OAK RIDGE National Laboratory

SCTuner: An Autotuner Addressing Dynamic I/O Needs on Supercomputer I/O Subsystems

Houjun Tang[1], **Bing Xie**[2], Suren Byna[1], Phil Carns[3], Quincey Koziol[1], Sudarsun Kannan[4], Jay Lofstead[5], Sarp Oral[1]

[1] Lawrence Berkeley National Laboratory, [2]Oak Ridge National Laboratory
[3]Argonne National Laboratory, [4]Rutgers University, [5] Sandia National Laboratories

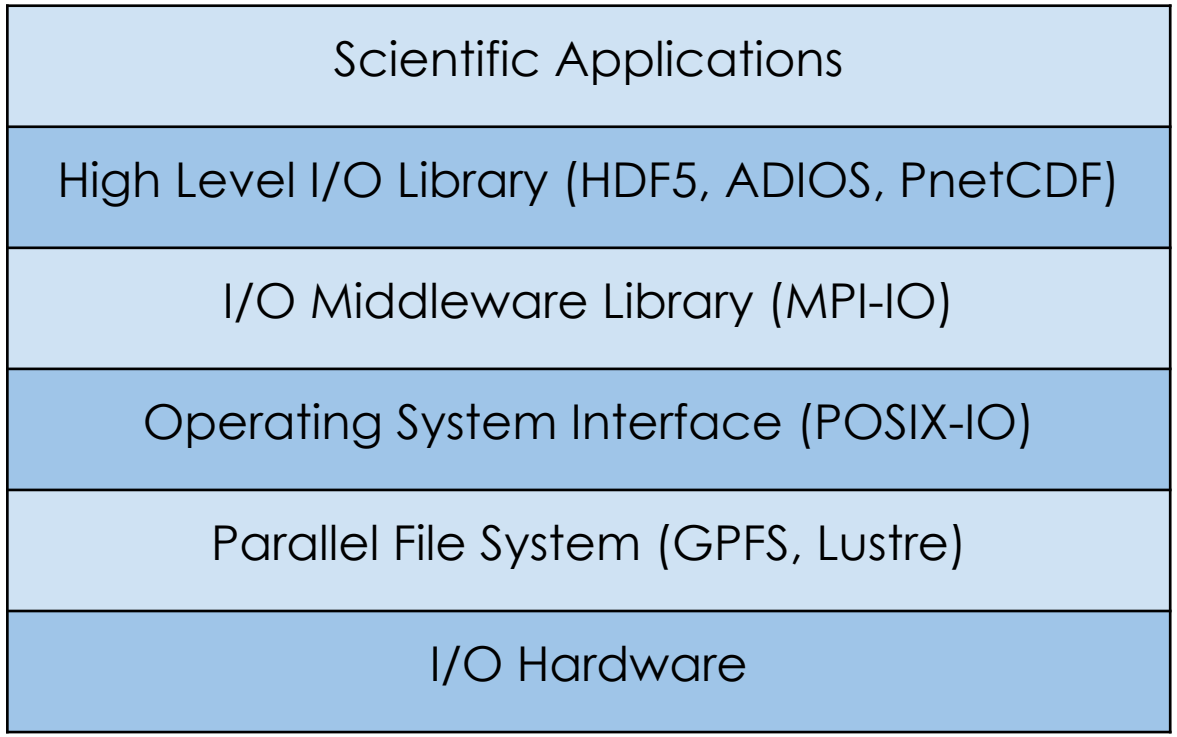
ORNL is managed by UT-Battelle, LLC for the US Department of Energy

U.S. DEPARTMENT OF ENERGY

Agenda

- Scientific I/O on Supercomputer Platforms
  - HPC I/O stack
  - HPC I/O middleware libraries, such like HDF5
  - Our previous HDF5 tuning works
- SCTuner: An Autotuner Addressing Dynamic I/O Needs at Application Runtime
  - Motivations
  - Architecture, methodology, and current implementation
- Preliminary Results

Scientific I/O on Supercomputer Platforms



HDF5 tuning parameters: alignment, metadata cache, ...

MPI tuning parameters: collective/independent, ...

File system tuning parameters: stripe size and count, ...

