14th ANNUAL WORKSHOP 2018

THE STORAGE PERFORMANCE DEVELOPMENT KIT AND NVME-OF

Paul Luse

Intel Corporation

Apr 2018

AGENDA

Storage Performance Development Kit

- What is SPDK?
- The SPDK Community
- Why are so many storage companies using it?
- How is it being used?

WHAT IS SPDK?

WHAT IS SPDK?

Storage Performance Development Kit

Scalable and Efficient Software Ingredients

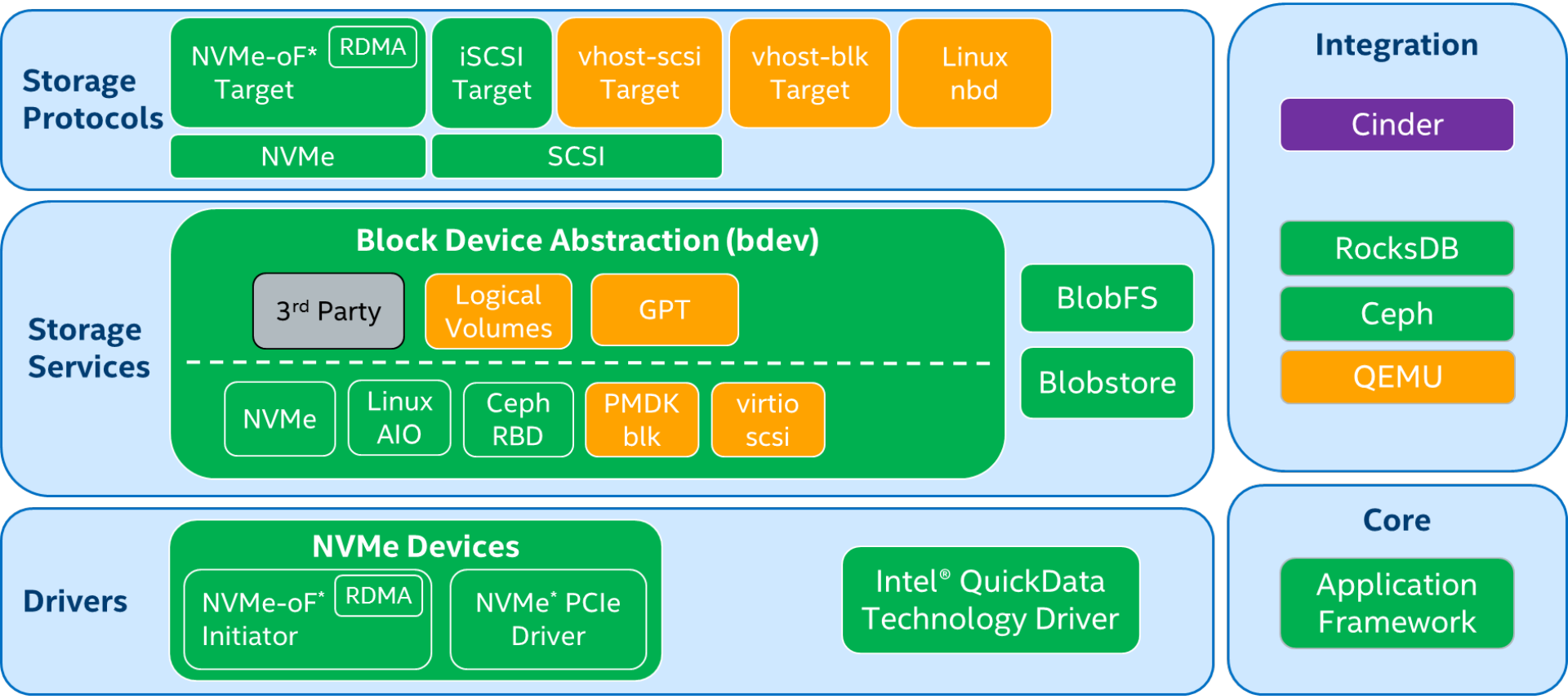
- User space, lockless, polled-mode components
- Up to millions of IOPS per core
- Designed to extract maximum performance from non-volatile media

Storage Reference Architecture

- Optimized for latest generation CPUs and SSDs
- Open source composable building blocks (BSD licensed)

