SDC
STORAGE DEVELOPER CONFERENCE
SNIA SANTA CLARA, 2016

SPDK: Building Blocks For Scalable, High Performance Storage Applications

Benjamin Walker
Intel Corporation

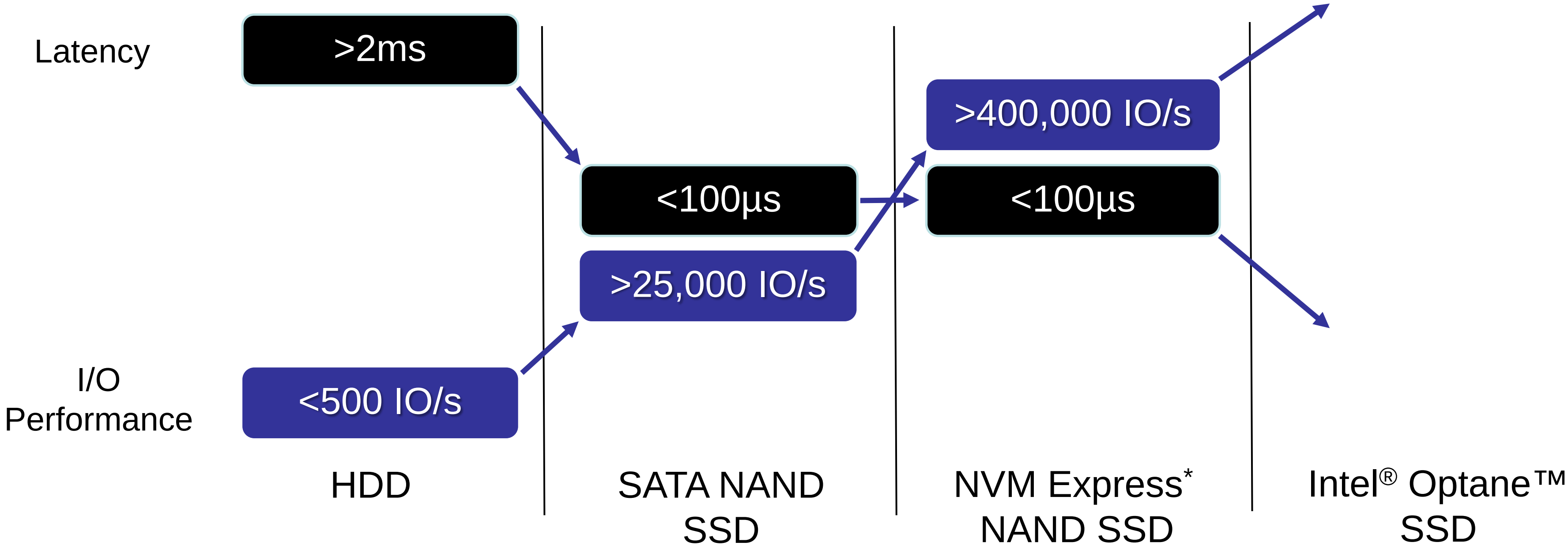
Agenda

- What is the Storage Performance Development Kit (SPDK)?
- How did SPDK get started?
- What are the benefits of an NVM Express* (NVMe) polled mode driver?
- How does SPDK support protocols like NVMe over Fabrics?
- What are some of the future areas of development for SPDK?
- Summary and Next Steps

Agenda

- What is the Storage Performance Development Kit (SPDK)?
- How did SPDK get started?
- What are the benefits of an NVM Express* (NVMe) polled mode driver?
- How does SPDK support protocols like NVMe over Fabrics?
- What are some of the future areas of development for SPDK?
- Summary and Next Steps

The Problem: Software is becoming the bottleneck



The Opportunity: Use Intel software ingredients to unlock the potential of new media

Storage Performance Development Kit

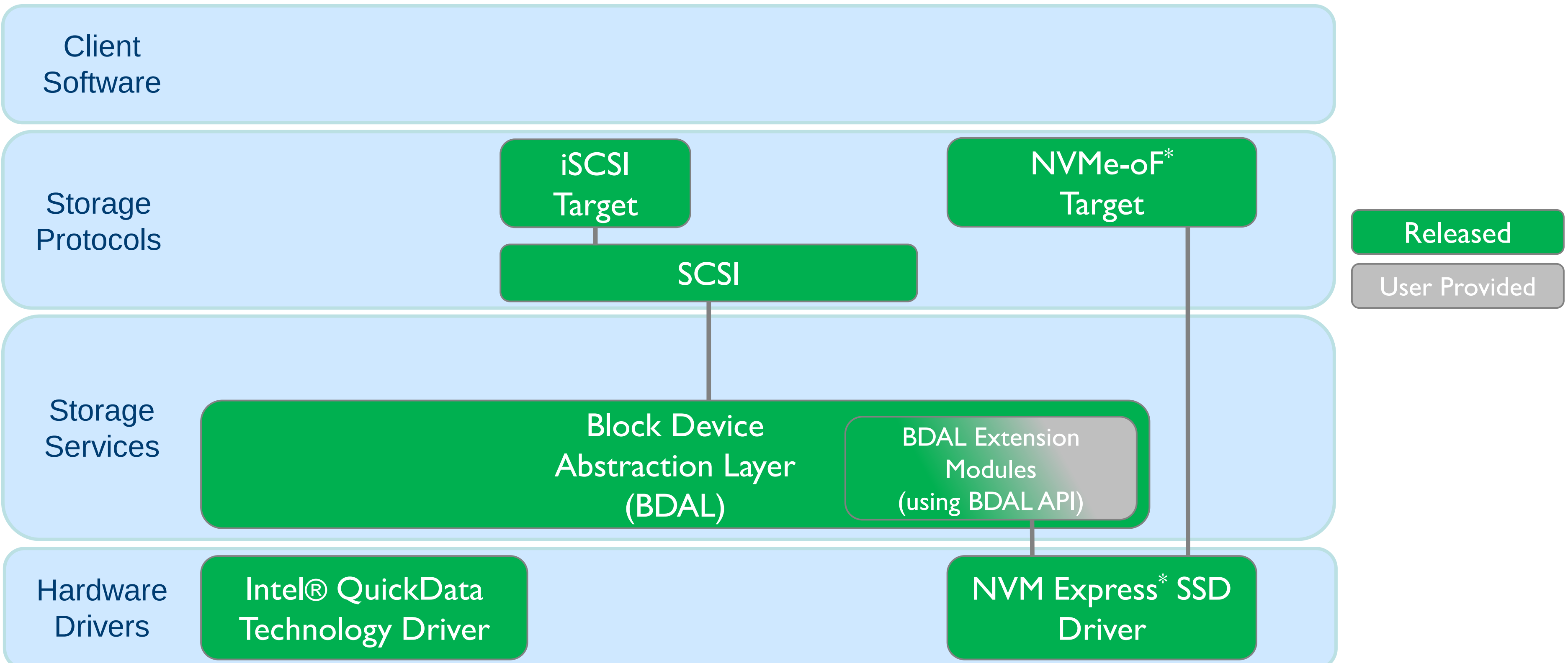
Intel® Platform Storage Reference Architecture