Default Configurations in I/O Libraries Are Not Optimal

- **HDF5, ADIOS, PnetCDF are widely used by HPC applications**
  - Provide support for heterogeneous data, cross platform, and portability.
  - Hide the low level details of MPI-IO, POSIX-IO and large-scale parallel file systems.

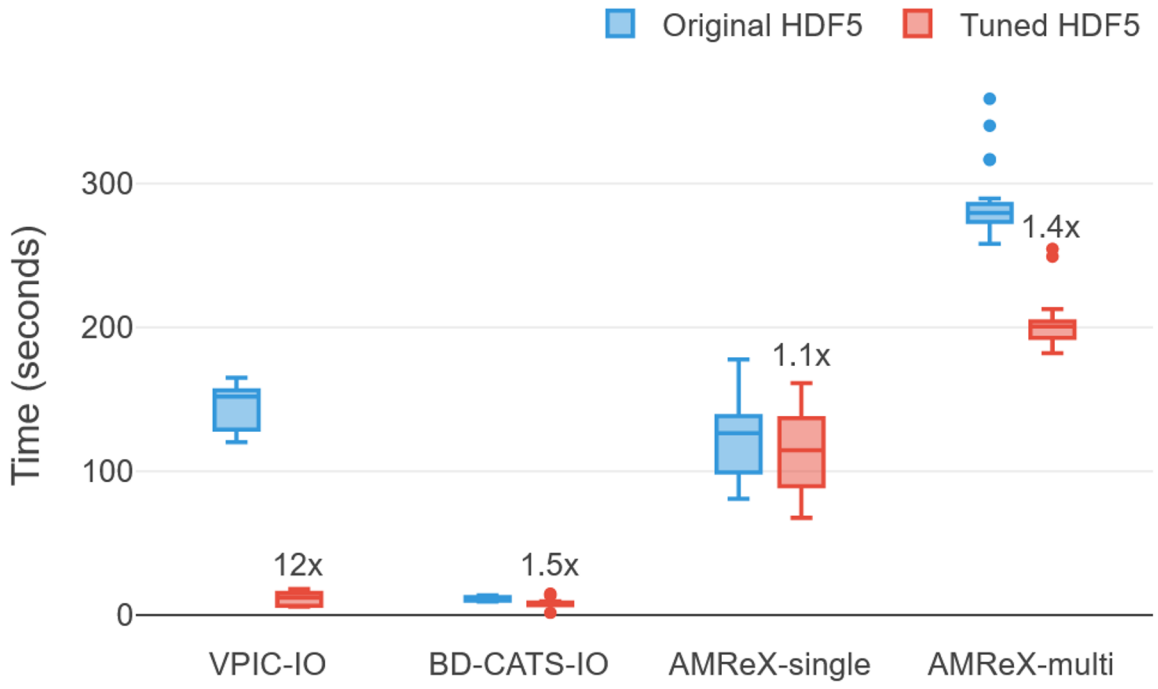
- **HDF5 has a large set of performance-tuning parameters**
  - HDF5 file internal metadata (e.g. object header, B-tree, etc.) management.
  - Parallel I/O in HDF5 are built on MPI-IO that has various parameters.
  - Parallel file system parameters (e.g. data layout properties).

- **Default HDF5 configurations are not optimized for specific systems**
  - HPC systems such as Summit@OLCF and Cori@NERSC have different software and hardware with different performance characteristics.
  - Performance gap between the default and optimized setting can be significant.