http://SPDK.IO

SPDK ARCHITECTURE



THE SPDK COMMUNITY

THE SPDK COMMUNITY

Full Transparency

Email Discussions

http://SPDK.IO

Main Web Presence

Real Time Chat w/
Development Community

Weekly Calls
Multiple Annual Meetups

Backlog and
Ideas for Things to Do

GerritHub™

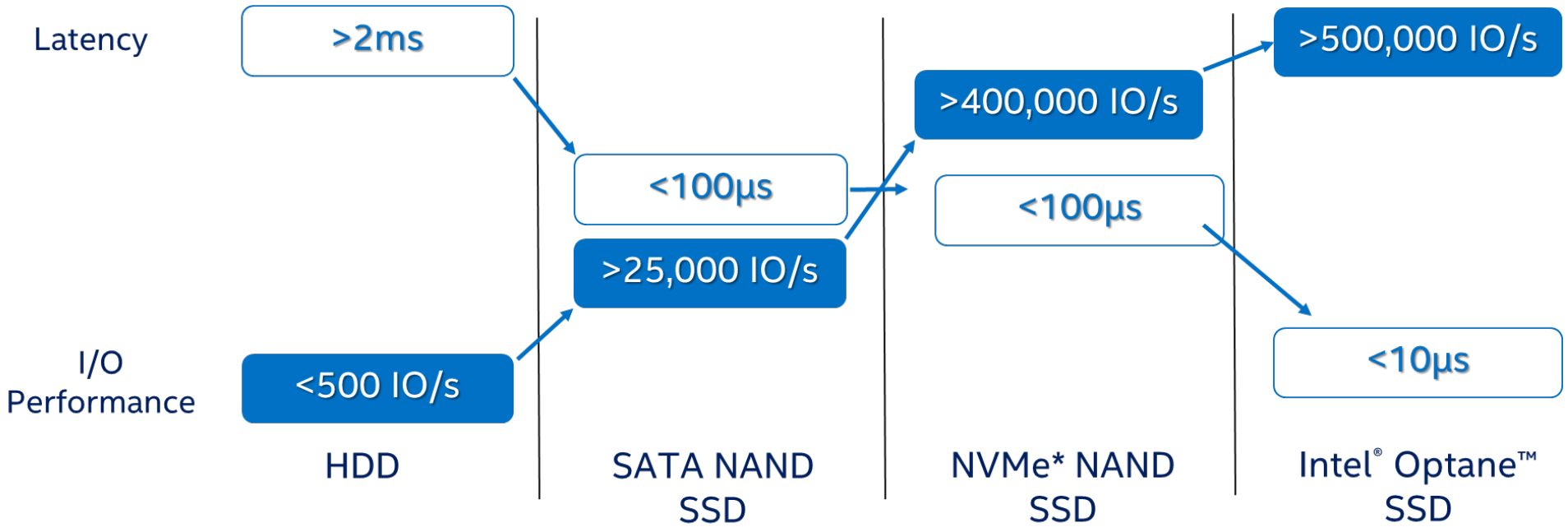
GitHub

Code Reviews & Repo

Continuous
Integration

WHY ARE SO MANY STORAGE COMPANIES USING SPDK?

SOFTWARE IS BECOMING THE BOTTLENECK



SPDK Unlocks New Media Potential

SPDK BENEFITS

Storage Performance Development Kit

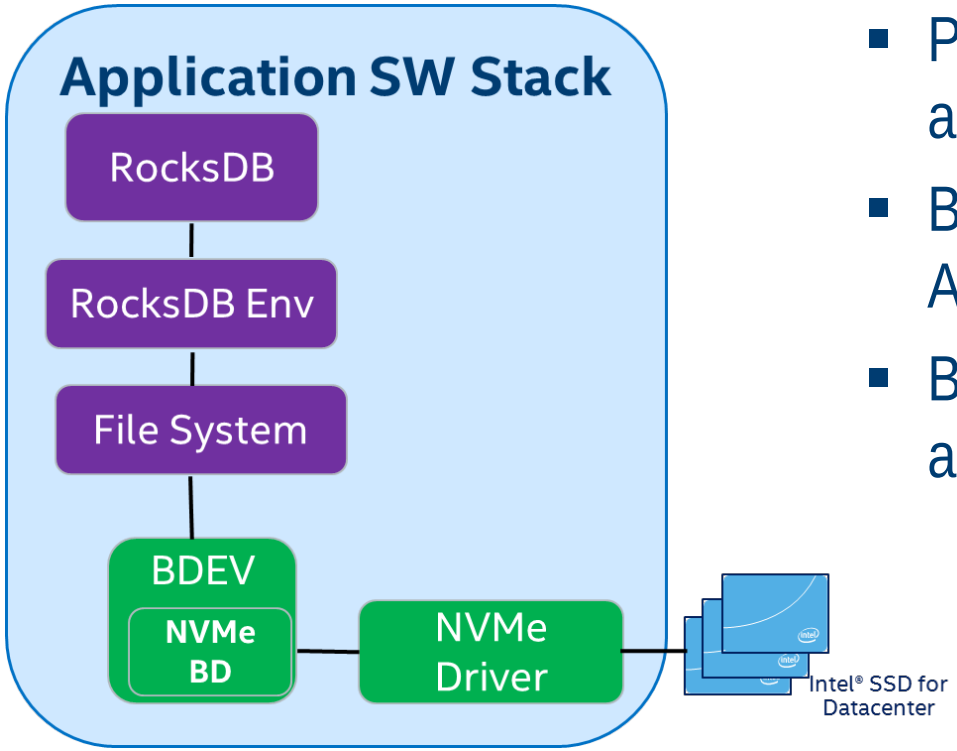
- Up to **10X MORE** IOPS/core for NVMe-oF* vs kernel
- Up to **8X MORE** IOPS/core for NVMe vs kernel
- Up to **50% BETTER** tail latency w/RocksDB workloads
- **FASTER TTM** w/less **RESOURCES** vs from scratch
- Provides **FUTURE PROOFING** as technologies evolve

INDUSTRY INVOLVEMENT

And many more!

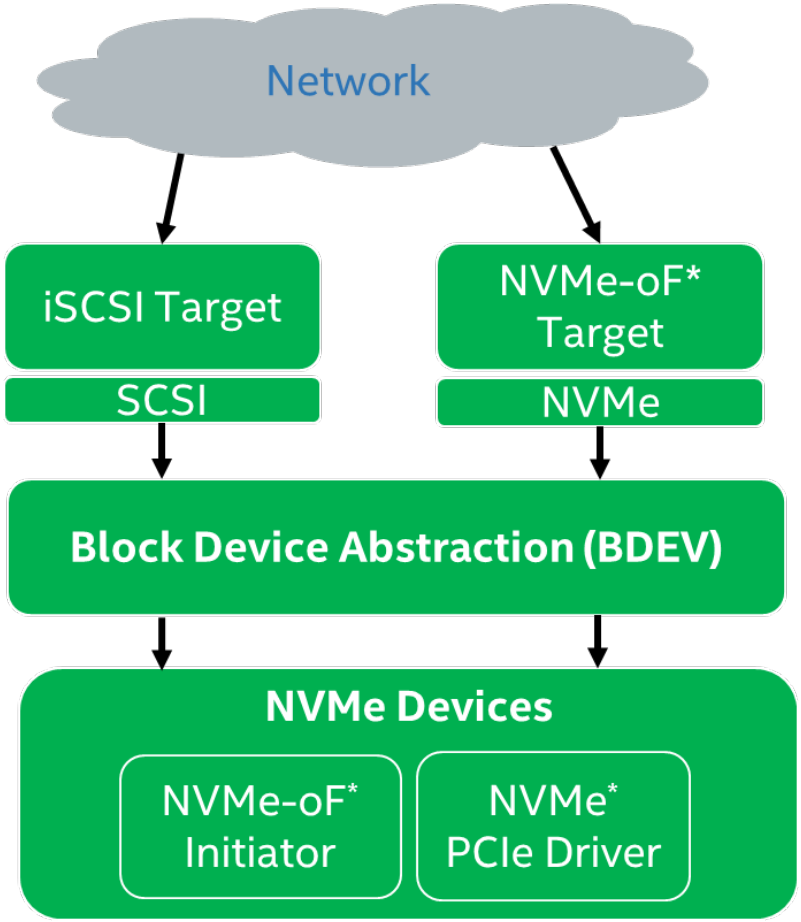
HOW IS SPDK BEING USED?