- Optimized for *Intel platform* characteristics
- Open source building blocks (BSD licensed)
- Available via github.com/spdk or spdk.io

Scalable and Efficient Software Ingredients

- User space, lockless, polled-mode components
- Up to millions of IOPS per core
- Designed for Intel Optane™ technology latencies

Storage Performance Development Kit (SPDK)



All products, computer systems, dates and figures specified are preliminary based on current expectations, and are subject to change without notice. The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request. This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps. Statements in this document that refer to Intel's plans and expectations for the quarter, the year, and the future, are forward-looking statements that involve a number of risks and uncertainties. A detailed discussion of the factors that could affect Intel's results and plans is included in Intel's SEC filings, including the annual report on Form 10-K.

Agenda

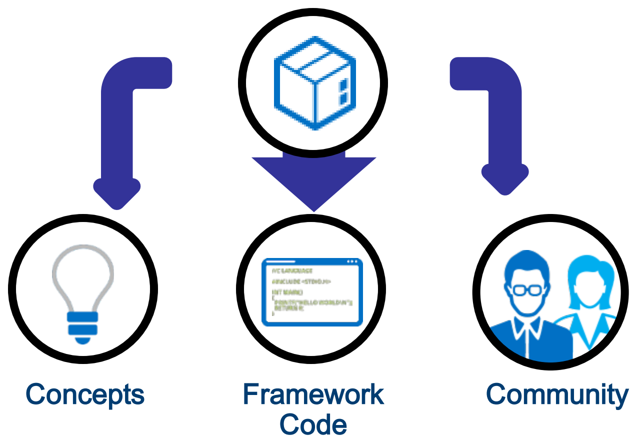
- What is the Storage Performance Development Kit (SPDK)?
- How did SPDK get started?
- What are the benefits of an NVM Express* (NVMe) polled mode driver?
- How does SPDK support protocols like NVMe over Fabrics?
- What are some of the future areas of development for SPDK?
- Summary and Next Steps

Data Plane Development Kit (DPDK)

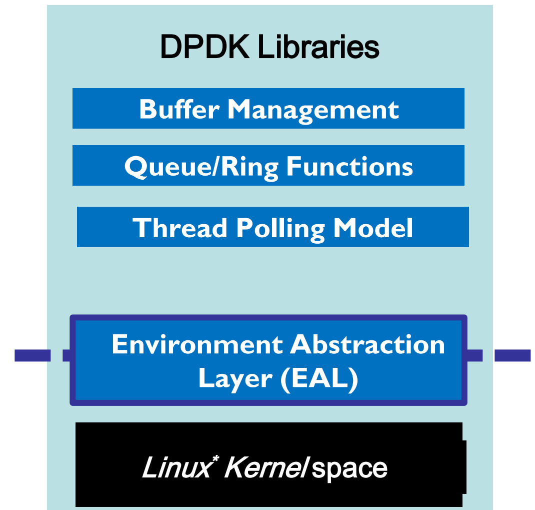
Software solution for accelerating Packet Processing workloads

- Optimized for IA platforms
- Vibrant community support
- Free, Open Source, BSD License
- Website: dpdk.org

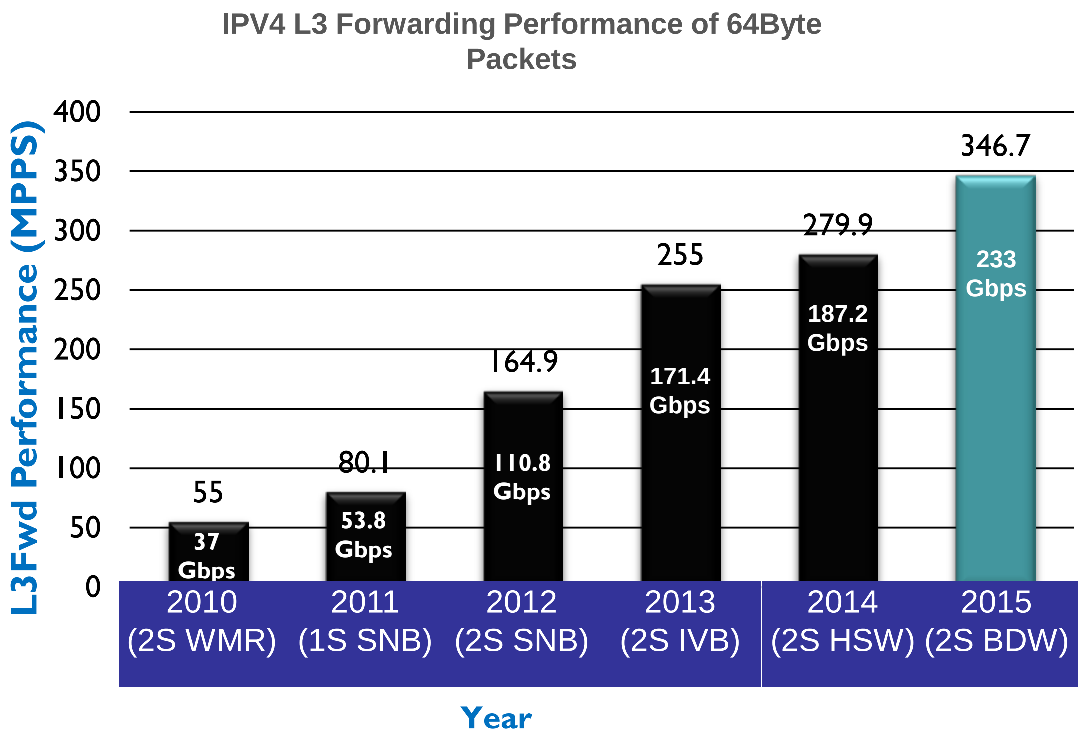
What does SPDK share with DPDK?



What DPDK Primitives Does SPDK Use?



