SDC 18
September 24-27, 2018
Santa Clara, CA
www.storagedeveloper.org

NVMe Over Fabrics: Scaling Up With The Storage Performance Development Kit

Ben Walker
Data Center Group
Intel Corporation

Notices and disclaimers

- Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at intel.com, or from the OEM or retailer.
- Some results have been estimated or simulated using internal Intel analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance..
- Intel processors of the same SKU may vary in frequency or power as a result of natural variability in the production process.
- 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.
- Optimization Notice: 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. Notice Revision #20110804.
- The benchmark results may need to be revised as additional testing is conducted. The results depend on the specific platform configurations and workloads utilized in the testing, and may not be applicable to any particular user's components, computer system or workloads. The results are not necessarily representative of other benchmarks and other benchmark results may show greater or lesser impact from mitigations.
- 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 www.intel.com/benchmarks.
- Results have been estimated or simulated using internal Intel analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance.
- 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 www.intel.com.
- The cost reduction scenarios described are intended to enable you to get a better understanding of how the purchase of a given Intel based product, combined with a number of situation-specific variables, might affect future costs and savings. Circumstances will vary and there may be unaccounted-for costs related to the use and deployment of a given product. Nothing in this document should be interpreted as either a promise of or contract for a given level of costs or cost reduction.
- No computer system can be absolutely secure.
-
- *Other names and brands may be claimed as the property of others.

Agenda

- Background
- Design Overview
- Benchmarking
  - Connections
  - Memory
  - CPU cores

Background

- Storage Performance Development Kit
  - BSD Licensed collection of C libraries
  - User-space drivers, storage targets
  - http://www.spdk.io

Requirements

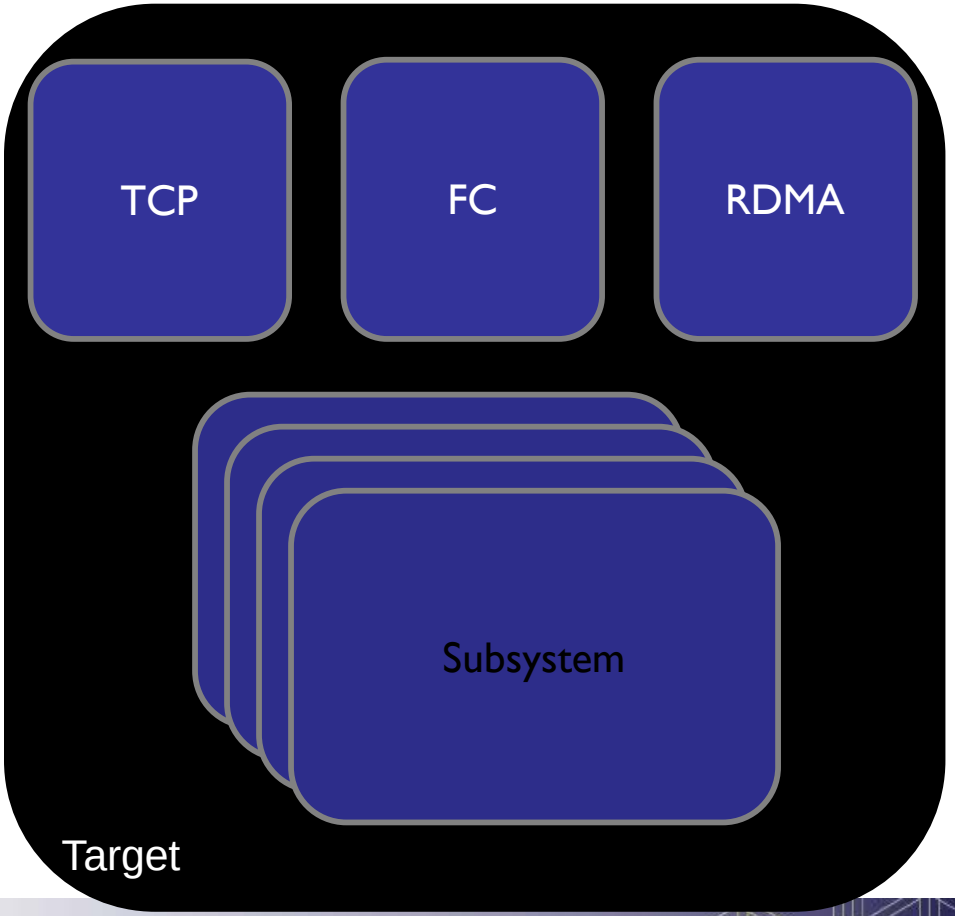
- User-space NVMe-oF Target
  - Leverage SPDK user-space drivers
- Zero copy
- Polled-mode
- Linear scaling (w/ CPU, network, storage)

