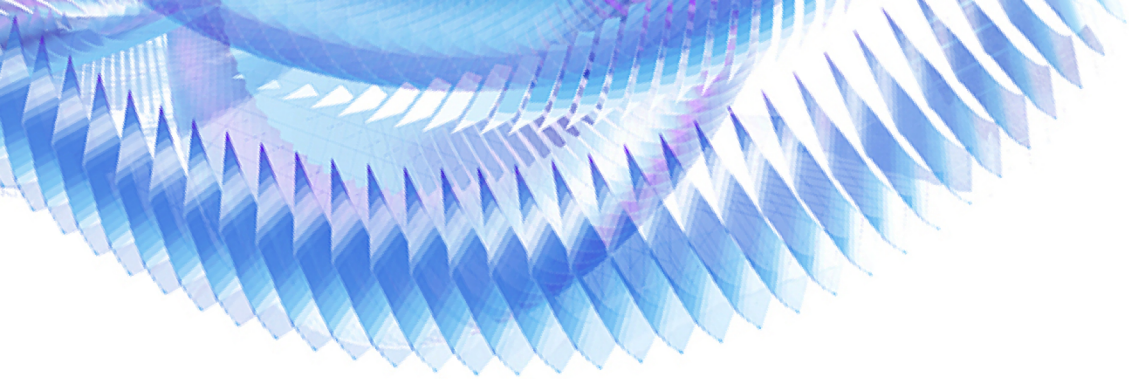


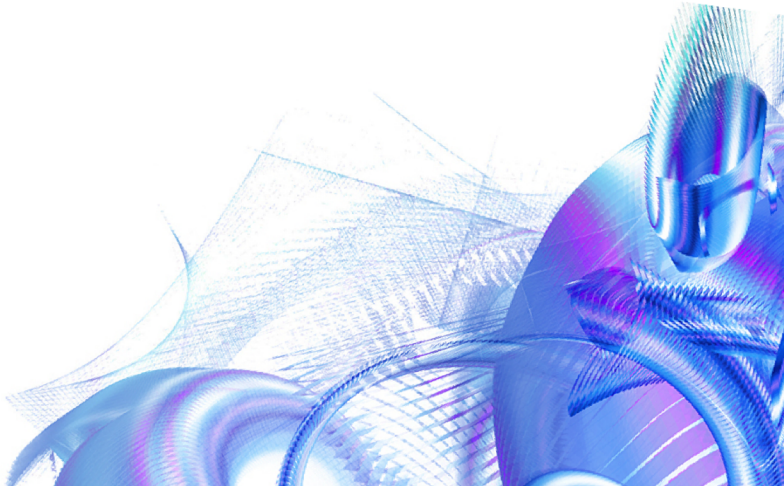
SC19

Denver, CO | hpc is now.

DAOS & IO-500

Mohamad Chaarawi, Intel

November, 2019



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 product or component can be absolutely secure.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. For more complete information about performance and benchmark results, visit http://www.intel.com/benchmarks .

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/benchmarks .

Intel Advanced Vector Extensions (Intel AVX) provides higher throughput to certain processor operations. Due to varying processor power characteristics, utilizing AVX instructions may cause a) some parts to operate at less than the rated frequency and b) some parts with Intel® Turbo Boost Technology 2.0 to not achieve any or maximum turbo frequencies. Performance varies depending on hardware, software, and system configuration and you can learn more at http://www.intel.com/go/turbo.

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.

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.

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.

