

SPDK Vhost Performance Report

Release 21.07

Testing Date: August 2021

Performed by:

Karol Latecki (karol.latecki@intel.com)

Maciej Wawryk (maciejx.wawryk@intel.com)

Acknowledgments:

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

Contents

Contents	2
Audience and Purpose.....	3
Test setup	4
Hardware configuration	4
BIOS Settings	5
Virtual Machine Settings.....	5
Kernel & BIOS Spectre-Meltdown information	6
Introduction to the SPDK Vhost target	7
SPDK Vhost target architecture	7
Test Case 1: SPDK Vhost Core Scaling	9
4KB Random Read Results	11
4KB Random Write Results	12
4KB Random Read-Write Results.....	13
Logical Volumes performance impact	14
LTO performance impact	15
Packed Ring performance impact.....	16
Test Case 2: Rate Limiting IOPS per VM.....	18
Test Case 2 Results	20
Conclusions	22
Test Case 3: Performance per NVMe drive	23
Test Case 3 results	24
Conclusions	27
Summary	28
List of Tables	29
List of Figures	30

Audience and Purpose


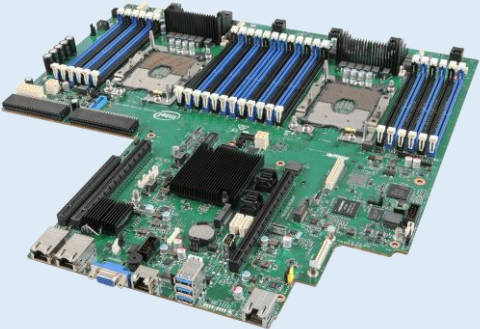
This report is intended for people who are interested in looking at SPDK Vhost scsi and blk stack performance and comparison to its Linux kernel equivalents. It provides performance and efficiency comparisons between SPDK Vhost-scsi and Linux Kernel Vhost-scsi software stacks under various test cases.

The purpose of this report is not to imply a single correct approach, but rather to provide a baseline of well-tested configurations and procedures that produce repeatable and reproducible results. This report can also be viewed as information regarding best known method when performance testing SPDK Vhost-scsi and Vhost-blk stacks.

Test setup

Hardware configuration

Table 1: Hardware setup configuration

Item	Description														
Server Platform	<p>Intel WolfPass R2224WFTZS</p>  <p>Server board S2600WFT</p> 														
Motherboard	S2600WFT														
CPU	<p>2 CPU sockets, Intel(R) Xeon(R) Gold 6230N CPU @ 2.30GHz</p> <p>Number of cores 20 per socket, number of threads 40 per socket</p> <p>Both sockets populated</p> <p>Microcode: 0x4003003</p>														
Memory	<p>12 x 32GB Micron DDR4 36ASF4G72PZ-2G9E2</p> <p>Total 384 GBs</p> <p>Memory channel population:</p> <table border="1"> <thead> <tr> <th>P1</th> <th>P2</th> </tr> </thead> <tbody> <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> </tbody> </table>	P1	P2	CPU1_DIMM_A1	CPU2_DIMM_A1	CPU1_DIMM_B1	CPU2_DIMM_B1	CPU1_DIMM_C1	CPU2_DIMM_C1	CPU1_DIMM_D1	CPU2_DIMM_D1	CPU1_DIMM_E1	CPU2_DIMM_E1	CPU1_DIMM_F1	CPU2_DIMM_F1
P1	P2														
CPU1_DIMM_A1	CPU2_DIMM_A1														
CPU1_DIMM_B1	CPU2_DIMM_B1														
CPU1_DIMM_C1	CPU2_DIMM_C1														
CPU1_DIMM_D1	CPU2_DIMM_D1														
CPU1_DIMM_E1	CPU2_DIMM_E1														
CPU1_DIMM_F1	CPU2_DIMM_F1														
Operating System	Fedora 33														
BIOS	SE5C620.86B.02.01.0013.121520200651														

Linux kernel version	5.11.20-200.fc33.x86_64
SPDK version	SPDK 21.04
Qemu version	QEMU emulator version 5.1.0 (qemu-5.1.0-9.fc33)
Storage	<p>OS: 1x 120GB Intel SSDSC2BB120G4</p> <p>Storage: 24x Intel® P4610™ 1.6TBs (FW: VDV10170) (6 on CPU NUMA Node 0, 18 on CPU NUMA Node 1)</p>

BIOS Settings

Table 2: Test platform BIOS settings

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

Virtual Machine Settings

Table 3: Guest VM configuration

Item	Description
CPU	<p>2vCPU, pass through from physical host server.</p> <p>Explicit core usage enforced using “taskset –a –c” command.</p> <p>QEMU arguments for starting the VM: -cpu host -smp 1</p>
Memory	<p>4 GB RAM. Memory is pre-allocated for each VM using Hugepages on host system and used from appropriate NUMA node, to match the CPU which was passed to the VM.</p> <p>QEMU arguments: -m 4096 -object memory-backend-file,id=mem,size=4096M,mem-path=/dev/hugepages,share=on,prealloc=yes,host-nodes=0,policy=bind</p>
Operating System	Fedora 33
Linux kernel version	5.9.16-200.fc33.x86_64
Additional boot options in /etc/default/grub	<ul style="list-style-type: none"> Multi queue enabled: scsi_mod.use_blk_mq=1

Kernel & BIOS Spectre-Meltdown information

Host server and guest OS systems use 5.11.20 and 5.9.16 kernel versions respectively, available from the DNF repository. The default Spectre-Meltdown mitigation patches for these kernel versions have been left enabled.

Introduction to the SPDK Vhost target

SPDK Vhost is a userspace target designed to extend the performance efficiencies of SPDK into QEMU/KVM virtualization environments. The SPDK Vhost-scsi target presents a broad range of SPDK-managed block devices into virtual machines. SPDK team has leveraged existing SPDK SCSI layer, DPDK Vhost library, QEMU Vhost-scsi and Vhost-user functionality in order to create the high performance SPDK userspace Vhost target.

SPDK Vhost target architecture

QEMU setups Vhost target via UNIX domain socket. The Vhost target transfers data to/from the guest VM via shared memory. QEMU pre-allocates huge pages for the guest VM to enable DMA by the Vhost target. The guest VM submits I/O directly to the Vhost target via virtqueues in shared memory as shown in Figure 1 on example of virtio-scsi. It should be noted that there is no QEMU intervention during the I/O submission process. The Vhost target then completes I/O to the guest VM via virtqueues in shared memory. There is a completion interrupt sent using eventfd which requires a system call and a guest VM exit.

