Flash Memory Summit

Accelerating NVMe-oF* for VMs with the Storage Performance Development Kit

Jim Harris

Principal Software Engineer

Intel Data Center Group

Notices and Disclaimers

- Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at intel.com, or from the OEM or retailer.
- Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.
- No computer system can be absolutely secure.
- Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit http://www.intel.com/performance.
- Intel, the Intel logo, Xeon, and others are trademarks of Intel Corporation in the U.S. and/or other countries. *Other names and brands may be claimed as the property of others.
- *Other names and brands may be claimed as the property of others.
- © 2017 Intel Corporation.

Flash Memory Summit

NVMe over Fabrics* Software Overhead

- NVMe Specification enables highly optimized drivers
  - Multiple I/O queues allows lockless submission from CPU cores in parallel
- But even the best kernel mode drivers have non-trivial software overhead
  - 3-5us of software overhead per I/O
  - 500K+ IO/s per SSD, 4-24 SSDs per server, 100Gb+ RDMA
  - <10us latency with latest media (i.e. Intel Optane$^{\text{TM}}$ SSD)
- Virtualization adds additional overhead
  - NVMe-oF typically not configured in virtual machine
- Enter the Storage Performance Development Kit
  - Includes polled-mode and user-space drivers for NVMe and NVMe-oF

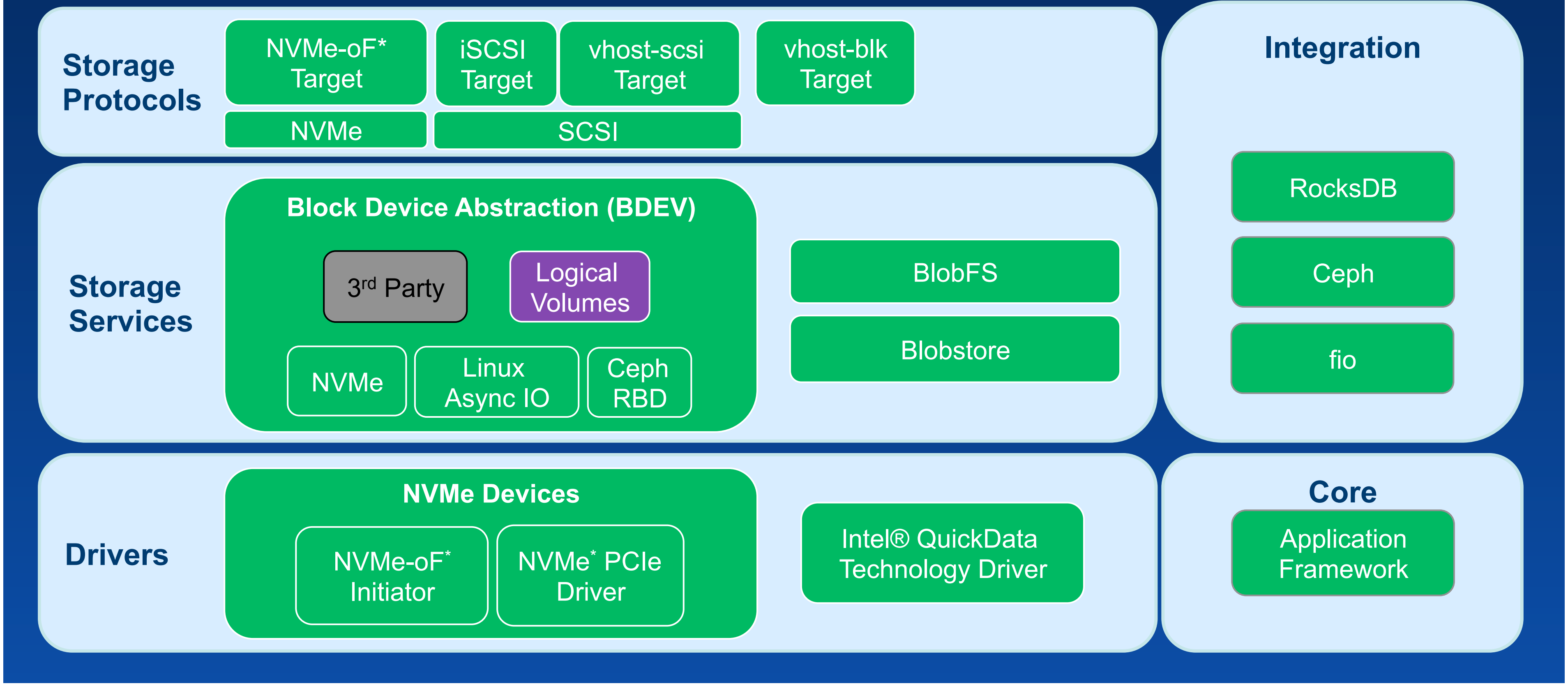
Storage Performance Development Kit (SPDK)

