SNIA NSF | NETWORKING STORAGE

Where Does SPDK Fit in the NVMe-oF™ Landscape?

Live Webcast
January 9, 2020
10:00 am PT

Today's Presenters

Moderator:
Tim Lustig
Mellanox

Presenter:
Jim Harris
Intel

Presenter:
Ben Walker
Intel

SNIA-At-A-Glance

SNIA NSF | NETWORKING STORAGE

SNIA-at-a-Glance

185
industry leading organizations

2,000
active contributing members

50,000
IT end users & storage pros worldwide

Learn more: **snia.org/technical** **@SNIA**

Technologies We Cover

- Ethernet
- iSCSI
- NVMe-oF
- InfiniBand
- Fibre Channel, FCoE
- Hyperconverged (HCI)
- Storage protocols (block, file, object)
- Virtualized storage
- Software-defined storage

SNIA NSF | NETWORKING STORAGE

SNIA Legal Notice

- The material contained in this presentation is copyrighted by the SNIA unless otherwise noted.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
  - Any slide or slides used must be reproduced in their entirety without modification
  - The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

Agenda

- Overview of the SPDK Project
- Key NVMe-oF Use Cases with SPDK
- SPDK NVMe-oF Architecture and Design
- Performance Data
- Q&A

SNIA NSF | NETWORKING STORAGE

Overview of the SPDK Project

What is SPDK?