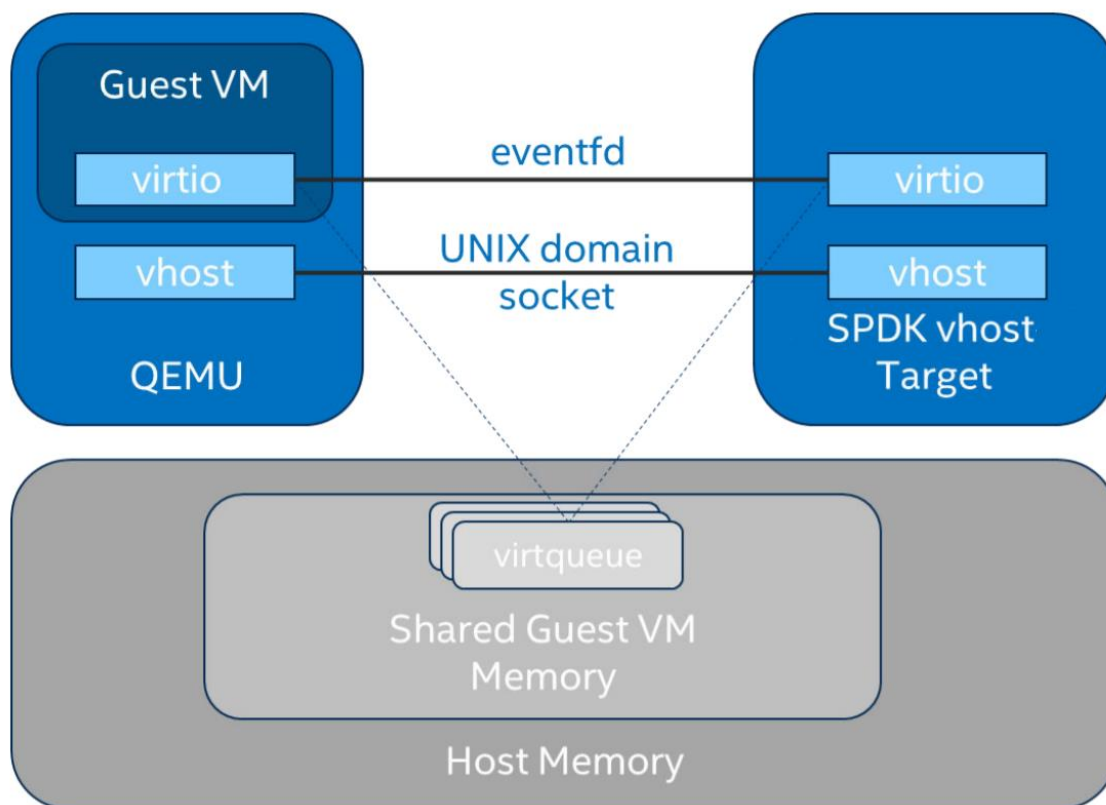


Figure 1: SPDK Vhost-scsi architecture

This report shows the performance comparisons between the traditional interrupt-driven kernel Vhost-scsi and the accelerated polled-mode driven SPDK Vhost-scsi under 3 different test cases using local

NVMe storage. Additionally, the SPDK Vhost-blk stack is included in the report for further comparison with the SCSI stack.

Test Case 1: SPDK Vhost Core Scaling

This test case was performed in order to understand aggregate VM performance with SPDK Vhost I/O core scaling. We ran up to 36 virtual machines, each running following FIO workloads:

- 4KB 100% Random Read
- 4KB 100% Random Write
- 4KB Random 70% Read / 30 % Write

We increased the number of CPU cores used by SPDK Vhost target to process I/O from 1 up to 12 and measured the throughput (in IOPS) and latency. The number of VMs between test runs was not constant and was increased by 6 for each Vhost CPU added, up to a maximum of 36 VMs. VM number was not increased beyond 36 because of the platform capabilities in terms of available CPU cores.

FIO was run in client-server mode. FIO client was run on the host machine and distributed jobs to FIO servers run on each VM. This allowed us to start the FIO jobs across all VMs at the same time. The `gtod_reduce=1` option was used to disable FIO latency measurements which allowed better IOPS and bandwidth results.

Results in the table and charts represent aggregate performance (IOPS and average latency) seen across all the VMs. The results are average of 3 runs.

Table 4: SPDK Vhost Core Scaling test configuration

Item	Description
Test case	Test SPDK Vhost target I/O core scaling performance
Test configuration	<p>FIO Version: fio-3.19</p> <p>VM Configuration:</p> <ul style="list-style-type: none">• Common settings are described in the Virtual Machine Settings chapter.• Number of VMs: variable (6 VMs per 1 Vhost CPU core, up to 36 VMs max).• Each VM has a single Vhost device as a target for the FIO workload. This is achieved by sharing SPDK NVMe bdevs by using either a Split NVMe vbdev or Logical Volume bdev configuration. <p>SPDK Vhost target configuration:</p> <ul style="list-style-type: none">• Test were run with both the Vhost-scsi and Vhost-blk stacks.• The Vhost-scsi stack was run with Split NVMe bdevs and Logical Volume bdevs.• Vhost-blk stack was run with Logical Volume bdevs.• Tests were ran with 1, 2, 4, 6, 8, 10 and 12 cores for each stack-bdev combination. <p>Kernel Vhost target configuration:</p> <ul style="list-style-type: none">- N/A

FIO configuration	<pre>[global] ioengine=libaio direct=1 thread=1 norandommap=1 time_based=1 gtod_reduce=1 ramp_time=60s runtime=240s numjobs=1 bs=4k rw=randrw rwmixread=100 (100% reads), 70 (70% reads, 30% writes), 0 (100% writes) iodepth={1, 32, 64}</pre>
--------------------------	---

4KB Random Read Results

Table 5: SPDK Vhost core scaling results, 4KB 100% Random Reads IOPS, QD=64

# of CPU cores	# of VMs	Stack / Backend	IOPS (millions)
1	6	SCSI / Split NVMe Bdev	1.83
		SCSI / Lvol Bdev	1.53
		BLK / Lvol Bdev	1.57
2	12	SCSI / Split NVMe Bdev	2.94
		SCSI / Lvol Bdev	2.55
		BLK / Lvol Bdev	2.74
4	24	SCSI / Split NVMe Bdev	5.03
		SCSI / Lvol Bdev	3.90
		BLK / Lvol Bdev	4.10
6	36	SCSI / Split NVMe Bdev	6.62
		SCSI / Lvol Bdev	5.24
		BLK / Lvol Bdev	5.58
8	36	SCSI / Split NVMe Bdev	7.20
		SCSI / Lvol Bdev	6.56
		BLK / Lvol Bdev	6.94
10	36	SCSI / Split NVMe Bdev	7.12
		SCSI / Lvol Bdev	6.70
		BLK / Lvol Bdev	7.44
12	36	SCSI / Split NVMe Bdev	7.42
		SCSI / Lvol Bdev	7.30
		BLK / Lvol Bdev	8.00