APPLICATION ACCELERATION
(LOCAL STORAGE)



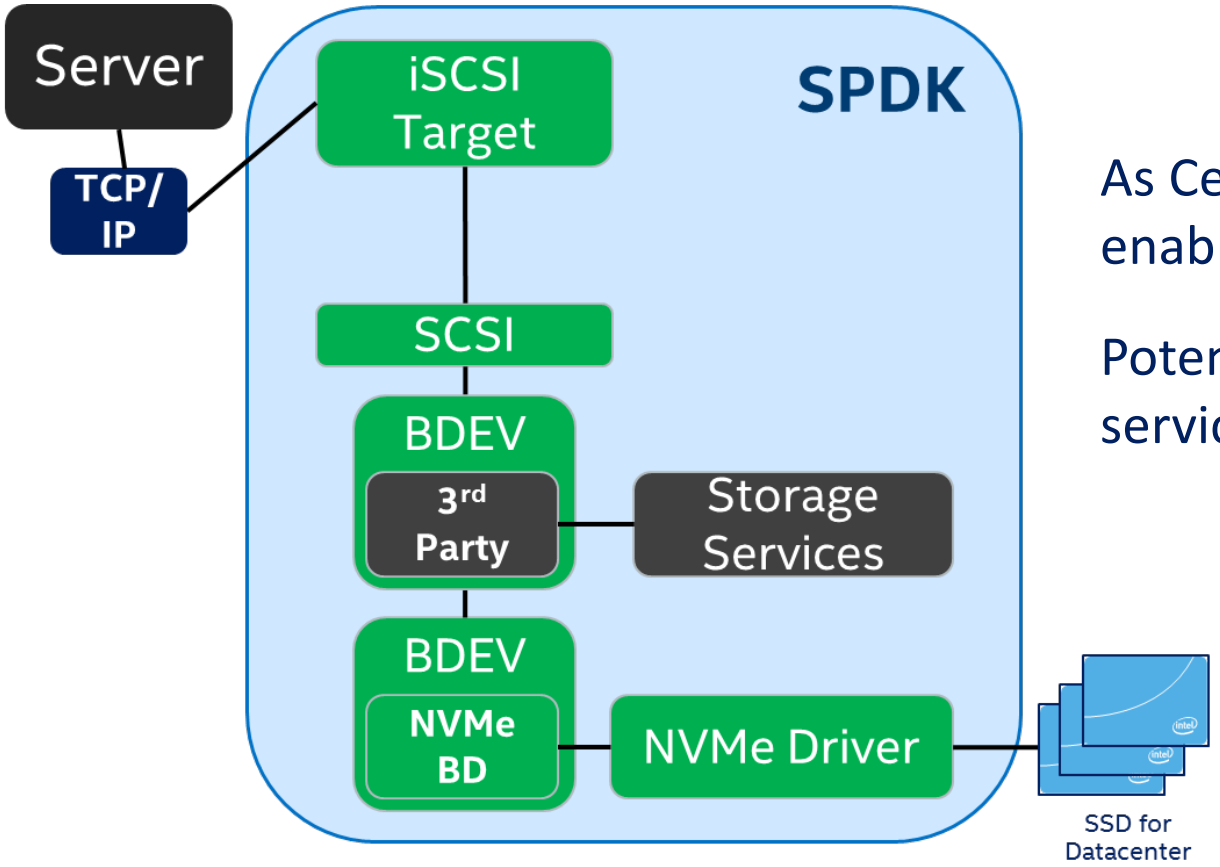
- Provides direct access from application to media
- BDEV abstraction provides consistent API to various types of block devices
- Benefits: dramatically reduces latency and improves IO consistency

REMOTE ACCESS TO STORAGE



- NVMe-oF supports different fabrics:
  - RDMA (iWARP, RoCE)
  - InfiniBand™
  - Fibre Channel
  - Intel® Omni-Path Architecture
  - TCP (coming soon)
- Unified interface for the NVMe PCIe driver and the NVMe-oF initiator
- Libraries & applications

