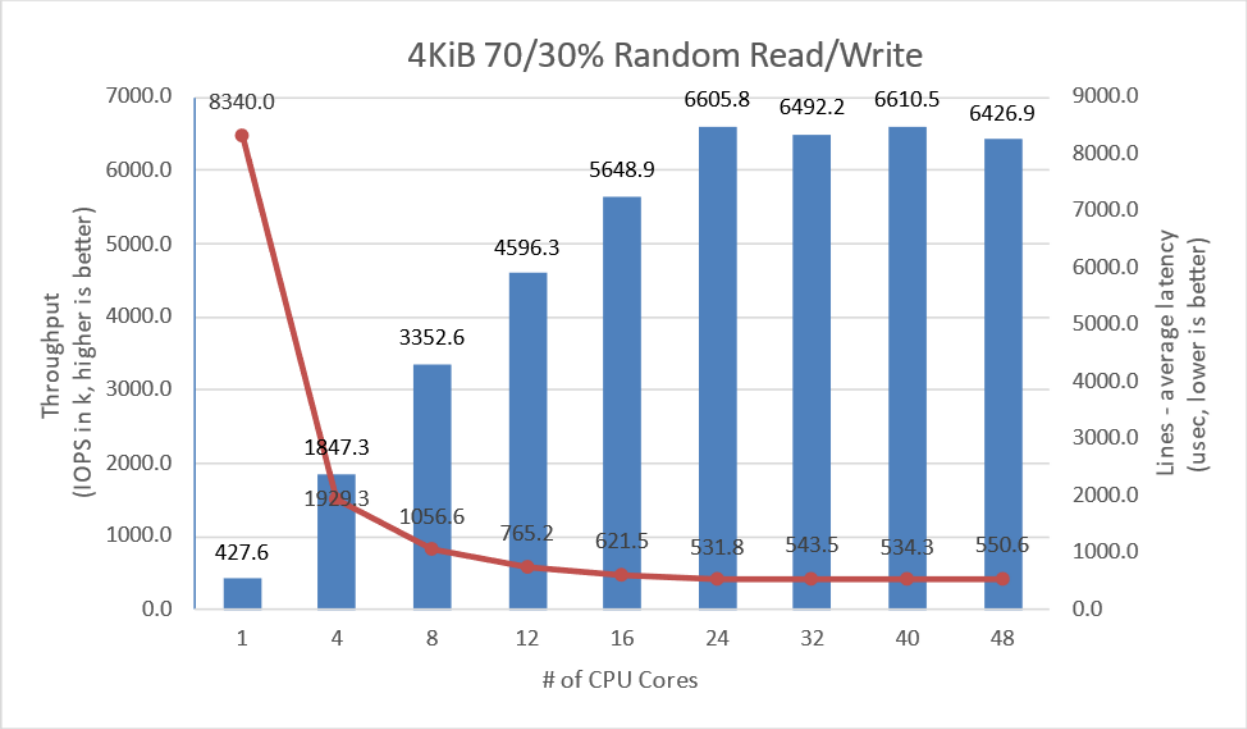

Figure 10: SPDK NVMe-oF TCP Initiator I/O core scaling: IOPS vs. Latency while running 4KiB Random 70% Read 30% Write Workload at QD=256

Conclusions

1. For the 4KiB Random Read workload, the SPDK NVMe-oF TCP Initiator performance scales linearly up to 12 CPU cores. Increasing the number of CPU cores beyond 16 CPU results in non-linear performance improvement, which peaks at 32 CPU cores with 5.69 million IOPS, reaching 200 Gbps network link saturation.

2. In case of 4KiB Random Write workload, performance scaling was non-linear. The IOPS rapidly increased to 2.8 million IOPS at 4 CPU cores reaching 100 Gbps network link saturation. Increasing the number of CPU cores resulted in non-linear performance change, that peaked at 12 CPU cores at 3.87 million IOPS. The 200 Gbps network link was not saturated.

3. Mixed Random Read-Write workload performance scales linearly up to 16 CPU cores, reaching 5.6 million IOPS. Increasing the number of cores further results in non-linear performance improvement that peaks at 6.61 million IOPS.

Test Case 3: Linux Kernel vs. SPDK NVMe-oF TCP Latency

This test case was designed to understand latency characteristics of SPDK NVMe-oF TCP Target and Initiator vs. the Linux Kernel NVMe-oF TCP Target and Initiator implementations on a single NVMe-oF subsystem. The average I/O latency and p99 latency was compared between SPDK NVMe-oF (Target/Initiator) vs. Linux Kernel (Target/Initiator). Both SPDK and Kernel NVMe-oF Targets were configured to run on a single core, with a single NVMe-oF subsystem on top of a *Null Block Device*. The null block device (bdev) was chosen as the backend block device to eliminate the media latency during these tests.

For this test case a Linux Kernel feature called aRFS (Accelerated Receive Flow Steering) was used. With RFS, network packets are forwarded depending on the location of the application thread processing the socket that includes those packets. This forwarding is done entirely in software. But for the accelerated RFS – or aRFS – NICs with proper flow steering support can do this forwarding in hardware instead. The key benefit to aRFS is ensuring that Rx packets for a given TCP flow are directed to an Rx queue that is processed on the same CPU core as the thread processing that TCP flow. This helps ensure caching effects of the Rx path and the user space processing happen on the same CPU core. Both RFS and aRFS are available in most Linux distributions but need to be configured before using. Steps for enabling aRFS in this test are described in table below.

Table 16: Linux Kernel vs. SPDK NVMe-oF TCP Latency test configuration

Item	Description
Test Case	Linux Kernel vs. SPDK NVMe-oF Latency
Test configuration	

SPDK NVMe-oF Target configuration	All below commands are executed with spdk/scripts/rpc.py script.
	
	Set iobuf buffer pool options:
	iobuf_set_options --small-pool-count 32767 --large-pool-count 16383
	
	nvmf_create_transport -t TCP
	(creates TCP transport layer with default values:
	{
	"trtype": "TCP",
	"max_queue_depth": 128,
	"max_io_qpairs_per_ctrlr": 127,
	"in_capsule_data_size": 4096,
	"max_io_size": 131072,
	"io_unit_size": 131072,
	"max_aq_depth": 128,
	"num_shared_buffers": 8192,
	"buf_cache_size": 32,
	"dif_insert_or_strip": false,
	"c2h_success": true,
	"sock_priority": 0,
	"abort_timeout_sec": 1
	}
	bdev_null_create Nvme0n1 10240 4096
	nvmf_subsystem_create nqn.2018-09.io.spdk:cnode1 -s SPDK001 -a -m 8
	nvmf_subsystem_add_ns nqn.2018-09.io.spdk:cnode1 Nvme0n1
	nvmf_subsystem_add_listener nqn.2018-09.io.spdk:cnode1 -t tcp -f ipv4 -s 4420 -a 20.0.0.1

Kernel NVMe-oF Target configuration	Target configuration file loaded using nvmet-cli tool.
	{
	"ports": [
	{
	"addr": {
	"adrfam": "ipv4",
	"traddr": "20.0.0.1",
	"trsvcid": "4420",
	"trtype": "tcp"
	},
	"portid": 1,
	"referrals": [],
	"subsystems": [
	"nqn.2018-09.io.spdk:cnode1"
	]
	}
	],
	"hosts": [],
	"subsystems": [
	{
	"allowed_hosts": [],
	"attr": {
	"allow_any_host": "1",
	"version": "1.3"
	},
	"namespaces": [
	{
	"device": {
	"path": "/dev/nullb0",
	"uuid": "621e25d2-8334-4c1a-8532-b6454390b8f9"
	},
	"enable": 1,
	"nsid": 1
	}
	],
	"nqn": "nqn.2018-09.io.spdk:cnode1"
	}
	]
	}

fio configuration	
SPDK NVMe-oF Initiator fio plugin configuration	**BDEV.conf** See appendix C. **fio.conf** [global] ioengine=/tmp/spdk/examples/bdev/fio_plugin/fio_plugin spdk_conf=/tmp/spdk/bdev.conf thread=1 group_reporting=1 direct=1 norandommap=1 rw=randrw rwmixread={100, 70, 0} bs=4k iodepth=1 time_based=1 ramp_time=30 runtime=30 [filename0]

	filename=Nvme0n1
Kernel initiator configuration	**Device config** Done using nvme-cli tool. modprobe nvme-fabrics nvme connect –n nqn.2018-09.io.spdk:cnode1 –t tcp –a 20.0.0.1 –s 4420 **fio.conf** [global] ioengine=io_uring thread=1 group_reporting=1 direct=1 norandommap=1 rw=randrw rwmixread={100, 70, 0} bs=4k iodepth=1 time_based=1 numjobs=1 ramp_time=30 runtime=30 [filename0] filename=/dev/nvme0n1
aRFS Configuration	

aRFS Configuration	**Enable ntuple feature in the NIC driver and check its status:**
	$ ethtool -K eth3 ntuple on
	