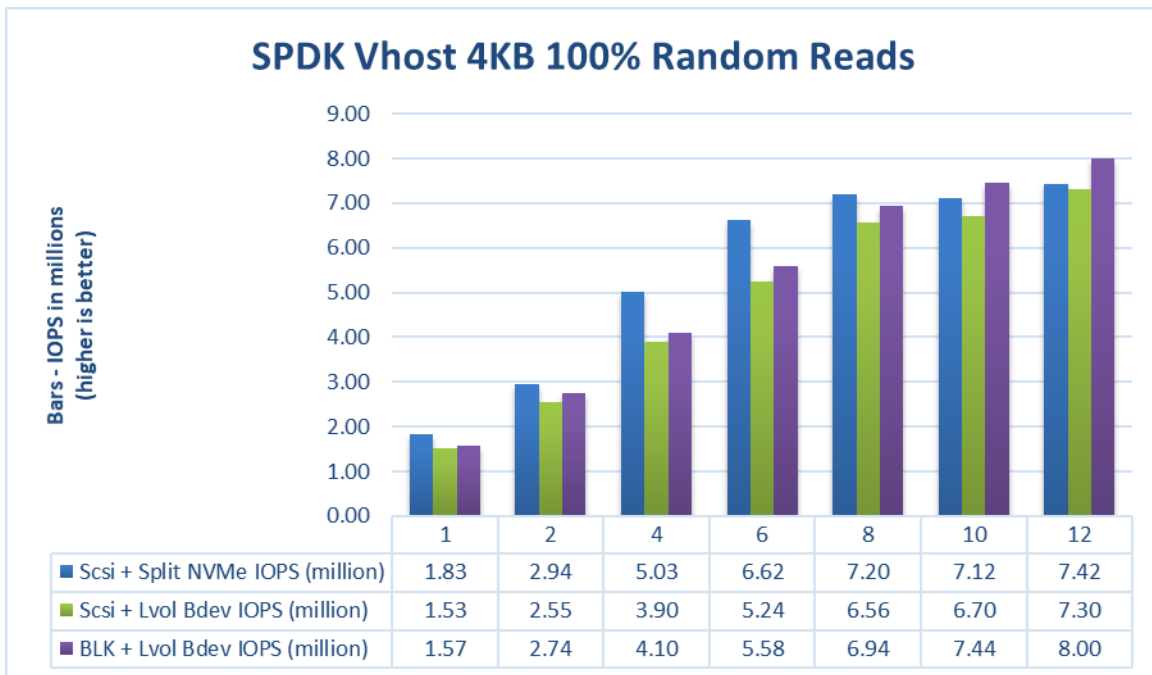


Figure 2: Comparison of performance between various SPDK Vhost stack-bdev combinations for 4KB Random Read QD=64 workload

4KB Random Write Results

Table 6: SPDK Vhost core scaling results, 4KB 100% Random Write IOPS, QD=32

# of CPU cores	# of VMs	Stack / Backend	IOPS (millions)
1	6	SCSI / Split NVMe Bdev	1.67
		SCSI / Lvol Bdev	1.51
		BLK / Lvol Bdev	1.58
2	12	SCSI / Split NVMe Bdev	2.95
		SCSI / Lvol Bdev	2.63
		BLK / Lvol Bdev	2.93
4	24	SCSI / Split NVMe Bdev	5.01
		SCSI / Lvol Bdev	4.35
		BLK / Lvol Bdev	4.68
6	36	SCSI / Split NVMe Bdev	6.02
		SCSI / Lvol Bdev	5.52
		BLK / Lvol Bdev	5.90
8	36	SCSI / Split NVMe Bdev	6.78
		SCSI / Lvol Bdev	6.49
		BLK / Lvol Bdev	6.91
10	36	SCSI / Split NVMe Bdev	6.51
		SCSI / Lvol Bdev	6.49
		BLK / Lvol Bdev	7.13
12	36	SCSI / Split NVMe Bdev	6.81
		SCSI / Lvol Bdev	7.01
		BLK / Lvol Bdev	7.56

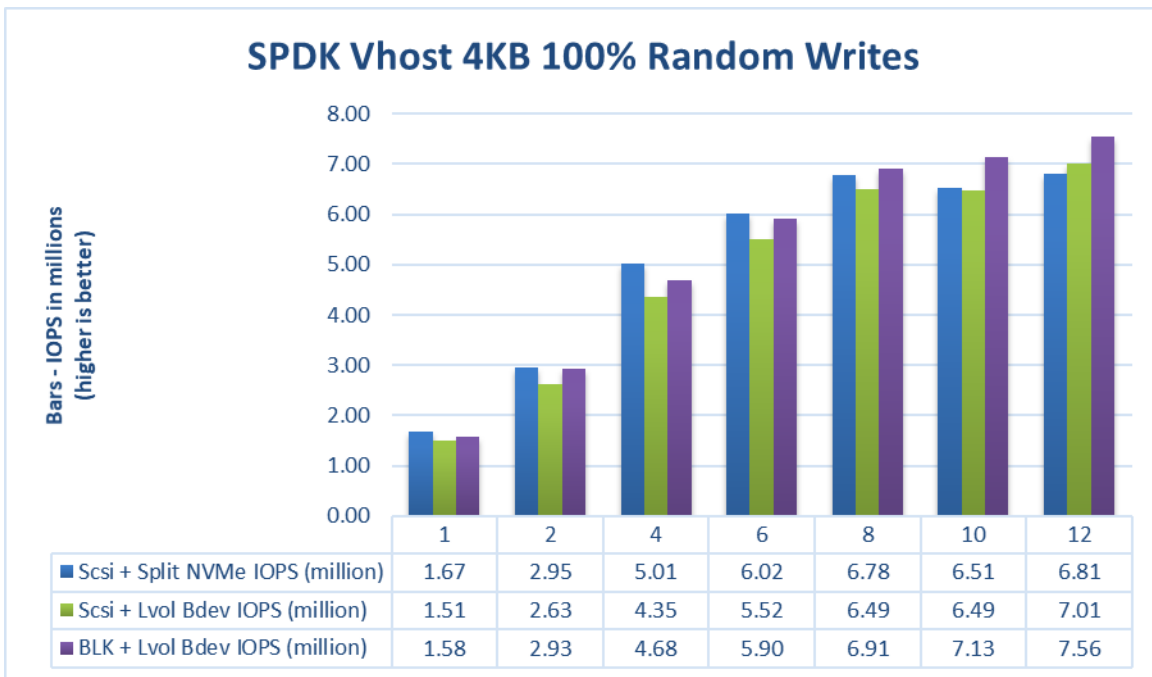


Figure 3: Comparison of performance between various SPDK Vhost stack-bdev combinations for 4KB Random Write QD=32 workload

4KB Random Read-Write Results

Table 7: SPDK Vhost core scaling results, 4KB Random 70% Read 30% Write IOPS, QD=64

# of CPU cores	# of VMs	Stack / Backend	IOPS (millions)
1	6	SCSI / Split NVMe Bdev	1.70
		SCSI / Lvol Bdev	1.48
		BLK / Lvol Bdev	1.53
2	12	SCSI / Split NVMe Bdev	2.78
		SCSI / Lvol Bdev	2.46
		BLK / Lvol Bdev	2.69
4	24	SCSI / Split NVMe Bdev	4.77
		SCSI / Lvol Bdev	3.98
		BLK / Lvol Bdev	4.41
6	36	SCSI / Split NVMe Bdev	6.30
		SCSI / Lvol Bdev	5.24
		BLK / Lvol Bdev	5.64
8	36	SCSI / Split NVMe Bdev	6.54
		SCSI / Lvol Bdev	6.17
		BLK / Lvol Bdev	6.64
10	36	SCSI / Split NVMe Bdev	6.12
		SCSI / Lvol Bdev	6.04
		BLK / Lvol Bdev	6.63
12	36	SCSI / Split NVMe Bdev	6.37
		SCSI / Lvol Bdev	6.47
		BLK / Lvol Bdev	7.02

