SDC

STORAGE DEVELOPER CONFERENCE

SNIA SANTA CLARA, 2017

Accelerate block service built on Ceph via SPDK

Ziye Yang
Intel

Agenda

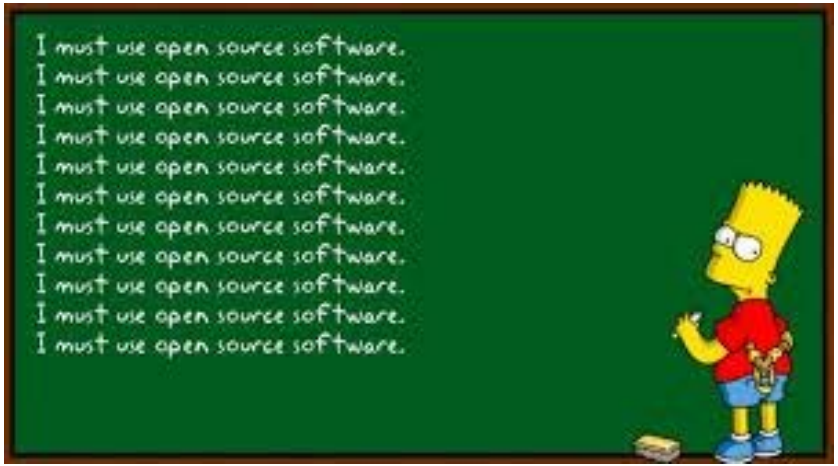
- SPDK Introduction
- Accelerate block service built on Ceph
- SPDK support in Ceph bluestore
- Summary

Agenda

- SPDK Introduction
- Accelerate block service built on Ceph
- SPDK support in Ceph bluestore
- Summary

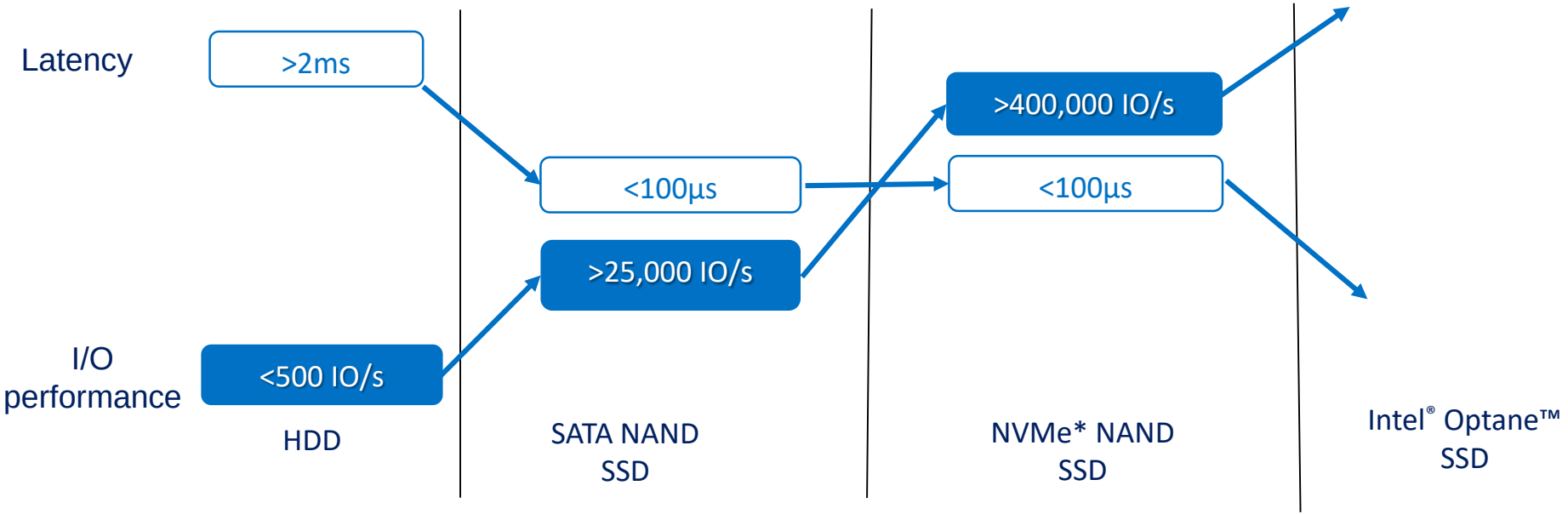
What?
Storage Performance Development Kit