IO-500

The creation of a suite of I/O benchmarks to compare facilities and storage systems. Sub goals of the benchmark are:

- Capture user-experienced performance
- Reported performance is representative for:
  - applications with well optimized I/O patterns
  - applications with random-like workloads
  - workloads involving metadata small/objects
- 2 Lists:
  - 10 node challenge: there must be exactly 10 physical compute nodes and at least one benchmark process must run on each
  - Open list: any number of compute nodes
- https://www.vi4io.org/std/io500/start

IO-500 Benchmarks

IOR

- Easy: any IOR pattern to show best-case performance without any explicit caching
- Hard: single shared file with transfer 47008 bytes!
- Separate Write and Read/verify runs.

mdtest

- Easy: private directory per process with empty files
- Hard: shared directory with 3901-byte files
- Separate write, read, stat, and delete runs

Find

- scan namespace created with IOR and mdtest

IO-500 DAOS Testbed

Remote cluster in Rio Rancho (NM)

- 10x & 26x compute nodes
  - 31 ranks per node for 10 node challenge
  - 28 ranks per node for open challenge
- 24x storage nodes
- Dual-rail Omni-Path® fabric

Compute node (CN) specifications

- 2x BDW CPU
  - Xeon® E5-2699 v4 @2.2GHz
  - 22 cores per CPU
- 2x Intel® Omni-Path® 100 adaptors

Storage node (SN) specifications

- 2x CLX CPU
  - Xeon® Platinum 8260L @ 2.4GHz
  - 24 cores per CPU
- 12x Optane® DC Persistent Memory DIMMs
  - 500GB each for a total of 3TB
  - Configured in app-direct/interleaved mode
- 2x Intel® Omni-Path® 100 adaptors

Steps to Reproduce

- Install DAOS (and dependencies)
- Install IOR & mdtest enabling DAOS backends:
  - https://github.com/hpc/ior (check README_DAOS)
- Install mpifileutils:
  - https://github.com/mchaarawi/mpifileutils
  - Fork of hpc/mpifileutils to add a dfs backend for find tool
- IO-500 script modifications:
  - Set DFS backend (with pool/cont etc.) for ior and mdtest runs
  - Set ior/mdtest parameters to increase run time to 5 mins
  - Update find command to use the dfind from mpifileutils and parameters for DFS backend
  - See io-500 script from DAOS submission.

DAOS & IO-500 – Main List Scores

Cluster Size

384 nodes?