	$ ethtool -k eth3| grep ntuple
	ntuple-filters: on
	
	Disable Linux Kernel IRQ balancer
	$ service irqbalance stop
	
	Ensure that NIC's IRQ affinity is spread across all cores:
	$ set_irq_affinity.sh eth3
	$ show_irq_affinity.sh eth3
	(Mellanox utility scripts available on Github.com)
	
	Configure the RFS global and per-queue flow table entries. This needs to be done for every NIC interface taking part in the test.
	echo 32768 > /proc/sys/net/core/rps_sock_flow_entries
	for r in /sys/class/net/eth3/queues/rx-*/rps_flow_cnt;
	do echo 512 > $r
	done

SPDK vs Kernel NVMe-oF Target Results

This following data was collected using the Linux Kernel initiator against both SPDK & Linux Kernel NVMe-oF TCP target.

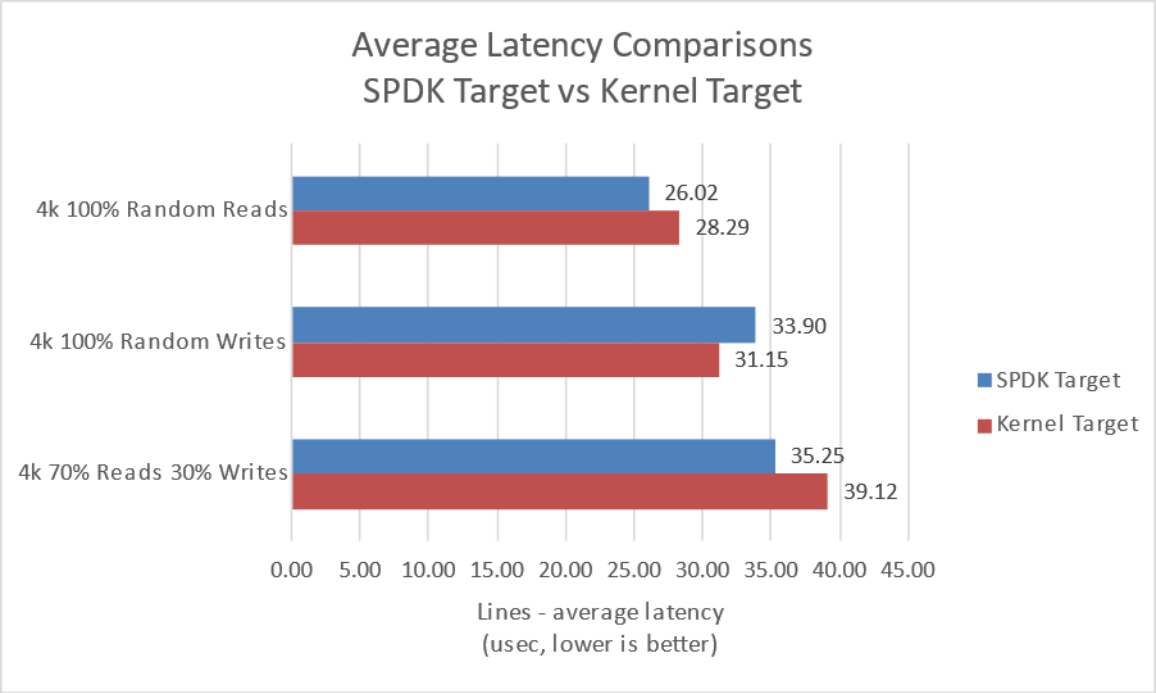

Figure 11: SPDK vs. Kernel NVMe-oF TCP Target Average I/O Latency for various workloads run using the Kernel Initiator

Table 17: SPDK NVMe-oF Target Latency and IOPS at QD=1, Null Block Device

Access Pattern	Avg. Latency (usec)	IOPS	p99 (usec)	p99.9 (usec)	p99.99 (usec)	p99.999 (usec)
4KiB 100% Random Reads IOPS	26.02	37873	37.8	72.2	235.2	402.1
4KiB 100% Random Writes IOPS	33.90	29176	139.6	179.9	267.9	428.0
4KiB 100% Random 70% Reads 30% Writes IOPS	35.25	28019	143.6	240.7	339.8	429.5

Table 18: Linux Kernel NVMe-oF Target Latency and IOPS at QD=1, Null Block Device

Access Pattern	Avg. Latency (usec)	IOPS	p99 (usec)	p99.9 (usec)	p99.99 (usec)	p99.999 (usec)
4KiB 100% Random Reads IOPS	28.29	34884	52.1	65.8	207.2	297.0
4KiB 100% Random Writes IOPS	31.15	31806	54.7	72.2	145.7	287.4
4KiB 100% Random 70% Reads 30% Writes IOPS	39.12	25265	154.6	253.2	344.9	374.1

SPDK vs Kernel NVMe-oF TCP Initiator Results

This following data was collected using Kernel & SPDK initiator against an SPDK target.

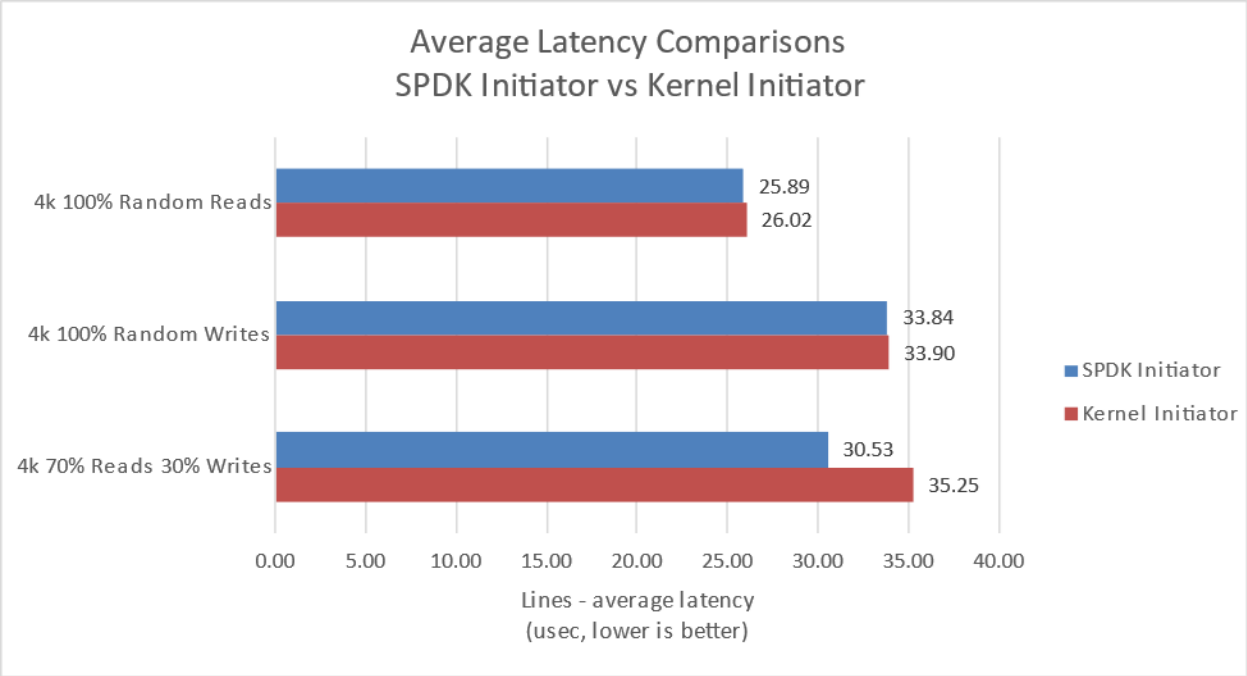

Figure 12: SPDK vs. Kernel NVMe-oF TCP Initiator Average I/O Latency for various workloads against SPDK Target