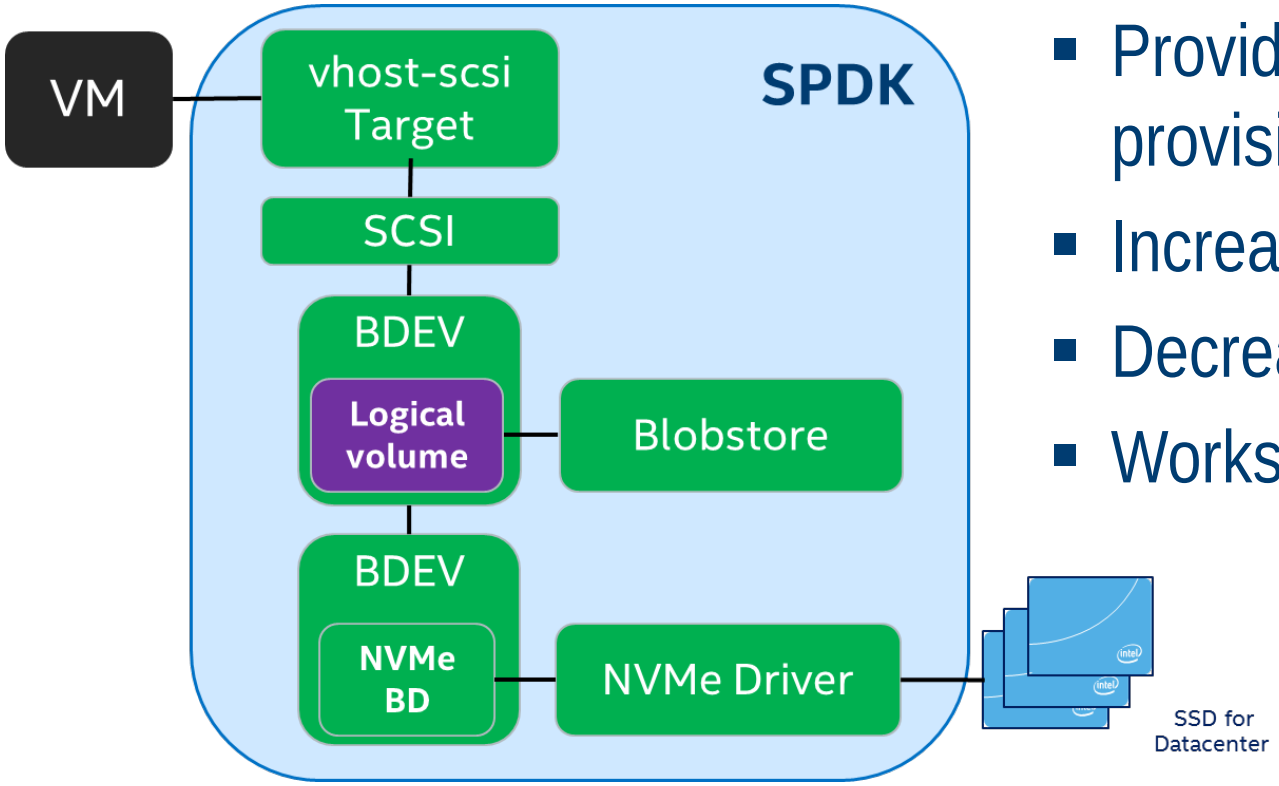
ISCSI TARGET ACCESS



As Ceph performance matures, enable higher efficiency access

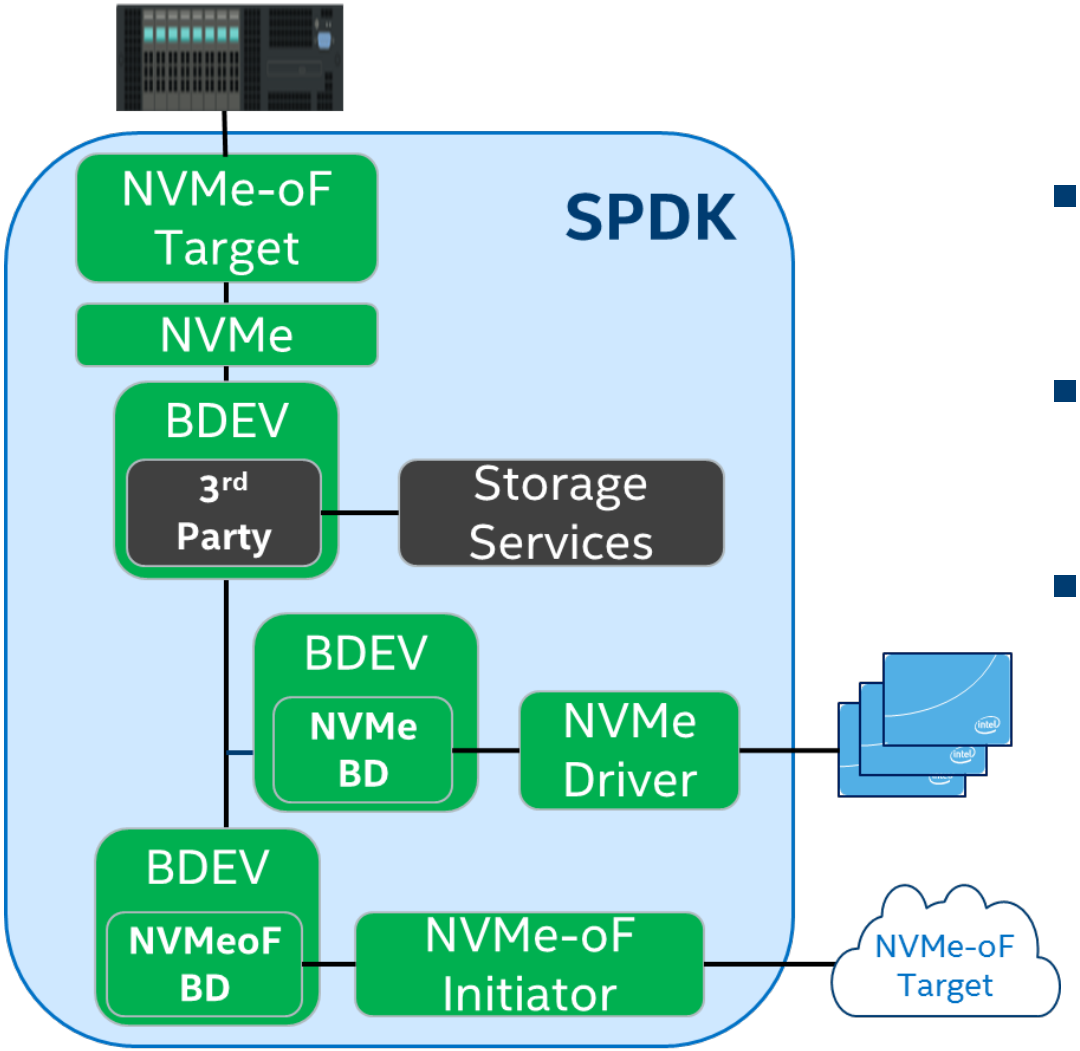
Potential for innovation in data services (e.g. cache, dedup...)

VIRTUAL MACHINE ACCELERATION



- Provides dynamic block provisioning
- Increases VM density
- Decreases guest latency
- Works with KVM/QEMU