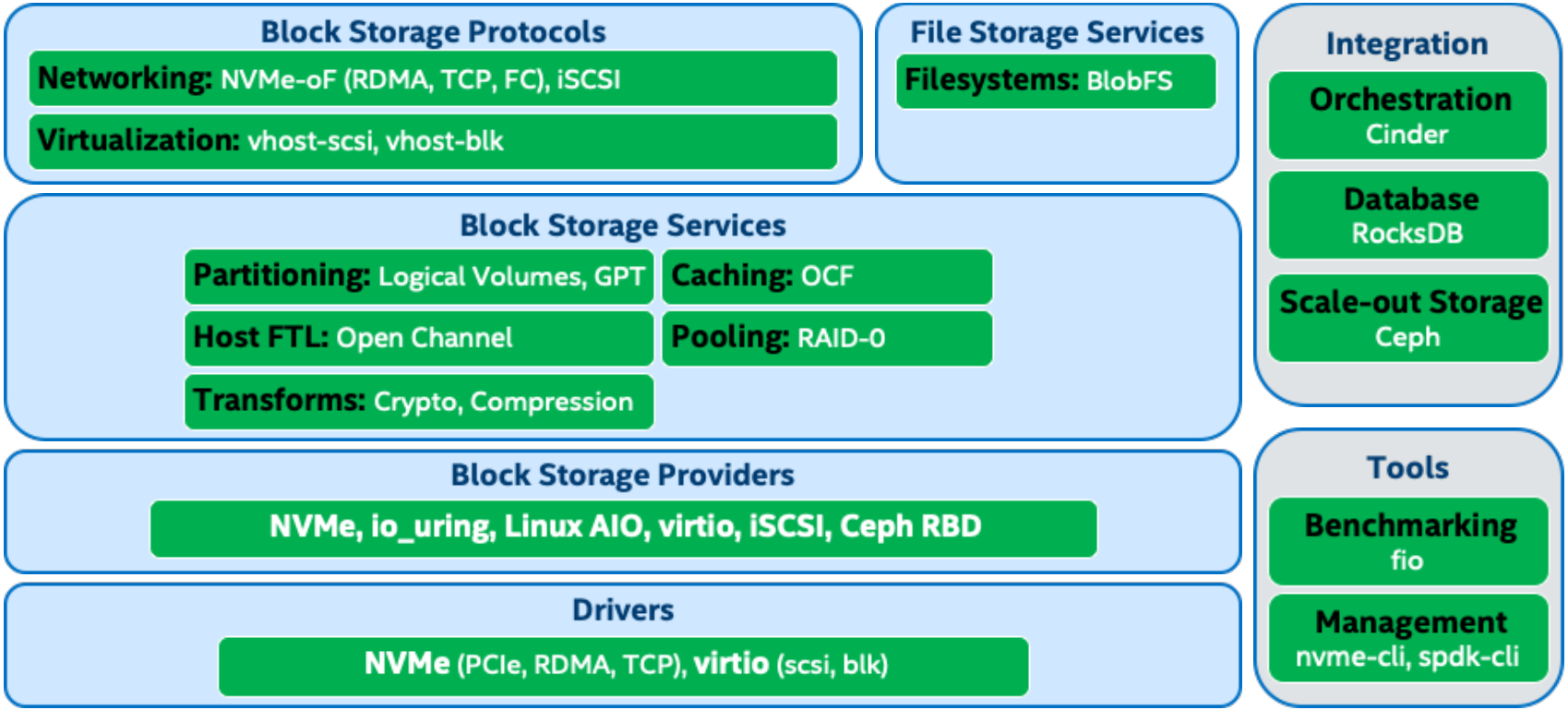
SNIA NSF | NETWORKING STORAGE

- Storage Performance Development Kit
  - Open source project
  - Framework for highly performant and efficient storage software
  - Userspace and polled mode programming model
  - Special focus on NVM Express (and NVMe over Fabrics!)
  - Includes storage networking and storage virtualization
  - Discrete libraries and fully-functional applications

What is SPDK?

- Project History
  - 2013: SPDK starts as an internal project at Intel
  - 2015: NVMe driver released on GitHub
  - 2016: First contributor outside of Intel
  - 2017: First core maintainer outside of Intel
  - 2018: NVMe/TCP support released in-step with specification
  - 2019: 700+ patches from 50+ contributors outside of Intel

Architecture Diagram



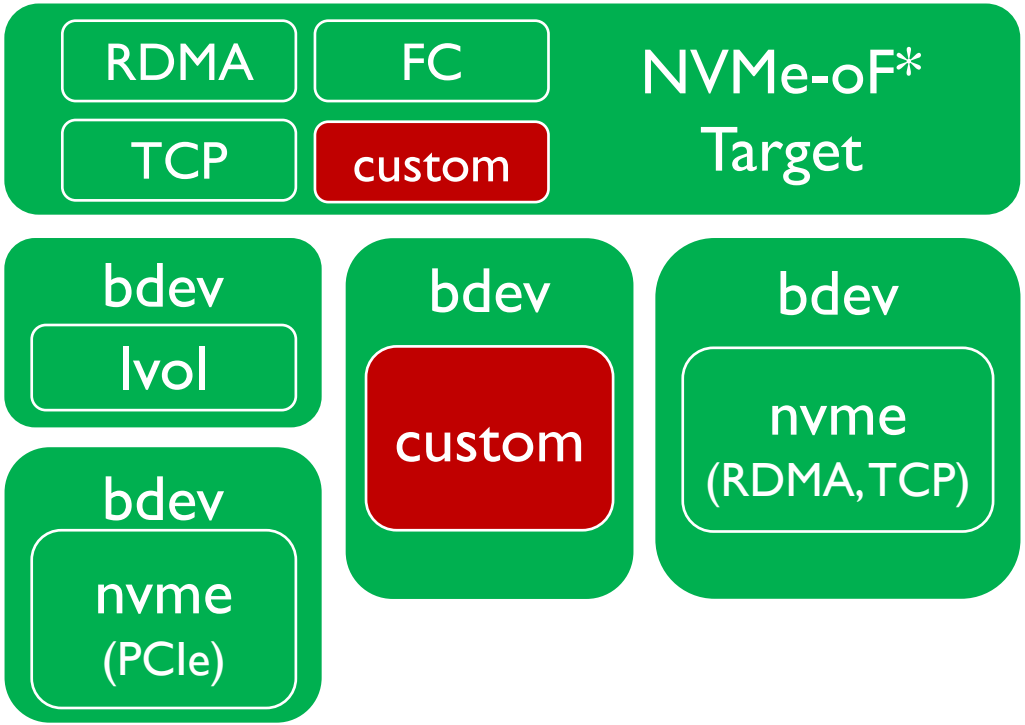
SNIA NSF | NETWORKING STORAGE

Key NVMe-oF Use Cases with SPDK

Determining Factors

- Performance and efficiency requirements
  - SPDK capable of up to millions of IOPs per CPU core
- Integration with existing software
  - SPDK provides well-defined APIs for integrating custom modules
- Customization
  - SPDK enables use of optional or vendor-specific NVMe features with little or no performance impact
- Licensing
  - SPDK is BSD licensed

Use Cases



- NVMe-oF target
- Basic block services
- Custom block services
  - Including integrating existing block storage stacks
- Custom transports
- Polled mode access to remote storage