- Open Source Software Project
  - BSD licensed
  - Source code: http://github.com/spdk
  - Project website: http://spdk.io
- Set of software building blocks for scalable efficient storage applications
  - Polled-mode and user-space drivers and protocol libraries
- Designed for current and next generation NVM media latencies

Architecture

🟩 Released

🟪 Q4'17



Storage Protocols

NVMe-oF* Target

iSCSI Target

vhost-scsi Target

vhost-blk Target

NVMe

SCSI

Storage Services

Block Device Abstraction (BDEV)

3rd Party

Logical Volumes

NVMe

Linux Async IO

Ceph RBD

BlobFS

Blobstore

Integration

RocksDB

Ceph

fio

Drivers

NVMe Devices

NVMe-oF* Initiator

NVMe* PCIe Driver

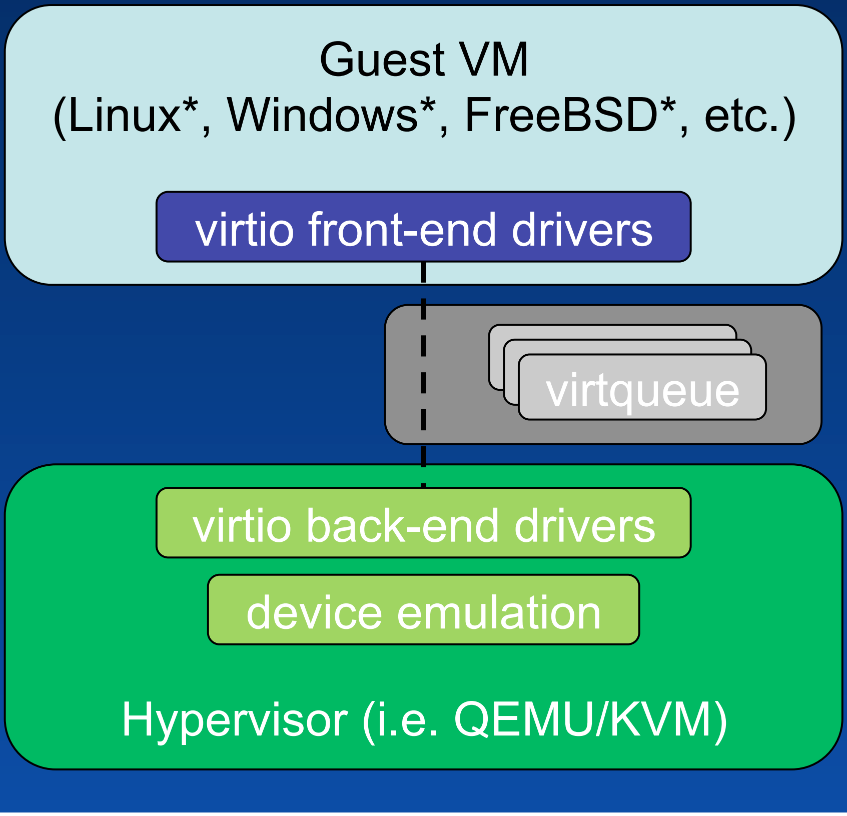
Intel® QuickData Technology Driver

Core

Application Framework

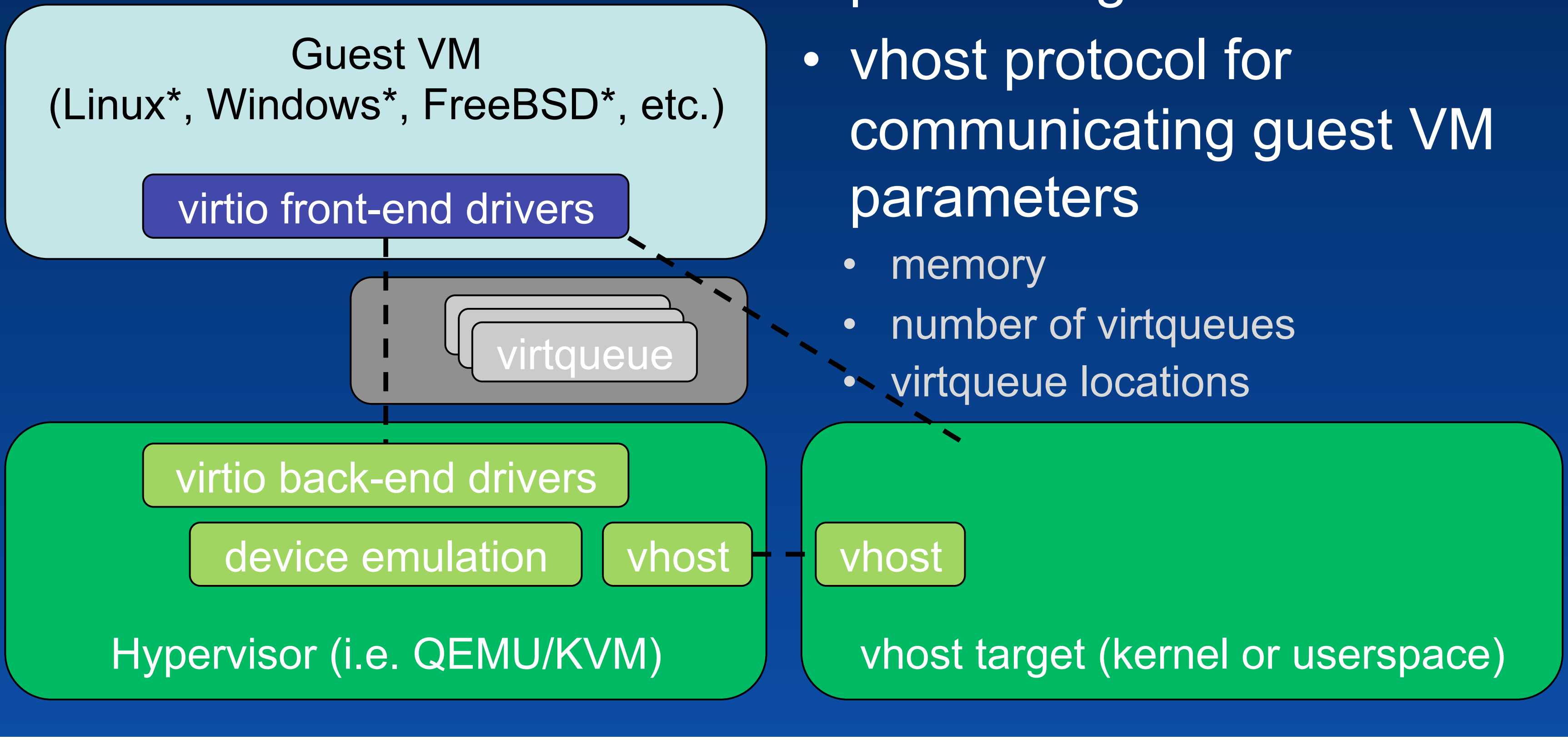
Flash Memory Summit

virtio



- Paravirtualized driver specification
- Common mechanisms and layouts for device discovery, I/O queues, etc.
- virtio device types include:
  - virtio-net
  - virtio-blk
  - virtio-scsi
  - virtio-gpu
  - virtio-rng
  - virtio-crypto