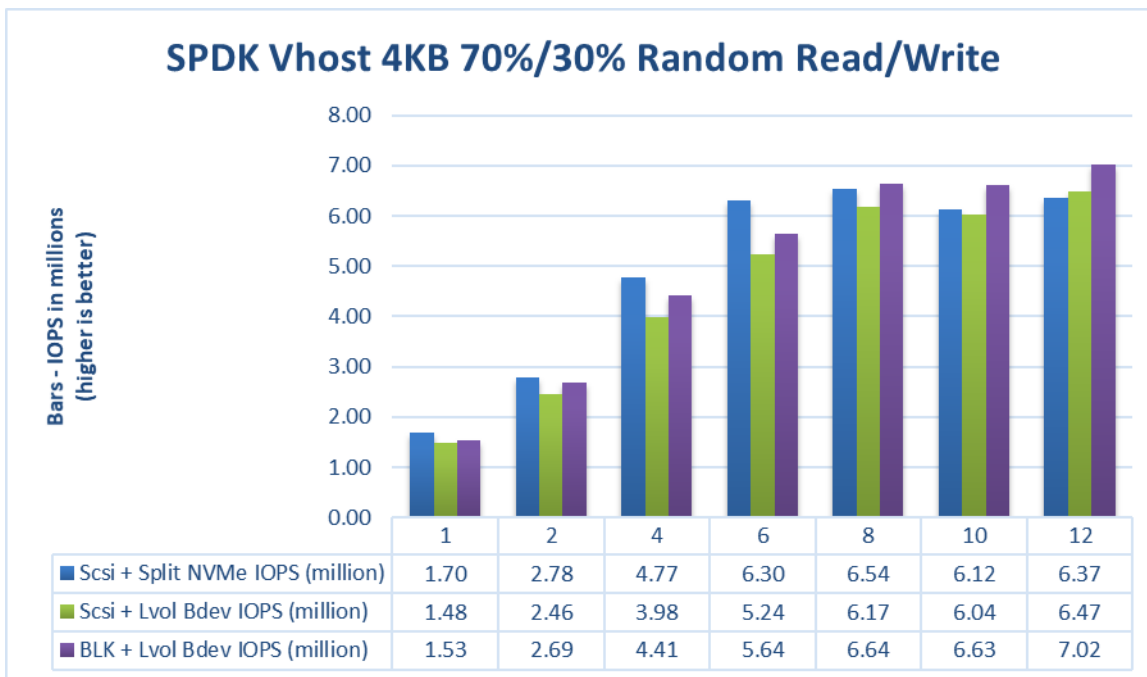


Figure 4: Comparison of performance between various SPDK Vhost stack-bdev combinations for 4KB Random 70% Read 30% Write QD=64 workload

Logical Volumes performance impact

The SPDK Vhost SCSI tests were run using two bdev backends – Split NVMe and Logical Volumes. Both “Split NVMe Bdevs” and “Logical Volume Bdevs” allow to logically partition NVMe SSDs, the latter being more flexible in configuration. Here we measure the overhead of extra flexibility afforded by Logical Volumes.

Table 8: Logical Volumes performance impact for SPDK Vhost SCSI

Workload	# of CPU cores	# of VMs	Vhost SCSI + Split NVMe IOPS (millions)	Vhost SCSI + Lvol IOPS (millions)	Lvol Impact (%)
4KB 100% Random Read	1	6	1.83	1.53	-16.31%
	2	12	2.94	2.55	-13.34%
	4	24	5.03	3.90	-22.53%
	6	36	6.62	5.24	-20.77%
	8	36	7.20	6.56	-8.96%
	10	36	7.12	6.70	-5.81%
	12	36	7.42	7.30	-1.54%
4KB 100% Random Write	1	6	1.67	1.51	-9.69%
	2	12	2.95	2.63	-10.96%
	4	24	5.01	4.35	-13.22%
	6	36	6.02	5.52	-8.30%
	8	36	6.78	6.49	-4.23%
	10	36	6.51	6.49	-0.44%
	12	36	6.81	7.01	3.01%
4KB 70% Random Read 30% Random Write	1	6	1.70	1.48	-12.68%
	2	12	2.78	2.46	-11.71%
	4	24	4.77	3.98	-16.46%
	6	36	6.30	5.24	-16.94%
	8	36	6.54	6.17	-5.73%
	10	36	6.12	6.04	-1.33%
	12	36	6.37	6.47	1.62%

LTO performance impact

Selected test cases were re-run with LTO (Link Time Optimization) enabled for SPDK compilation. This should positively impact overall SPDK performance. The following comparison was done using SPDK Vhost SCSI with Logical Volume bdevs.

Table 9: LTO performance SPDK Vhost SCSI with Logical Volume bdevs

Workload	# of CPU cores	# of VMs	IOPS (millions) LTO Disabled	IOPS (millions) LTO Enabled	LTO Impact (%)
4KB 100% Random Read	1	6	1.53	1.68	10.05%
	2	12	2.55	2.77	8.63%
	4	24	3.90	4.42	13.34%
	6	36	5.24	5.72	9.02%
	8	36	6.56	6.83	4.18%
	10	36	6.70	6.76	0.85%
	12	36	7.30	7.22	-1.18%
4KB 100% Random Write	1	6	1.51	1.59	5.60%
	2	12	2.63	2.82	7.12%
	4	24	4.35	4.65	7.01%
	6	36	5.52	5.80	5.19%
	8	36	6.49	6.78	4.45%
	10	36	6.49	6.48	-0.03%
	12	36	7.01	6.85	-2.42%
4KB 70% Random Read 30% Random Write	1	6	1.48	1.58	6.72%
	2	12	2.46	2.65	7.97%
	4	24	3.98	4.33	8.64%
	6	36	5.24	5.73	9.35%
	8	36	6.17	6.30	2.14%
	10	36	6.04	6.09	0.92%
	12	36	6.47	6.38	-1.43%

Packed Ring performance impact

Selected test cases were re-run to show benefits of using Packed Rings as an option when configuring SPDK Vhost BLK controllers. For this, an optional parameter “`—packed_ring`” must be used when creating a SPDK Vhost BLK controller. Packed Ring feature requires QEMU 4.2.0 or later.

Following results show comparison of running SPDK Vhost-Blk with Packed Ring enabled with fio latency measurements both enabled and disabled. Because other QEMU version was used, base results (Split Ring) were run again to produce a fresh base to compare to.

Table 10: Packed Ring performance impact on SPDK Vhost BLK controllers. *Fio gtd_reduce=disabled*

Workload	# of CPU cores	# of VMs	IOPS (millions) Split Ring	IOPS (millions) Packed Ring	Avg. Latency (usec) Split Ring	Avg. Latency (usec) Packed Ring	Packed Ring IOPS impact (%)	Packed Ring Avg. Latency impact (%)
4KB 100% Random Read QD=64	1	6	1.57	1.55	243.70	247.02	-1.29%	1.37%
	2	12	2.74	2.75	279.43	278.30	0.31%	-0.41%
	4	24	4.22	4.29	362.86	358.13	1.59%	-1.31%
	6	36	5.56	5.57	412.51	412.93	0.31%	0.10%
	8	36	6.76	6.86	339.84	334.32	1.41%	-1.62%
	10	36	7.07	7.11	321.73	329.08	0.59%	2.29%
	12	36	7.55	7.57	302.78	304.18	0.20%	0.46%
4KB 100% Random Write QD=32	1	6	1.50	1.50	129.98	129.63	0.07%	-0.26%
	2	12	2.89	2.93	131.67	130.67	1.38%	-0.76%
	4	24	4.65	4.69	164.59	163.35	0.98%	-0.76%
	6	36	5.70	5.72	205.18	201.49	0.32%	-1.80%
	8	36	6.79	6.70	169.79	172.33	-1.25%	1.49%
	10	36	6.71	6.90	171.23	169.23	2.76%	-1.17%
	12	36	7.10	7.38	161.93	155.83	3.91%	-3.77%
4KB 70% Random Read 30% Random Write QD=64	1	6	1.55	1.55	251.47	247.36	0.30%	-1.63%
	2	12	2.67	2.74	287.26	277.12	2.40%	-3.53%
	4	24	4.16	4.27	368.01	356.85	2.63%	-3.03%
	6	36	5.62	5.65	408.62	411.45	0.53%	0.69%
	8	36	6.40	6.33	360.61	361.55	-1.05%	0.26%
	10	36	6.25	6.36	368.40	372.17	1.66%	1.02%
	12	36	6.56	6.60	349.81	350.15	0.60%	0.10%