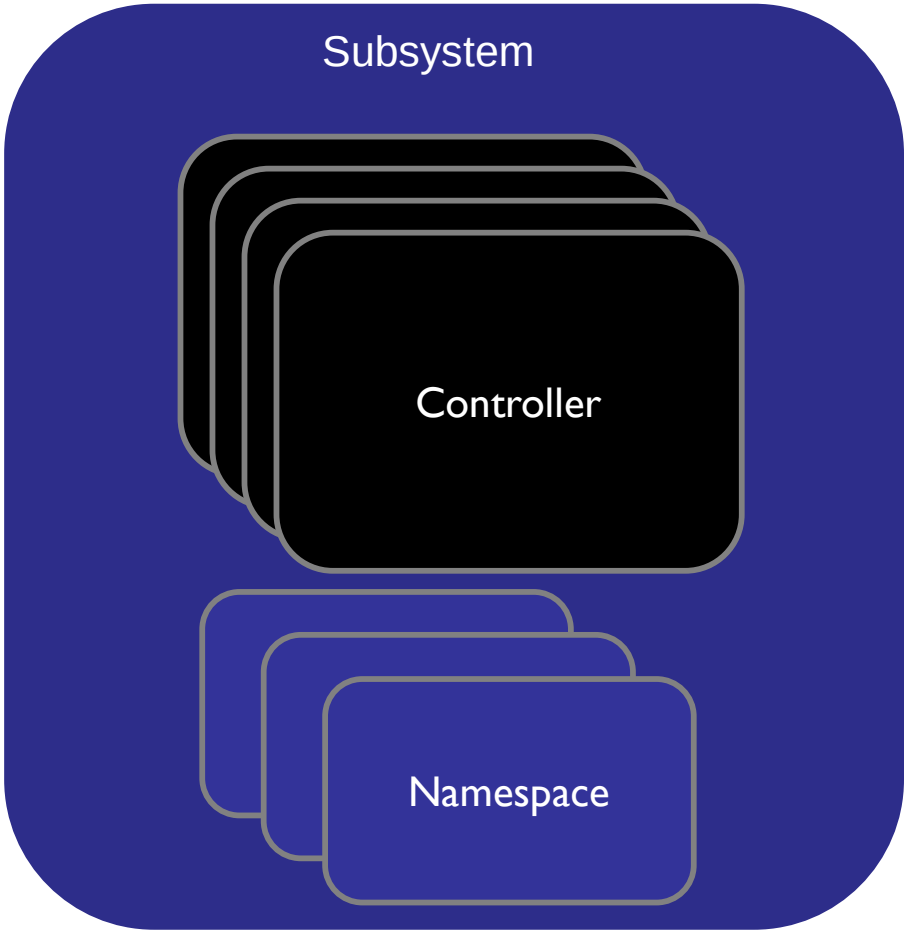
Linear Scaling

- Avoid locks and atomics
  - Ok to make management operations more time consuming to avoid interrupting I/O path
- Avoid cache contention
  - Keep each core focused on an independent job, as much as possible
- NUMA awareness