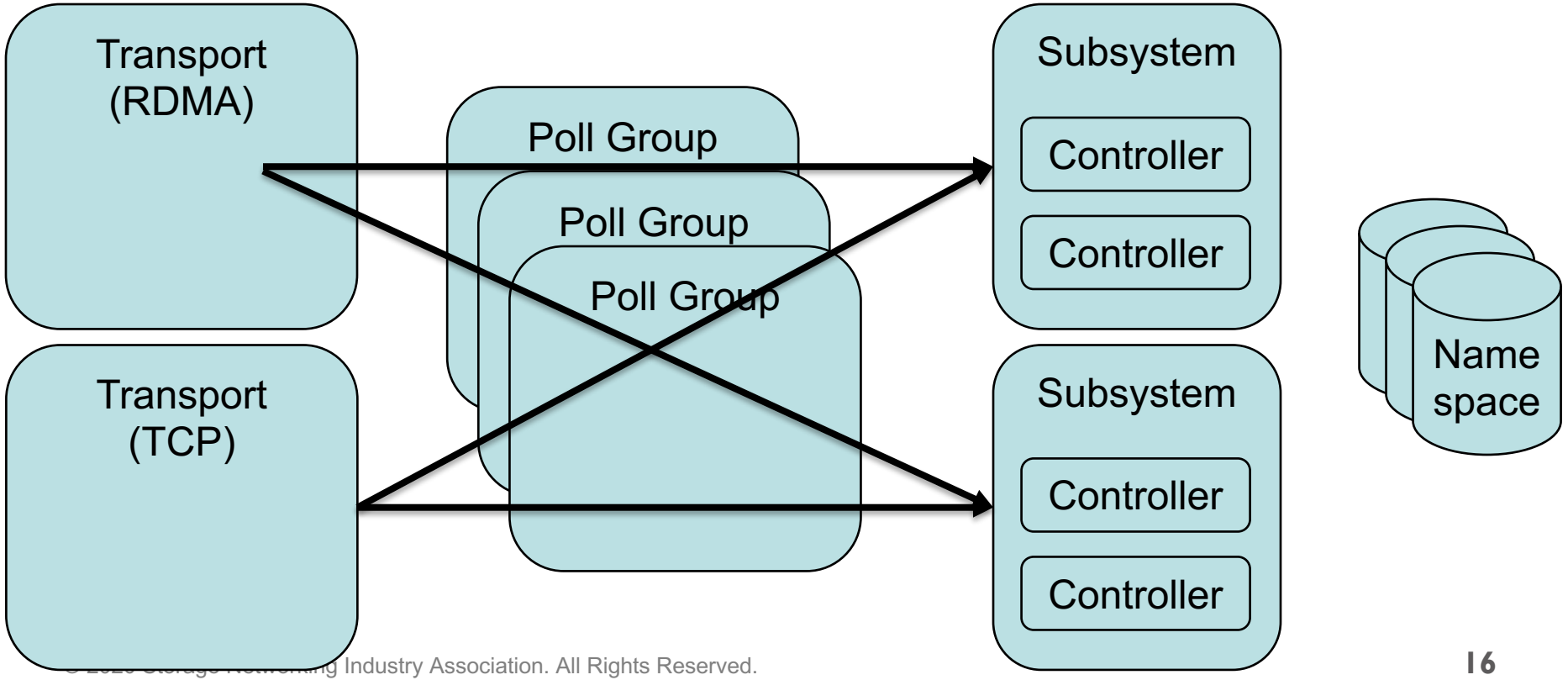
Where is SPDK not suited?

- Reduced performance requirements
  - Kernel-based interrupt-driven storage software is typically sufficient for lower IOPs workloads
- Integration with legacy software
  - SPDK APIs designed for asynchronous operation with relatively fixed number of threads
- Support requirements
  - Kernel-based solutions may provide paid support options that are not available with SPDK
- General purpose filesystem requirements