Conclusions

1. For SPDK Vhost SCSI performance with split NVMe bdevs, we measured 1.67 million IOPS on one Vhost core for the 4KB 100% Random Write workload. The single Vhost core IOPS for 4 KB Random Read and 4KB Random 70/30 Read/Write were 1.83 million and 1.70 million IOPS respectively. For all workloads, the IOPS scaled near linearly with addition of I/O processing cores up to 6 CPU cores. Peak performance was achieved at 8 CPU cores for all workloads. Further increasing the number of cores does not result in performance improvement or it is not significant.
2. For SPDK Vhost SCSI with Logical Volume backend devices, we measured about 1.5 million IOPS on one Vhost core for all 3 workloads. Performance scaled near linearly with addition of I/O processing cores up to 8 CPU cores. Increasing the number of I/O processing cores further results in non-linear IOPS gains. Peak performance is reached at 12 CPU cores for all workloads.
3. For SPDK Vhost BLK with Logical Volume backend devices, we measured about 1.5 million IOPS on one Vhost core for all workloads. Performance scaled near linearly up to 8 CPU cores. Increasing the number of cores improves performance further, but the gains are not linear.
4. Using Logical Volumes has a noticeable impact on the overall performance. For Vhost tests using 6 or less CPUs (when Vhost is saturated with IO traffic from VMs) performance impact of Logical Volumes is between 10-22%. Further increasing SPDK Vhost CPU cores allow Logical Volumes to perform better and their performance impact is on par with Split NVMe Bdevs (less than 10% difference).
5. LTO compilation option increased SPDK Vhost performance by up to 13% percent in the scaling phase (8 Vhost CPU cores or less) for Random Read and Random Read/Write workloads. Beyond 8 vhost cores, LTO benefit is negligible. The reason for this behavior is described in point 6.
6. For some workloads there is a slight performance drop when Vhost is run with 10 or 12 CPU cores. The platform has 80 CPU threads available, and 36 VMs use 72 CPU threads. Therefore, when 10 or 12 are used for the Vhost process there is not enough left to accommodate all the VMs. Some of the VMs share CPU threads, thus becoming less efficient.
7. Using Packed Ring option instead of default Split Ring mode for SPDK Vhost BLK controllers results in minor performance improvement.

Test Case 2: Rate Limiting IOPS per VM

This test case was geared towards understanding how many VMs can be supported at a pre-defined Quality of Service of IOPS per Vhost device. Both read and write IOPS were rate limited for each Vhost device on each of the VMs and then VM density was compared between SPDK & the Linux Kernel. 10K IOPS were chosen as the rate limiter using linux cgroups2.

Each individual VM was running FIO with the following workloads:

- 4KB 100% Random Read
- 4KB 100% Random Write

The results in tables are average of 3 runs.

Table 11: Rate Limiting IOPS per VM test case configuration

Item	Description
Test case	Test rate limiting IOPS/VM to 10000 IOPS
Test configuration	<p>FIO Version: fio-3.19</p> <p>VM Configuration:</p> <ul style="list-style-type: none"> • Common settings are described in the Virtual Machine Settings chapter. • Number of VMs: 24 / 48 / 72 • Each VM has a single Vhost device which is one of equal partitions of NVMe drive. Total number of partitions depends on run test case. <ul style="list-style-type: none"> ○ For 24 VMs: 24xNVMe * 1 partition per NVMe = 24 partitions ○ For 48 VMs: 24xNVMe * 2 partitions per NVMe = 48 partitions ○ For 72 VMs: 24xNVMe * 3 partitions per NVMe = 72 partitions • Devices on VMs were throttled to run at a maximum of 10k IOPS (read and write) <p>SPDK Vhost target configuration:</p> <ul style="list-style-type: none"> • Test were run with both Vhost-scsi and Vhost-blk stacks. • The Vhost-scsi stack was run with Split NVMe bdevs and Logical Volume bdevs. • The Vhost-blk stack was run with Logical Volume bdevs. • Test were run with Vhost using 4 CPU cores (NUMA optimized). <p>Kernel Vhost-scsi configuration:</p> <ul style="list-style-type: none"> • Cgroups were used to limit the Vhost process to 4 cores. • NUMA optimization were not explored.
FIO configuration run on each VM	<pre>[global] ioengine=libaio direct=1 rw=randrw</pre>

	<pre>rwmixread=100 (100% reads), 0 (100% writes) thread=1 norandommap=1 time_based=1 runtime=300s ramp_time=10s bs=4k iodepth=1 numjobs=1</pre>
--	---

Test Case 2 Results

Table 12: 4KB 100% Random Reads QD=1 rate limiting test results

# of VMs	Stack	Backend bdev	IOPS (k)	Avg Lat. (usec)
24 VMs	SPDK-SCSI	Split NVMe	239.98	99.43
	SPDK-SCSI	Logical Volume	239.98	99.44
	SPDK-BLK	Logical Volume	239.97	99.46
	Kernel-SCSI	Partitioned NVMe	208.00	114.74
48 VMs	SPDK-SCSI	Split NVMe	479.84	99.09
	SPDK-SCSI	Logical Volume	479.77	99.10
	SPDK-BLK	Logical Volume	479.88	99.11
	Kernel-SCSI	Partitioned NVMe	235.73	206.96
72 VMs	SPDK-SCSI	Split NVMe	684.70	104.08
	SPDK-SCSI	Logical Volume	669.48	106.48
	SPDK-BLK	Logical Volume	685.55	103.92
	Kernel-SCSI	Partitioned NVMe	316.11	227.19