NVMe-oF Primitives

- `spdk_nvmf_tgt`
  - `spdk_nvmf_subsystem`
  - `spdk_nvmf_transport`





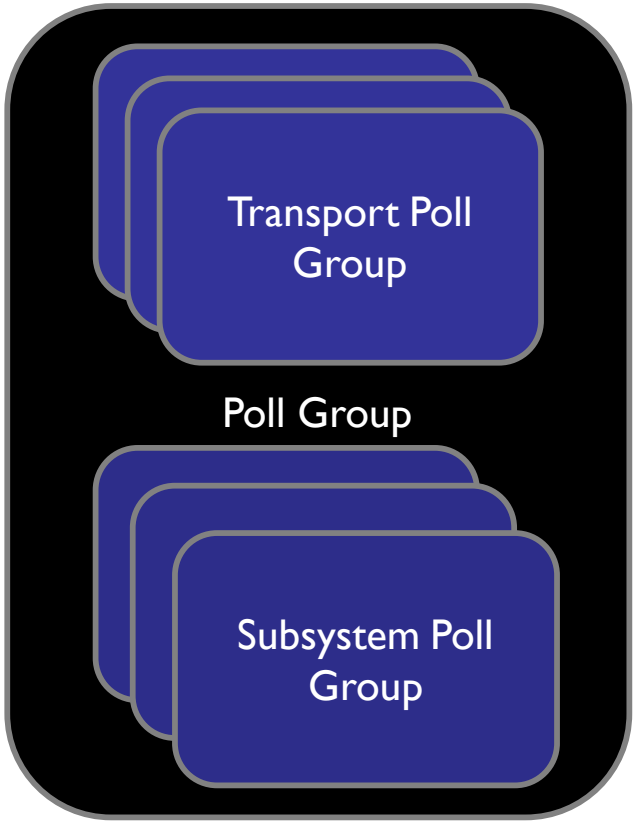
NVMe-oF Subsystems

- Subsystems are **global**
- Subsystems have states
  - `Inactive`
  - `Paused`
  - `Active`
- `spdk_nvmf_subsystem` may only be modified while not in the active state.
- Contains controllers and namespaces

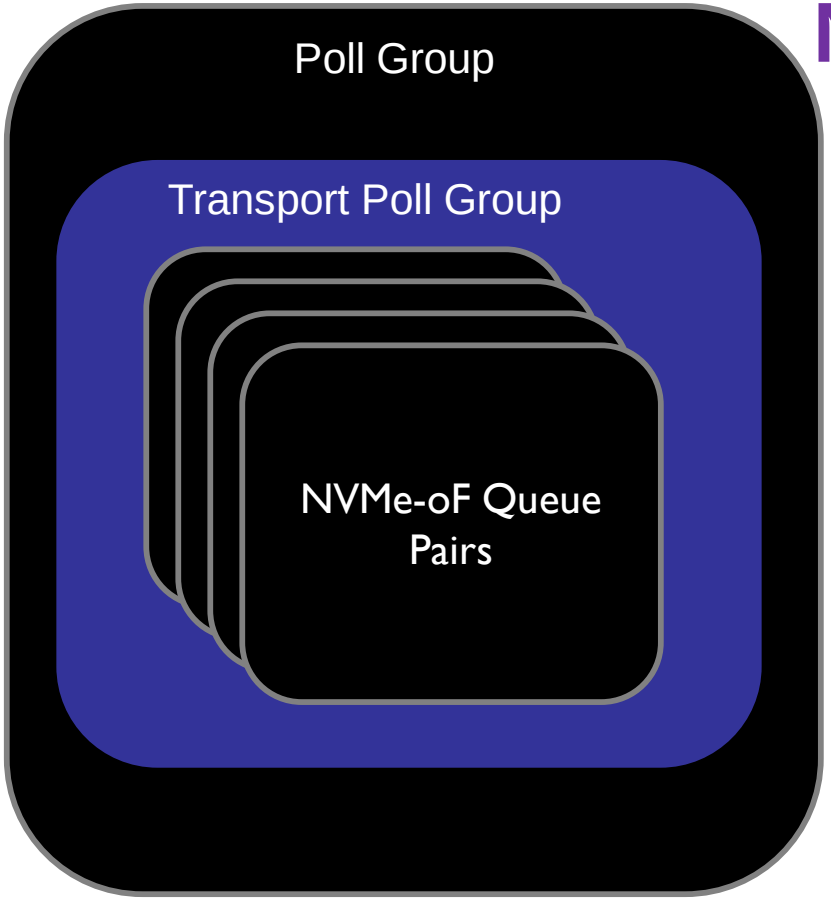
NVMe-oF Primitives

- `spdk_nvmf_poll_group`
  - `spdk_nvmf_subsystem_poll_group`
  - `spdk_nvmf_transport_poll_group`