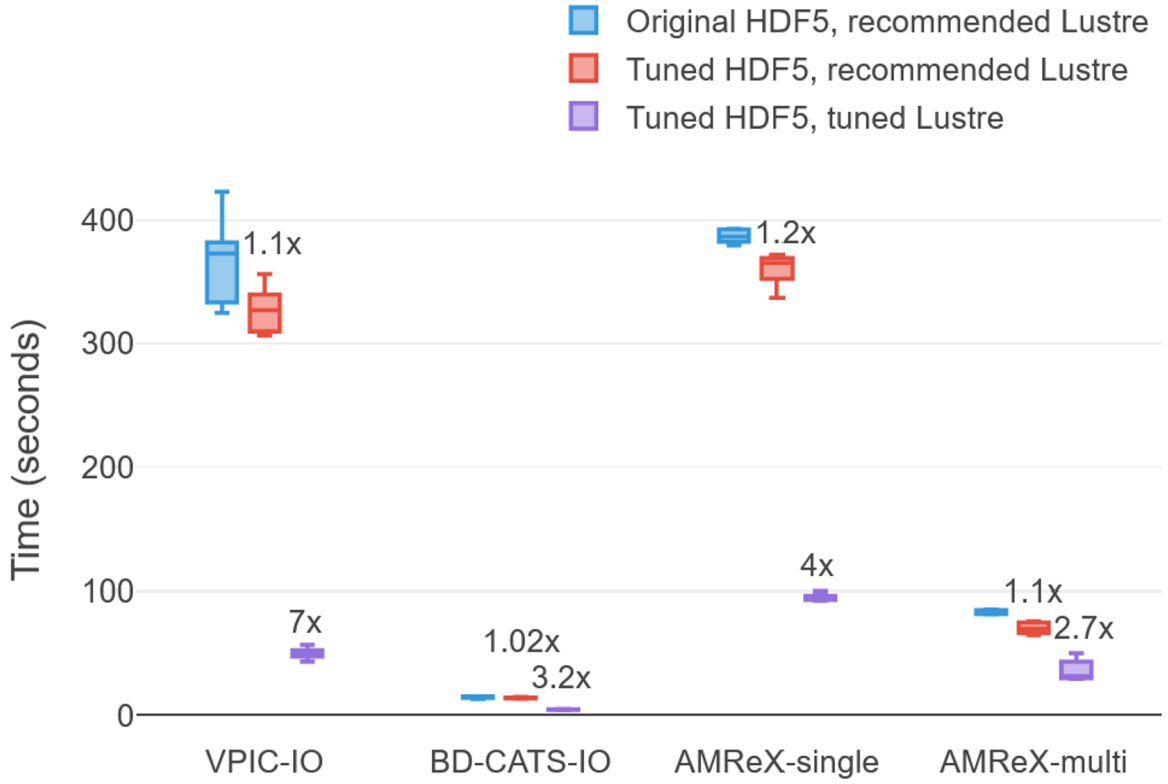
Identifying Optimal Parameters via Benchmarking

- **Our CCGrid'21 work**
  - Battle of the Defaults: Extracting Performance Characteristics of HDF5 under Production Load
- **Approach: collecting a large set of repeated benchmarking results with various configurations**
  - Benchmarking on two production HPC systems: Cori and Summit
  - Using IOR with 3328 different I/O configurations.
  - Measuring the performance of POSIX-IO, MPI-IO, and HDF5.
  - Repeating each set for at least 100 times.
- **Analysis: identifying a set of parameters that consistently provide good I/O performance in the benchmarking results**
  - HDF5 alignment setting and metadata management are crucial to overall performance.
  - Performance difference can be 10x better than the default.
- **HDF5 file layout and interaction with file system is crucial to performance**