- Software Building Blocks
- Open Source
- BSD Licensed
- Userspace and Polled Mode



http://spdk.io

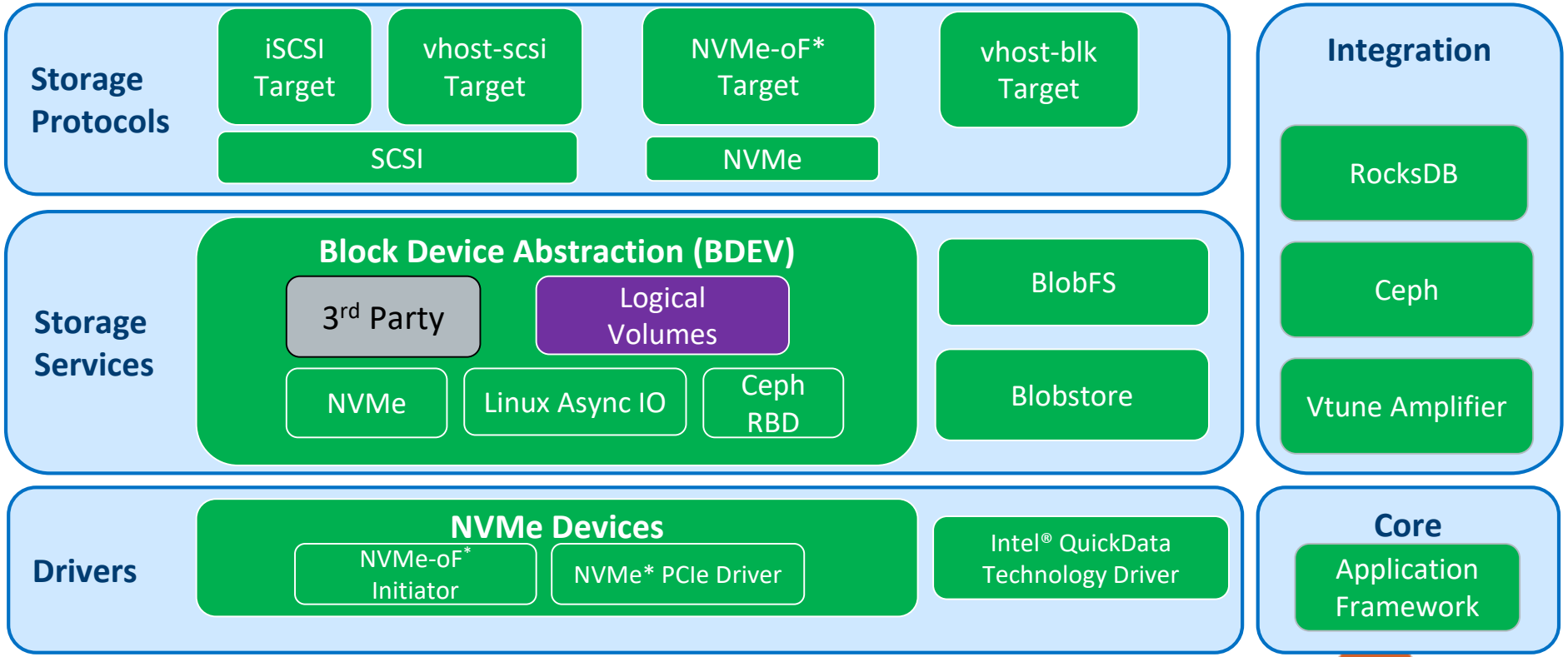
The problem: Software is becoming the bottleneck



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

Architecture

Released
Q4'17



Why? Efficiency & Performance

SPDK

more performance from Intel CPUs, non-volatile media, and networking

Up to **10X MORE** IOPS/core for NVMe-oF* vs. Linux kernel

Up to **8X MORE** IOPS/core for NVMe vs. Linux kernel

Up to **350%** **BETTER** Tail Latency for RocksDB workloads

FASTER TTM/
LESS RESOURCES
than developing components from scratch

Provides Future Proofing as NVM technologies increase in 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 information go to http://www.intel.com/performance