Table 19: SPDK NVMe-oF Initiator Latency and IOPS at QD=1, Null Block Device

Access Pattern	Avg. Latency (usec)	IOPS	p99 (usec)	p99.9 (usec)	p99.99 (usec)	p99.999 (usec)
4KiB 100% Random Reads IOPS	25.89	38874	76.3	109.4	152.6	384.3
4KiB 100% Random Writes IOPS	33.84	30171	126.5	149.0	203.1	402.1
4KiB 100% Random 70% Reads 30% Writes IOPS	30.53	32477	94.1	150.1	211.3	395.0

Table 20: Linux Kernel NVMe-oF Initiator Latency and IOPS at QD=1, Null Block Device

Access Pattern	Avg. Latency (usec)	IOPS	p99 (usec)	p99.9 (usec)	p99.99 (usec)	p99.999 (usec)
4KiB 100% Random Reads IOPS	26.02	37873	37.8	72.2	235.2	402.1
4KiB 100% Random Writes IOPS	33.90	29176	139.6	179.9	267.9	428.0
4KiB 100% Random 70% Reads 30% Writes IOPS	35.25	28019	143.6	240.7	339.8	429.5

SPDK vs Kernel NVMe-oF Kernel + Initiator Results

Following data was collected using SPDK Target with SPDK Initiator and Linux Target with Linux Initiator.

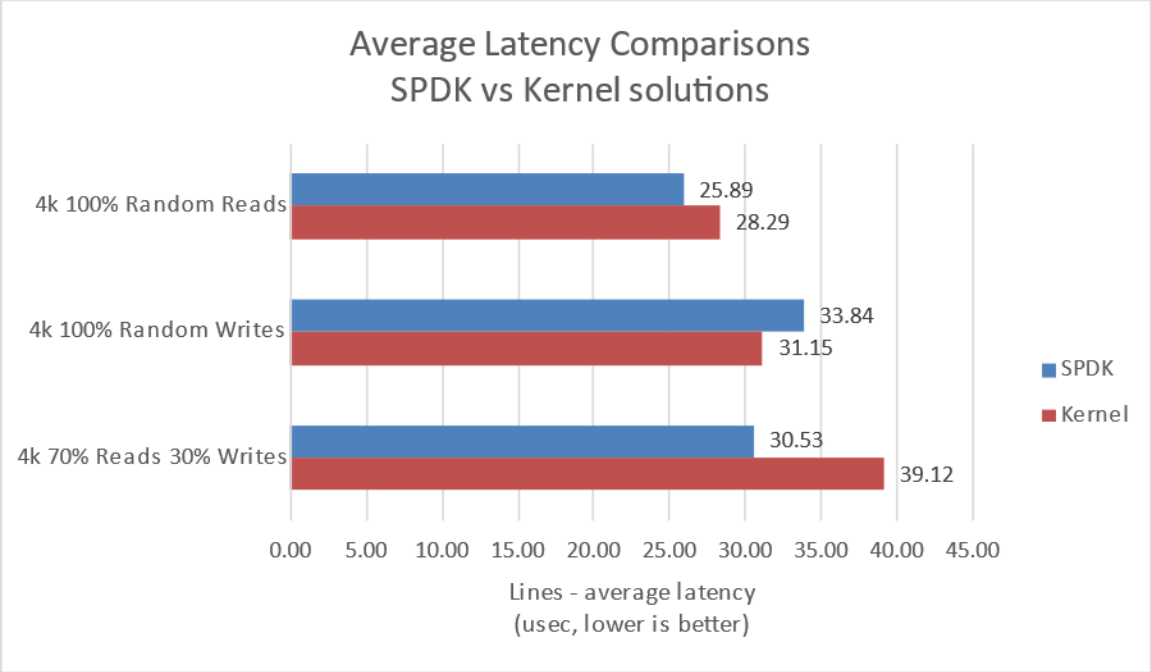

Figure 13: SPDK vs. Kernel NVMe-oF TCP solutions Average I/O Latency for various workloads

Table 21: SPDK NVMe-oF Latency and IOPS at QD=1, Null Block Device

Access Pattern	Avg. Latency (usec)	IOPS	p99 (usec)	p99.9 (usec)	p99.99 (usec)	p99.999 (usec)
4KiB 100% Random Reads IOPS	25.89	38874	76.3	109.4	152.6	384.3
4KiB 100% Random Writes IOPS	33.84	30171	126.5	149.0	203.1	402.1
4KiB 100% Random 70% Reads 30% Writes IOPS	30.53	32477	94.1	150.1	211.3	395.0

Table 22: Linux Kernel NVMe-oF Latency and IOPS at QD=1, Null Block Device

Access Pattern	Avg. Latency (usec)	IOPS	p99 (usec)	p99.9 (usec)	p99.99 (usec)	p99.999 (usec)
4KiB 100% Random Reads IOPS	28.29	34884	52.14	65.79	207.19	296.96
4KiB 100% Random Writes IOPS	31.15	31806	54.70	72.19	145.75	287.40
4KiB 100% Random 70% Reads 30% Writes IOPS	39.12	25265	154.62	253.24	344.88	374.10

Conclusions

1. SPDK NVMe-oF TCP Initiator reduces latency by up to 3.87 usec. vs. Linux Kernel NVMe-oF TCP Initiator, which eliminates up to 9.9% of software overhead.
2. SPDK NVMe-oF TCP Initiator reduces the average latency by up to 4.72 usec. vs. Linux Kernel NVMe-oF TCP Initiator, which eliminates up to 13.4% of software overhead.
3. SPDK NVMe-oF TCP Target and Initiator reduce the average latency by up to 8.59 usec. vs. Linux Kernel NVMe-oF TCP Target and Initiator, which eliminates up to 21.97% of software overhead.

Test Case 4: NVMe-oF Performance with increasing # of connections

This test case was performed to understand throughput and latency capabilities of SPDK NVMe-oF Target vs. Linux Kernel NVMe-oF Target under increasing number of connections per subsystem. The number of connections (or I/O queue pairs) per NVMe-oF subsystem were varied and corresponding aggregated IOPS and number of CPU cores metrics were reported. The number of CPU cores metric was calculated from % CPU utilization measured using sar (systat package in Linux). The SPDK NVMe-oF Target was configured to run on 24 cores, 14 NVMe-oF subsystems (1 per Kioxia NVMe SSD) and 2 initiators were used both running I/Os to 7 separate subsystems using Kernel NVMe-oF initiator. We ran the following workloads on the host systems:

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

Table 23: NVMe-oF Performance with increasing number of connections test configuration

Item	Description
Test Case	NVMe-oF Target performance under varying # of connections
SPDK NVMe-oF Target configuration	Same as in Test Case #1, using 24 CPU cores.
Kernel NVMe-oF Target configuration	Target configuration file loaded using nvmet-cli tool. For detail configuration file contents please see Appendix D.
Kernel NVMe-oF Initiator #1	**Device config** Performed using nvme-cli tool. modprobe nvme-fabrics nvme connect –n nqn.2018-09.io.spdk:cnode1 –t tcp –a 20.0.0.1 –s 4420 nvme connect –n nqn.2018-09.io.spdk:cnode2 –t tcp –a 20.0.0.1 –s 4420 nvme connect –n nqn.2018-09.io.spdk:cnode3 –t tcp –a 20.0.0.1 –s 4420 nvme connect –n nqn.2018-09.io.spdk:cnode4 –t tcp –a 20.0.0.1 –s 4420 nvme connect –n nqn.2018-09.io.spdk:cnode5 –t tcp –a 20.0.1.1 –s 4420 nvme connect –n nqn.2018-09.io.spdk:cnode6 –t tcp –a 20.0.1.1 –s 4420 nvme connect –n nqn.2018-09.io.spdk:cnode7 –t tcp –a 20.0.1.1 –s 4420
Kernel NVMe-oF Initiator #2	**Device config** Performed using nvme-cli tool. modprobe nvme-fabrics nvme connect –n nqn.2018-09.io.spdk:cnode8 –t tcp –a 10.0.0.1 –s 4420 nvme connect –n nqn.2018-09.io.spdk:cnode9 –t tcp –a 10.0.0.1 –s 4420 nvme connect –n nqn.2018-09.io.spdk:cnode10 –t tcp –a 10.0.0.1 –s 4420 nvme connect –n nqn.2018-09.io.spdk:cnode11 –t tcp –a 10.0.0.1 –s 4420 nvme connect –n nqn.2018-09.io.spdk:cnode12 –t tcp –a 10.0.1.1 –s 4420 nvme connect –n nqn.2018-09.io.spdk:cnode13 –t tcp –a 10.0.1.1 –s 4420 nvme connect –n nqn.2018-09.io.spdk:cnode14 –t tcp –a 10.0.1.1 –s 4420
fio configuration (used on both initiators)	**fio.conf** [global] ioengine=io_uring thread=1