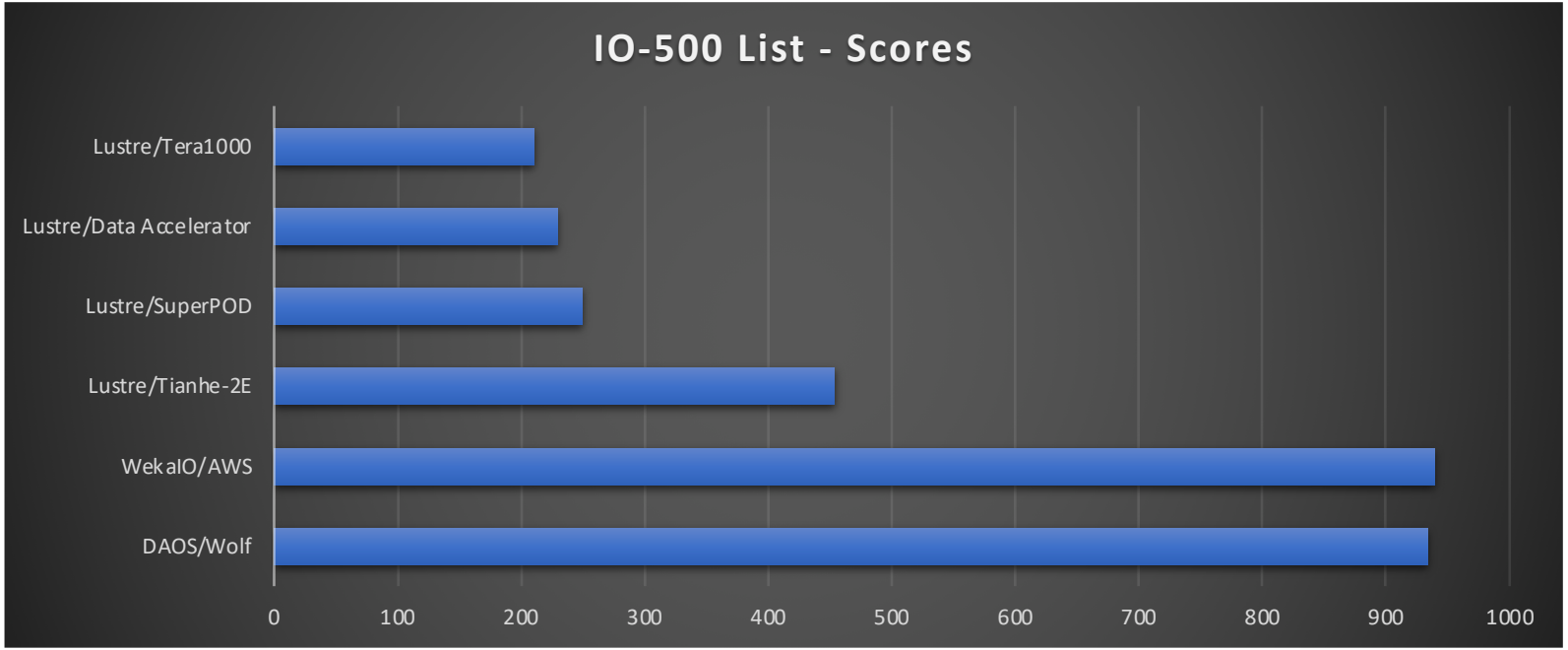
152 nodes

30 nodes

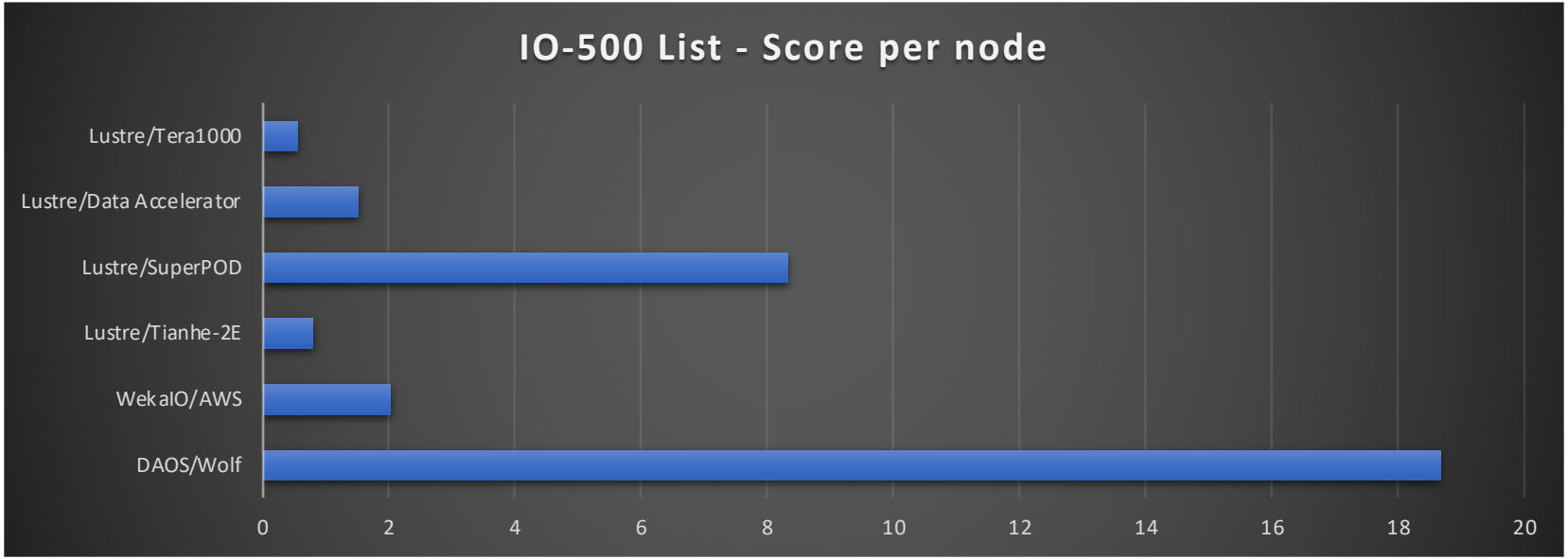