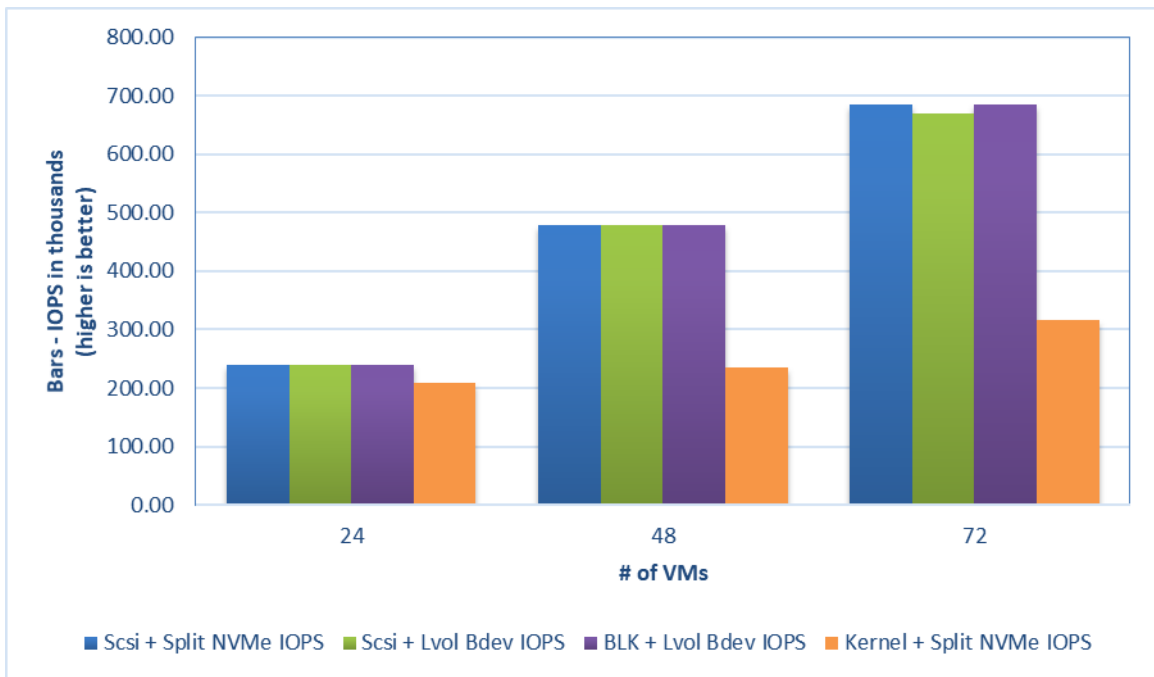


Figure 5: 4KB 100% Random Reads IOPS, QD=1, throttling = 10k IOPS

Table 13: 4KB 100% Random Writes QD=1 rate limiting test results

# of VMs	Stack	Backend bdev	IOPS (k)	Avg Lat. (usec)
24 VMs	SPDK-SCSI	Split NVMe	240.00	99.45
	SPDK-SCSI	Logical Volume	240.00	99.44
	SPDK-BLK	Logical Volume	240.00	99.47
	Kernel-SCSI	Partitioned NVMe	197.24	124.45
48 VMs	SPDK-SCSI	Split NVMe	479.96	99.24
	SPDK-SCSI	Logical Volume	479.98	99.16
	SPDK-BLK	Logical Volume	479.99	99.23
	Kernel-SCSI	Partitioned NVMe	222.23	221.07
72 VMs	SPDK-SCSI	Split NVMe	719.97	99.08
	SPDK-SCSI	Logical Volume	719.96	99.05
	SPDK-BLK	Logical Volume	719.93	99.08
	Kernel-SCSI	Partitioned NVMe	343.04	209.04

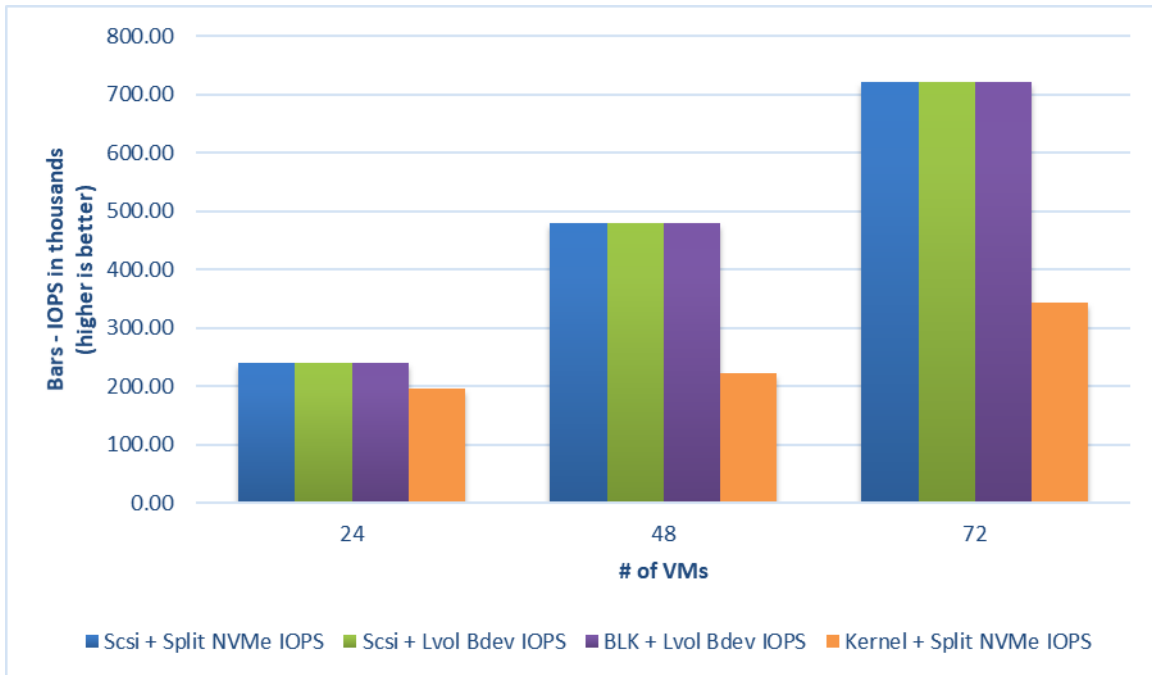


Figure 6: 4KB 100% Random Writes IOPS, QD=1, throttling = 10k IOPS

Conclusions

1. Using just 4 I/O processing cores, the SPDK vhost served 10,000 IOPS/VM to up to 48 VMs for 4 KB random read workload and 72 VMs for the 4 KB random write workload.
2. The Kernel Vhost was not able to serve IO at 10K IOPS/VM with just 4 I/O processing cores.
3. Average latencies were up to 2.18x times better for Random Read and up to 2.23x times better for Random Write workloads with the SPDK Vhost when compared to Kernel Vhost.

Note: The Kernel-Vhost process was not NUMA-optimized for this scenario.

Test Case 3: Performance per NVMe drive

This test case was performed in order to understand performance and efficiency of the Vhost scsi and blk process using SPDK vs. Linux Kernel with a single NVMe drive on 2 VMs. Each VM had a single Vhost device which is one of two equal partitions of an NVMe drive. Results in the table represent performance (IOPS, avg. latency & CPU %) seen from the VM. The VM was running FIO with the following workloads:

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

The results in tables are average of 3 runs.

Table 14: Performance per NVMe drive test case configuration

Item	Description
Test case	Test SPDK Vhost target I/O core scaling performance
Test configuration	<p>FIO Version: fio-3.19</p> <p>VM Configuration:</p> <ul style="list-style-type: none"> • Common settings are described in the Virtual Machine Settings chapter. • 2 VMs were tested • Each VM had a single Vhost device which was one of two equal partitions of a single NVMe drive. <p>SPDK Vhost target configuration:</p> <ul style="list-style-type: none"> • The SPDK Vhost process was run on a single, physical CPU core. • The Vhost-scsi stack was run with Split NVMe bdevs and Logical Volume bdevs. • The Vhost-blk stack was run with Logical Volume bdevs. <p>Kernel Vhost target configuration:</p> <ul style="list-style-type: none"> • The Vhost process was run on a single, physical CPU core using cgroups.
FIO configuration	<pre>[global] ioengine=libaio direct=1 rw=randrw rwmixread=100 (100% reads), 70 (70% reads, 30% writes), 0 (100% writes) thread=1 norandommap=1 time_based=1 runtime=240s ramp_time=60s bs=4k iodepth=1 / 8 / 32 / 64 numjobs=1</pre>

Test Case 3 results

SPDK Vhost-Scsi

Table 15: Performance per NVMe drive IOPS and latency results, SPDK SCSI stack

Access pattern	Backend	QD	Throughput (IOPS)	Avg. latency (usec)
4k 100% Random Reads	Split NVMe	1	24828.53	79.99
4k 100% Random Reads	Split NVMe	8	171426.45	92.57
4k 100% Random Reads	Split NVMe	32	444640.40	143.60
4k 100% Random Reads	Split NVMe	64	563590.26	224.74
4k 100% Random Reads	Lvol	1	24820.75	80.13
4k 100% Random Reads	Lvol	8	170962.74	93.11
4k 100% Random Reads	Lvol	32	444761.04	143.69
4k 100% Random Reads	Lvol	64	572688.77	222.67
4k 100% Random Writes	Split NVMe	1	116671.05	16.80
4k 100% Random Writes	Split NVMe	8	406291.18	40.10
4k 100% Random Writes	Split NVMe	32	451173.15	141.56
4k 100% Random Writes	Split NVMe	64	451821.80	283.06
4k 100% Random Writes	Lvol	1	114770.52	16.99
4k 100% Random Writes	Lvol	8	387797.84	42.02
4k 100% Random Writes	Lvol	32	464796.36	137.81
4k 100% Random Writes	Lvol	64	468299.69	273.68
4k 70%/30% Random Read Writes	Split NVMe	1	31530.95	63.09
4k 70%/30% Random Read Writes	Split NVMe	8	174266.16	92.65
4k 70%/30% Random Read Writes	Split NVMe	32	319680.76	195.98
4k 70%/30% Random Read Writes	Split NVMe	64	419539.10	305.65
4k 70%/30% Random Read Writes	Lvol	1	31464.73	63.09
4k 70%/30% Random Read Writes	Lvol	8	174627.95	89.60
4k 70%/30% Random Read Writes	Lvol	32	327826.04	194.54
4k 70%/30% Random Read Writes	Lvol	64	397437.78	322.34