How? SPDK Community

- **Github :** https://github.com/spdk/spdk
- **Trello** : https://trello.com/spdk
- **GerritHub** : https://review.gerrithub.io/#/q/project:spdk/spdk+status:open
- **IRC:** https://freenode.net/ we're on #spdk
- **Home Page:** http://www.spdk.io/

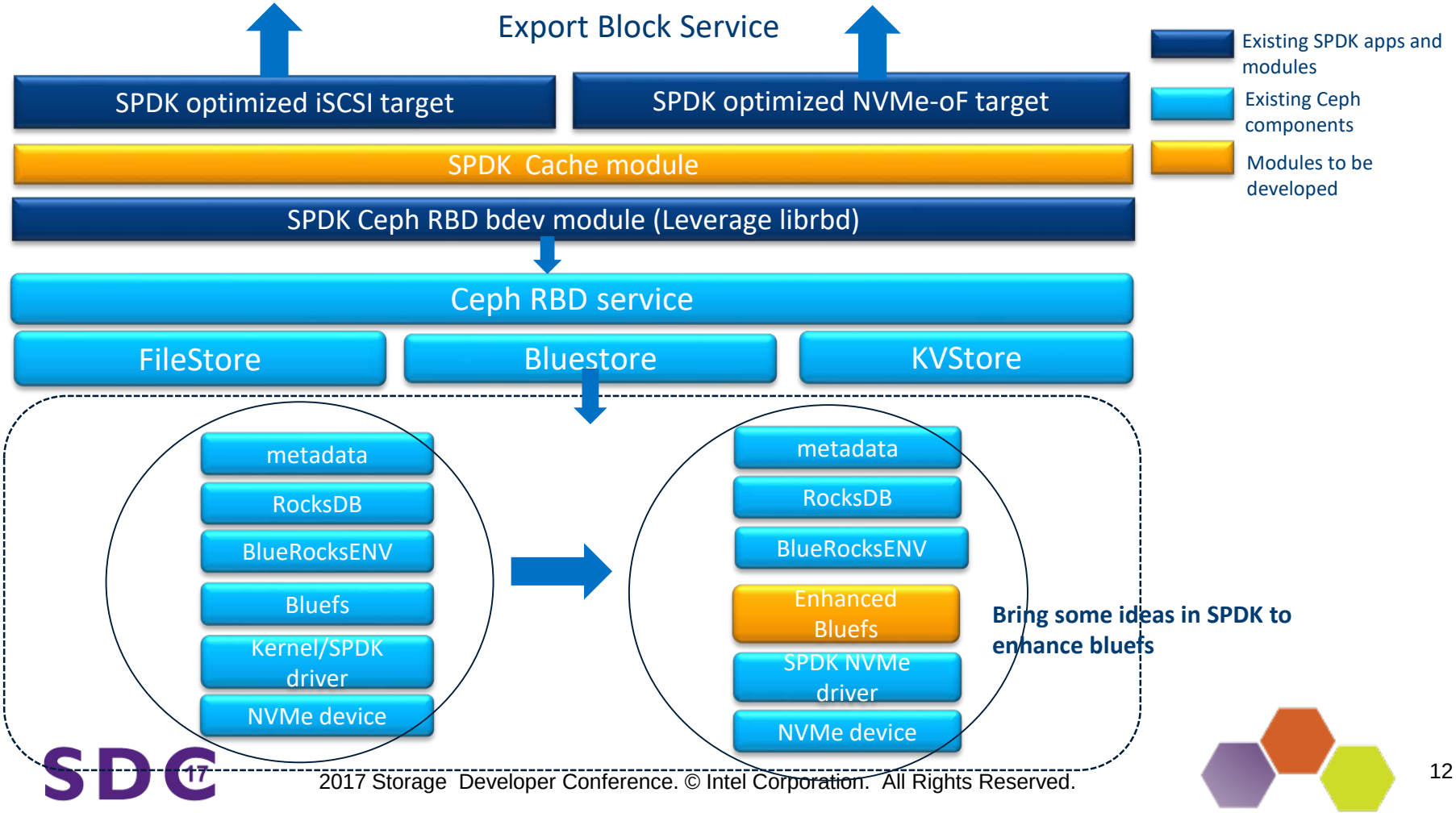
1st SPDK Hackathon!! Nov 6-8 2017, Phoenix