SNIA NSF | NETWORKING STORAGE

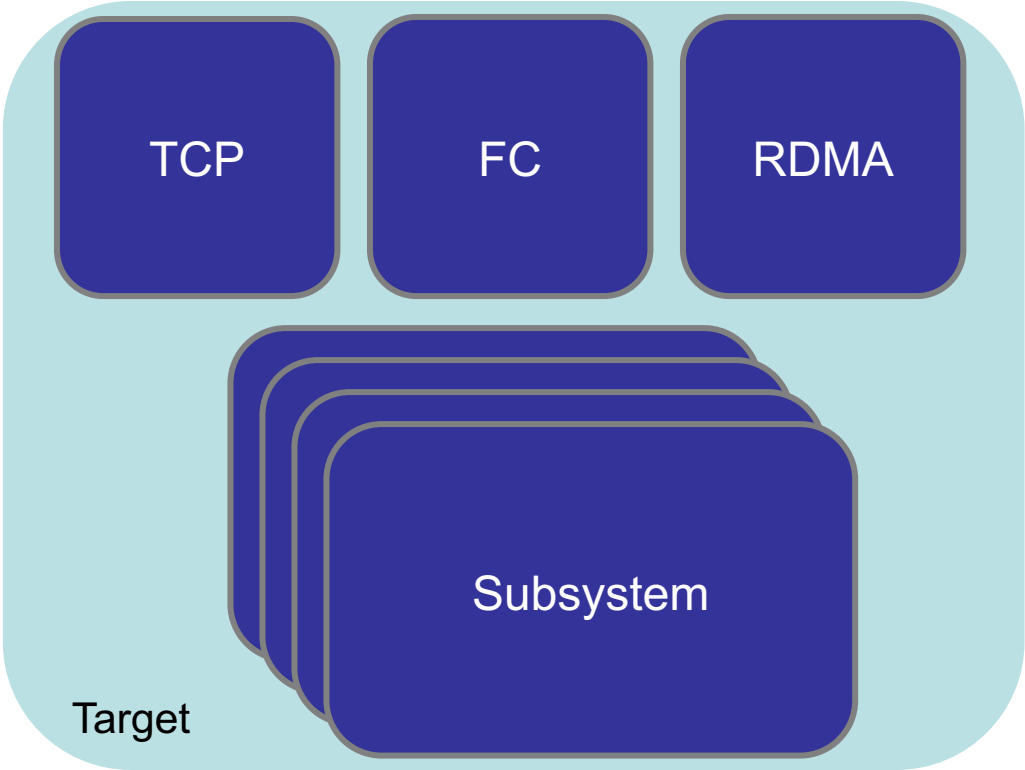
SPDK NVMe-oF Architecture and Design

NVMe-oF Target Architecture



NVMe-oF Primitives

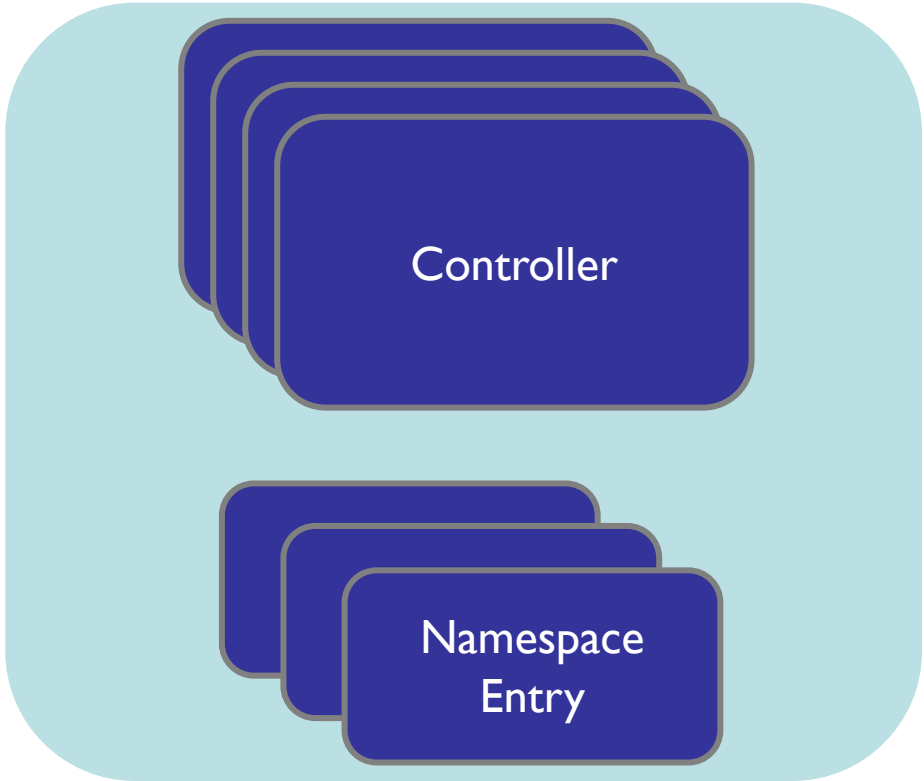
- spdk_nvmf_tgt
  - spdk_nvmf_subsystem
  - spdk_nvmf_transport



NVMe-oF Subsystems

- Subsystems are **global**
- Controller – Network session
- Namespace – Set of logical blocks