	group_reporting=1 direct=1 norandommap=1 rw=randrw rwmixread={100, 70, 0} bs=4k iodepth={128, 192, 384} time_based=1 ramp_time=60 runtime=300 numjobs={1, 4, 8, 12, 16} [filename1] filename=/dev/nvme0n1 [filename2] filename=/dev/nvme1n1 [filename3] filename=/dev/nvme2n1 [filename4] filename=/dev/nvme3n1 [filename5] filename=/dev/nvme4n1 [filename6] filename=/dev/nvme5n1 [filename7] filename=/dev/nvme6n1

The number of CPU cores used while running the SPDK NVMe-oF target was 24, whereas for the case of Linux Kernel NVMe-oF target there was no CPU core limitation applied.

The metrics in the graph represent relative efficiency in IOPS/core which was calculated based on total aggregate IOPS divided by total CPU cores used while running that specific workload. For the case of Kernel NVMe-oF target, total CPU cores was calculated from % CPU utilization which was measured using sar utility in Linux.

4KiB Random Read Results

Figure 14: Relative Efficiency Comparison of Linux Kernel vs. SPDK NVMe-oF Target IOPS/Core for 4KiB 100% Random Reads QD=384 using the Kernel Initiator

Table 24: Linux Kernel NVMe-oF TCP Target: 4KiB 100% Random Reads, QD=384

Connections per subsystem	Bandwidth (Gbps)	Throughput (IOPS k)	Avg. Latency (usec)	# CPU Cores
1	128.79	3930.4	1597.3	17.0
4	268.67	8199.0	768.4	29.7
8	319.27	9743.4	653.1	40.8
12	297.72	9085.6	591.0	44.0
16	304.63	9296.5	699.4	44.6

Table 25: SPDK NVMe-oF TCP Target: 4KiB 100% Random Reads, QD=384

Connections per subsystem	Bandwidth (Gbps)	Throughput (IOPS k)	Avg. Latency (usec)	# CPU Cores
1	129.45	3950.4	1588.7	26.2
4	272.62	8319.7	767.1	27.5
8	294.56	8989.4	731.4	28.4
12	258.33	7883.7	681.6	29.3
16	265.51	8102.8	827.7	29.3

4KiB Random Write Results

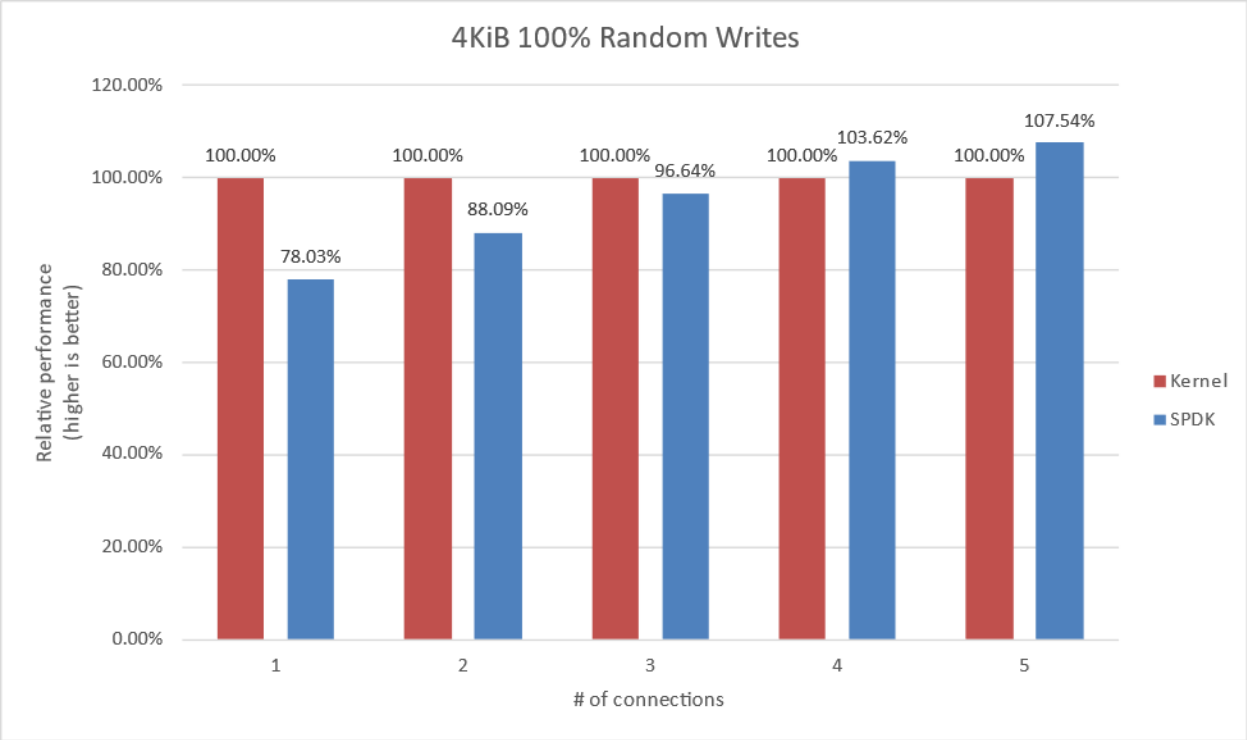

Figure 15: Relative Efficiency Comparison of Linux Kernel vs. SPDK NVMe-oF Target IOPS/Core for 4KiB 100% Random Writes QD=256 using the Kernel Initiator

Note: Drives were not pre-conditioned while running 100% Random write I/O Test

Table 26: Linux Kernel NVMe-oF TCP Target: 4KiB 100% Random Writes, QD=256

Connections per subsystem	Bandwidth (Gbps)	Throughput (IOPS k)	Avg. Latency (usec)	# CPU Cores
1	136.07	4152.4	864.8	26.8
4	183.92	5612.8	641.3	32.3
8	225.33	6876.6	524.6	43.1
12	231.70	7070.9	588.8	47.4
16	227.66	6947.7	520.0	49.4

Table 27: SPDK NVMe-oF TCP Target: 4KiB 100% Random Writes, QD=256

Connections per subsystem	Bandwidth (Gbps)	Throughput (IOPS k)	Avg. Latency (usec)	# CPU Cores
1	111.29	3396.1	1055.1	28.1
4	140.93	4300.8	833.1	28.1
8	146.41	4468.2	801.8	29.0
12	147.09	4489.0	854.3	29.0
16	145.12	4428.7	809.0	29.3

4KiB Random Read-Write Results

Figure 16: Relative Efficiency Comparison of Linux Kernel vs. SPDK NVMe-oF Target IOPS/Core for 4KiB Random 70% Reads 30% Writes QD=192 using Kernel Initiator

Table 28: Linux Kernel NVMe-oF TCP Target: 4KiB 70% Random Read 30% Random Write, QD=192

Connections per subsystem	Bandwidth (Gbps)	Throughput (IOPS k)	Avg. Latency (usec)	# CPU Cores
1	126.82	3870.3	925.7	19.9
4	208.49	6362.5	563.0	29.6
8	253.45	7734.6	462.9	42.9
12	256.15	7817.1	556.6	44.3
16	211.62	6458.0	554.4	43.0