Experiments: Application Performance with Tuned Parameters



512 Summit nodes



512 Cori nodes

Tuned HDF5 achieves **1.1X to 12X** I/O performance improvement

6

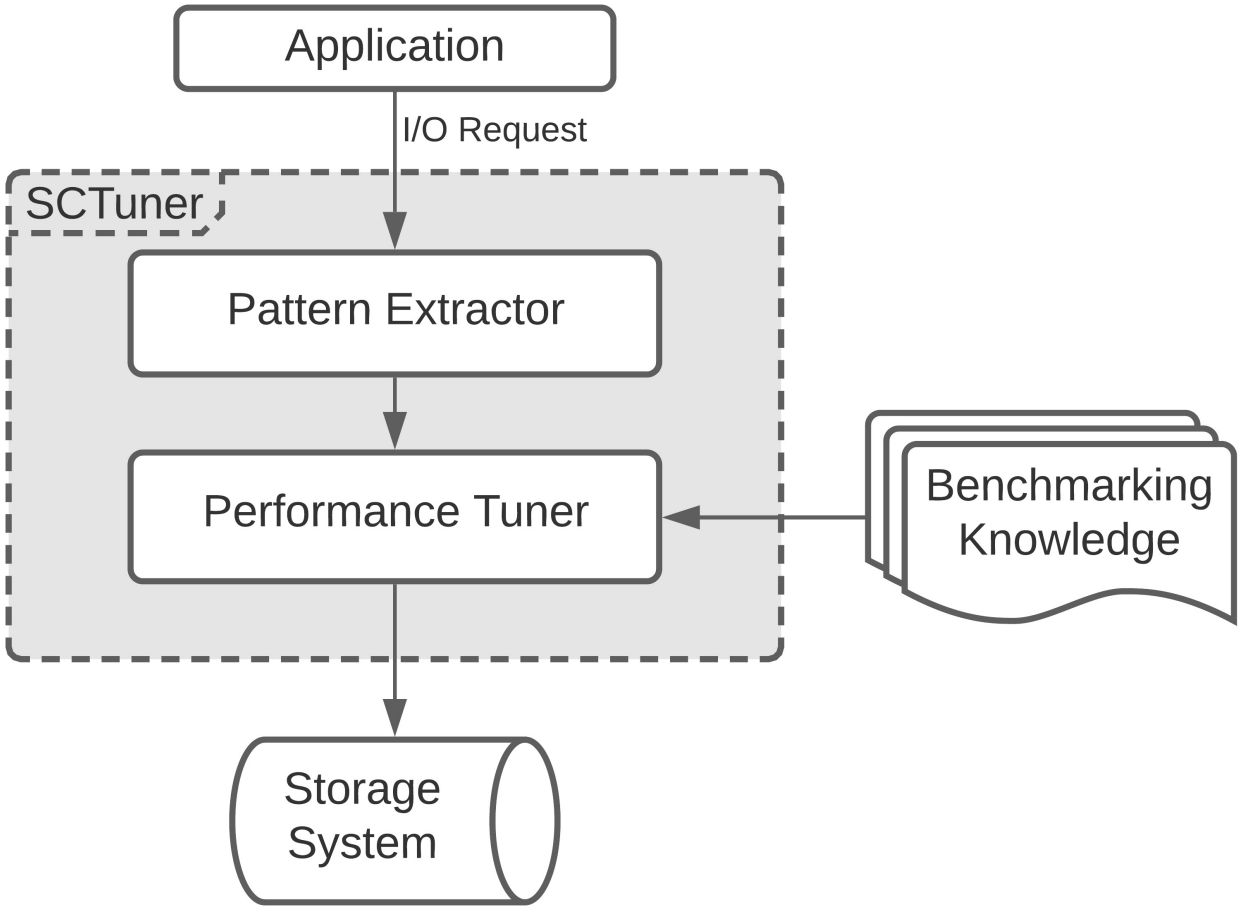
SCTuner: An Autotuner Addressing Dynamic I/O Needs at Application Runtime

- Motivations
  - Different application I/O exist on supercomputers, exhibiting different performance needs at application runtime.
  - Different file systems have different hardware, are deployed with different file system software, and configured differently.
  - HPC I/O middleware libraries empower users to customize configurations on different layers of I/O stack, but users stick to the default configurations with poor I/O performance.

SCTuner: An Autotuner Addressing Dynamic I/O Needs at Application Runtime

- Design principles
  - To address dissimilarity in the target file systems, we design benchmarking experiments to profile the I/O behaviors of individual systems.
  - To address I/O patterns, we use IOR as an I/O pattern generator that covers a wide range of burst sizes and I/O scales.
  - To capture the consistent behaviors from noise and randomness, we repeat the experiments and characterize the results by a five-number summary and clustering.

