SDC 18

September 24-27, 2018
Santa Clara, CA

www.storagedeveloper.org

SPDK NVMe
In-depth Look at its Architecture and Design

Jim Harris

Intel

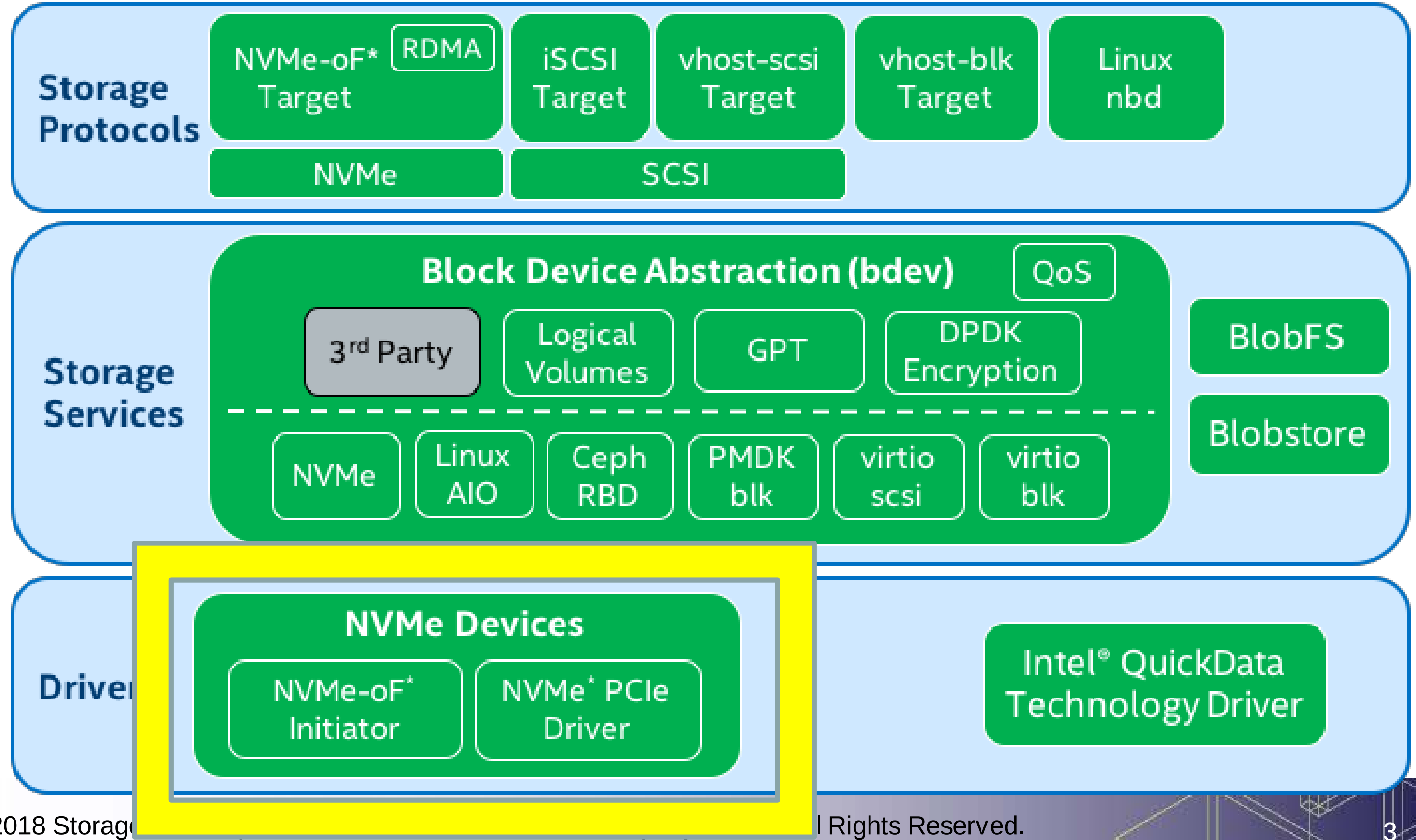
What is SPDK?

- Storage Performance Development Kit
- Open Source, BSD Licensed
- http://spdk.io
- User-space Drivers and Libraries for Storage, Storage Networking and Storage Virtualization
  - This talk focused specifically on the SPDK NVMe userspace driver

2018 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

SPDK Block Diagram

Focus for today's talk



2018 Storage ... l Rights Reserved.

SPDK and Kernel

- Better performance and efficiency compared to traditional interrupt-driven approaches
- BUT...
- SPDK is not a general-purpose solution
  - covers some use cases very well – others not at all (or at least not well)
- Polled mode design and userspace implementation drove much of the SPDK design