Table 29: SPDK NVMe-oF TCP Target: 4KiB 70% Random Read 30% Random Write, QD=192

Connections per subsystem	Bandwidth (Gbps)	Throughput (IOPS k)	Avg. Latency (usec)	# CPU Cores
1	127.86	3901.8	918.2	27.3
4	184.99	5645.3	634.5	28.0
8	188.18	5742.8	623.7	29.2
12	195.07	5953.1	677.6	29.6
16	172.85	5275.0	679.3	30.2

Low Connections Results

During testing, it was initially observed that the relative efficiency of SPDK Target was about 50-60% of Kernel Target. This was primarily because SPDK traditionally used a fixed number of CPU cores configured at startup, without an intrinsic mechanism to decrease the number of I/O cores on-the-fly if the SPDK target does not need all of the CPU resources.
However, with the implementation of the dynamic scheduler, SPDK now can adjust CPU resource allocation based on demand. The dynamic scheduler allows SPDK to dynamically reduce or increase the number of I/O cores in response to changing workloads, thereby optimizing CPU resource utilization. This significantly enhances SPDK's efficiency and performance, ensuring it does not underutilize or overcommit CPU resources. For detailed information, please refer to the section on the dynamic scheduler in the documentation (appendix A).

The test cases with 1 connection per subsystems were re-run with SPDK using only 4 CPU cores.

Table 30: SPDK & Kernel NVMe-oF TCP Target relative efficiency comparison for various workloads, QD=192, 1 connection per subsystem

Workload	Target	Bandwidth (Gbps)	Throughput (IOPS k)	Avg. Latency (usec)	# CPU Cores
Random Read	Linux	119.45	3645.4	858.5	14.7
	SPDK	68.15	2079.8	1490.6	5.6
Random Write	Linux	139.41	4254.3	732.6	24.9
	SPDK	39.23	1197.2	2607.5	5.9
Random Read/Write	Linux	122.85	3749.1	835.1	18.0
	SPDK	50.43	1539.0	2007.4	5.9

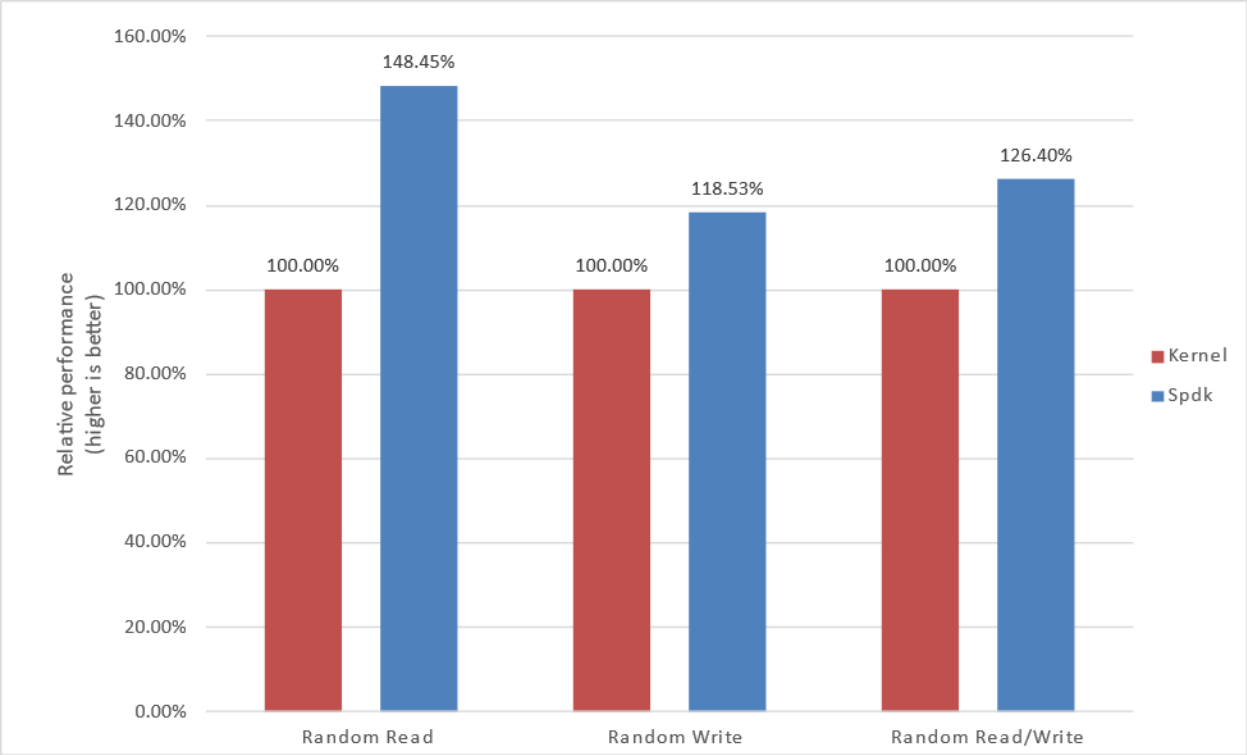

Figure 17: Relative Efficiency Comparison of Linux Kernel vs. SPDK NVMe-oF Target IOPS/Core for various workloads, 1 connection per subsystem and reduced number of SPDK Target CPU Cores (4)

Conclusions

1. The Linux Kernel NVMe-oF TCP target relative efficiency in IOPS/Core was better than SPDK when there was low number of connections per subsystem because the SPDK NVMe-oF target uses a fixed number of CPU cores when configured with the static scheduler. Therefore, we re-run the test cases with 1 connection per subsystem but lowered the number of I/O cores used by the SPDK Target to 4 and added the results to the tables which show a relative performance better than Linux Kernel NVMe-oF TCP target up to 1.48x for Random Read, 1.18x times for Random Write and 1.26x times for Random Read/Write workloads.

2. The performance peaked for all workloads at 8 connections per subsystem for both SPDK and Kernel NVMe-oF TCP Target for Random Read and at 12 connections for Random Write and Random Read/Write workloads.

3. The SPDK NVMe-oF TCP target relative efficiency in IOPS/Core was up to 1.32x, 1.07x and 1.13x times better than the Linux Kernel NVMe-oF target for Random Read, Random Write and Random Read/Write workloads respectively.

Summary

This report showcased performance results with SPDK NVMe-oF TCP target and initiator under various test cases, including:

- I/O core scaling
- Average I/O latency
- Performance with increasing number of connections per subsystems

It compared performance results while running Linux Kernel NVMe-oF (Target/Initiator) against the accelerated polled-mode driven SPDK NVMe-oF (Target/Initiator) implementation.

Throughput scales up and latency decreases almost linearly with the scaling of SPDK NVMe-oF target I/O cores when serving 4KiB random workloads. The SPDK NVMe-oF target saturates a 400 Gbps network link using 32 CPU cores for the 4KiB Random Read and 200 Gbps for Random Write workload at 40 CPU cores. The IOPS scalability remains close to linear for all workloads until the results are close to saturating network link (or NVMe drives throughput in case of Random Write workload).

For the SPDK NVMe-oF TCP Initiator running Random Read and Random Read/Write workloads the IOPS throughput scales almost linearly with addition of CPU cores until the network was almost saturated. Further increasing the number of CPU cores results in performance degradation. A single initiator was able to saturate a 200Gb link for these workloads.

The SPDK NVMe-oF TCP Target performed up to 1.48 times better w.r.t IOPS/core than Linux Kernel NVMe-oF target while running 4KiB 100% Random Read workload with increasing number of connections per NVMe-oF subsystem. Performance in case of 4KiB 100% Random Write workload was comparable between SPDK and Kernel Targets.

This report provides information regarding methodologies and practices while benchmarking NVMe-oF using SPDK, as well as the Linux Kernel. It should be noted that the performance data showcased in this report is based on specific hardware and software configurations and that performance results may vary depending on the hardware and software configurations.