Agenda

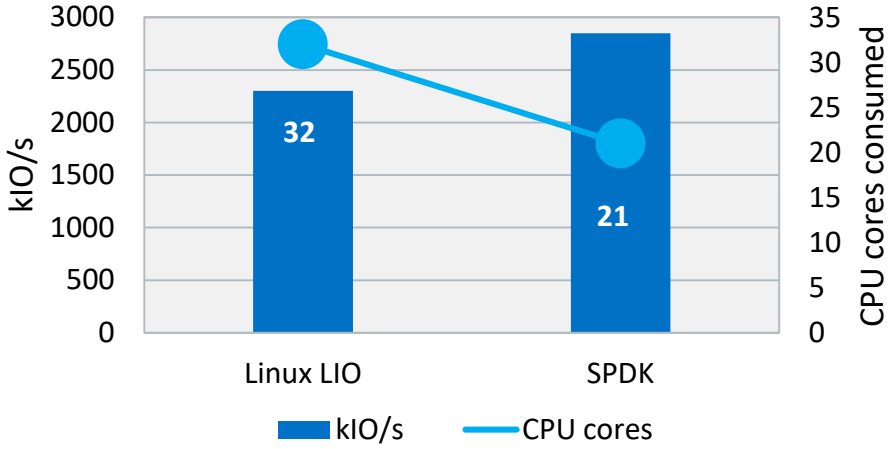
- SPDK Introduction
- Accelerate block service built on Ceph
- SPDK support in Ceph bluestore
- Summary

Leverage SPDK to accelerate the block service built on Ceph

- Block service daemon optimization outside Ceph
  - Use optimized Block service daemon, e.g., SPDK iSCSI target or NVMe-oF target
  - Introduce Proper Cache policy in optimized block service daemon.
- OSD Optimization inside Ceph
  - Use SPDK's user space NVMe driver instead of Kernel NVMe driver in bluestore (already have)
  - Bring some ideas from SPDK Blobfs/Blobstore into Ceph Bluefs/Bluestore
  - Network optimization (e.g., Leverage user space stack on DPDK or RDMA, will not be discussed in this topic)



SPDK iSCSI target and LIO performance comparison for local detached storage



- 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

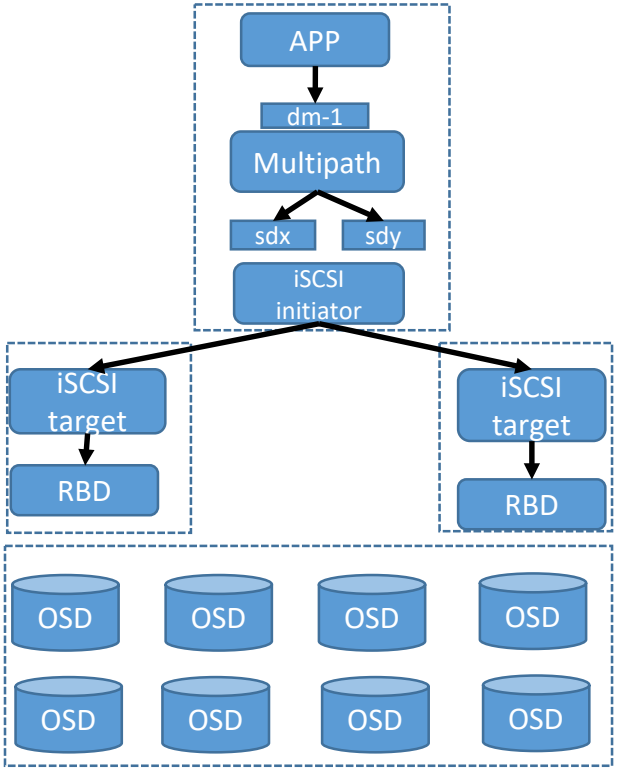
SPDK improves efficiency almost 2x

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

Agenda

- SPDK Introduction
- Accelerate block service built on Ceph
  - Case study: Accelerate iSCSI service exported by Ceph (From iStuary's talk in SPDK meetup 2016)
- SPDK support in Ceph bluestore
- Summary

Block service exported by Ceph via iSCSI protocol



Client

iSCSI gateway

Ceph cluster

- Cloud service providers which provision VM service can use iSCSI.
- If Ceph could export block service with good performance, it would be easy to glue those providers to Ceph cluster solution.