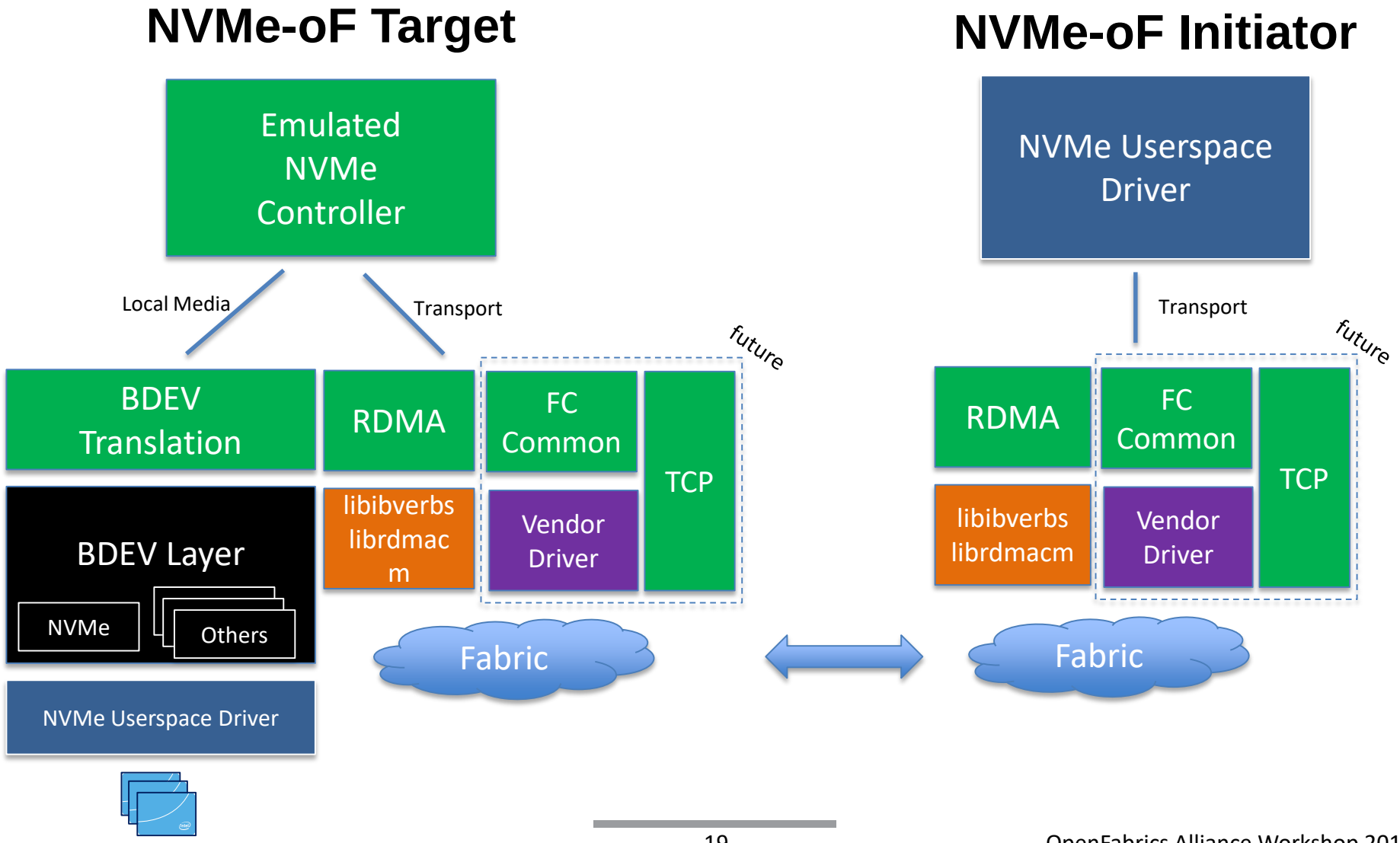
3RD PARTY BLOCK SERVICES



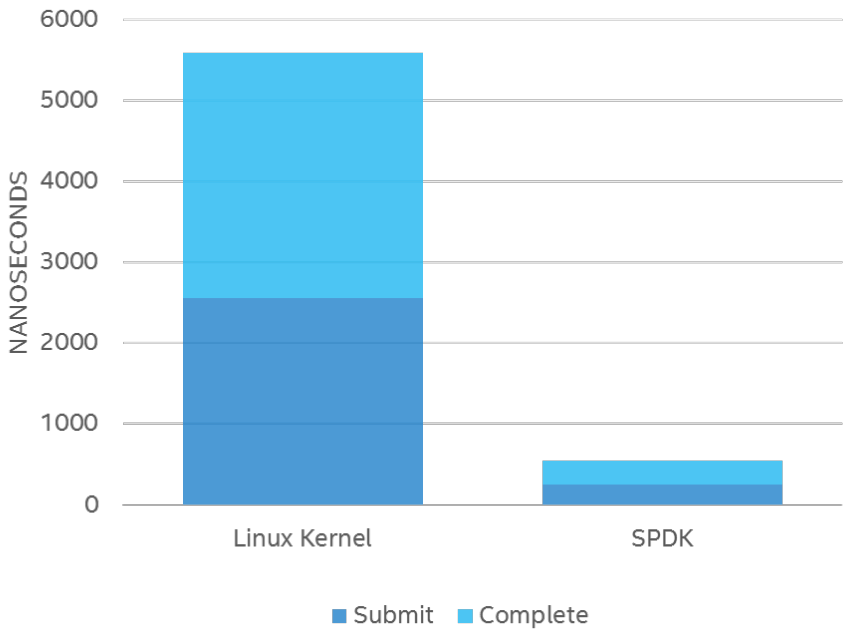
- Blobstore enables SSD virtualization
- BDEV enables stackable SW
- BDEV enables innovation

NVME-OF

SPDK NVME-OF


NVMe-oF Target
Emulated NVMe Controller
Local Media
Transport
BDEV Translation
RDMA
FC Common
TCP
future
BDEV Layer
NVMe
Others
libibverbs librdmacm
Vendor Driver
NVMe Userspace Driver
Fabric
NVMe-oF Initiator
NVMe Userspace Driver
Transport
future
RDMA
FC Common
TCP
libibverbs librdmacm
Vendor Driver
Fabric