DPDK Generational Performance

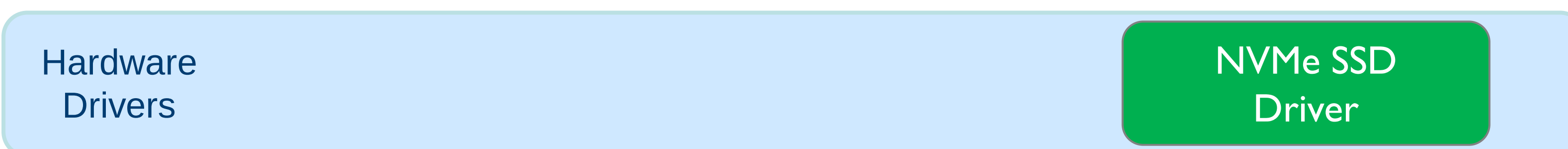


Broadwell EP System Configuration	
Hardware	
Platform	SuperMicro* - X10DRX
CPU	Intel® Xeon® Processor E5-2658 v4
Chipset	Intel® C612 chipset
Sockets	2
Cores per Socket	14 (28 threads)
LL CACHE	30 MB
QPI/DMI	9.6GT/s
PCIe	Gen3x8
MEMORY	DDR4 2400 MHz, 1Rx4 8GB (total 64GB), 4 Channel per Socket
NIC	10 x Intel® Ethernet CNA XL710-QDA2PCI-Express* Gen3 x8 Dual Port 40 GbE Ethernet NIC (1x40G/card)
NIC Mbps	40,000
BIOS	BIOS version: 1.0c (02/12/2015)
Software	
OS	Debian* 8.0
Kernel version	3.18.2
Other	DPDK2.2.0

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. For more complete information visit http://www.intel.com/performance.
Data captured by Intel with DPDK l3fwd (Layer 3 forwarding) sample application. Packet generator: Ixia IxNetwork 8.03 EA.

Agenda

- What is the Storage Performance Development Kit (SPDK)?
- How did SPDK get started?
- What are the benefits of an NVM Express* (NVMe) polled mode driver?
- How does SPDK support protocols like NVMe over Fabrics?
- What are some of the future areas of development for SPDK?
- Summary and Next Steps

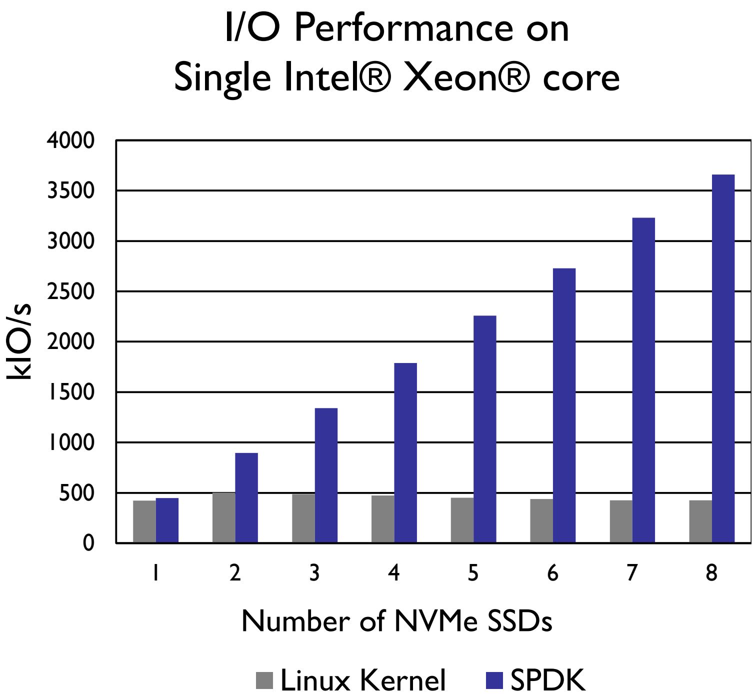


NVM Express* Driver Key Characteristics

- Supports NVM Express* (NVMe) 1.2 spec-compliant devices
- Userspace Asynchronous Polled Mode operation
- Application owns I/O queue allocation and synchronization

Feature	Description
End-to-end Data Protection	Integrity from host to drive with T10-DIF/DIX
Scatter-Gather Lists (SGL)	Eliminates buffer copies
Reservations	For dual port NVMe usage models
Namespace Management	Support multiple dynamic NVMe namespaces
Weighted Round Robin	Quality of Service for NVMe I/O queues

NVM Express* Driver Throughput Scalability

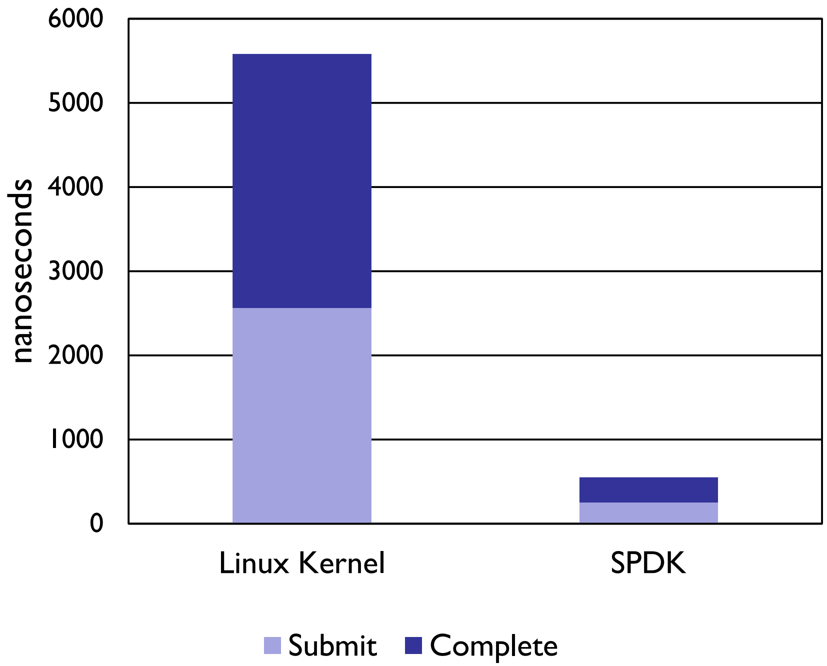


- Systems with multiple NVM Express* (NVMe) SSDs capable of millions of I/O per second
- Results in many cores of software overhead with kernel-based interrupt-driven driver model
- SPDK enables:
  - more CPU cycles for storage services
  - lower I/O latency

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 enabled, 8x 8GB DDR4 2133 MT/s, 1 DIMM per channel, CentOS* Linux* 7.2, Linux kernel 4.7.0-rc1, 8x Intel® P3700 NVMe SSD (800GB), 4x per CPU socket, FW 8DV10102, 4KB Random Read I/O, Queue Depth: 32 per SSD. Performance measured by Intel using SPDK perf tool, 4KB Random Read I/O, Queue Depth: 128/SSD