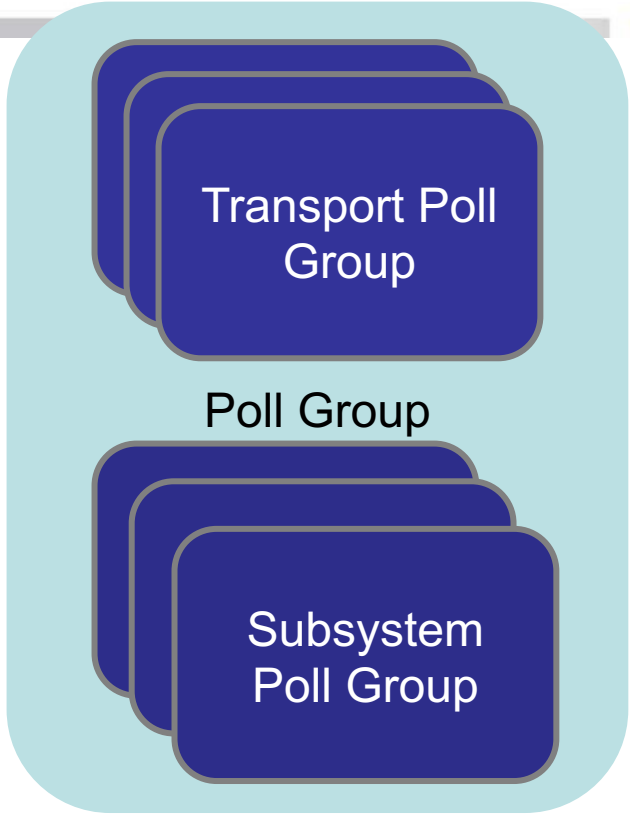
- Subsystems are **access control lists**



NVMe-oF Primitives

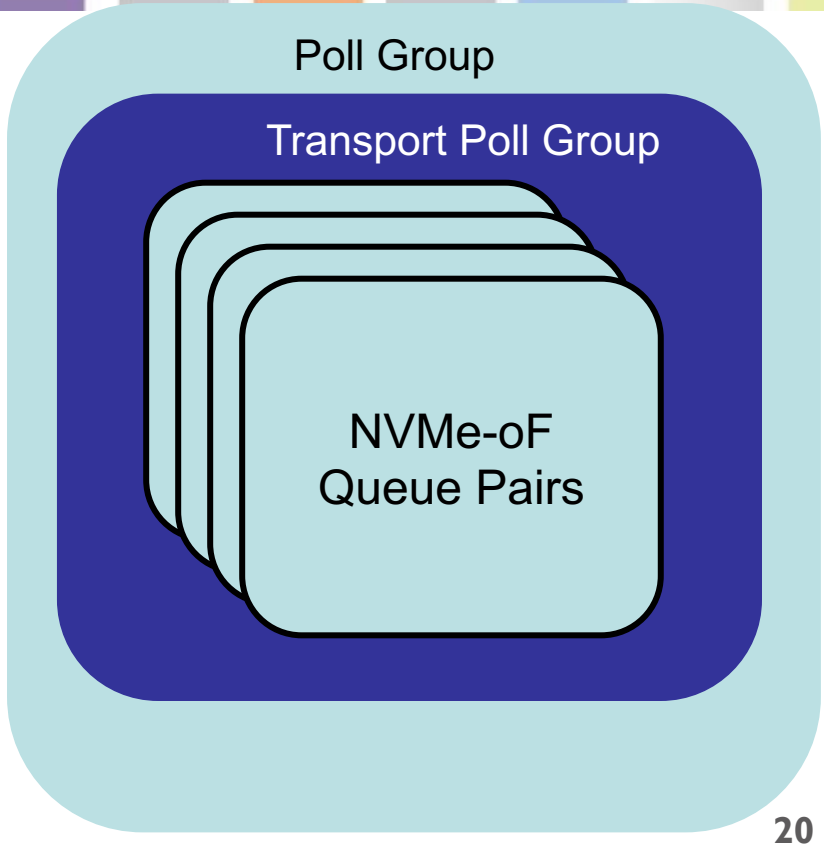
- spdk_nvmf_poll_group
  - spdk_nvmf_subsystem_poll_group
  - spdk_nvmf_transport_poll_group