Architecture of SCTuner



Done: benchmarking experiments

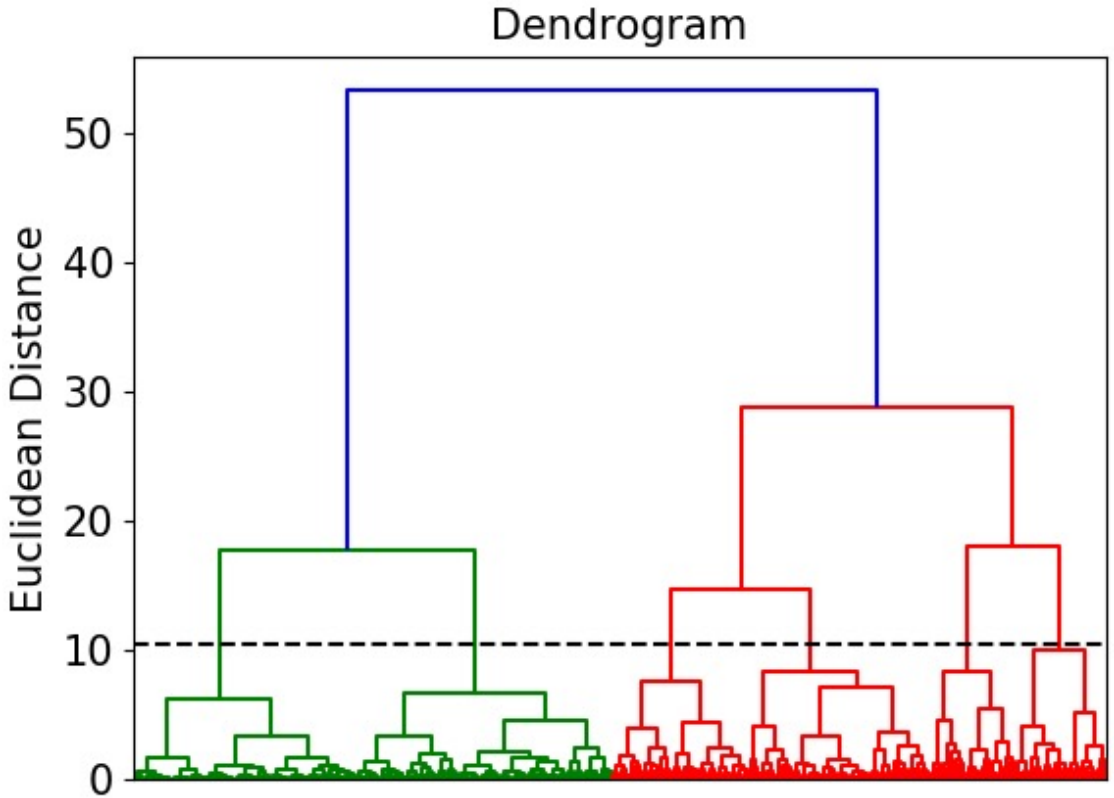
Done: pattern extractor built in HDF5

Todo: performance tuner

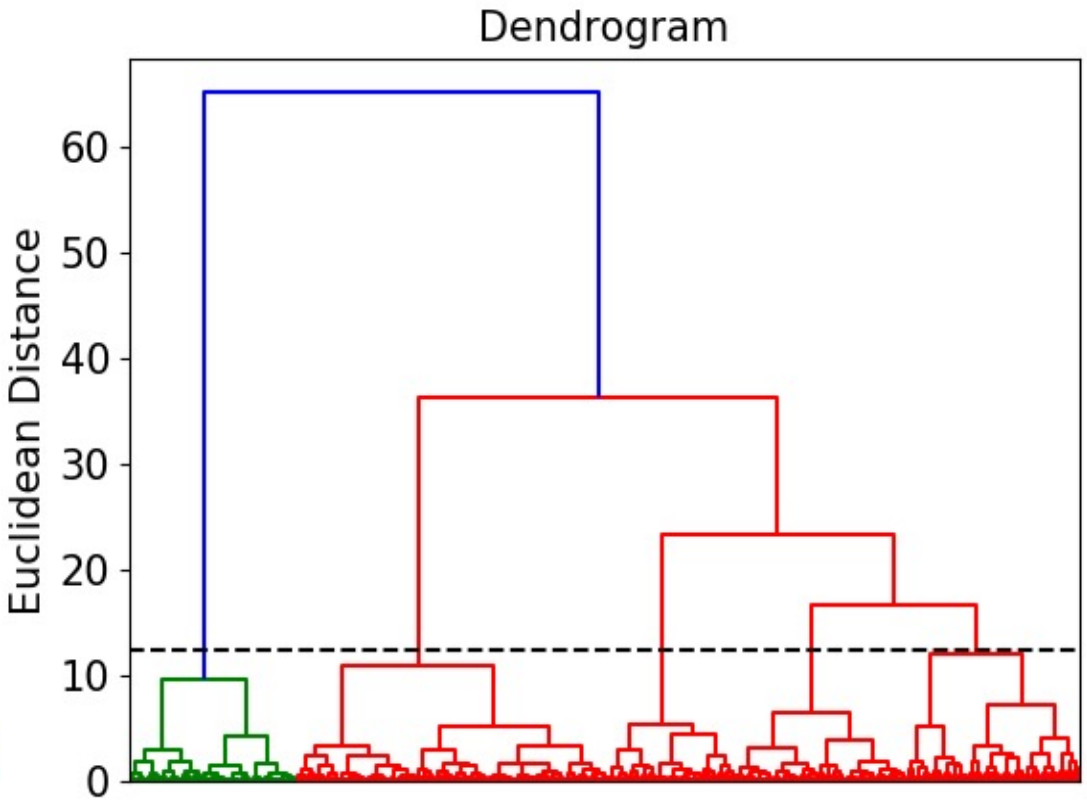
Benchmarking with IOR

- **Summit@OLCF**
- **Varying the values of the parameters on IOR, HDF5**
  - Number of total MPI processes / compute nodes.
  - Number of processes per node.
  - Individual I/O size.
  - Number of aggregators in MPI-IO.
  - Burst Size in MPI-IO.
- **Five-number summary, hierarchical clustering**