Per-thread Scope



NVMe-oF Transport Poll Groups

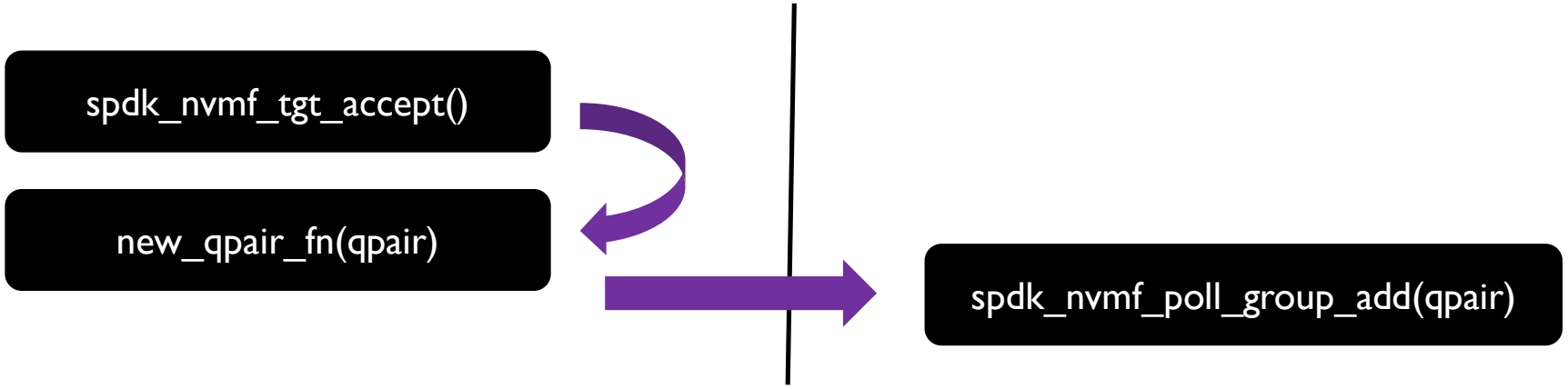


- Per-thread collection of transport data
- Uses a transport-specific mechanism to efficiently poll the group
  - RDMA: Shared completion queue
  - FC: Shared hardware queue pair
  - TCP: epoll/kqueue
- The queue pairs are not necessarily related to one another

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.

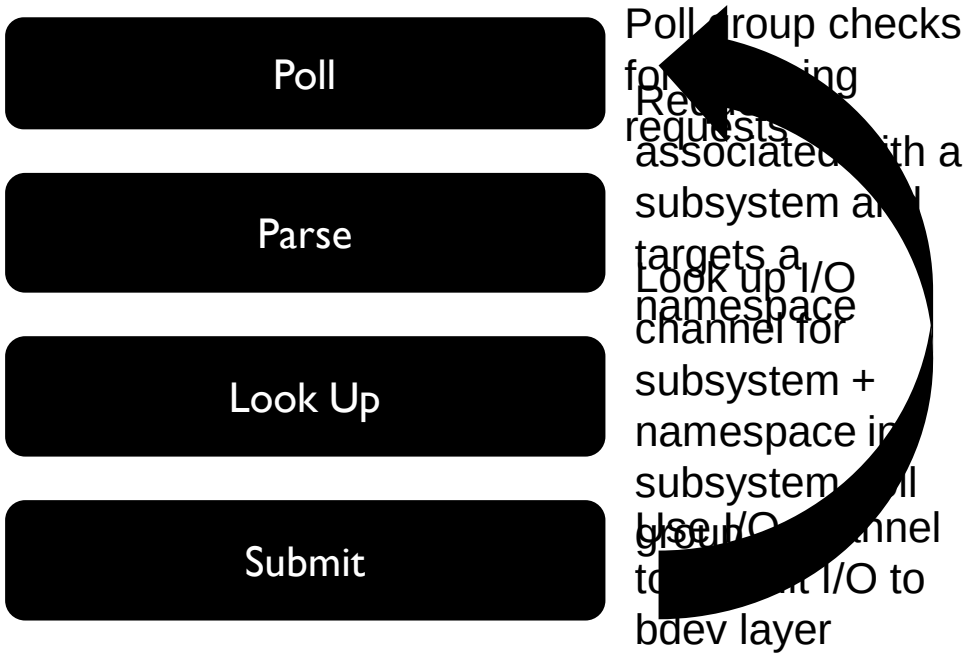
Accepting a New Connection



When does a queue pair identify which subsystem it belongs to?