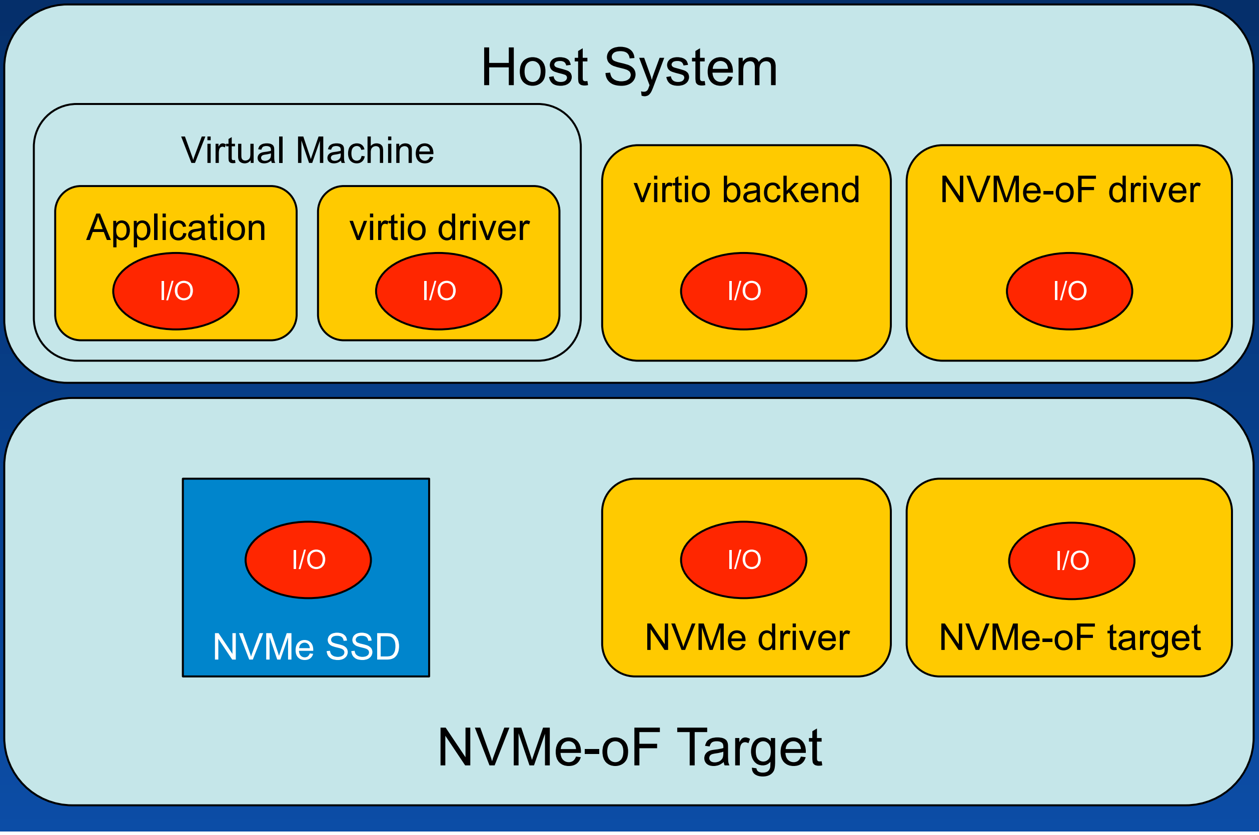
vhost


Guest VM
(Linux*, Windows*, FreeBSD*, etc.)
virtio front-end drivers
virtqueue
virtio back-end drivers
device emulation
vhost
vhost
Hypervisor (i.e. QEMU/KVM)
vhost target (kernel or userspace)

- Separate process for I/O processing
- vhost protocol for communicating guest VM parameters
  - memory
  - number of virtqueues
  - virtqueue locations