NVM Express* Driver Software Overhead



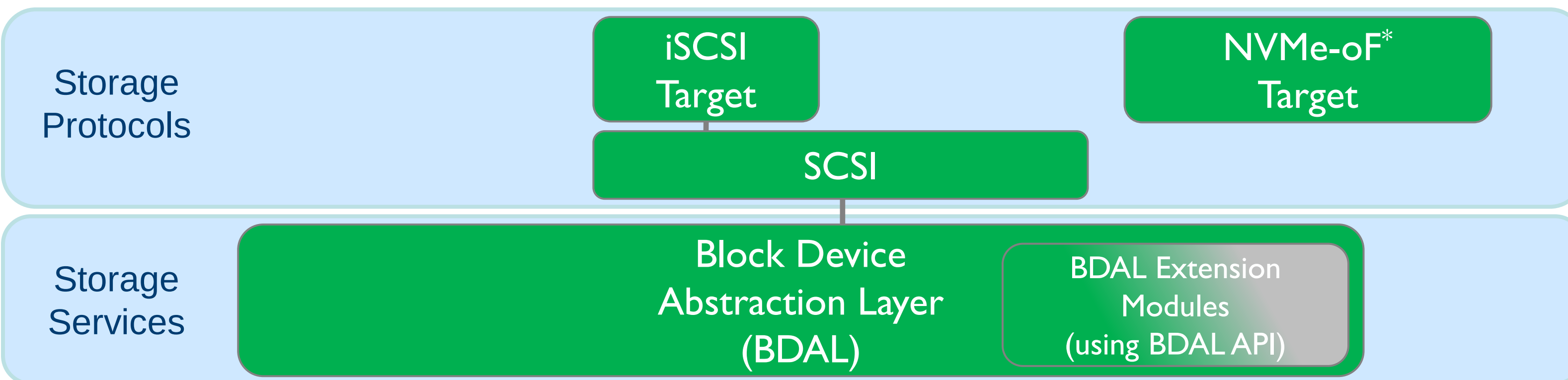
Kernel Source of Overhead	SPDK Approach
Interrupts	Asynchronous Polled Mode
Synchronization	Lockless
System Calls	Userspace Hardware Access
DMA Mapping	Hugepages
Generic Block Layer	Specific for Flash Latencies

SPDK reduces NVM Express* (NVMe) software overhead up to 10x!

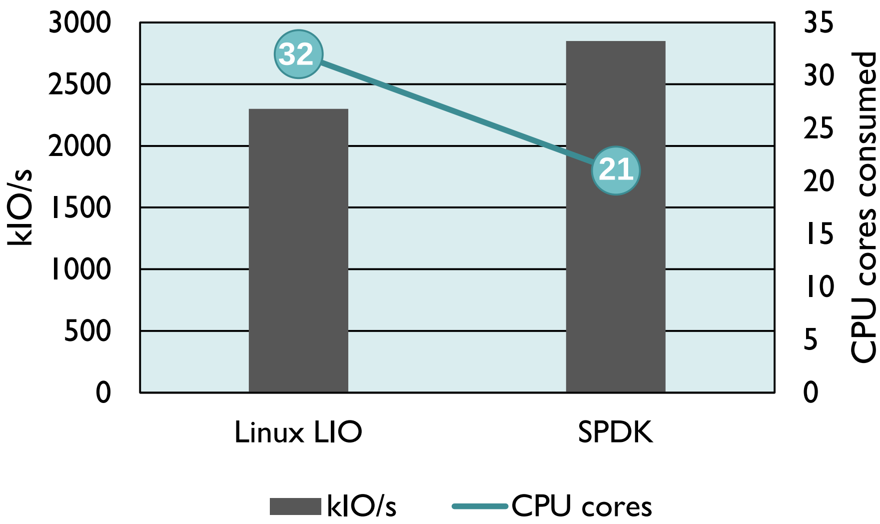
System Configuration: 2x Intel® Xeon® E5-2695v4 (HT off), Intel® Speed Step enabled, Intel® Turbo Boost Technology disabled, 8x 8GB DDR4 2133 MT/s, 1 DIMM per channel, CentOS* Linux* 7.2, Linux kernel 4.7.0-rc1, 1x Intel® P3700 NVMe SSD (800GB), 4x per CPU socket, FW 8DV10102, I/O workload 4KB random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

Agenda

- What is the Storage Performance Development Kit (SPDK)?
- How did SPDK get started?
- What are the benefits of an NVM Express* (NVMe) polled mode driver?
- How does SPDK support protocols like NVMe over Fabrics?
- What are some of the future areas of development for SPDK?
- Summary and Next Steps



iSCSI Performance



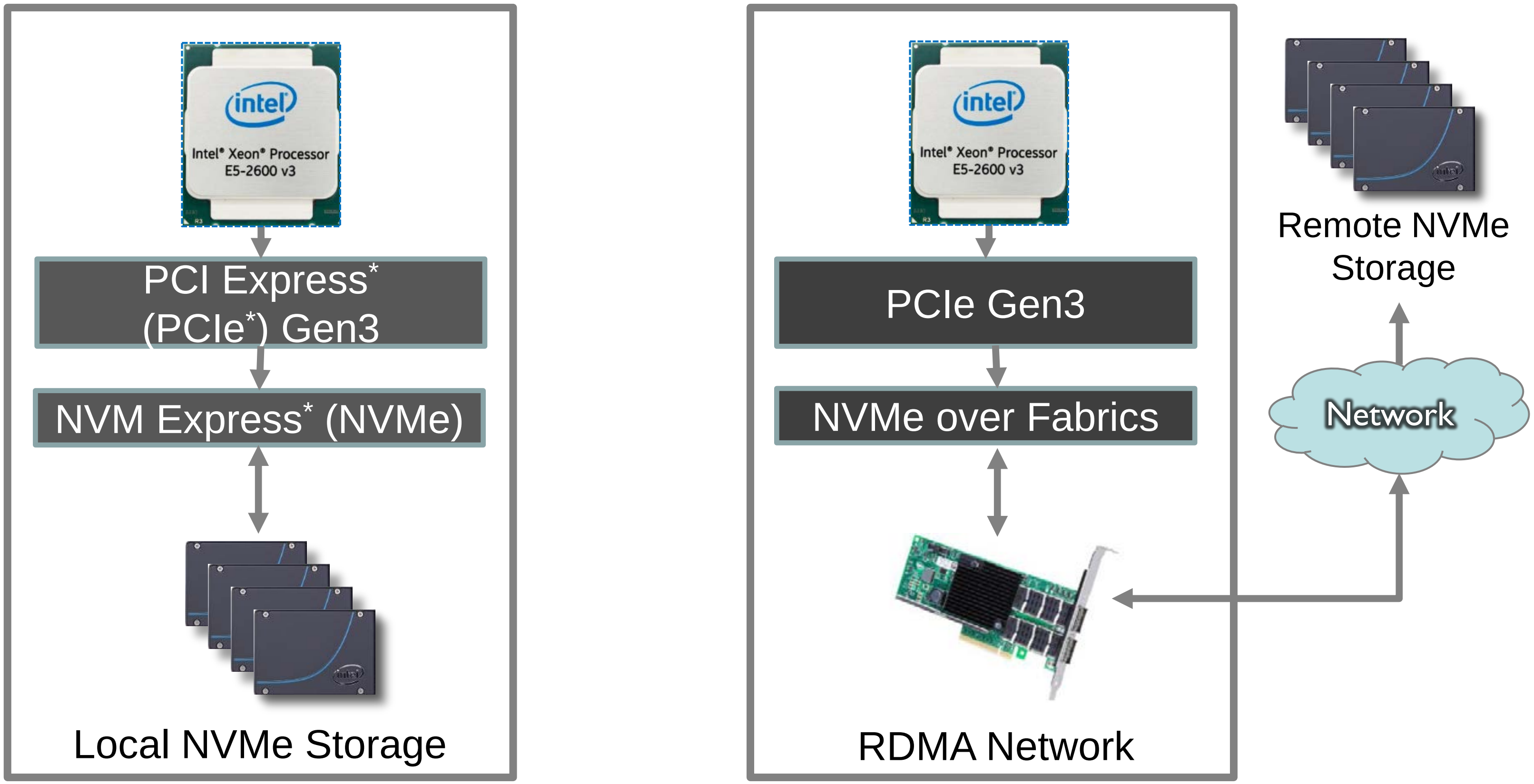
- iSCSI Target improvements stem from:
  - Non-blocking TCP sockets
  - Pinned iSCSI connections
  - SPDK storage access model
- TCP processing is limiting factor
  - 70%+ CPU cycles consumed in kernel network stack
  - Userspace polled mode TCP required for more improvement