A WORD ON THE SPDK NVME DRIVER

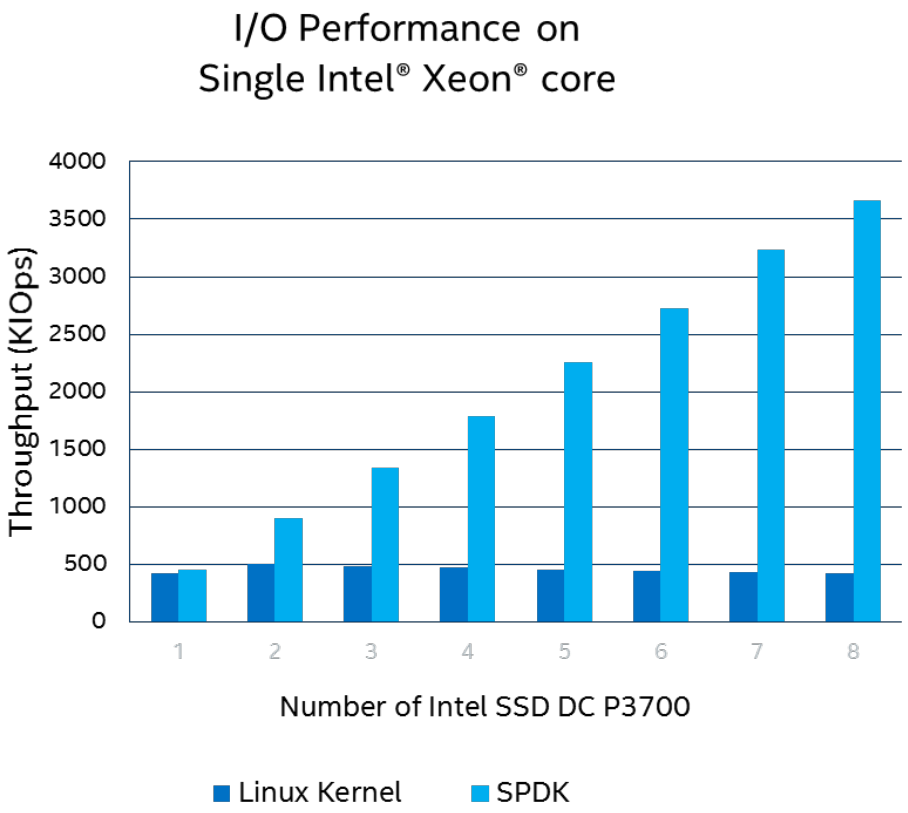


Kernel Source of Overhead	SPDK Approach
Interrupts	Asynchronous Polled Mode
Synchronization	Lockless
System Calls	User Space Hardware Access
DMA Mapping	Hugepages
Generic Block Layer	Specific for Flash Latencies

System Configuration: 2x Intel® Xeon® E5-2695v4 (HT off), Intel® Speed Step enabled, Intel® Turbo Boost Technology disabled, 8x 8GB DDR4 2133 MT/s, 1 DIMM per channel, CentOS* Linux* 7.2, Linux kernel 4.7.0-rc1, 1x Intel® P3700 NVMe SSD (800GB), 4x per CPU socket, FW 8DV10102, I/O workload 4KB random read, Queue Depth: 1 per SSD, Performance measured by Intel using SPDK overhead tool, Linux kernel data using Linux AIO

SPDK reduces NVMe software overhead up to 10x!

A WORD ON THE SPDK NVME DRIVER

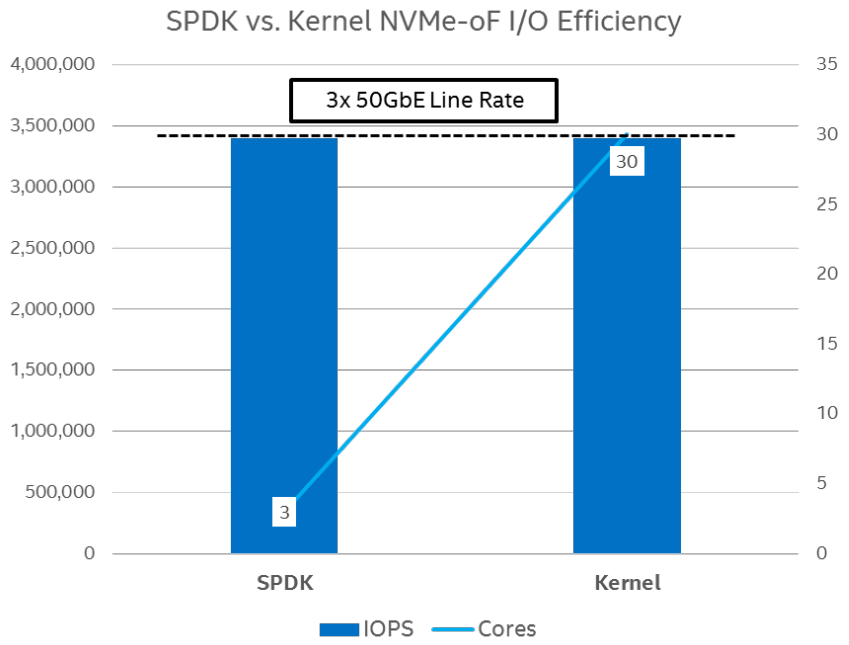


- Systems with multiple NVMe SSDs capable of millions of IOPS
- SPDK enables:
  - more CPU cycles for storage services
  - lower I/O latency

SPDK saturates 8 NVMe SSDs with a single CPU core!

System Configuration: 2x Intel® Xeon® E5-2695v4 (HT off), Intel® Speed Step enabled, Intel® Turbo Boost Technology disabled, 8x 8GB DDR4 2133 MT/s, 1 DIMM per channel, CentOS* Linux* 7.2, Linux kernel 4.7.0-rc1, 1x Intel® P3700 NVMe SSD (800GB), 4x per CPU socket, FW 8DV10102, I/O workload 4KB random read, Queue Depth: 1 per SSD, Performance measured by Intel using SPDK overhead tool, Linux kernel data using Linux AIO