Per-thread Scope



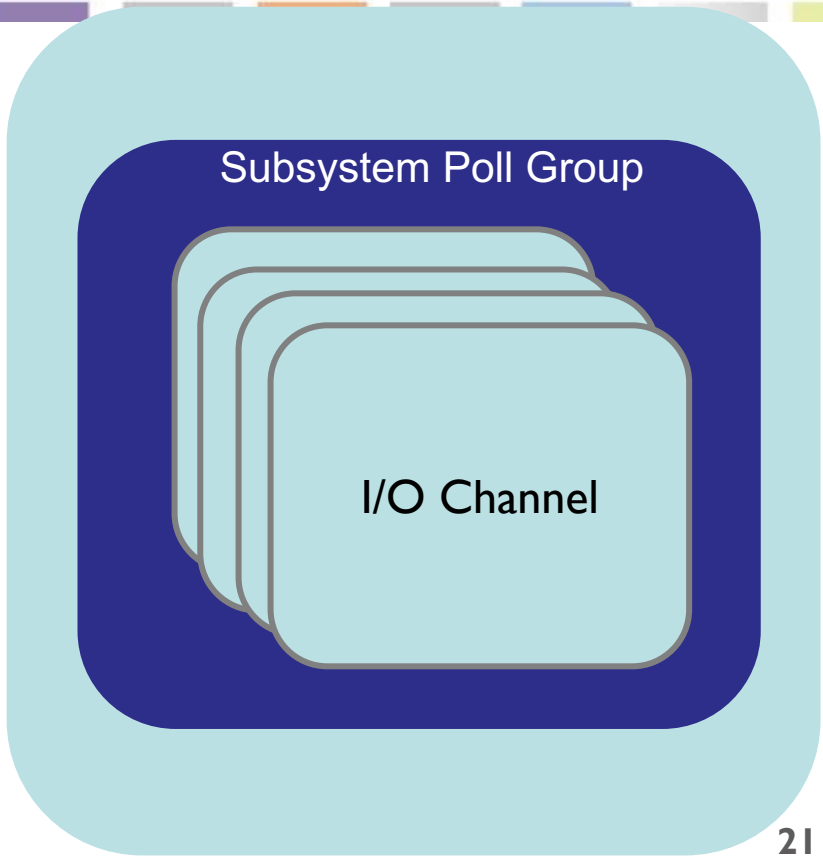
NVMe-oF Transport Poll Groups

- Per-thread collection of transport data
- TCP: NVMe-oF queue pair is a socket
- Uses a transport-specific mechanism to efficiently poll the group
  - TCP: epoll/kqueue
- The queue pairs are not necessarily for the same controller/subsystem/host

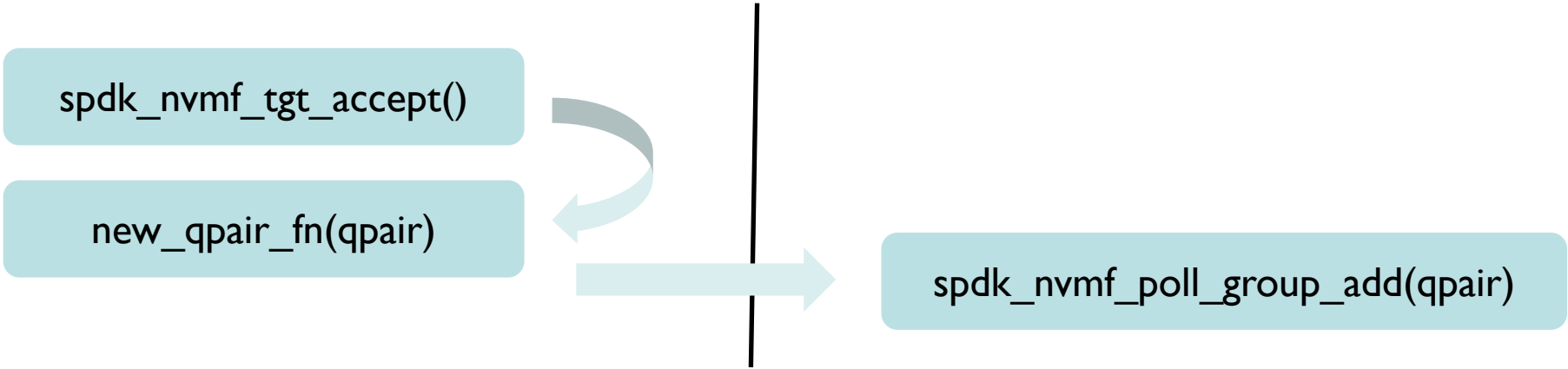


NVMe-oF Subsystem Poll Groups

- Per-thread collection of subsystem data
- Contains thread-unique I/O channels for each namespace in the subsystem.
- Think of an I/O channel as an NVMe queue pair for the local device.