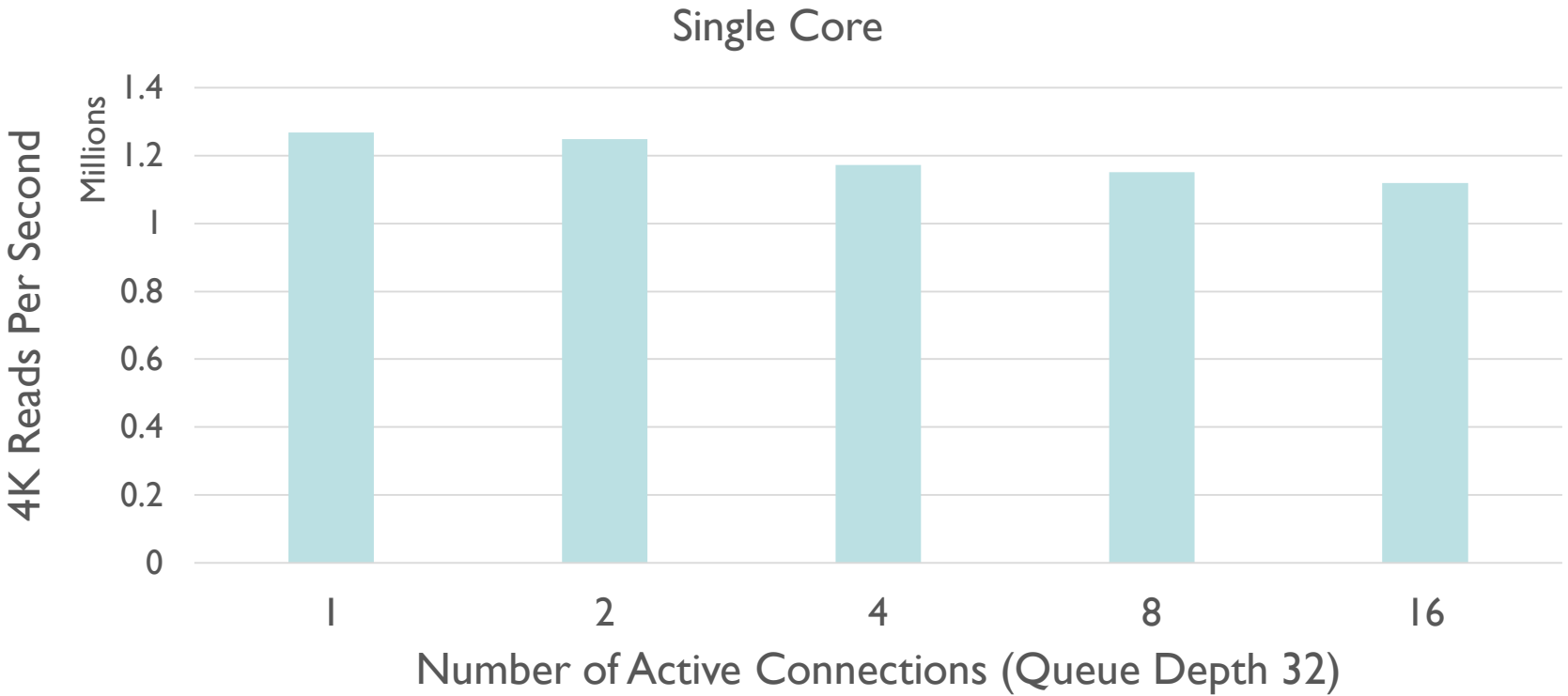
Performing an I/O

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



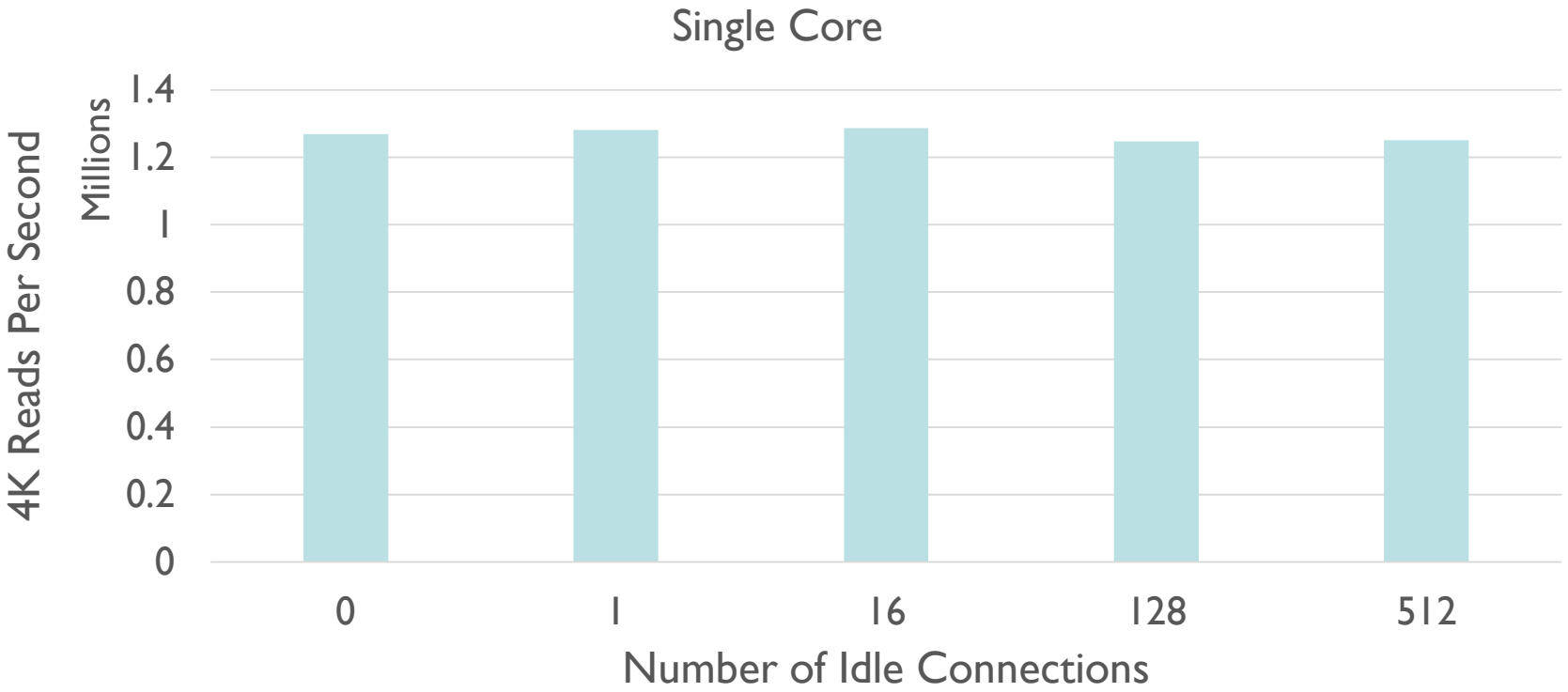
Benchmarks