List of Figures

List of Tables

Appendix A – Test Case 1 SPDK NVMe-oF Initiator bdev configuration

Initiator system 1

```
{
  "subsystems": [
    {
      "subsystem": "bdev",
      "config": [
        {
          "method": "bdev_nvme_attach_controller",
          "params": {
            "name": "Nvme0",
            "trtype": "tcp",
            "traddr": "20.0.0.1",
            "trsvcid": "4420",
            "subnqn": "nqn.2018-09.io.spdk:cnode0",
            "adrfam": "IPv4"
          }
        },
        {
          "method": "bdev_nvme_attach_controller",
          "params": {
            "name": "Nvme1",
            "trtype": "tcp",
            "traddr": "20.0.0.1",
            "trsvcid": "4420",
            "subnqn": "nqn.2018-09.io.spdk:cnode1",
            "adrfam": "IPv4"
          }
        },
        {
          "method": "bdev_nvme_attach_controller",
          "params": {
            "name": "Nvme2",
            "trtype": "tcp",
            "traddr": "20.0.0.1",
            "trsvcid": "4420",
            "subnqn": "nqn.2018-09.io.spdk:cnode2",
            "adrfam": "IPv4"
          }
        },
        {
          "method": "bdev_nvme_attach_controller",
          "params": {
            "name": "Nvme3",
            "trtype": "tcp",
            "traddr": "20.0.0.1",
            "trsvcid": "4420",
            "subnqn": "nqn.2018-09.io.spdk:cnode3",
            "adrfam": "IPv4"
          }
        },
```

```
      {
        "method": "bdev_nvme_attach_controller",
        "params": {
          "name": "Nvme4",
          "trtype": "tcp",
          "traddr": "20.0.1.1",
          "trsvcid": "4420",
          "subnqn": "nqn.2018-09.io.spdk:cnode4",
          "adrfam": "IPv4"
        }
      },
      {
        "method": "bdev_nvme_attach_controller",
        "params": {
          "name": "Nvme5",
          "trtype": "tcp",
          "traddr": "20.0.1.1",
          "trsvcid": "4420",
          "subnqn": "nqn.2018-09.io.spdk:cnode5",
          "adrfam": "IPv4"
        }
      },
      {
        "method": "bdev_nvme_attach_controller",
        "params": {
          "name": "Nvme6",
          "trtype": "tcp",
          "traddr": "20.0.1.1",
          "trsvcid": "4420",
          "subnqn": "nqn.2018-09.io.spdk:cnode6",
          "adrfam": "IPv4"
        }
      },
      {
        "method": "bdev_nvme_attach_controller",
        "params": {
          "name": "Nvme7",
          "trtype": "tcp",
          "traddr": "20.0.1.1",
          "trsvcid": "4420",
          "subnqn": "nqn.2018-09.io.spdk:cnode7",
          "adrfam": "IPv4"
        }
      }
    ]
  },
  {
    "subsystem": "iobuf",
    "config": [
      {
        "method": "iobuf_set_options",
        "params": {
          "small_pool_count": 32768,
          "large_pool_count": 16384
        }
      }
    ]
```

```
    }
  ]
}
```

Initiator system 2

```
{
  "subsystems": [
    {
      "subsystem": "bdev",
      "config": [
        {
          "method": "bdev_nvme_attach_controller",
          "params": {
            "name": "Nvme0",
            "trtype": "tcp",
            "traddr": "10.0.0.1",
            "trsvcid": "4420",
            "subnqn": "nqn.2018-09.io.spdk:cnode0",
            "adrfam": "IPv4"
          }
        },
        {
          "method": "bdev_nvme_attach_controller",
          "params": {
            "name": "Nvme1",
            "trtype": "tcp",
            "traddr": "10.0.0.1",
            "trsvcid": "4420",
            "subnqn": "nqn.2018-09.io.spdk:cnode1",
            "adrfam": "IPv4"
          }
        },
        {
          "method": "bdev_nvme_attach_controller",
          "params": {
            "name": "Nvme2",
            "trtype": "tcp",
            "traddr": "10.0.0.1",
            "trsvcid": "4420",
            "subnqn": "nqn.2018-09.io.spdk:cnode2",
            "adrfam": "IPv4"
          }
        },
        {
          "method": "bdev_nvme_attach_controller",
          "params": {
            "name": "Nvme3",
            "trtype": "tcp",
            "traddr": "10.0.0.1",
            "trsvcid": "4420",
            "subnqn": "nqn.2018-09.io.spdk:cnode3",
            "adrfam": "IPv4"
          }
        },
        {
          "method": "bdev_nvme_attach_controller",
```

```
          "params": {
            "name": "Nvme4",
            "trtype": "tcp",
            "traddr": "10.0.1.1",
            "trsvcid": "4420",
            "subnqn": "nqn.2018-09.io.spdk:cnode4",
            "adrfam": "IPv4"
          }
        },
        {
          "method": "bdev_nvme_attach_controller",
          "params": {
            "name": "Nvme5",
            "trtype": "tcp",
            "traddr": "10.0.1.1",
            "trsvcid": "4420",
            "subnqn": "nqn.2018-09.io.spdk:cnode5",
            "adrfam": "IPv4"
          }
        },
        {
          "method": "bdev_nvme_attach_controller",
          "params": {
            "name": "Nvme6",
            "trtype": "tcp",
            "traddr": "10.0.1.1",
            "trsvcid": "4420",
            "subnqn": "nqn.2018-09.io.spdk:cnode6",
            "adrfam": "IPv4"
          }
        },
        {
          "method": "bdev_nvme_attach_controller",
          "params": {
            "name": "Nvme7",
            "trtype": "tcp",
            "traddr": "10.0.1.1",
            "trsvcid": "4420",
            "subnqn": "nqn.2018-09.io.spdk:cnode7",
            "adrfam": "IPv4"
          }
        }
      ]
    },
    {
      "subsystem": "iobuf",
      "config": [
        {
          "method": "iobuf_set_options",
          "params": {
            "small_pool_count": 32768,
            "large_pool_count": 16384
          }
        }
      ]
    }
  ]
```

```
}
```