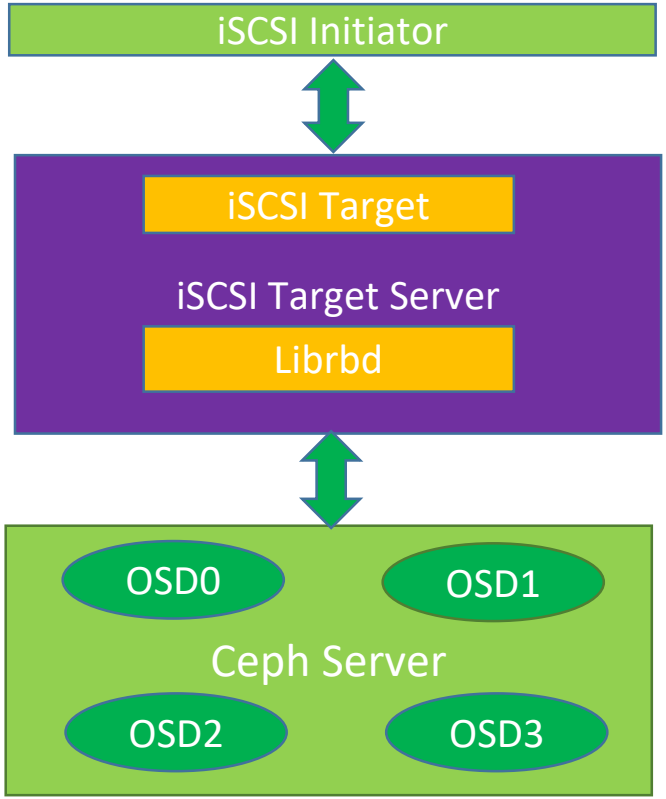
iSCSI + RBD Gateway

Ceph server

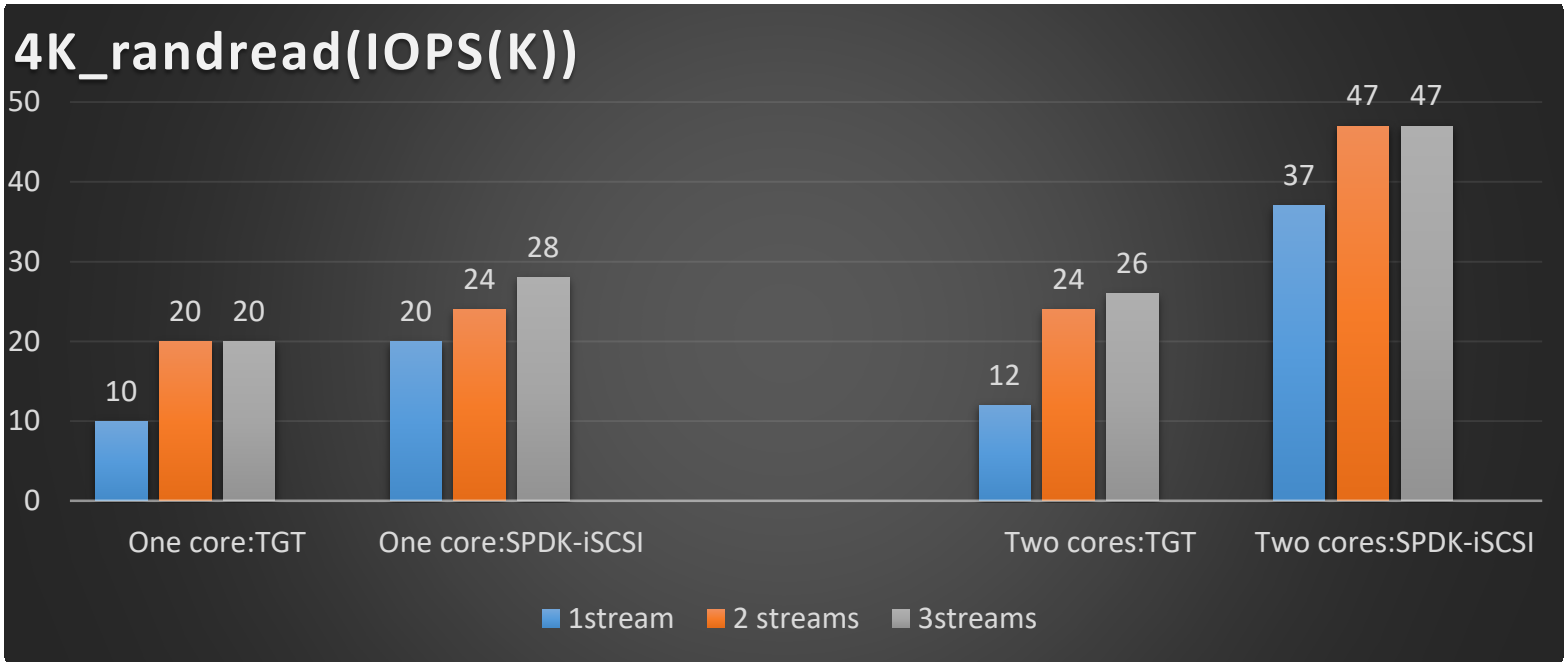
- CPU: Intel(R) Xeon(R) CPU E5-2660 v4 @2.00GHz
- Four intel P3700 SSDs
- One OSD on each SSD, total 4 osds
- 4 pools PG number 512, one 10G image in one pool