572 nodes

465 nodes

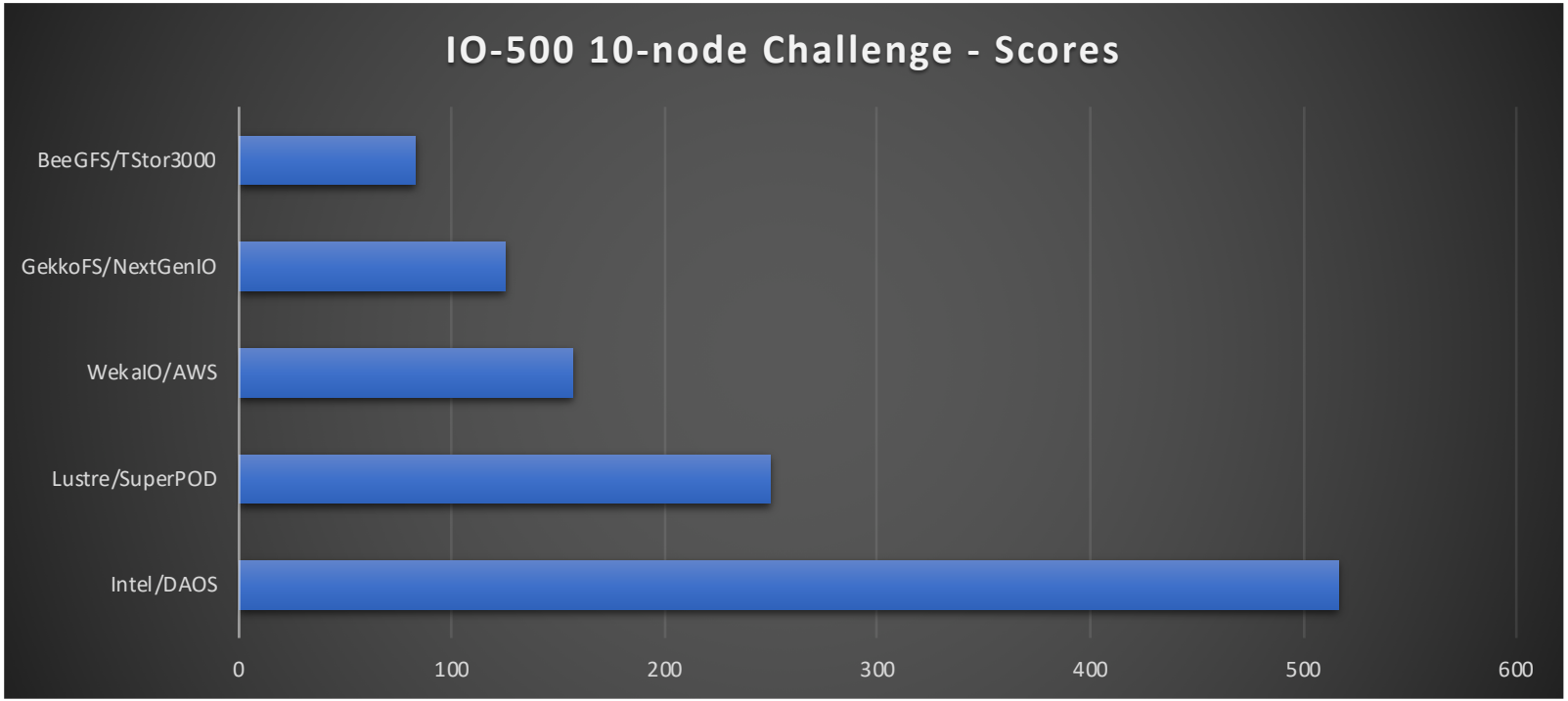
50 nodes



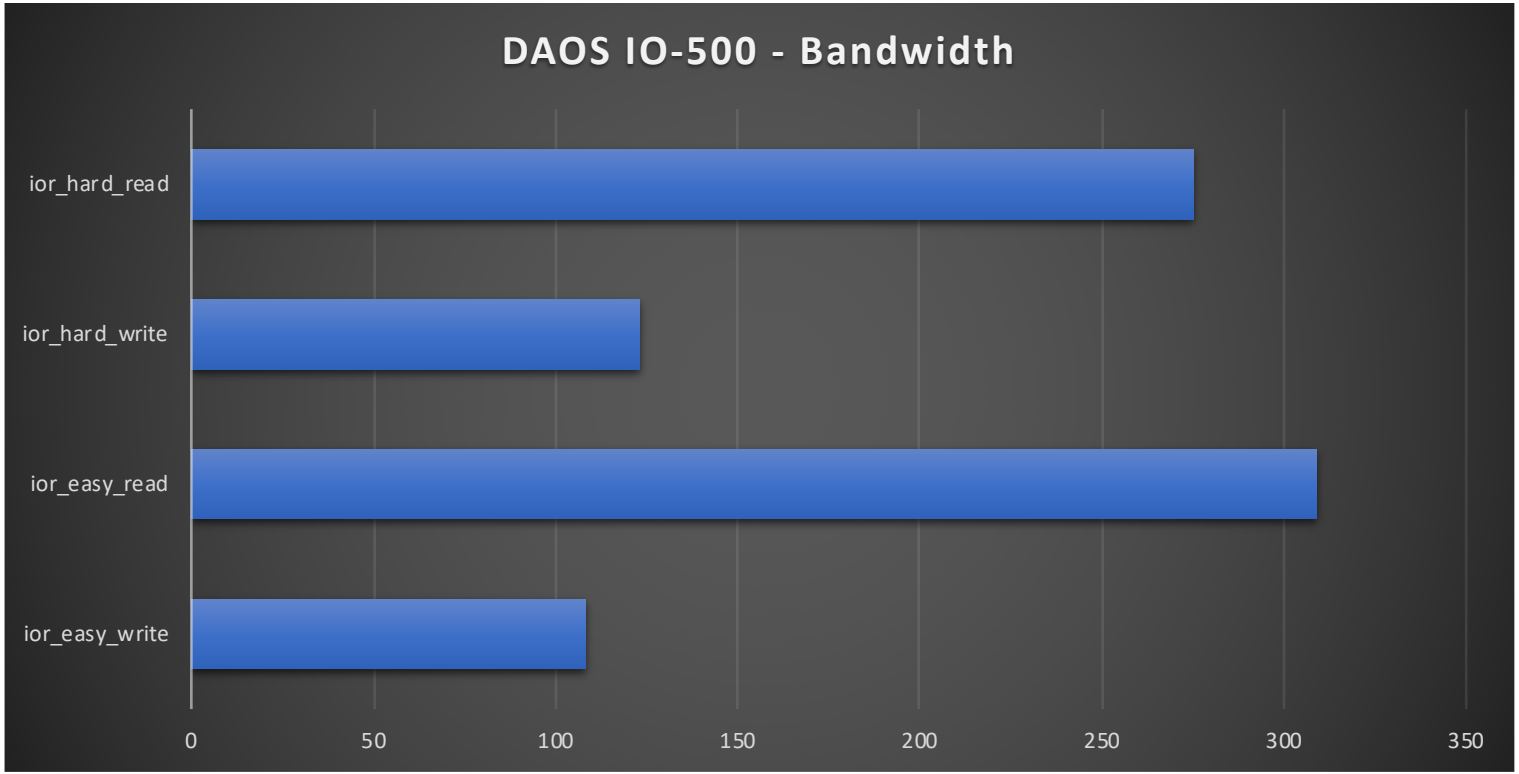