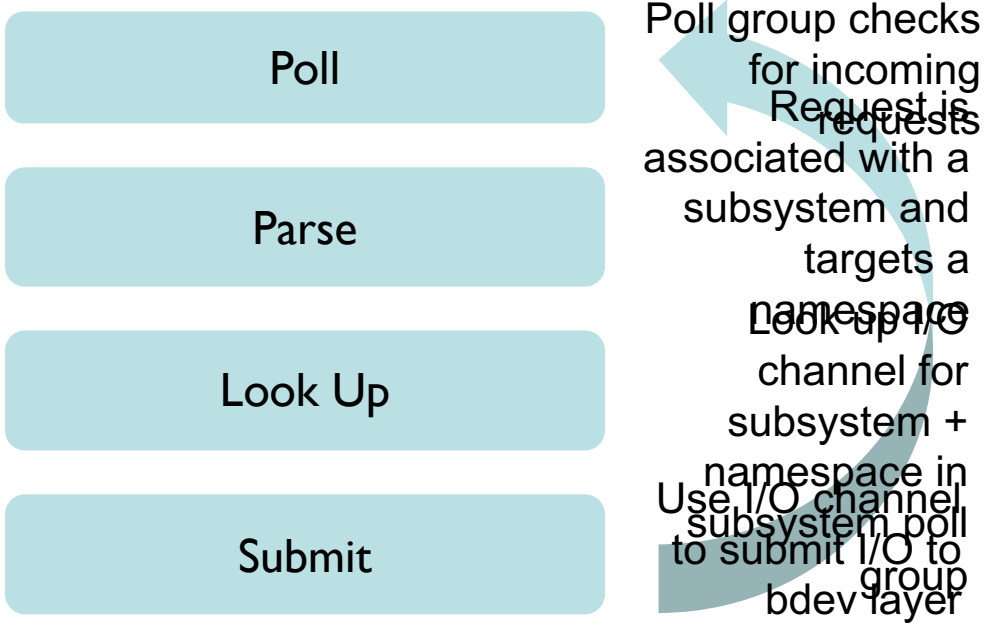


Accepting a New Connection



Performing an I/O

- No Locks!
- Touches only thread-local data (cache friendly)!
  - Lookups are all array math!



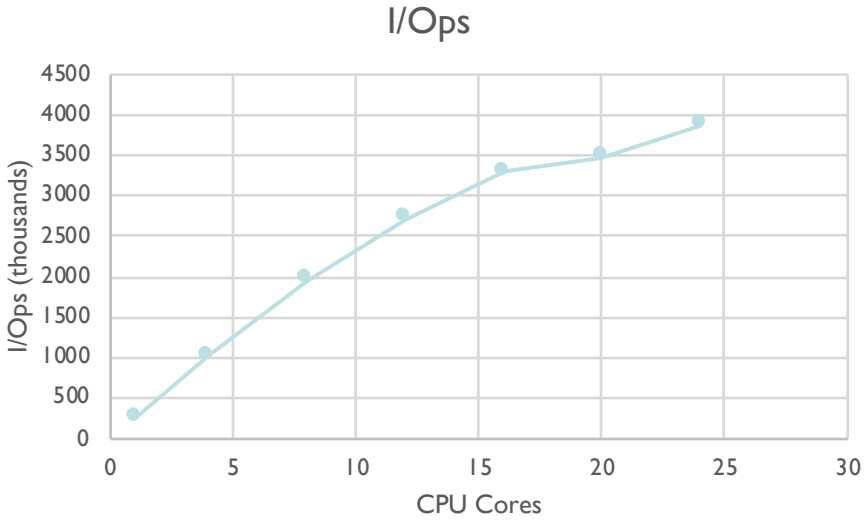
NVMe-oF Host Architecture

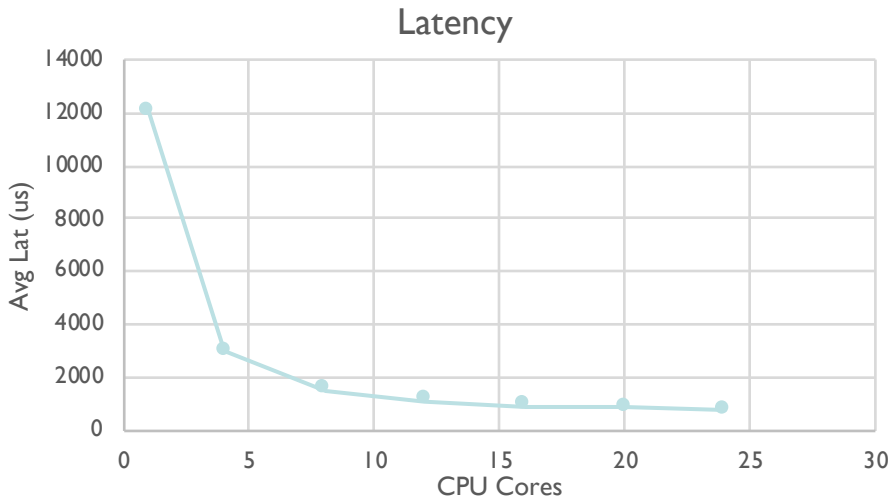
- Same library/API as local PCIe NVMe driver
- Pluggable Transports
- No poll groups
  - Doing a spdk_nvme_connect() creates an spdk_nvme_ctrlr (network session) which includes the admin qpair.
  - I/O qpairs are polled directly