Scaling: Active Connections



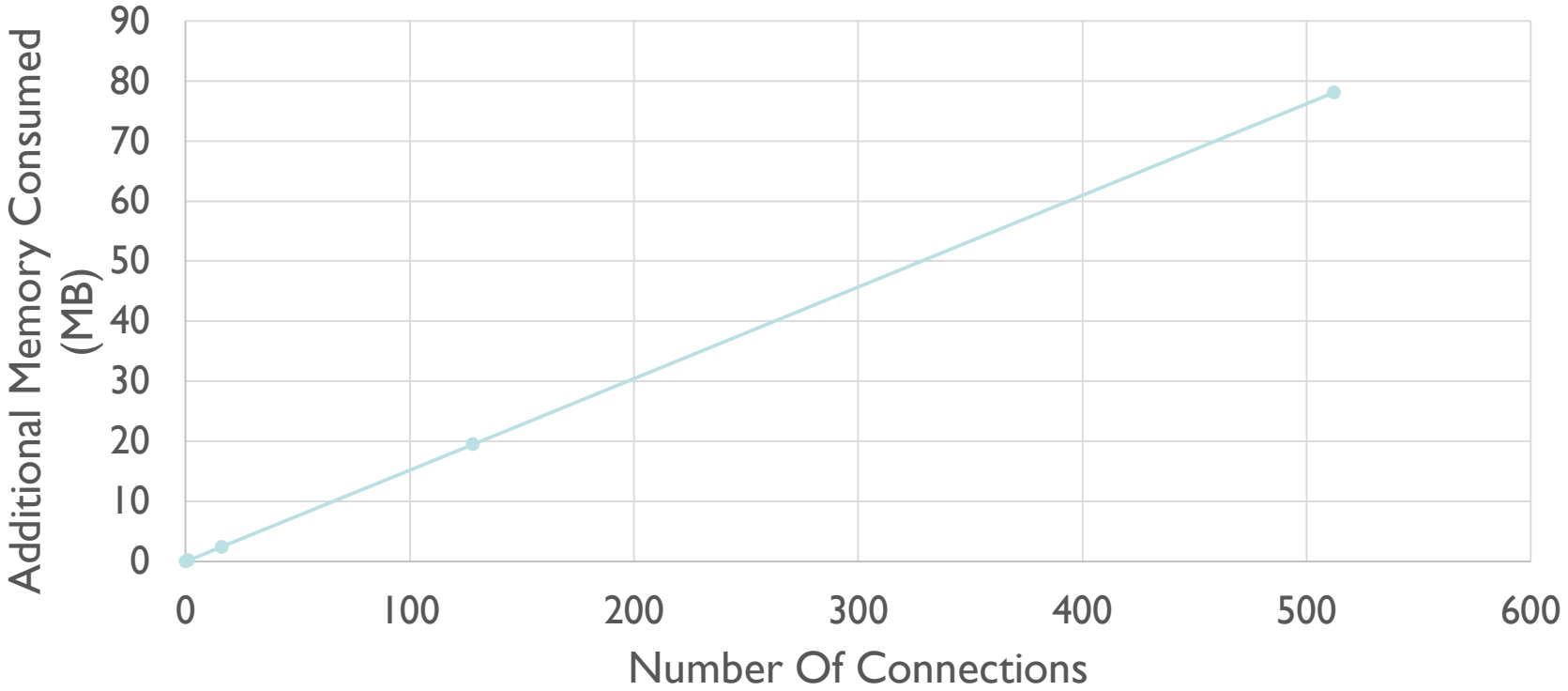
System Configuration: 2x Intel® Xeon® Platinum 8180 CPU @ 2.50 GHz, Intel® Speed Step enabled, Intel® Turbo Boost Technology enabled, 4x 2GB DDR4 2666 MT/s, 1 DIMM per channel, Ubuntu* Linux 17.10, Linux kernel 4.13.0, SPDK 18.04, DPDK 18.01, Mellanox® ConnectX-4 MT27700