SPDK improves efficiency almost 2x

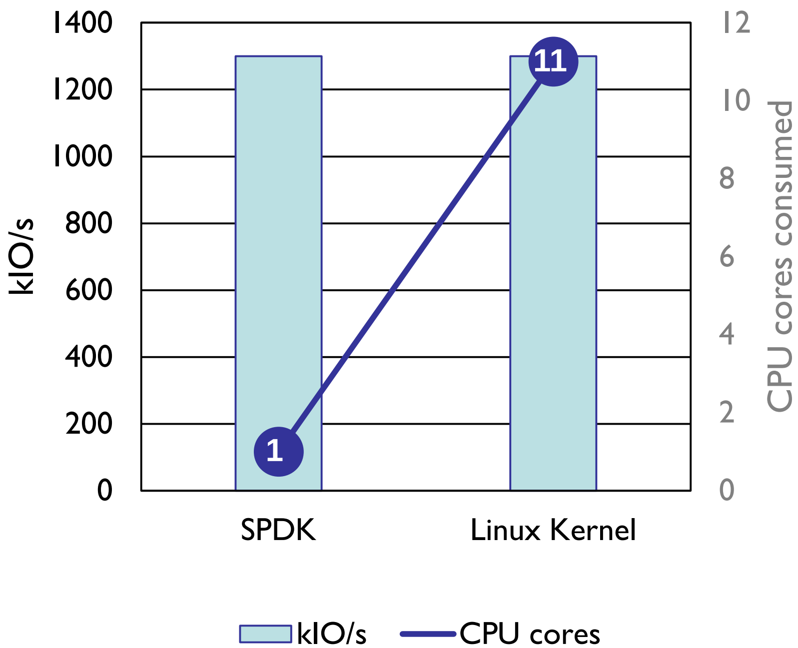
System Configuration: 2S Intel® Xeon® E5-2699v3: 18C, 2.3GHz (HT off), Intel® Speed Step enabled, Intel® Turbo Boost Technology disabled, 8x4GB DDR4 2133 MT/s, 1 DIMM per channel, Ubuntu* Server 14.10, 3.16.0-30-generic kernel, Ethernet Controller XL710 for 40GbE, 8x Intel® P3700 NVM Express* SSD – 800GB (4 per CPU socket), FW 8DV10102

As measured by: fio – Direct=Yes, 4KB random read I/O, QueueDepth=32, Ramp Time=30s, Run Time=180s, Norandommap=1, I/O Engine = libaio, Numjobs=1

Why NVM Express* over Fabrics?



NVM Express* over Fabrics 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: Target system: 2x Intel® Xeon® E5-2695v4 (HT off), Intel® Speed Step enabled, Intel® Turbo Boost Technology enabled, 8x 8GB DDR4 2133 MT/s, 1 DIMM per channel, 8x Intel® P3700 NVMe SSD (800GB), 4x per CPU socket, FW 8DV10102, Network: Mellanox* ConnectX-4 100Gb RDMA, direct connection between initiator and target; Initiator OS: CentOS* Linux* 7.2, Linux kernel 4.7.0-rc2, Target OS (SPDK): CentOS Linux 7.2, Linux kernel 3.10.0-327.el7.x86_64, Target OS (Linux kernel): CentOS Linux 7.2, Linux kernel 4.7.0-rc2 Performance as measured by: fio, 4KB Random Read I/O, 2 RDMA QP per remote SSD, Numjobs=4 per SSD, Queue Depth: 32/job

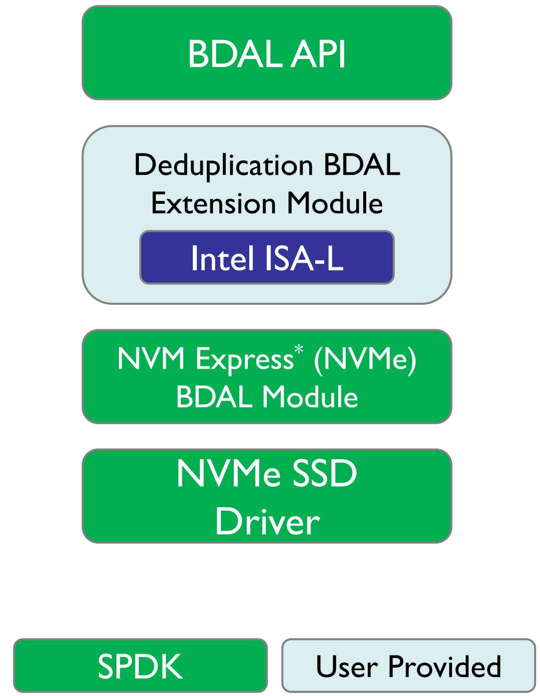
Block Device Abstraction Layer (BDAL)

- Block layer optimized for SPDK programming model
  - Lockless, event driven API
  - BDAL API for creating new BDAL drivers
  - Stackable
- Several BDAL modules available today
  - NVM Express* (NVMe) – SPDK NVMe polled mode driver
  - AIO – Linux libaio
  - malloc – Userspace ramdisk

BDAL Extension Modules – Example #1

Intel® Intelligent Storage Acceleration Library (Intel® ISA-L)