Performance Data

NVMe-oF Performance: TCP

Random read/write 70/30 @ 4K QD=64

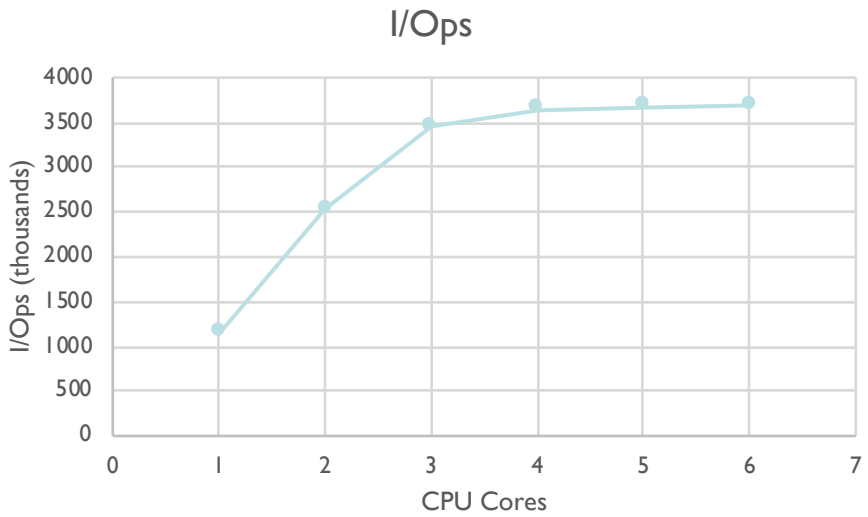


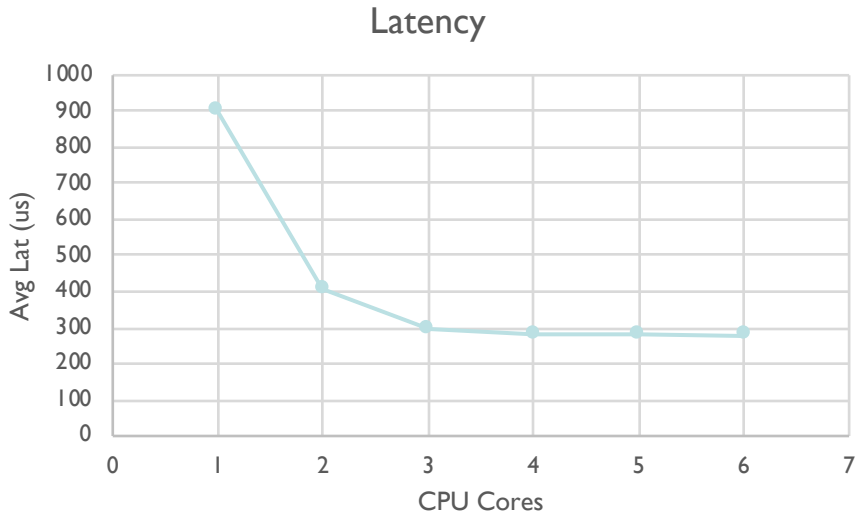


https://dqtibwqq6s6ux.cloudfront.net/download/performance-reports/SPDK_vhost_perf_report_1910.pdf

NVMe-oF Performance: RDMA

Random read/write 70/30 @ 4K QD=64





https://dqtibwqq6s6ux.cloudfront.net/download/performance-reports/SPDK_19.04_NVMeOF_RDMA_benchmark_report.pdf

Q&A

After This Webcast

- Please rate this webcast and provide us with feedback
- This webcast and a PDF of the slides will be posted to the SNIA Networking Storage Forum (NSF) website and available on-demand at www.snia.org/forums/nsf/knowledge/webcasts
- A full Q&A from this webcast, including answers to questions we couldn't get to today, will be posted to the SNIA-NSF blog: sniansfblog.org
- Follow us on Twitter @SNIANSF

More Resources

- Let's Talk Fabrics – NVMe over Fabrics
  SNIAVideo YouTube: https://youtu.be/HfcZwkPzj4w
- What NVMe/TCP Means for Networked Storage
  On-Demand Webcast: https://www.brighttalk.com/webcast/663/344698
- Under the Hood with NVMe over Fabrics
  On-Demand Webcast: https://www.brighttalk.com/webcast/663/175515
- What's New in NVM Express:
  SNIAVideo YouTube: https://youtu.be/m8nq2BzawNk
- The Performance Impact of NVMe and NVMe over Fabrics
  On-Demand Webcast: https://www.brighttalk.com/webcast/663/132761
- SPDK Performance Reports
  - https://spdk.io/doc
- Links to SPDK Summit presentations not covered today
  - https://spdk.io/blog

Thank You