Appendix B – Test Case 2 SPDK NVMe-oF Initiator bdev configuration

```
{
  "subsystems": [
    {
      "subsystem": "bdev",
      "config": [
        {
          "method": "bdev_nvme_attach_controller",
          "params": {
            "name": "Nvme0",
            "trtype": "tcp",
            "traddr": "20.0.0.1",
            "trsvcid": "4420",
            "subnqn": "nqn.2018-09.io.spdk:cnode0",
            "adrfam": "IPv4"
          }
        },
        {
          "method": "bdev_nvme_attach_controller",
          "params": {
            "name": "Nvme1",
            "trtype": "tcp",
            "traddr": "20.0.0.1",
            "trsvcid": "4420",
            "subnqn": "nqn.2018-09.io.spdk:cnode1",
            "adrfam": "IPv4"
          }
        },
        {
          "method": "bdev_nvme_attach_controller",
          "params": {
            "name": "Nvme2",
            "trtype": "tcp",
            "traddr": "20.0.0.1",
            "trsvcid": "4420",
            "subnqn": "nqn.2018-09.io.spdk:cnode2",
            "adrfam": "IPv4"
          }
        },
        {
          "method": "bdev_nvme_attach_controller",
          "params": {
            "name": "Nvme3",
            "trtype": "tcp",
            "traddr": "20.0.0.1",
            "trsvcid": "4420",
            "subnqn": "nqn.2018-09.io.spdk:cnode3",
            "adrfam": "IPv4"
          }
        },
```

```
{
  "method": "bdev_nvme_attach_controller",
  "params": {
    "name": "Nvme4",
    "trtype": "tcp",
    "traddr": "20.0.0.1",
    "trsvcid": "4420",
    "subnqn": "nqn.2018-09.io.spdk:cnode4",
    "adrfam": "IPv4"
  }
},
{
  "method": "bdev_nvme_attach_controller",
  "params": {
    "name": "Nvme5",
    "trtype": "tcp",
    "traddr": "20.0.0.1",
    "trsvcid": "4420",
    "subnqn": "nqn.2018-09.io.spdk:cnode5",
    "adrfam": "IPv4"
  }
},
{
  "method": "bdev_nvme_attach_controller",
  "params": {
    "name": "Nvme6",
    "trtype": "tcp",
    "traddr": "20.0.0.1",
    "trsvcid": "4420",
    "subnqn": "nqn.2018-09.io.spdk:cnode6",
    "adrfam": "IPv4"
  }
},
{
  "method": "bdev_nvme_attach_controller",
  "params": {
    "name": "Nvme7",
    "trtype": "tcp",
    "traddr": "20.0.0.1",
    "trsvcid": "4420",
    "subnqn": "nqn.2018-09.io.spdk:cnode7",
    "adrfam": "IPv4"
  }
},
{
  "method": "bdev_nvme_attach_controller",
  "params": {
    "name": "Nvme8",
    "trtype": "tcp",
    "traddr": "20.0.1.1",
    "trsvcid": "4420",
    "subnqn": "nqn.2018-09.io.spdk:cnode8",
    "adrfam": "IPv4"
  }
},
{
  "method": "bdev_nvme_attach_controller",
```

```
      "params": {
        "name": "Nvme9",
        "trtype": "tcp",
        "traddr": "20.0.1.1",
        "trsvcid": "4420",
        "subnqn": "nqn.2018-09.io.spdk:cnode9",
        "adrfam": "IPv4"
      }
    },
    {
      "method": "bdev_nvme_attach_controller",
      "params": {
        "name": "Nvme10",
        "trtype": "tcp",
        "traddr": "20.0.1.1",
        "trsvcid": "4420",
        "subnqn": "nqn.2018-09.io.spdk:cnode10",
        "adrfam": "IPv4"
      }
    },
    {
      "method": "bdev_nvme_attach_controller",
      "params": {
        "name": "Nvme11",
        "trtype": "tcp",
        "traddr": "20.0.1.1",
        "trsvcid": "4420",
        "subnqn": "nqn.2018-09.io.spdk:cnode11",
        "adrfam": "IPv4"
      }
    },
    {
      "method": "bdev_nvme_attach_controller",
      "params": {
        "name": "Nvme12",
        "trtype": "tcp",
        "traddr": "20.0.1.1",
        "trsvcid": "4420",
        "subnqn": "nqn.2018-09.io.spdk:cnode12",
        "adrfam": "IPv4"
      }
    },
    {
      "method": "bdev_nvme_attach_controller",
      "params": {
        "name": "Nvme13",
        "trtype": "tcp",
        "traddr": "20.0.1.1",
        "trsvcid": "4420",
        "subnqn": "nqn.2018-09.io.spdk:cnode13",
        "adrfam": "IPv4"
      }
    },
    {
      "method": "bdev_nvme_attach_controller",
      "params": {
        "name": "Nvme14",
```

```
            "trtype": "tcp",
            "traddr": "20.0.1.1",
            "trsvcid": "4420",
            "subnqn": "nqn.2018-09.io.spdk:cnode14",
            "adrfam": "IPv4"
          }
        },
        {
          "method": "bdev_nvme_attach_controller",
          "params": {
            "name": "Nvme15",
            "trtype": "tcp",
            "traddr": "20.0.1.1",
            "trsvcid": "4420",
            "subnqn": "nqn.2018-09.io.spdk:cnode15",
            "adrfam": "IPv4"
          }
        }
      ]
    },
    {
      "subsystem": "iobuf",
      "config": [
        {
          "method": "iobuf_set_options",
          "params": {
            "small_pool_count": 32768,
            "large_pool_count": 16384
          }
        }
      ]
    }
  ]
}
```


Appendix C – Test Case 3 SPDK NVMe-oF Initiator bdev configuration


```
{
  "subsystems": [
    {
      "subsystem": "bdev",
      "config": [
        {
          "method": "bdev_nvme_attach_controller",
          "params": {
            "name": "Nvme0",
            "trtype": "tcp",
            "traddr": "20.0.0.1",
            "trsvcid": "4420",
            "subnqn": "nqn.2018-09.io.spdk:cnode0",
            "adrfam": "IPv4"
          }
```

```
        }
      ]
    },
    {
      "subsystem": "iobuf",
      "config": [
        {
          "method": "iobuf_set_options",
          "params": {
            "small_pool_count": 32768,
            "large_pool_count": 16384
          }
        }
      ]
    }
  ]
}
```