SPDK Vhost-Blk

Table 16: Performance per NVMe drive IOPS and latency results, SPDK BLK stack

Access pattern	Backend	QD	Throughput (IOPS)	Avg. latency (usec)
4k 100% Random Reads	Lvol	1	25063.29	79.24
4k 100% Random Reads	Lvol	8	173263.80	91.76
4k 100% Random Reads	Lvol	32	450064.23	141.74
4k 100% Random Reads	Lvol	64	579842.64	219.83
4k 100% Random Writes	Lvol	1	123431.38	15.89
4k 100% Random Writes	Lvol	8	405988.43	40.25
4k 100% Random Writes	Lvol	32	464315.55	137.79
4k 100% Random Writes	Lvol	64	460442.91	278.46
4k 70%/30% Random Read Writes	Lvol	1	31768.27	62.34
4k 70%/30% Random Read Writes	Lvol	8	170953.17	93.04
4k 70%/30% Random Read Writes	Lvol	32	337441.22	190.57
4k 70%/30% Random Read Writes	Lvol	64	408736.60	316.25

Kernel Vhost-Scsi

Table 17: Performance per NVMe drive IOPS and latency results, Kernel Vhost-Scsi

Access pattern	Backend	QD	Throughput (IOPS)	Avg. latency (usec)
4k 100% Random Reads	NVMe	1	22754.32	87.08
4k 100% Random Reads	NVMe	8	154318.09	103.41
4k 100% Random Reads	NVMe	32	365991.97	174.09
4k 100% Random Reads	NVMe	64	442703.23	286.71
4k 100% Random Writes	NVMe	1	71954.73	27.18
4k 100% Random Writes	NVMe	8	258969.89	61.93
4k 100% Random Writes	NVMe	32	410521.96	156.21
4k 100% Random Writes	NVMe	64	461278.46	277.23
4k 70%/30% Random Read Writes	NVMe	1	26375.63	75.40
4k 70%/30% Random Read Writes	NVMe	8	161021.53	97.88
4k 70%/30% Random Read Writes	NVMe	32	291720.70	219.86
4k 70%/30% Random Read Writes	NVMe	64	359975.32	354.88