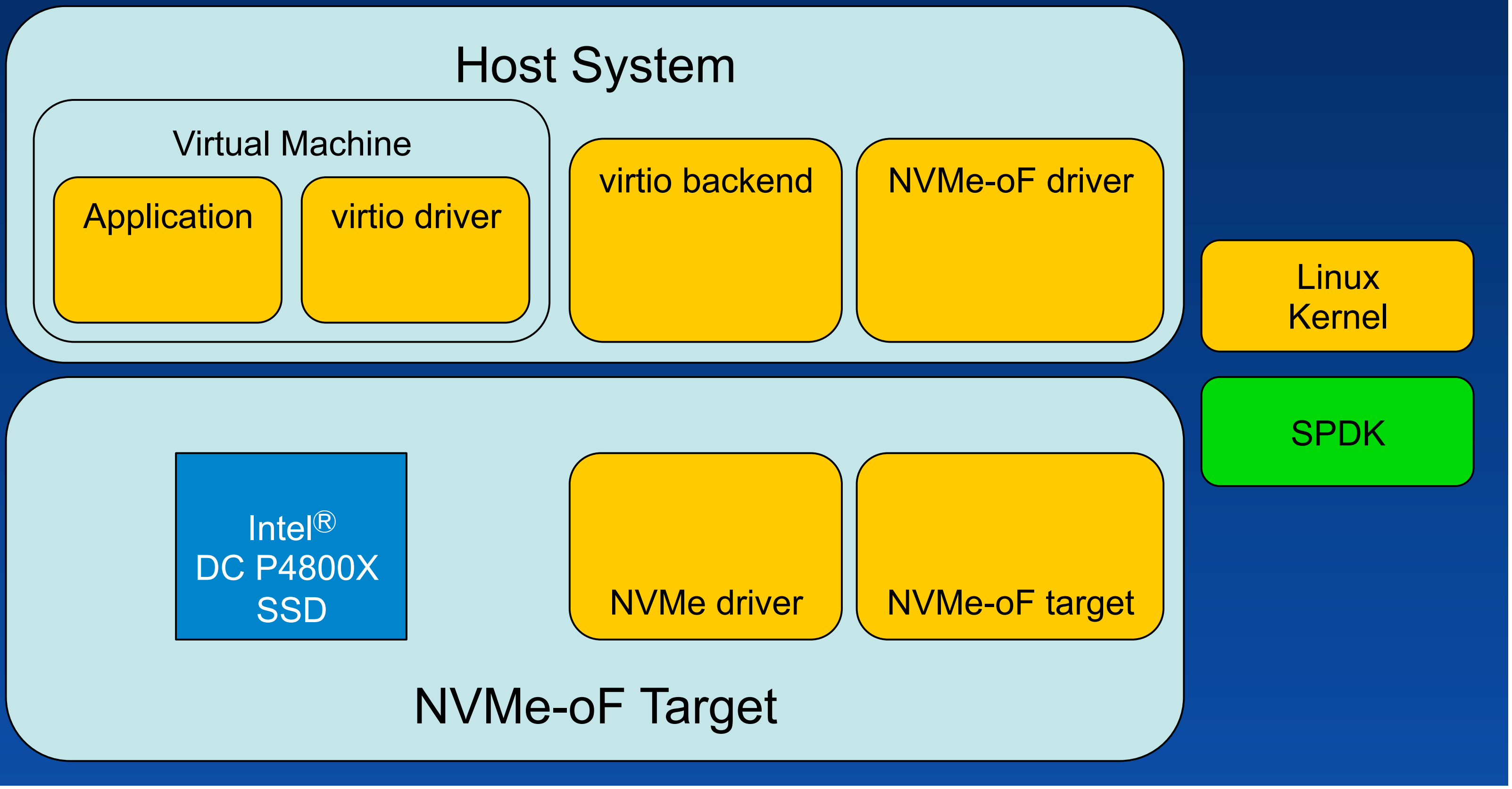
Flash Memory Summit

I/O Path



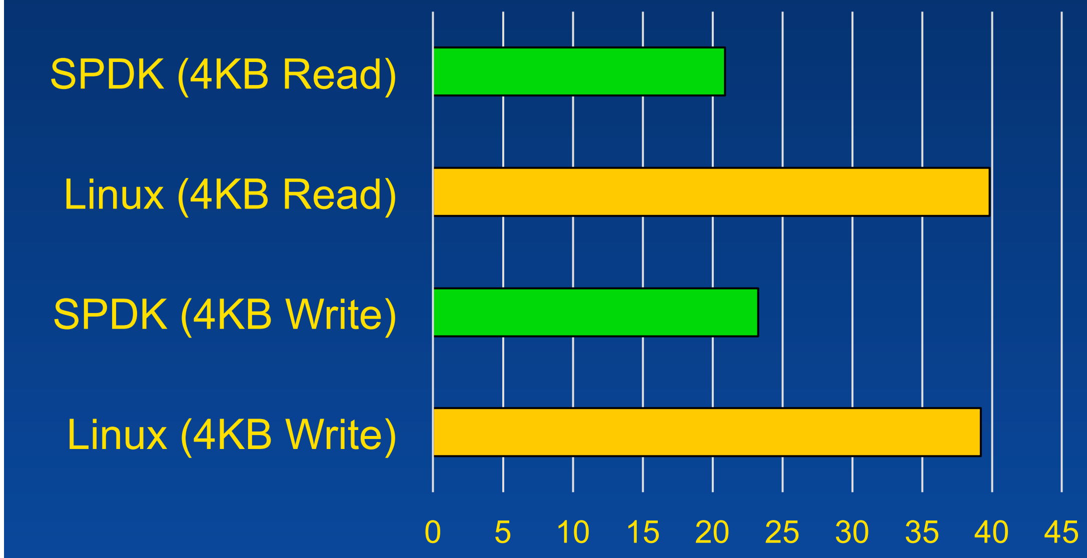
Flash Memory Summit

Performance Comparison



Performance Comparison

QD=1 Latency (microseconds)



- Configuration
  - 4KB Random I/O
  - Queue Depth 1
  - Intel® P4800X SSD
  - Mellanox® LX-710 25Gb
  - Measured from VM (fio)

SPDK cuts latency almost in half!