Appendix D – Kernel NVMe-oF TCP Target configuration

Example Kernel NVMe-oF TCP Target configuration for Test Case 4.

```
{
  "ports": [
    {
      "addr": {
        "adrfam": "ipv4",
        "traddr": "20.0.0.1",
        "trsvcid": "4420",
        "trtype": "tcp"
      },
      "portid": 1,
      "referrals": [],
      "subsystems": [
        "nqn.2018-09.io.spdk:cnode1"
      ]
    },
    {
      "addr": {
        "adrfam": "ipv4",
        "traddr": "20.0.0.1",
        "trsvcid": "4421",
        "trtype": "tcp"
      },
      "portid": 2,
      "referrals": [],
      "subsystems": [
        "nqn.2018-09.io.spdk:cnode2"
      ]
    },
    {
      "addr": {
        "adrfam": "ipv4",
        "traddr": "20.0.0.1",
        "trsvcid": "4422",
        "trtype": "tcp"
```

```
      },
      "portid": 3,
      "referrals": [],
      "subsystems": [
        "nqn.2018-09.io.spdk:cnode3"
      ]
    },
    {
      "addr": {
        "adrfam": "ipv4",
        "traddr": "20.0.0.1",
        "trsvcid": "4423",
        "trtype": "tcp"
      },
      "portid": 4,
      "referrals": [],
      "subsystems": [
        "nqn.2018-09.io.spdk:cnode4"
      ]
    },
    {
      "addr": {
        "adrfam": "ipv4",
        "traddr": "20.0.1.1",
        "trsvcid": "4424",
        "trtype": "tcp"
      },
      "portid": 5,
      "referrals": [],
      "subsystems": [
        "nqn.2018-09.io.spdk:cnode5"
      ]
    },
    {
      "addr": {
        "adrfam": "ipv4",
        "traddr": "20.0.1.1",
        "trsvcid": "4425",
        "trtype": "tcp"
      },
      "portid": 6,
      "referrals": [],
      "subsystems": [
        "nqn.2018-09.io.spdk:cnode6"
      ]
    },
    {
      "addr": {
        "adrfam": "ipv4",
        "traddr": "20.0.1.1",
        "trsvcid": "4426",
        "trtype": "tcp"
      },
      "portid": 7,
      "referrals": [],
      "subsystems": [
        "nqn.2018-09.io.spdk:cnode7"
```

```
      ]
    },
    {
      "addr": {
        "adrfam": "ipv4",
        "traddr": "20.0.1.1",
        "trsvcid": "4427",
        "trtype": "tcp"
      },
      "portid": 8,
      "referrals": [],
      "subsystems": [
        "nqn.2018-09.io.spdk:cnode8"
      ]
    },
    {
      "addr": {
        "adrfam": "ipv4",
        "traddr": "10.0.0.1",
        "trsvcid": "4428",
        "trtype": "tcp"
      },
      "portid": 9,
      "referrals": [],
      "subsystems": [
        "nqn.2018-09.io.spdk:cnode9"
      ]
    },
    {
      "addr": {
        "adrfam": "ipv4",
        "traddr": "10.0.0.1",
        "trsvcid": "4429",
        "trtype": "tcp"
      },
      "portid": 10,
      "referrals": [],
      "subsystems": [
        "nqn.2018-09.io.spdk:cnode10"
      ]
    },
    {
      "addr": {
        "adrfam": "ipv4",
        "traddr": "10.0.0.1",
        "trsvcid": "4430",
        "trtype": "tcp"
      },
      "portid": 11,
      "referrals": [],
      "subsystems": [
        "nqn.2018-09.io.spdk:cnode11"
      ]
    },
    {
      "addr": {
        "adrfam": "ipv4",
```

```
        "traddr": "10.0.0.1",
        "trsvcid": "4431",
        "trtype": "tcp"
      },
      "portid": 12,
      "referrals": [],
      "subsystems": [
        "nqn.2018-09.io.spdk:cnode12"
      ]
    },
{
      "addr": {
        "adrfam": "ipv4",
        "traddr": "10.0.1.1",
        "trsvcid": "4432",
        "trtype": "tcp"
      },
      "portid": 13,
      "referrals": [],
      "subsystems": [
        "nqn.2018-09.io.spdk:cnode13"
      ]
    },
    {
      "addr": {
        "adrfam": "ipv4",
        "traddr": "10.0.1.1",
        "trsvcid": "4433",
        "trtype": "tcp"
      },
      "portid": 14,
      "referrals": [],
      "subsystems": [
        "nqn.2018-09.io.spdk:cnode14"
      ]
    },
    {
      "addr": {
        "adrfam": "ipv4",
        "traddr": "10.0.1.1",
        "trsvcid": "4434",
        "trtype": "tcp"
      },
      "portid": 15,
      "referrals": [],
      "subsystems": [
        "nqn.2018-09.io.spdk:cnode15"
      ]
    },
    {
      "addr": {
        "adrfam": "ipv4",
        "traddr": "10.0.1.1",
        "trsvcid": "4435",
        "trtype": "tcp"
      },
      "portid": 16,
```

```
      "referrals": [],
      "subsystems": [
        "nqn.2018-09.io.spdk:cnode16"
      ]
    }
  ],
"hosts": [],
 "subsystems": [
    {
      "allowed_hosts": [],
      "attr": {
        "allow_any_host": "1",
        "version": "1.3"
      },
      "namespaces": [
        {
          "device": {
            "path": "/dev/nvme0n1",
            "uuid": "b53be81d-6f5c-4768-b3bd-203614d8cf20"
          },
          "enable": 1,
          "nsid": 1
        }
      ],
      "nqn": "nqn.2018-09.io.spdk:cnode1"
    },
    {
      "allowed_hosts": [],
      "attr": {
        "allow_any_host": "1",
        "version": "1.3"
      },
      "namespaces": [
        {
          "device": {
            "path": "/dev/nvme1n1",
            "uuid": "12fcf584-9c45-4b6b-abc9-63a763455cf7"
          },
          "enable": 1,
          "nsid": 2
        }
      ],
      "nqn": "nqn.2018-09.io.spdk:cnode2"
    },
    {
      "allowed_hosts": [],
      "attr": {
        "allow_any_host": "1",
        "version": "1.3"
      },
      "namespaces": [
        {
          "device": {
            "path": "/dev/nvme2n1",
            "uuid": "ceae8569-69e9-4831-8661-90725bdf768d"
          },
          "enable": 1,
```

```
        "nsid": 3
      }
    ],
    "nqn": "nqn.2018-09.io.spdk:cnode3"
  },
 {
   "allowed_hosts": [],
    "attr": {
      "allow_any_host": "1",
      "version": "1.3"
    },
    "namespaces": [
      {
        "device": {
          "path": "/dev/nvme3n1",
          "uuid": "39f36db4-2cd5-4f69-b37d-1192111d52a6"
        },
        "enable": 1,
        "nsid": 4
      }
    ],
    "nqn": "nqn.2018-09.io.spdk:cnode4"
  },
  {
    "allowed_hosts": [],
    "attr": {
      "allow_any_host": "1",
      "version": "1.3"
    },
    "namespaces": [
      {
        "device": {
          "path": "/dev/nvme4n1",
          "uuid": "984aed55-90ed-4517-ae36-d3afb92dd41f"
        },
        "enable": 1,
        "nsid": 5
      }
    ],
    "nqn": "nqn.2018-09.io.spdk:cnode5"
  },
  {
    "allowed_hosts": [],
    "attr": {
      "allow_any_host": "1",
      "version": "1.3"
    },
    "namespaces": [
      {
        "device": {
          "path": "/dev/nvme5n1",
          "uuid": "d6d16e74-378d-40ad-83e7-b8d8af3d06a6"
        },
        "enable": 1,
        "nsid": 6
      }
    ],
```

```
      "nqn": "nqn.2018-09.io.spdk:cnode6"
    },
    {
      "allowed_hosts": [],
      "attr": {
        "allow_any_host": "1",
        "version": "1.3"
      },
      "namespaces": [
        {
          "device": {
            "path": "/dev/nvme6n1",
            "uuid": "a65dc00e-d35c-4647-9db6-c2a8d90db5e8"
          },
          "enable": 1,
          "nsid": 7
        }
      ],
      "nqn": "nqn.2018-09.io.spdk:cnode7"
    },
    {
      "allowed_hosts": [],
      "attr": {
        "allow_any_host": "1",
        "version": "1.3"
      },
      "namespaces": [
        {
          "device": {
            "path": "/dev/nvme7n1",
            "uuid": "1b242cb7-8e47-4079-a233-83e2cd47c86c"
          },
          "enable": 1,
          "nsid": 8
        }
      ],
      "nqn": "nqn.2018-09.io.spdk:cnode8"
    },
    {
      "allowed_hosts": [],
      "attr": {
        "allow_any_host": "1",
        "version": "1.3"
      },
      "namespaces": [
        {
          "device": {
            "path": "/dev/nvme8n1",
            "uuid": "f12bb0c9-a2c6-4eef-a94f-5c4887bbf77f"
          },
          "enable": 1,
          "nsid": 9
        }
      ],
      "nqn": "nqn.2018-09.io.spdk:cnode9"
    },
    {
```

```
      "allowed_hosts": [],
      "attr": {
        "allow_any_host": "1",
        "version": "1.3"
      },
      "namespaces": [
        {
          "device": {
            "path": "/dev/nvme9n1",
            "uuid": "40fae536-227b-47d2-bd74-8ab76ec7603b"
          },
          "enable": 1,
          "nsid": 10
        }
      ],
      "nqn": "nqn.2018-09.io.spdk:cnode10"
    },
    {
      "allowed_hosts": [],
      "attr": {
        "allow_any_host": "1",
        "version": "1.3"
      },
      "namespaces": [
        {
          "device": {
            "path": "/dev/nvme10n1",
            "uuid": "b9756b86-263a-41cf-a68c-5cfb23c7a6eb"
          },
          "enable": 1,
          "nsid": 11
        }
      ],
      "nqn": "nqn.2018-09.io.spdk:cnode11"
    },
    {
      "allowed_hosts": [],
      "attr": {
        "allow_any_host": "1",
        "version": "1.3"
      },
      "namespaces": [
        {
          "device": {
            "path": "/dev/nvme11n1",
            "uuid": "9d7e74cc-97f3-40fb-8e90-f2d02b5fff4c"
          },
          "enable": 1,
          "nsid": 12
        }
      ],
      "nqn": "nqn.2018-09.io.spdk:cnode12"
    },
{
      "allowed_hosts": [],
      "attr": {
        "allow_any_host": "1",
```

```
        "version": "1.3"
      },
      "namespaces": [
        {
          "device": {
            "path": "/dev/nvme12n1",
            "uuid": "d3f4017b-4f7d-454d-94a9-ea75ffc7436d"
          },
          "enable": 1,
          "nsid": 13
        }
      ],
      "nqn": "nqn.2018-09.io.spdk:cnode13"
    },
    {
      "allowed_hosts": [],
      "attr": {
        "allow_any_host": "1",
        "version": "1.3"
      },
      "namespaces": [
        {
          "device": {
            "path": "/dev/nvme13n1",
            "uuid": "6b9a65a3-6557-4713-8bad-57d9c5cb17a9"
          },
          "enable": 1,
          "nsid": 14
        }
      ],
      "nqn": "nqn.2018-09.io.spdk:cnode14"
    },
    {
      "allowed_hosts": [],
      "attr": {
        "allow_any_host": "1",
        "version": "1.3"
      },
      "namespaces": [
        {
          "device": {
            "path": "/dev/nvme14n1",
            "uuid": "ed69ba4d-8727-43bd-894a-7b08ade4f1b1"
          },
          "enable": 1,
          "nsid": 15
        }
      ],
      "nqn": "nqn.2018-09.io.spdk:cnode15"
    },
  {
      "allowed_hosts": [],
      "attr": {
        "allow_any_host": "1",
        "version": "1.3"
      },
      "namespaces": [
```

```
        {
          "device": {
            "path": "/dev/nvme15n1",
            "uuid": "5b8e9af4-0ab4-47fb-968f-b13e4b607f4e"
          },
          "enable": 1,
          "nsid": 16
        }
      ],
      "nqn": "nqn.2018-09.io.spdk:cnode16"
    }
  ]
}
```

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.

§