Preliminary Results on Summit and Alpine

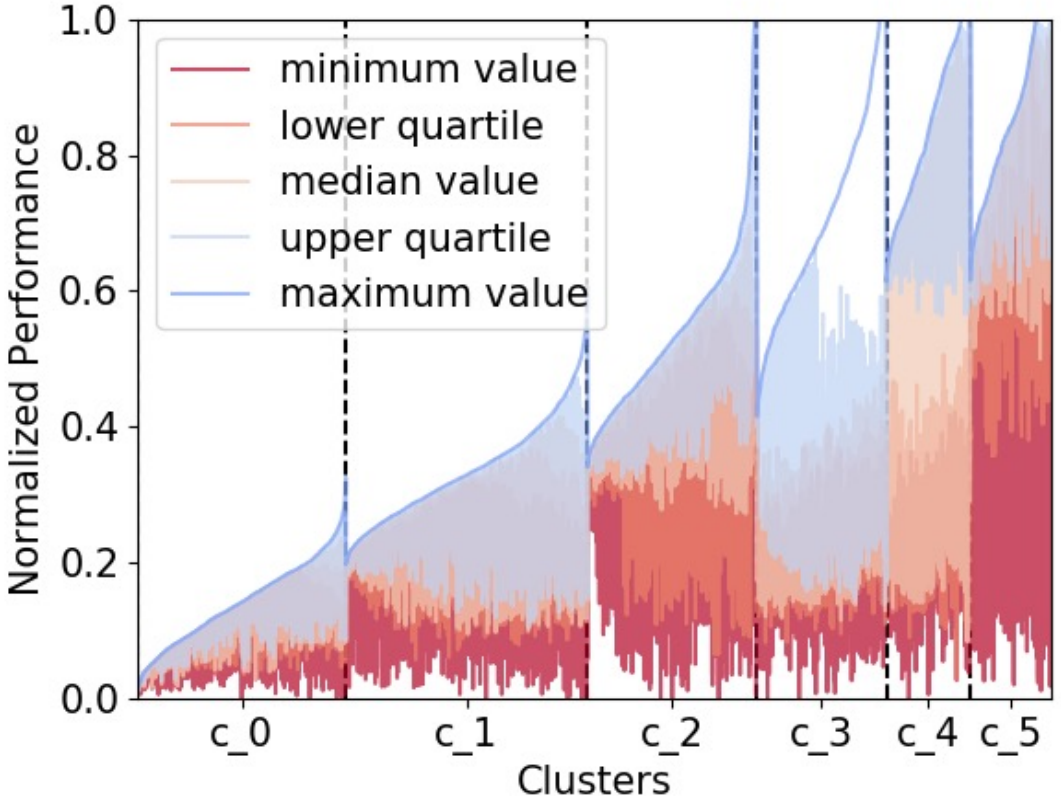


Clustered results for read: 6 groups

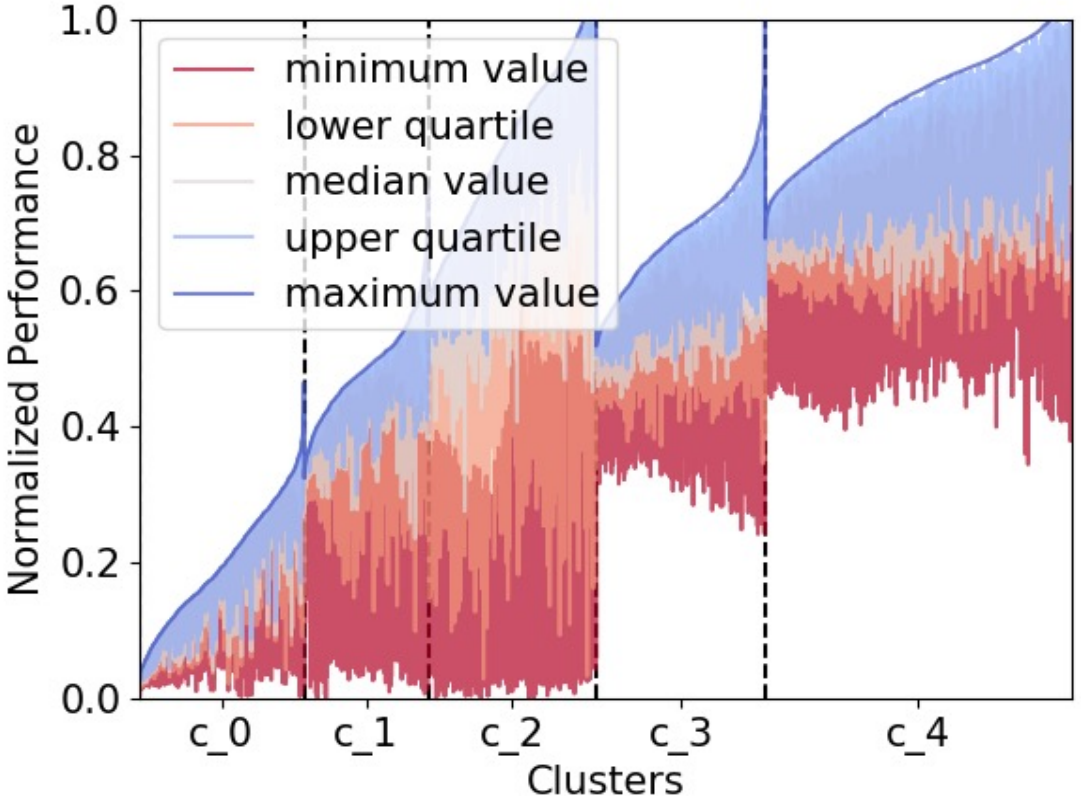


Clustered results for write: 5 groups

Preliminary Results on Summit and Alpine

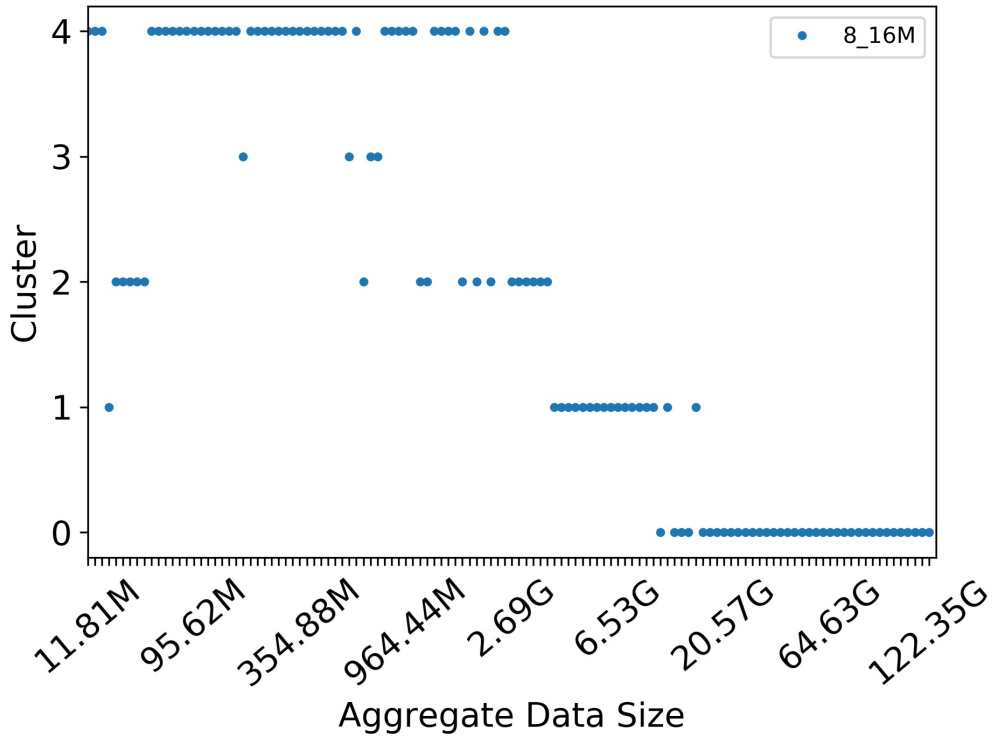


Clustered five-number summary for read

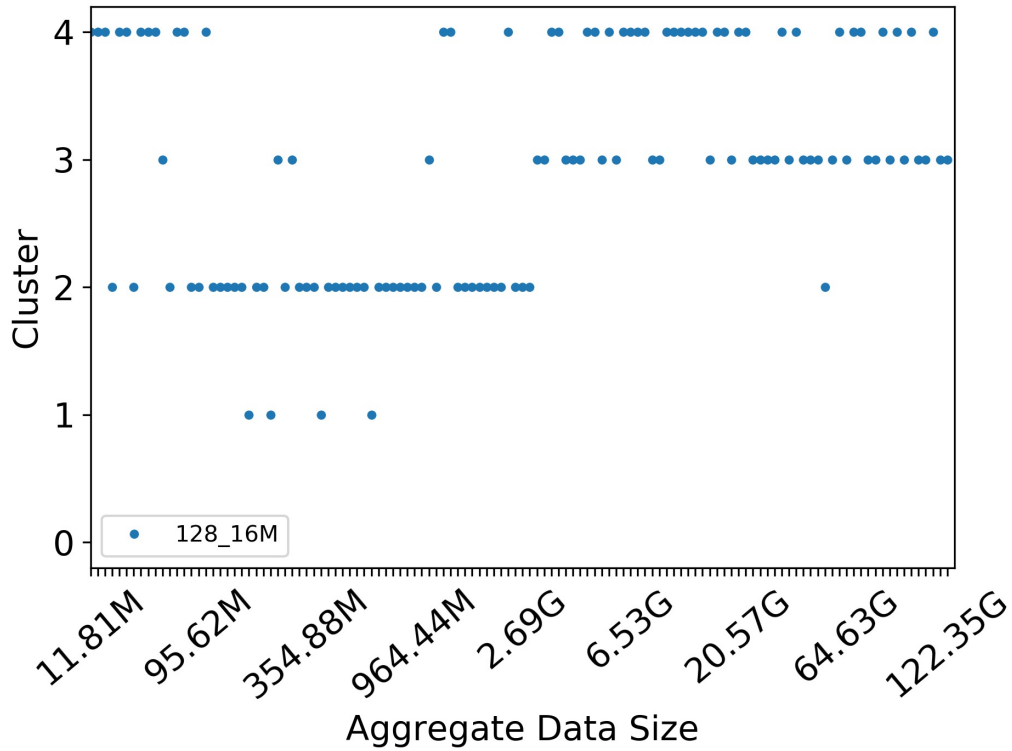


Clustered five-number summary for write

Preliminary Results on Summit and Alpine



Write performance for 8 aggregators, 64MB buffer size on 32 nodes



Write performance for 128 aggregators, 16MB buffer size on 32 nodes

Conclusions and Future Work

- A benchmarking analysis with controlled experiments to capture the consistent I/O behaviors under production loads
- Implemented I/O pattern extractor in HDF5
- Plan to realize online performance tuner in HDF5 asynchronous I/O VOL connector.

Acknowledgements

- This research is supported by the Director, Office of Science, Office of Advanced Scientific Computing Research, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.
- This work was supported by the U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research, under Contract DE-AC02-06CH11357.
- This work used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.
- Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525 (SAND2021-12186 C).
- Sudarsun Kannan was partially supported by NSF CNS 1850297 award.
- This material is based upon work supported by the U.S. Department of Energy , Office of Science, under contract DE-AC02-06CH11357.