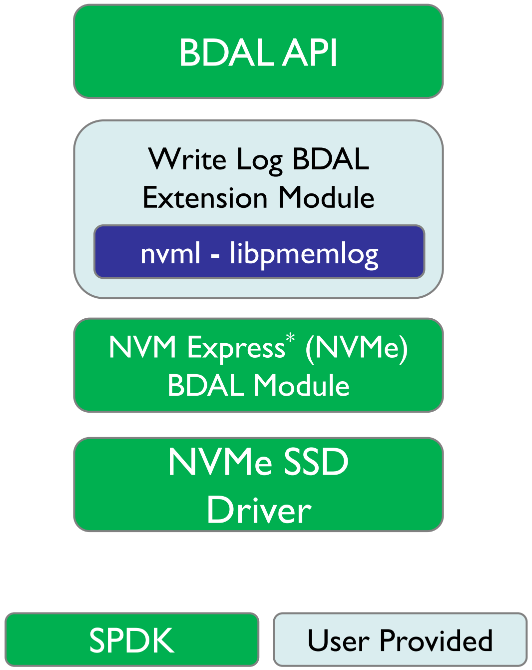
- Intel® Intelligent Storage Acceleration Library (Intel® ISA-L)
  - Optimized low-level functions targeting storage applications
  - Erasure coding, parity, CRC, compression, crypto, hashing
  - https://github.com/01org/isa-l
- Example:
  - User-provided deduplication extension module



BDAL Extension Modules – Example #2

nvml – Linux NVM Library

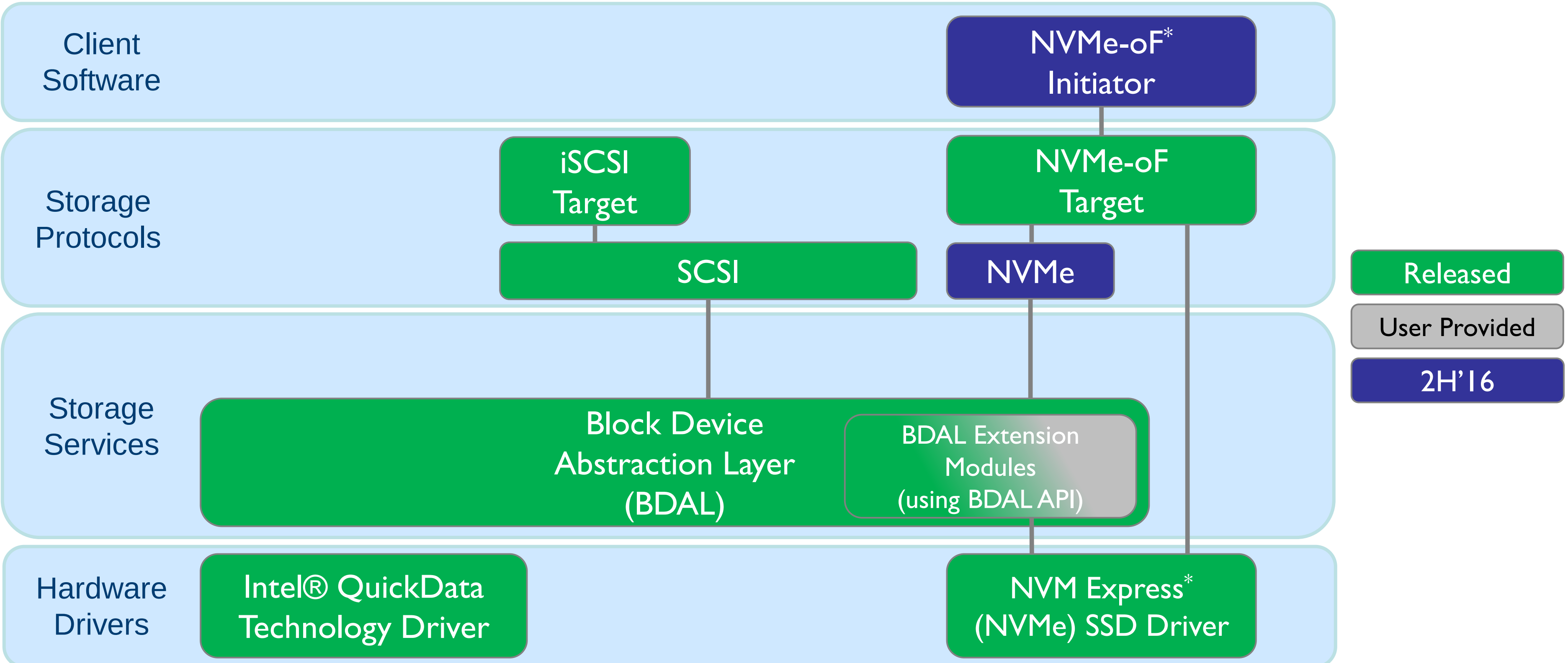
- Linux* NVM Library
  - Set of libraries to provide useful APIs for persistent memory server applications
  - Enables 3D XPoint™ memory
- Example:
  - User-provided write log



Agenda

- What is the Storage Performance Development Kit (SPDK)?
- How did SPDK get started?
- What are the benefits of an NVM Express* (NVMe) polled mode driver?
- How does SPDK support protocols like NVMe over Fabrics?
- What are some of the future areas of development for SPDK?
- Summary and Next Steps

Storage Performance Development Kit (SPDK)