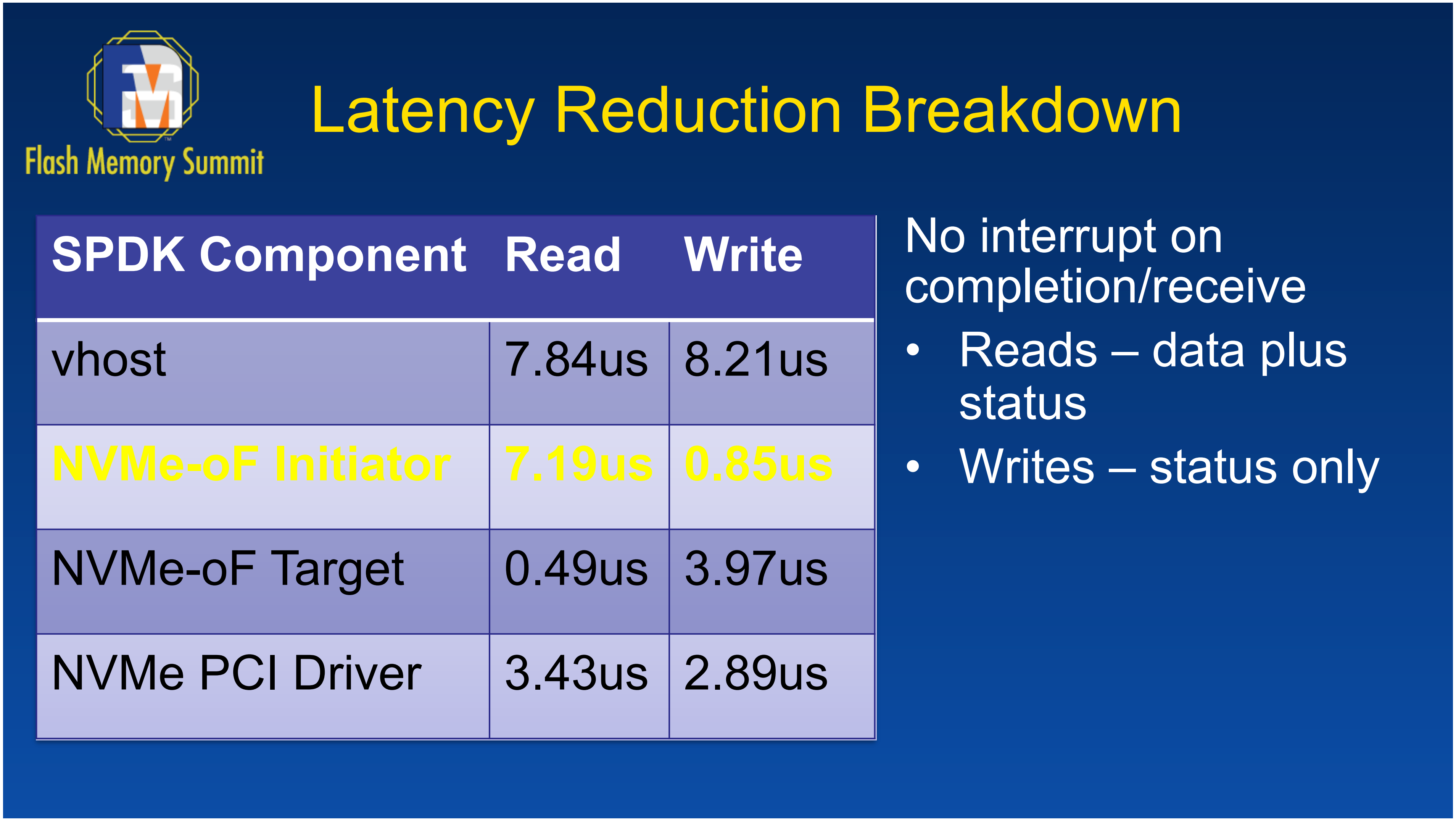
System Configuration: NVMe-oF Target: OS: Ubuntu 17.04 Linux Kernel: 4.10.0-19-generic x86_64, 187 GB Intel P4800X SSD, 2x E5-2695 v4 @ 2.10GHz, Hyperthreading disabled, 16x4GB DRAM, SPDK commit ID 7fedfb48, Mellanox LX-710 25Gb/s (SFP+) **VM Host System:** OS: Ubuntu 17.04 Linux Kernel: 4.10.0-19-generic x86_64, 2x E5-2699 v3 @2.30GHz Hyperthreading enabled, 16x8GB DRAM, SPDK commit 1e2aed0a, Mellanox LX-710 25Gb/s (SFP+) Link Speed **QEMU VM:** OS: CentOS Linux release 7.3.1611 Linux Kernel: 3.10.0-514.21.1.el7.x86_64