DAOS & IO-500 – Score per node



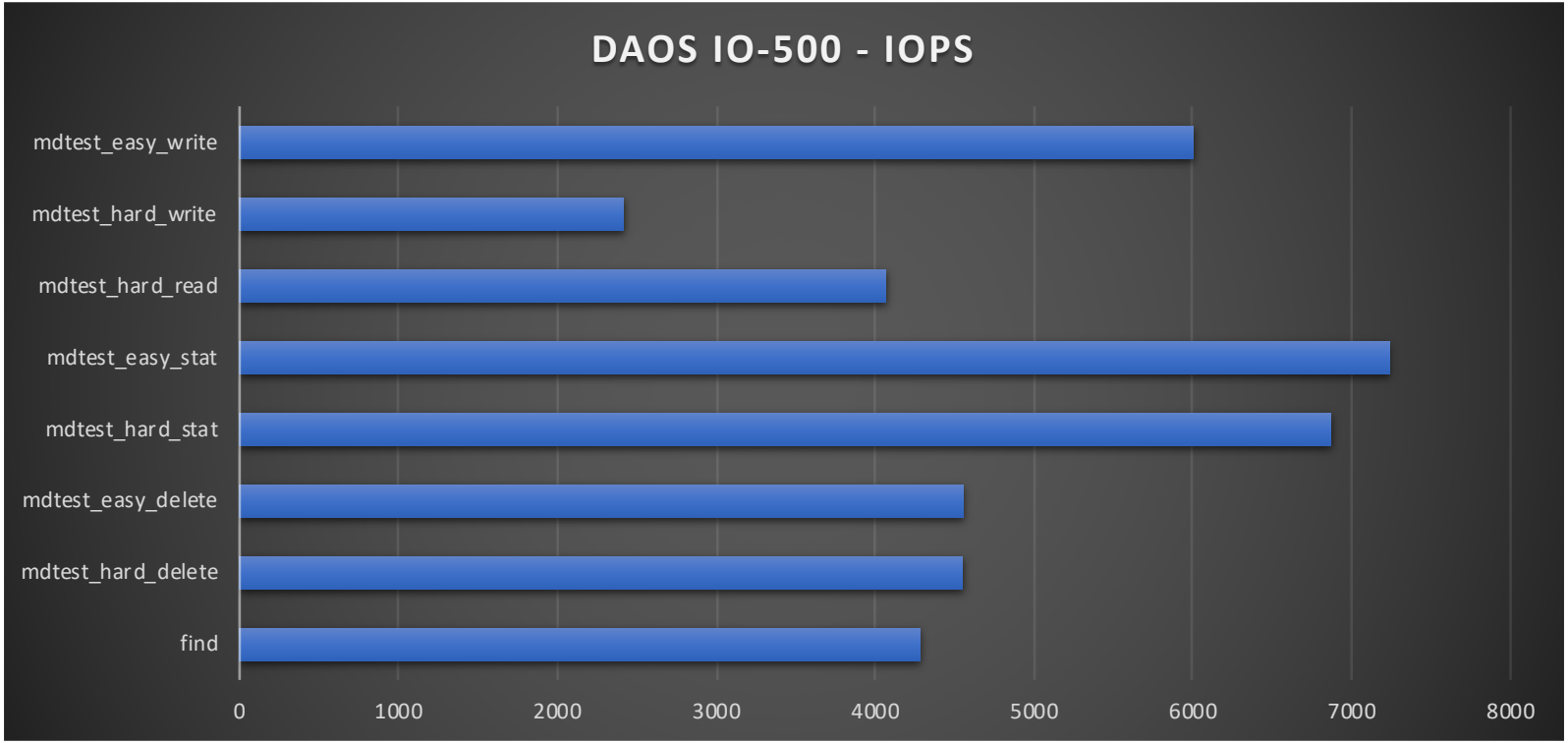
DAOS & IO-500 – 10-node Challenge

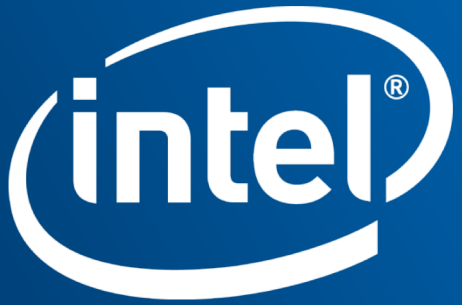


DAOS & IO-500 - Bandwidth



DAOS & IO-500 - IOPS




intel®