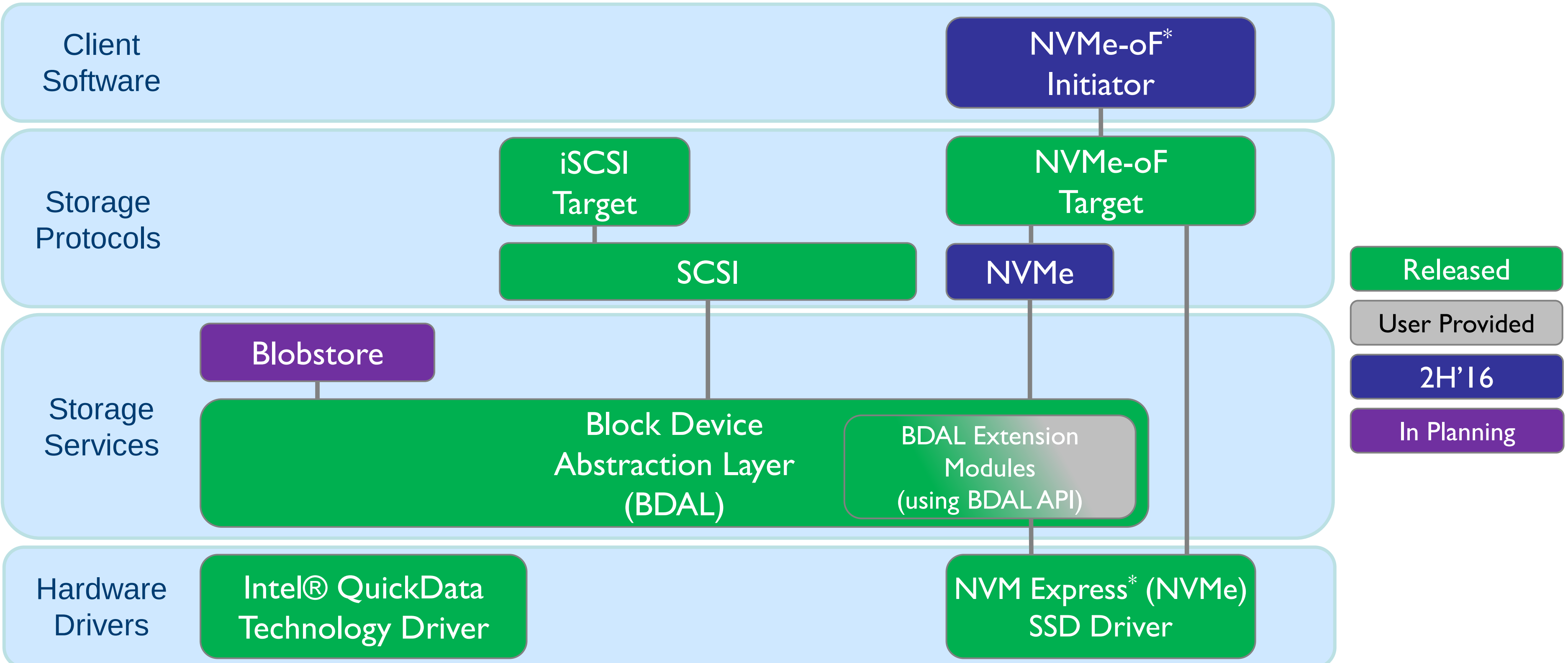


All products, computer systems, dates and figures specified are preliminary based on current expectations, and are subject to change without notice. The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request. This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps. Statements in this document that refer to Intel's plans and expectations for the quarter, the year, and the future, are forward-looking statements that involve a number of risks and uncertainties. A detailed discussion of the factors that could affect Intel's results and plans is included in Intel's SEC filings, including the annual report on Form 10-K.

NVM Express* over Fabrics Additions

- Initiator
  - Enable polled mode userspace access to remote NVM Express* (NVMe) devices
  - Same programming model as SPDK local NVMe access
- BDAL integration w/ NVMe over Fabrics target
  - Export SPDK block devices over NVMe over Fabrics
    - Similar to iSCSI
- Continued performance tuning
  - Scaling to more NVMe devices, more RDMA throughput

Storage Performance Development Kit (SPDK)

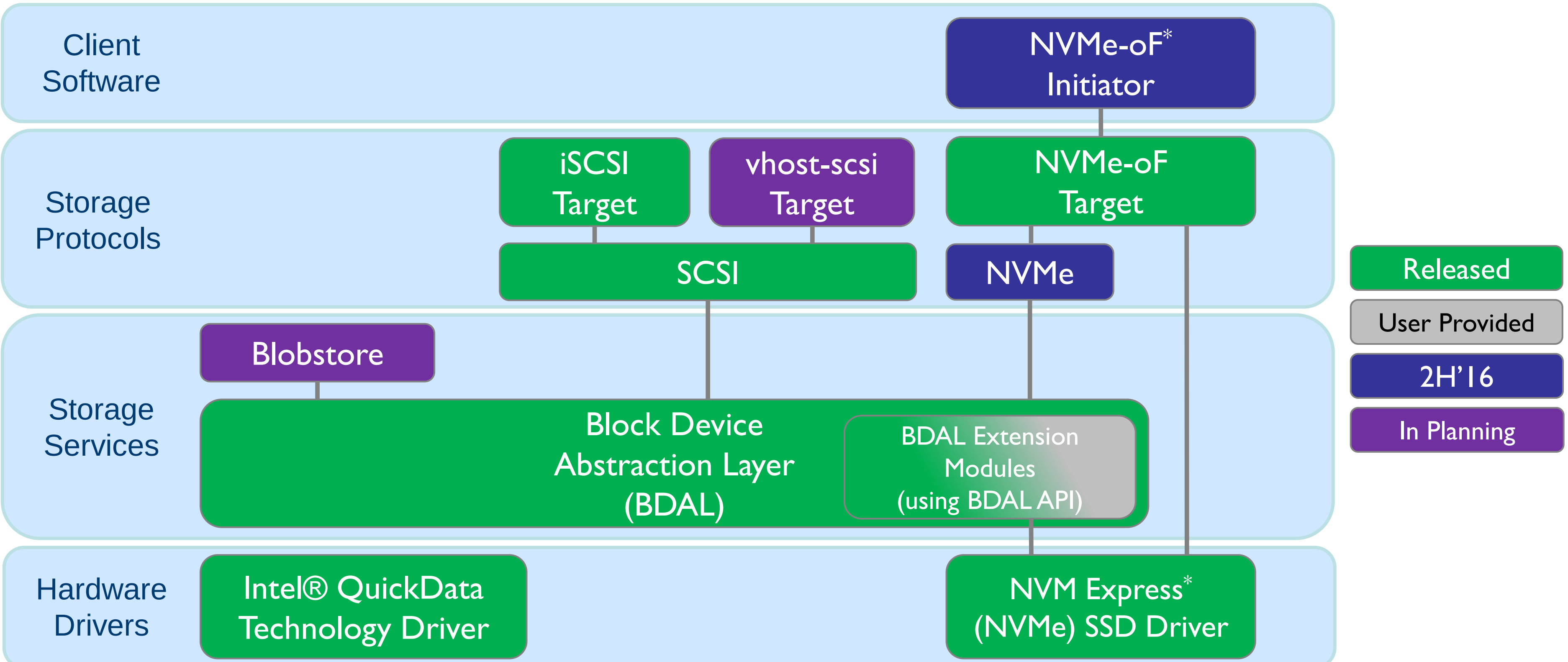


All products, computer systems, dates and figures specified are preliminary based on current expectations, and are subject to change without notice. The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request. This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps. Statements in this document that refer to Intel's plans and expectations for the quarter, the year, and the future, are forward-looking statements that involve a number of risks and uncertainties. A detailed discussion of the factors that could affect Intel's results and plans is included in Intel's SEC filings, including the annual report on Form 10-K.

What about a filesystem?