NVME-OF TARGET PERFORMANCE

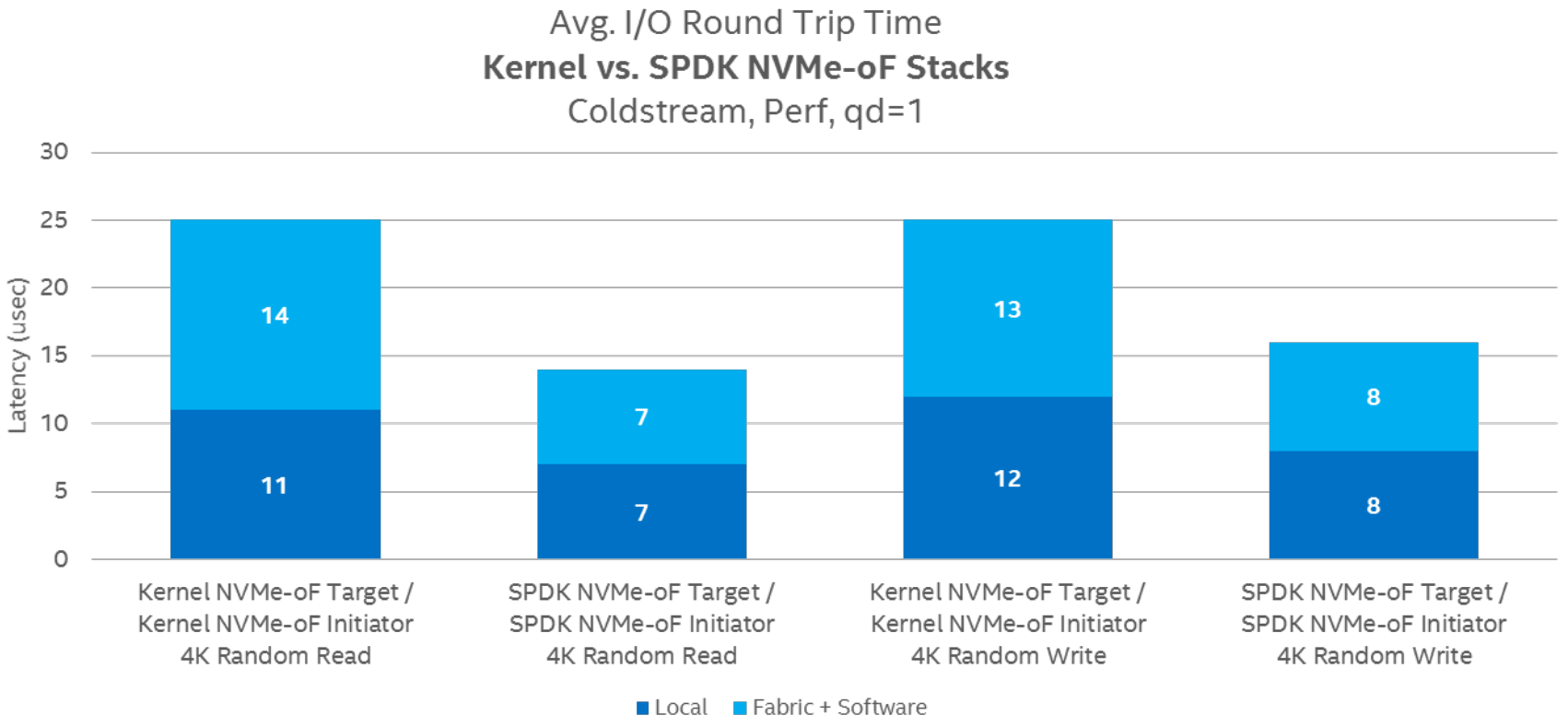


NVMe* over Fabrics Target Features	Realized Benefit
Utilizes NVM Express* (NVMe) Polled Mode Driver	Reduced overhead per NVMe I/O
RDMA Queue Pair Polling	No interrupt overhead
Connections pinned to CPU cores	No synchronization overhead

SPDK reduces NVMe over Fabrics software overhead up to 10x!

System Configuration: 2x Intel® Xeon® E5-2695v4 (HT off), Intel® Speed Step enabled, Intel® Turbo Boost Technology disabled, 8x 8GB DDR4 2133 MT/s, 1 DIMM per channel, CentOS* Linux* 7.2, Linux kernel 4.7.0-rc1, 1x Intel® P3700 NVMe SSD (800GB), 4x per CPU socket, FW 8DV10102, I/O workload 4KB random read, Queue Depth: 1 per SSD, Performance measured by Intel using SPDK overhead tool, Linux kernel data using Linux AIO

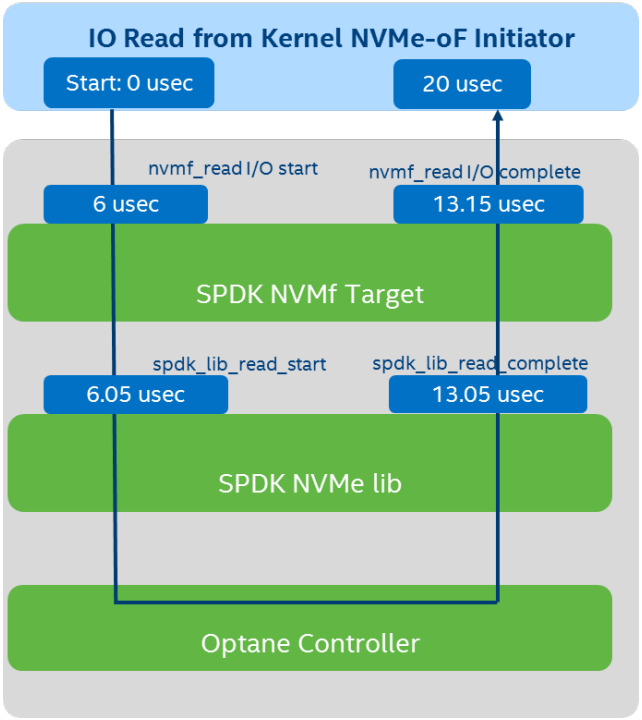
SPDK HOST+TARGET VS KERNEL HOST+TARGET

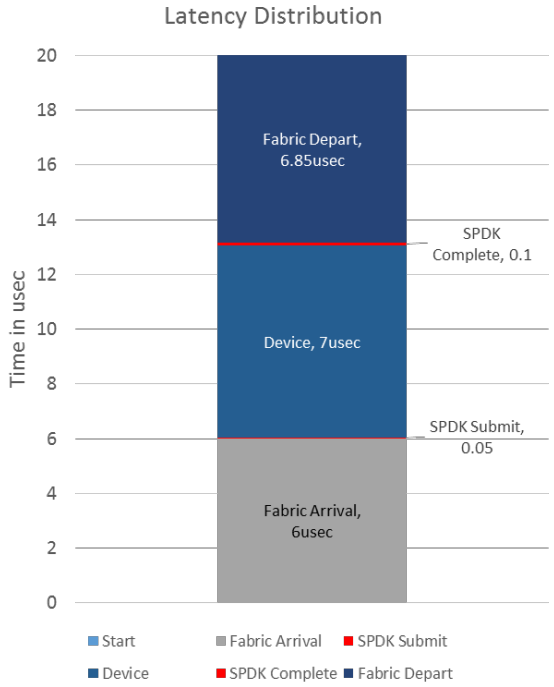


SPDK reduces Optane NVMe-oF latency by 44%, write latency by 32%!