Scaling: Idle Connections



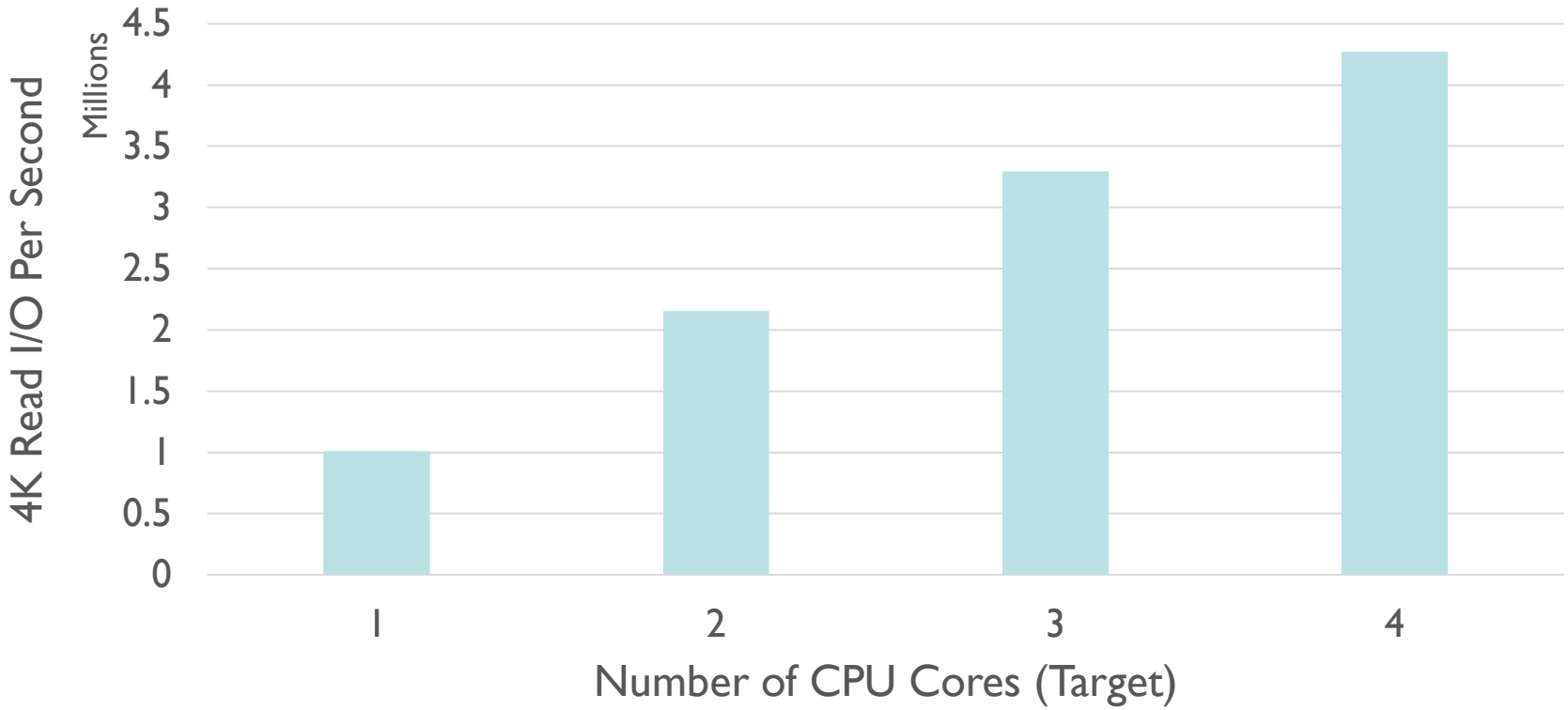
System Configuration: 2x Intel® Xeon® Platinum 8180 CPU @ 2.50 GHz, Intel® Speed Step enabled, Intel® Turbo Boost Technology enabled, 4x 2GB DDR4 2666 MT/s, 1 DIMM per channel, Ubuntu* Linux 17.10, Linux kernel 4.13.0, SPDK 18.04, DPDK 18.01, Mellanox® ConnectX-4 MT27700

Memory Usage vs Number of Connections



System Configuration: 2x Intel® Xeon® Platinum 8180 CPU @ 2.50 GHz, Intel® Speed Step enabled, Intel® Turbo Boost Technology enabled, 4x 2GB DDR4 2666 MT/s, 1 DIMM per channel, Ubuntu* Linux 17.10, Linux kernel 4.13.0, SPDK 18.04, DPDK 18.01, Mellanox® ConnectX-4 MT27700

Performance vs Number of CPU cores



System Configuration: 2x Intel® Xeon® Platinum 8180 CPU @ 2.50 GHz, Intel® Speed Step enabled, Intel® Turbo Boost Technology enabled, 4x 2GB DDR4 2666 MT/s, 1 DIMM per channel, Ubuntu* Linux 17.10, Linux kernel 4.13.0, SPDK 18.04, DPDK 18.01, Mellanox® ConnectX-4 MT27700