iSCSI target server (librbd+SPDK / librbd+tgt)

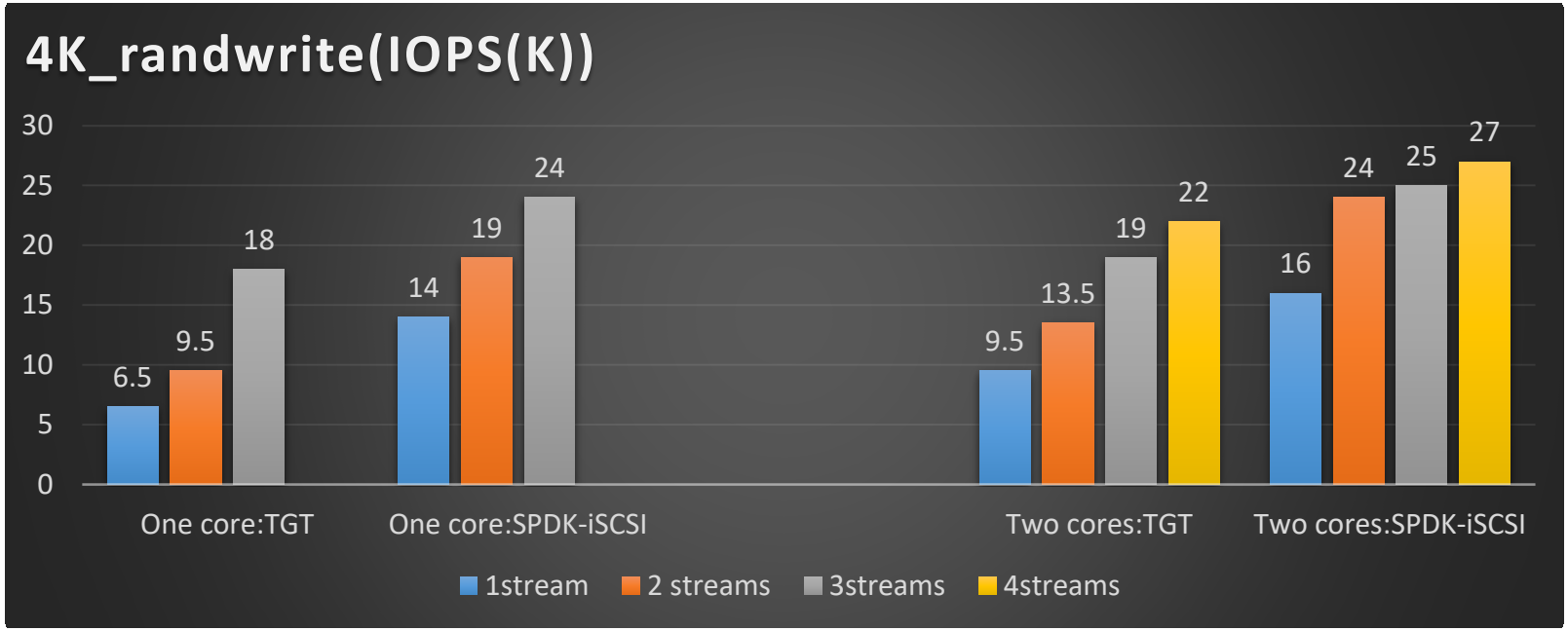
- CPU: Intel(R) Core(TM) i7-6700 CPU @ 3.40GHz
- Only one core enable iSCSI initiator
- CPU: Intel(R) Core(TM) i7-6700 CPU @ 3.40GHz