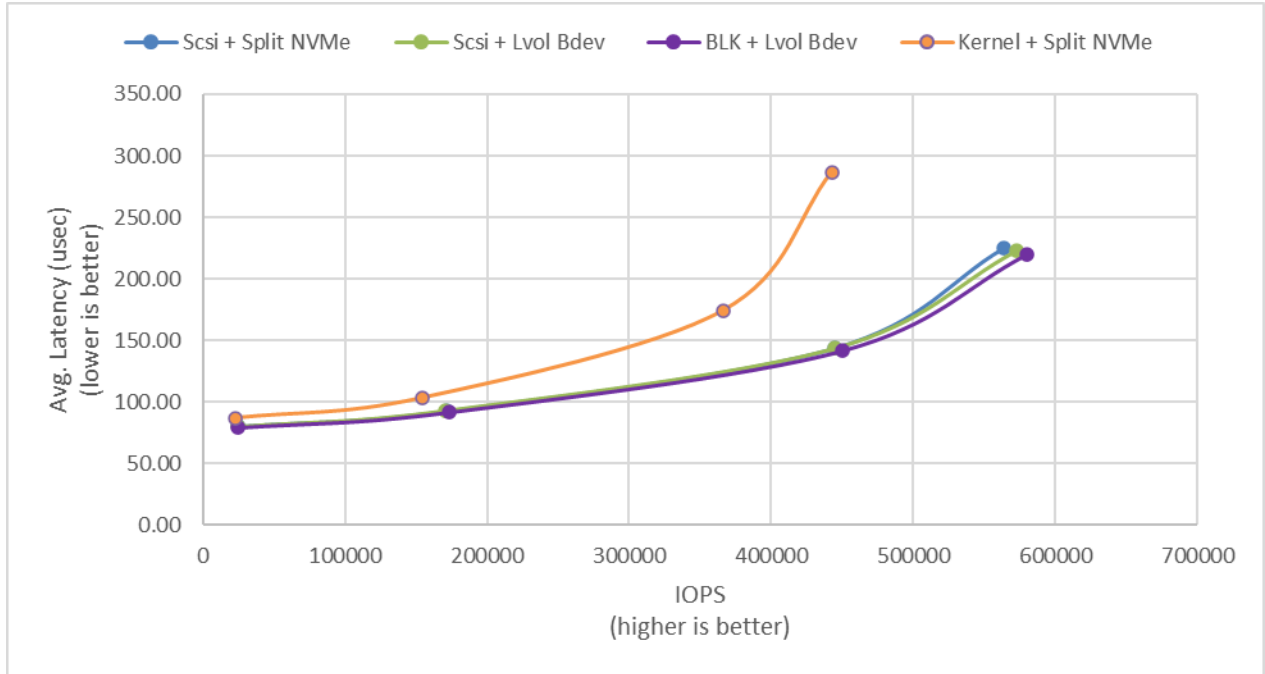


Figure 7: 4KB 100% Random Reads IOPS and latency

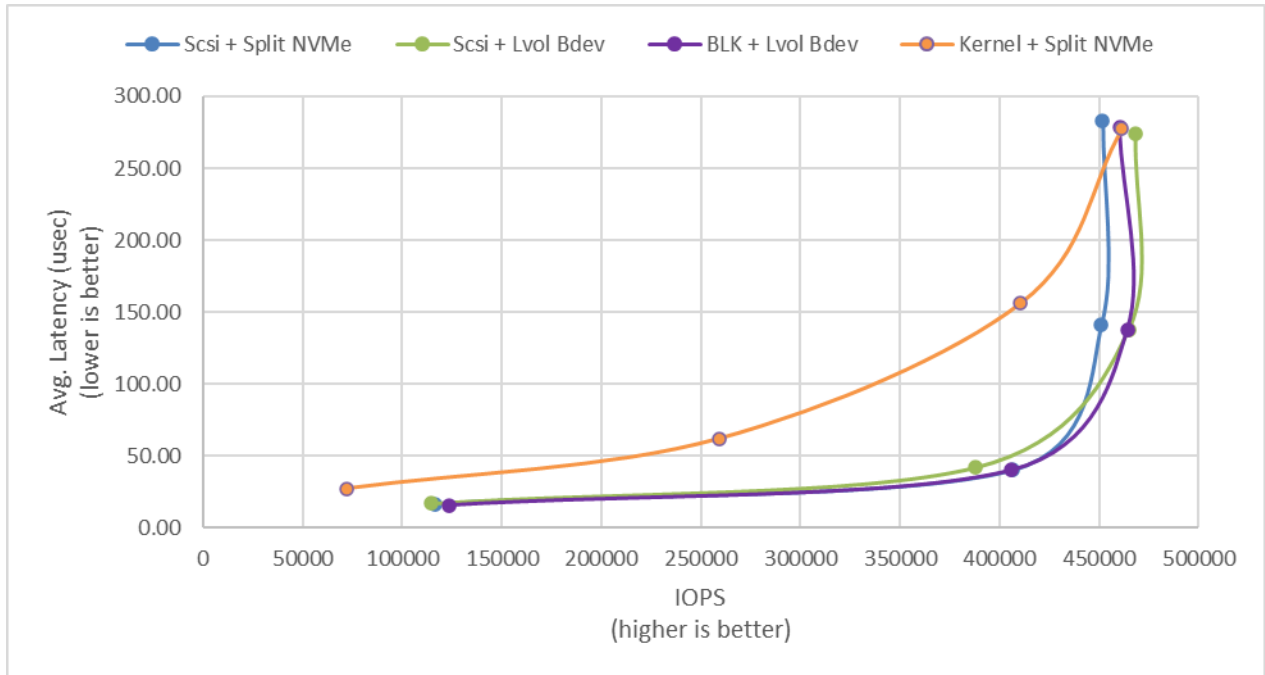


Figure 8: 4KB 100% Random Writes IOPS and latency

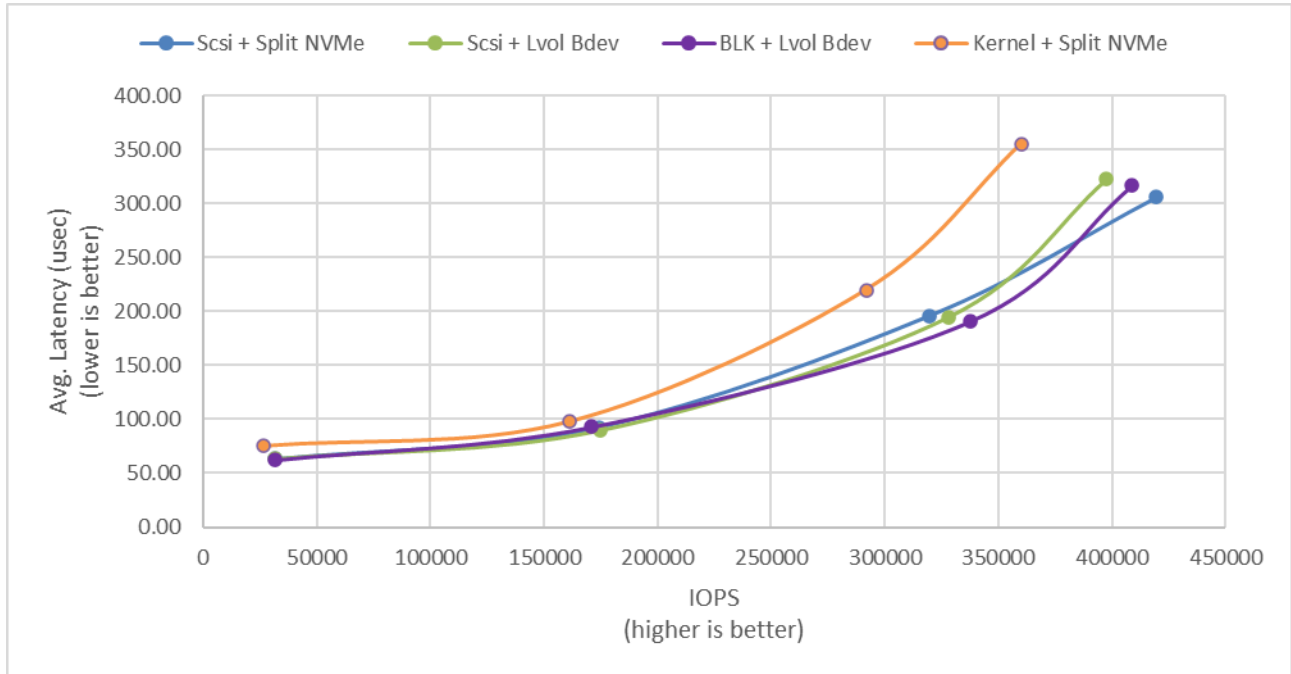


Figure 9: 4KB 70%/30% Random Read/Write IOPS and latency

Conclusions

1. SPDK Vhost-scsi with NVMe Split bdevs has lower latency and higher throughput than Kernel Vhost-scsi in majority of workload / queue depth combinations. Kernel Vhost-Scsi was able to match SPDK performance using Random Write QD=64 workload.

Summary

This report compared performance results while running Vhost-scsi using traditional interrupt-driven kernel Vhost-scsi against the accelerated polled-mode driven SPDK implementation. Various local ephemeral configurations were demonstrated, including rate limiting IOPS, performance per VM, and maximum performance from an underlying system when comparing kernel vs. SPDK Vhost-scsi target implementations.

In addition, performance impacts of using SPDK Logical Volume Bdevs and the SPDK Vhost-blk stack were presented.

This report provided information regarding methodologies and practices while benchmarking Vhost-scsi and Vhost-blk using both SPDK and 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 different hardware and software configurations.

List of Tables

Table 1: Hardware setup configuration	4
Table 2: Test platform BIOS settings	5
Table 3: Guest VM configuration	5
Table 4: SPDK Vhost Core Scaling test configuration	9
Table 5: SPDK Vhost core scaling results, 4KB 100% Random Reads IOPS, QD=64	11
Table 6: SPDK Vhost core scaling results, 4KB 100% Random Write IOPS, QD=32	12
Table 7: SPDK Vhost core scaling results, 4KB Random 70% Read 30% Write IOPS, QD=64	13
Table 8: Logical Volumes performance impact for SPDK Vhost SCSI	14
Table 9: LTO performance SPDK Vhost SCSI with Logical Volume bdevs	15
Table 10: Packed Ring performance impact on SPDK Vhost BLK controllers. Fio gtod_reduce=disabled	16
Table 11: Rate Limiting IOPS per VM test case configuration	18
Table 12: 4KB 100% Random Reads QD=1 rate limiting test results.....	20
Table 13: 4KB 100% Random Writes QD=1 rate limiting test results	21
Table 14: Performance per NVMe drive test case configuration	23
Table 15: Performance per NVMe drive IOPS and latency results, SPDK SCSI stack	24
Table 16: Performance per NVMe drive IOPS and latency results, SPDK BLK stack	25
Table 17: Performance per NVMe drive IOPS and latency results, Kernel Vhost-Scsi	25

List of Figures

Figure 1: SPDK Vhost-scsi architecture 7

Figure 2: Comparison of performance between various SPDK Vhost stack-bdev combinations for 4KB Random Read QD=64 workload11

Figure 3: Comparison of performance between various SPDK Vhost stack-bdev combinations for 4KB Random Write QD=32 workload12

Figure 4: Comparison of performance between various SPDK Vhost stack-bdev combinations for 4KB Random 70% Read 30% Write QD=64 workload13

Figure 5: 4KB 100% Random Reads IOPS, QD=1, throttling = 10k IOPS20

Figure 6: 4KB 100% Random Writes IOPS, QD=1, throttling = 10k IOPS21

Figure 7: 4KB 100% Random Reads IOPS and latency26

Figure 8: 4KB 100% Random Writes IOPS and latency26

Figure 9: 4KB 70%/30% Random Read/Write IOPS and latency27

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. See backup for configuration details. No product or component can be absolutely secure.

Your costs and results may vary.

No product or component can be absolutely secure.

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

© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.