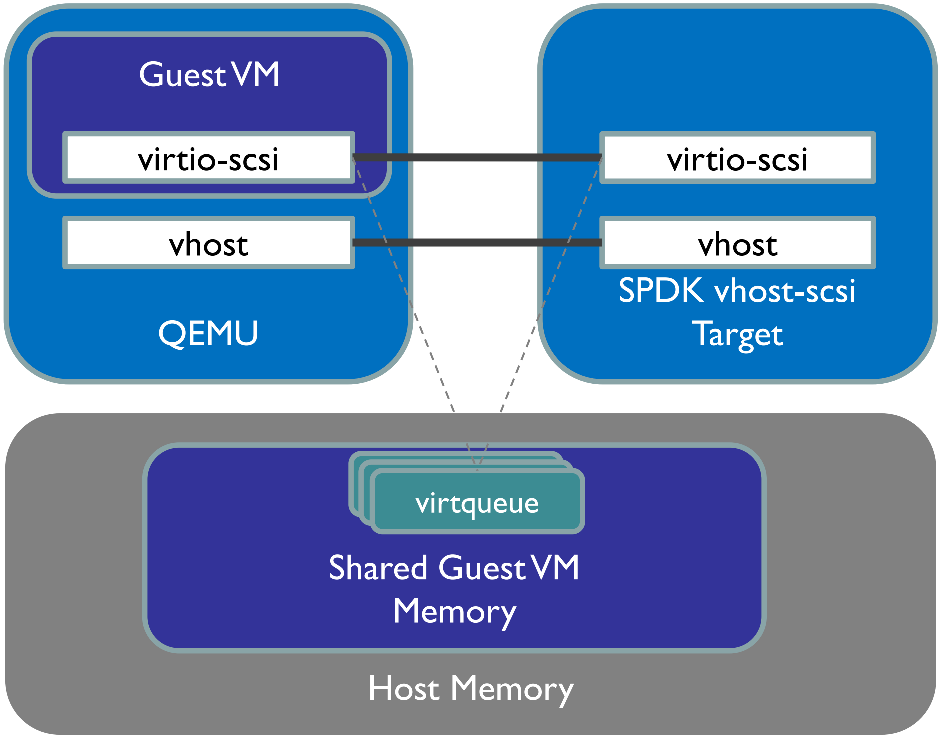
- Most applications want some level of file semantics
  - Example: databases, key/value stores – small number of files, flat hierarchy, no permissions
- Kernel filesystems not usable in SPDK programming model
  - They are in the kernel
  - They are based on POSIX synchronous file semantics
- Need framework for SPDK file-like semantics – an SPDK "Blobstore"
  - Asynchronous, polled-mode, lockless, event driven (i.e., not POSIX)
  - Framework for building higher order services
    - Lightweight filesystem, extent allocator, etc.

Storage Performance Development Kit (SPDK)



All products, computer systems, dates and figures specified are preliminary based on current expectations, and are subject to change without notice. The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request. This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps. Statements in this document that refer to Intel's plans and expectations for the quarter, the year, and the future, are forward-looking statements that involve a number of risks and uncertainties. A detailed discussion of the factors that could affect Intel's results and plans is included in Intel's SEC filings, including the annual report on Form 10-K.

SPDK vhost-scsi



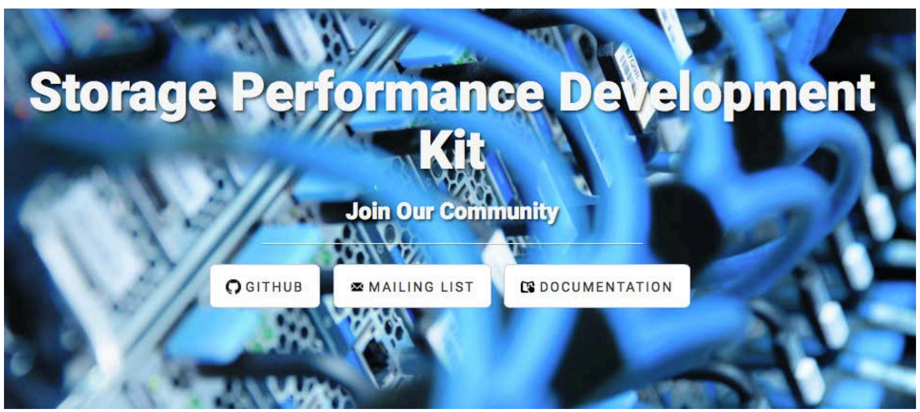
- Serve SPDK storage to local virtual machines
  - NVM Express* ephemeral storage
  - SPDK-based BDAL storage
- Leverage existing infrastructure for
  - QEMU vhost-scsi
  - QEMU/DPDK vhost-net user

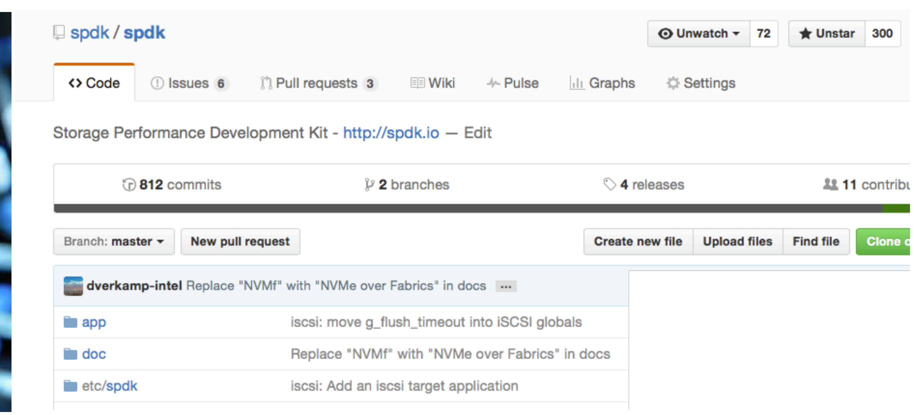
Agenda

- What is the Storage Performance Development Kit (SPDK)?
- How did SPDK get started?
- What are the benefits of an NVM Express* (NVMe) polled mode driver?
- How does SPDK support protocols like NVMe over Fabrics?
- What are some of the future areas of development for SPDK?
- Summary and Next Steps

Summary and Next Steps

- Fully realizing new media performance requires software optimizations
- SPDK positioned to enable developers to realize this performance
- SPDK available today via http://spdk.io
- Help us build SPDK as an open source community!





Q&A

Legal Notices and Disclaimers

- 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. Check with your system manufacturer or retailer or learn more at intel.com.
- Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit http://www.intel.com/performance.
- 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.
- 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.
- This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.
- No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.
- Statements in this document that refer to Intel's plans and expectations for the quarter, the year, and the future, are forward-looking statements that involve a number of risks and uncertainties. A detailed discussion of the factors that could affect Intel's results and plans is included in Intel's SEC filings, including the annual report on Form 10-K.
- All products, computer systems, dates and figures specified are preliminary based on current expectations, and are subject to change without notice. The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.
- 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.
-
-

Backup