2018 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

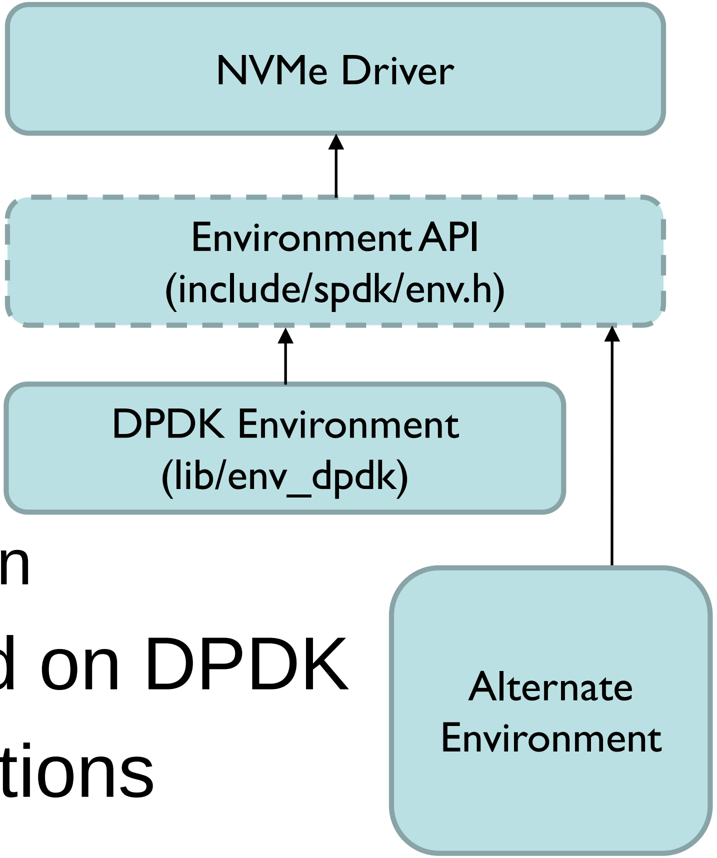
SPDK Paradigm

- SSD dedicated to single process
  - NVMe SR-IOV can alleviate this restriction
- Smallish, known number of threads
- Pre-allocated pinned memory

2018 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

Environment Abstraction

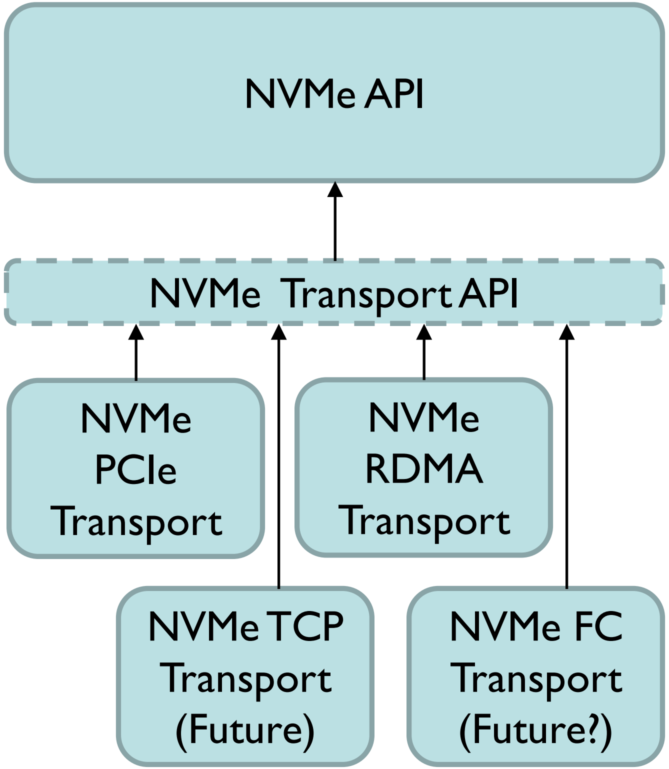
- Provides services for:
  - enumerating PCIe devices
  - polled mode thread creation
  - allocating/pinning memory
    - including VA to IOVA translation
- Default implementation is based on DPDK
- Enables alternative implementations



2018 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

Transport Abstraction

- Enables separate implementations for different transports
  - construct/destruct controller
  - set/get register value
  - create/delete I/O queue pair
  - submit request
  - process completions



2018 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

NVMe Probe and Attach

```
int spdk_nvme_probe(const struct spdk_nvme_transport_id *trid,
                    void *cb_ctx,
                    spdk_nvme_probe_cb probe_cb,
                    spdk_nvme_attach_cb attach_cb,
                    spdk_nvme_remove_cb remove_cb);
```

- Controller identification by transport ID
  - Transport ID used to identify local (PCIe) and remote controllers
  - NULL means “all PCIe”

2018 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

NVMe Probe and Attach (continued)

- probe_fn
  - decide whether to attach
    - based on PCIe ID when trid == NULL
    - based on NQN when trid = discovery controller identification by transport ID
  - specify options (struct spdk_nvme_ctrlr_opts)
    - includes: number of I/O queues, use CMB (if available), arbitration mechanism, KATO, host ID