Latency Reduction Breakdown

SPDK Component	Read	Write
vhost	**7.84us**	**8.21us**
NVMe-oF Initiator	7.19us	0.85us
NVMe-oF Target	0.49us	3.97us
NVMe PCI Driver	3.43us	2.89us

No VMEXIT on submission

No context switch to wake SPDK thread



Latency Reduction Breakdown

SPDK Component	Read	Write
vhost	7.84us	8.21us
NVMe-oF Initiator	7.19us	0.85us
NVMe-oF Target	0.49us	3.97us
NVMe PCI Driver	3.43us	2.89us

No interrupt on completion/receive

- Reads – data plus status
- Writes – status only

Flash Memory Summit

Latency Reduction Breakdown

SPDK Component	Read	Write
vhost	7.84us	8.21us
NVMe-oF Initiator	7.19us	0.85us
NVMe-oF Target	**0.49us**	**3.97us**
NVMe PCI Driver	3.43us	2.89us

No interrupt on submission/receive

- Reads – command only
- Writes – command plus data

Flash Memory Summit

Latency Reduction Breakdown

SPDK Component	Read	Write
vhost	7.84us	8.21us
NVMe-oF Initiator	7.19us	0.85us
NVMe-oF Target	0.49us	3.97us
NVMe PCI Driver	**3.43us**	**2.89us**

No interrupt on I/O completion

Pinned hugepages

Flash Memory Summit

Software Overhead

- Not just a latency improvement
- Reducing software overhead means:
  - Fewer I/O processing cores => More cores for VMs
  - Fewer VMEXITs in VMs => More cycles for application

SPDK vhost Hyper-Converged Demo

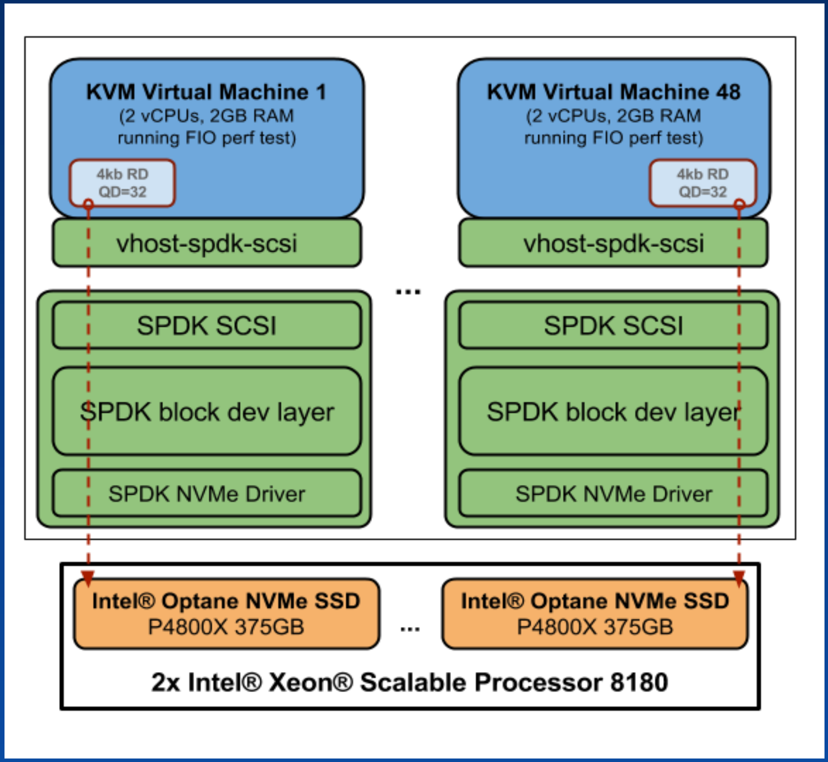
Use case

Software accelerated Virtual Machine Storage

Configuration

Hyper-converged Server Node

- Intel® Xeon® Scalable Processor node running 48 virtual machines with 24x direct-attached Intel® NVMe SSDs



KVM Virtual Machine 1
(2 vCPUs, 2GB RAM
running FIO perf test)

4kb RD
QD=32

vhost-spdk-scsi

SPDK SCSI

SPDK block dev layer

SPDK NVMe Driver

...