System Configuration: 2x Intel® Xeon® E5-2695v4 (HT off), Intel® Speed Step enabled, Intel® Turbo Boost Technology disabled, 8x 8GB DDR4 2133 MT/s, 1 DIMM per channel, CentOS* Linux* 7.2, Linux kernel 4.7.0-rc1, 1x Intel® P3700 NVMe SSD (800GB), 4x per CPU socket, FW 8DV10102, I/O workload 4KB random read, Queue Depth: 1 per SSD, Performance measured by Intel using SPDK overhead tool, Linux kernel data using Linux AIO

NVME-OF LATENCY 4KB RANDOM READ, INTEL OPTANE SSD, SPDK TGT & KERNEL INITIATOR





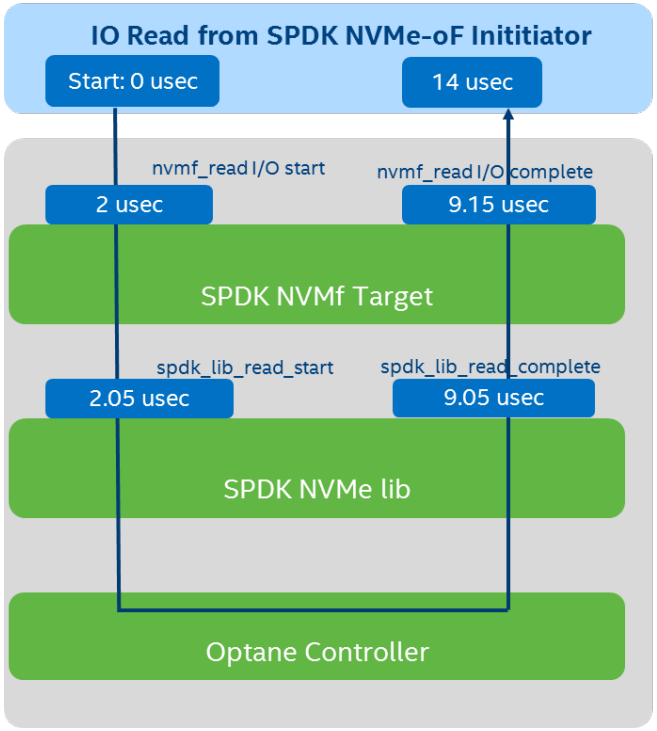
- 20 us round trip time measured from NVMe-oF initiator
- Out of 20usec, ~7 us spent in NVMe controller
- 12-13 usmeasured time in the fabric and kernel NVMe-oF initiator
- SPDK NVMf target adds just 100-200 nsto fabric overhead

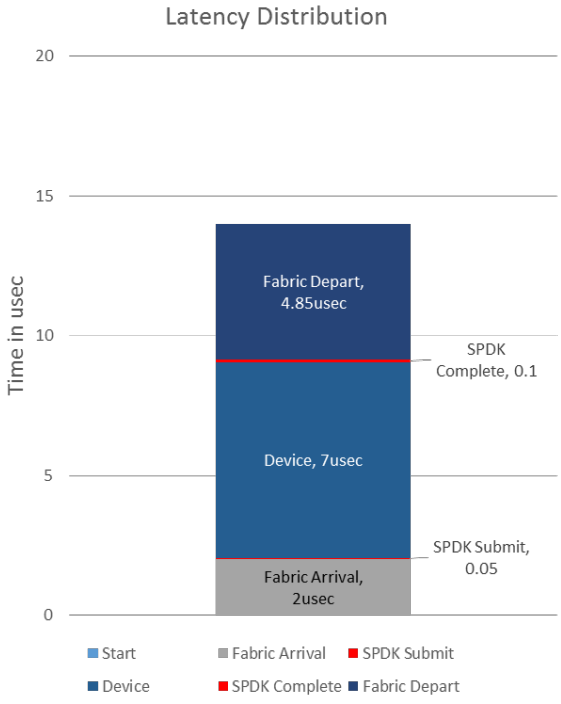
Disclaimer: Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those 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.

System Configuration: 2x Intel® Xeon® E5-2695v4 (HT on, Intel® Speed Step enabled, Intel® Turbo Boost Technology enabled, 64GB DDR4 Memory, 8x 8GB DDR4 2400 MT/s, Ubuntu 16.04.1, Linux kernel 4.10.1, 1x 25GbE Mellanox 2P CX-4, CX-4 FW= 14.16.1020, mlx5_core= 3.0-1 driver, 1 ColdStream, connected to socket 0, 4KB Random Read I/0 1 initiators, each initiator connected to bx NVMe-oF subsystems using 2P 25GbE Mellanox. Performance measured by Intel using SPDK perf tool, 4KB Random Read I/O, Queue Depth: 1/NVMe-oF subsystem.

numjobs 1, 300 sec runtime, direct=1, norandommap=1

NVME-OF LATENCY 4KB RANDOM READ, INTEL OPTANE SSD, SPDK TGT & SPDK INITIATOR





- 14 usround trip time measured from NVMf client
- Out of 14usec, ~7 usspent in NVMe controller
- 7 usmeasured time in the fabric and SPDK NVMe-oF initiator
- SPDK NVMf target adds just 100-200 nsec to fabric overhead

System Configuration: 2x Intel® Xeon® E5-2695v4 (HT on, Intel® Speed Step enabled, Intel® Turbo Boost Technology enabled, 64GB DDR4 Memory, 8x 8GB DDR4 2400 MT/s, Ubuntu 16.04.1, Linux kernel 4.10.1, 1x 25GbE Mellanox 2P CX-4, CX-4 FW= 14.16.1020, mlx5_core= 3.0-1 driver, 1 ColdStream, connected to socket 0, 4KB Random Read I/0 1 initiators, each initiator connected to bx NVMe-oF subsystems using 2P 25GbE Mellanox. Performance measured by Intel using SPDK perf tool, 4KB Random Read I/O, Queue Depth: 1/NVMe-oF subsystem.

numjobs 1, 300 sec runtime, direct=1, norandommap=1

14th ANNUAL WORKSHOP 2018

THANK YOU

Paul Luse

Intel Corporation

LEGAL DISCLAIMER

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation.
Performance varies depending on system configuration.
No computer system can be absolutely secure.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. For more complete information about performance and benchmark results, visit http://www.intel.com/benchmarks .
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/benchmarks .
Benchmark results were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown." Implementation of these updates may make these results inapplicable to your device or system.
Intel® Advanced Vector Extensions (Intel® AVX)* provides higher throughput to certain processor operations. Due to varying processor power characteristics, utilizing AVX instructions may cause a) some parts to operate at less than the rated frequency and b) some parts with Intel® Turbo Boost Technology 2.0 to not achieve any or maximum turbo frequencies. Performance varies depending on hardware, software, and system configuration and you can learn more at http://www.intel.com/go/turbo.
Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.