2018 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

NVMe Probe and Attach (continued)

- attach_fn
  - notification controller is ready for use
  - ”negotiated” controller options
- Synchronous but will attach multiple controllers in parallel (asynchronous version in next release)
- spdk_nvme_connect()
  - spdk_nvme_probe() but for specific controller

2018 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

Queue Pair Creation

- struct spdk_nvme_io_qpair_opts
  - Priority (for WRR)
  - I/O queue size, # I/O requests (discussed later)
- Submit admin commands
  - CREATE_CQ, CREATE_SQ
- Allocate memory for requests, trackers
- Synchronous API

2018 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

Command Submission

```
int spdk_nvme_ns_cmd_read(struct spdk_nvme_ns *ns, struct spdk_nvme_qpair *qpair,
                          void *payload, uint64_t lba, uint32_t lba_count,
                          spdk_nvme_cmd_cb cb_fn, void *cb_arg, uint32_t io_flags);
```

- *All* command submission APIs are async
  - cb_fn and cb_arg parameters
  - completion only by polling qpair
- Payload buffers are virtual addresses
- Not necessarily 1:1 with NVMe commands
  - i.e. I/O splitting

2018 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

Command Completions

```
int32_t spdk_nvme_qpair_process_completions(struct spdk_nvme_qpair *qpair,
                                            uint32_t max_completions);
```

- Application responsible for:
  - polling for completions
  - submit and poll on same thread
- Completion functions called by SPDK NVMe driver from spdk_nvme_qpair_process_completions() context
- Use max_completions to limit number of completions

2018 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

struct nvme_request

- Private, internal data structure
- Dedicated set of objects per qpair
  - Memory for objects allocated along with qpair
  - Count dictated by controller or qpair options
- Transport agnostic
- Can be queued
  - Depending on transport specifics

2018 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

I/O Splitting

- I/O may need to be split before sent to controller
  - Max Data Transfer Size (MDTS)
  - Namespace Optimal IO Boundary (NOIOB)
- Two possible approaches
  - 1) User must split I/O before calling SPDK API
  - 2) SPDK split I/O internally
- SPDK has implemented approach #2

2018 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

I/O Splitting in SPDK

- Simplify SPDK usage and avoid code duplication
- Ignoring NOIOB is functionally OK, but will result in unexpected performance issues
- Minimal (if any) impact on applications that require no I/O splitting
- struct nvme_request supports parent/child relationships
  - one allocated per child I/O

2018 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

Vectored I/O

```
int spdk_nvme_ns_cmd_readv(struct spdk_nvme_ns *ns, struct spdk_nvme_qpair *qpair,
                           uint64_t lba, uint32_t lba_count,
                           spdk_nvme_cmd_cb cb_fn, void *cb_arg, uint32_t io_flags,
                           spdk_nvme_req_reset_sgl_cb reset_sgl_fn,
                           spdk_nvme_req_next_sge_cb next_sge_fn);
```

- SGEs fetched by SPDK NVMe driver using callback functions
  - avoids application translating to struct iovec
  - avoids struct iovec => NVMe PRP translation

2018 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

VA-to-IOVA Translation

- SPDK maintains a userspace “page table”
  - describes the pre-allocated pinned memory
- Two-level page table
  - First level: 1GB granularity (0x0 => 256TB)
  - Second level: 2MB granularity
- Similar registration/translation scheme used for RDMA
  - Translates to MR instead of IOVA

2018 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

Timeout Callbacks

```
void spdk_nvme_ctrlr_register_timeout_callback(struct spdk_nvme_ctrlr *ctrlr,
                                               uint64_t timeout_us,
                                               spdk_nvme_timeout_cb cb_fn, void *cb_arg)
```

- Optional – can be set at controller level
- Timeouts without interrupts?
  - Expirations checked during completion polling
- Driver takes no action (i.e. abort or reset)
  - Application is notified and can take action
- Pending requests linked in submission order

2018 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

Benchmarking

- Two options
  - fio plugin
  - nvme/perf
- fio per I/O overhead significant at 3-4M IOPs/core
- nvme/perf – lower overhead but fewer features

2018 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

nvme/perf histograms

- Very granular (<1us) buckets
- Enabled via –LL option to nvme/perf

```
Latency histogram for INTEL SSDPE2KE020T7 from core 0:
==============================================================================
       Range in us     Cumulative    IO count
    4.429 -     4.460:    0.0001%  (        1)
    4.460 -     4.491:    0.0008%  (        6)
    4.491 -     4.521:    0.0018%  (        8)
    4.521 -     4.552:    0.0052%  (       28)
    4.552 -     4.582:    0.0125%  (       60)
    4.582 -     4.613:    0.0247%  (      101)
    4.613 -     4.643:    0.0560%  (      258)
    ...
```


2018 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

Error Injection

2018 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

Raw/Passthrough Commands

2018 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.

Extended LBA Formats

2018 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.