KVM Virtual Machine 48
(2 vCPUs, 2GB RAM
running FIO perf test)

4kb RD
QD=32

vhost-spdk-scsi

SPDK SCSI

SPDK block dev layer

SPDK NVMe Driver

Intel® Optane NVMe SSD
P4800X 375GB

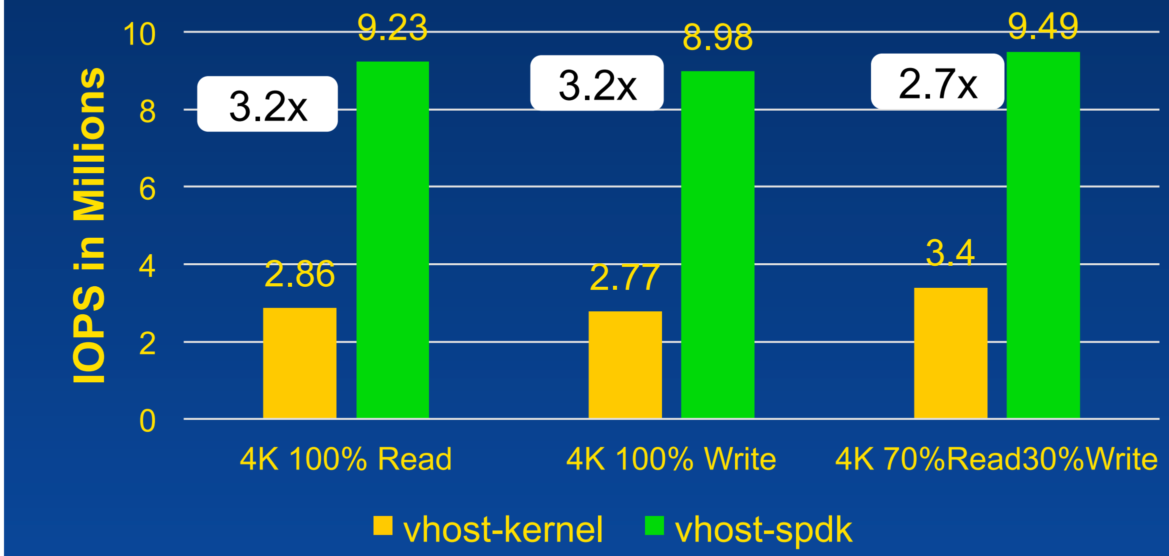
...

Intel® Optane NVMe SSD
P4800X 375GB

2x Intel® Xeon® Scalable Processor 8180

Flash Memory Summit

vhost-scsi performance – 48 VMs (SPDK vs. Kernel)



- 2x Intel Xeon Platinum 8180 Processor
- 24x Intel P4800x 375GB
- 10 vhost I/O processing cores

SPDK vhost yields up to 3.2x more IOPs

System Configuration:Intel Xeon Platinum 8180 @ 2.5GHz. 56 physical cores 6x 16GB, 2667 DDR4, 6 memory Channels, SSD: Intel P4800x 375GB x24 drives, Bios: HT disabled, p-states enabled, turbo enabled, Ubuntu 16.04.1 LTS, 4.11.0 x86_64 kernel, 48 VMs, number of partition: 2, VM config : 1core 1GB memory, VM OS: fedora 25, blk-mq enabled, Software packages: Qemu-2.9, libvirt-3.0.0, spdk (3bfecec994), IO distribution: 10 vhost-cores for SPDK / Kernel. Rest 46 cores for QEMU using cgroups, FIO-2.1.10 with SPDK plugin, io depth=1, 8, 32 numjobs=1, direct=1, block size 4k

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit http://www.intel.com/performance.

Flash Memory Summit

Summary

- Significant software overhead in virtualization usage models with NVMe-oF
- Software overhead impacts performance and CPU efficiency
- SPDK can reduce this software overhead by up to 20us per I/O
- Check out SPDK at http://spdk.io