Read performance comparison

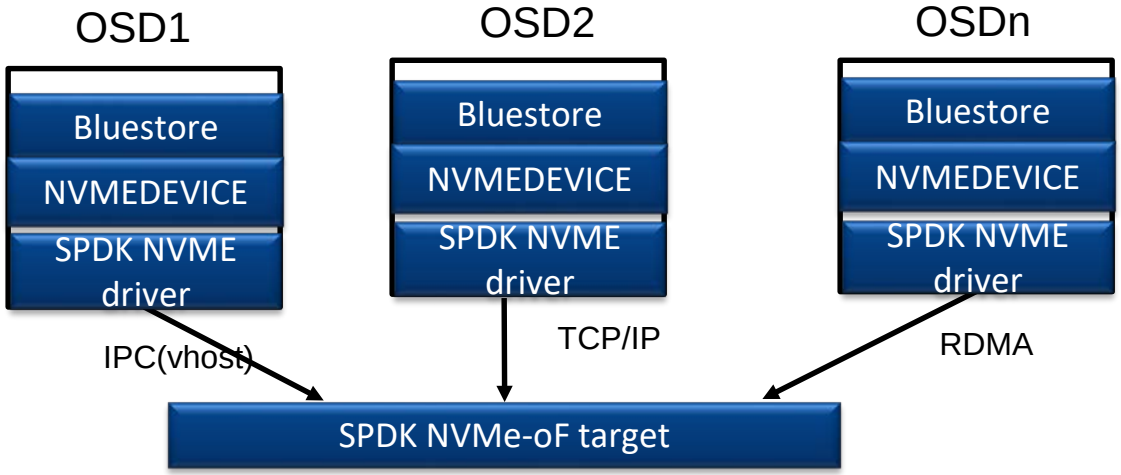


Write performance comparison



Proposals/opportunities for better leveraging SPDK in Ceph

- Multiple OSDs support on same NVMe Device by using SPDK.
  - Leverage SPDK's NVMe-oF target with NVMe driver.
  - Risks: Same with kernel, i.e., fail all OSDs on the device if the daemon crash.



Proposals/opportunities for better leveraging SPDK in Ceph

- Enhance cache support in NVMEDEVICE via using SPDK NVMe driver
  - Currently, No read/write cache while using SPDK NVMe driver.
  - Need better cache/buffer strategy for read/write performance improvement.
- Enable zero copy
  - Currently, there are memory copy in NVMEDEVICE while conducting I/O read/write
  - May need to eliminate the memory copy (Possible solution: Enable using DPDK memory while starting OSD)

Agenda

- SPDK Introduction
- Accelerate block service built on Ceph
- SPDK support in Ceph bluestore
- Summary

Current SPDK support in Ceph bluestore

- SPDK upgrade in Ceph:
  - Upgraded SPDK to 16.11 in Dec, 2016
  - Upgraded SPDK to 17.03 in April, 2017
  - Upgraded SPDK to 17.07 in August, 2017
- Stability
  - Several compilation issues, running time bugs are fixed in code base while using SPDK.

SPDK support for Ceph in future

- To make SPDK really useful in Ceph, we will still do the following works with partners:
  - **Continue stability maintenance**
    - Version upgrade, bug fixing in compilation/running time.
  - **Performance enhancement**
    - Continue optimizing NVMEDEVICE module according to customers or partners' feedback.
  - **New feature Development**
    - Occasionally pickup some common requirements/feedback in community and may upstream those features in NVMEDEVICE module

Agenda

- SPDK Introduction
- Accelerate block service built on Ceph
- SPDK support in Ceph bluestore
- Summary

Summary

- SPDK proves to be useful to explore the capability of fast storage devices (e.g., NVMe SSDs) in many scenarios.
- However it still needs extra development efforts to make SPDK useful for Bluestore in Ceph.
- Call for actions:
  - Call for participation in SPDK community
  - Welcome to leverage SPDK for Ceph optimization, and contact SPDK dev team for help and collaboration.

SDC

STORAGE DEVELOPER CONFERENCE

SNIA SANTA CLARA, 2017

Q & A