# Auto-tuning of Parallel I/O Parameters for HDF5 Applications

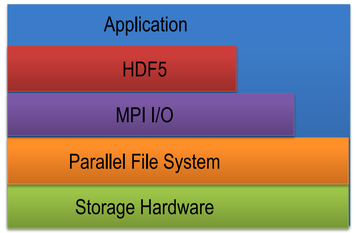

**Babak Behzad[1,3], Joey Huchette[2,4], Huong Luu[1,3],**
**Ruth Aydt[1], Suren Byna[2], Mohamad Chaarawi[1], Quincey Koziol[1], Prabhat[2], Yushu Yao[2]**

[1] The HDF Group [2] Lawrence Berkeley National Laboratory [3] University of Illinois at Urbana-Champaign [4] Rice University


## MOTIVATION

- I/O performance is a major bottleneck for many HPC applications.
- I/O tuning is a very difficult yet important problem.
- Significant improvement can be achieved when parameters are set appropriately.

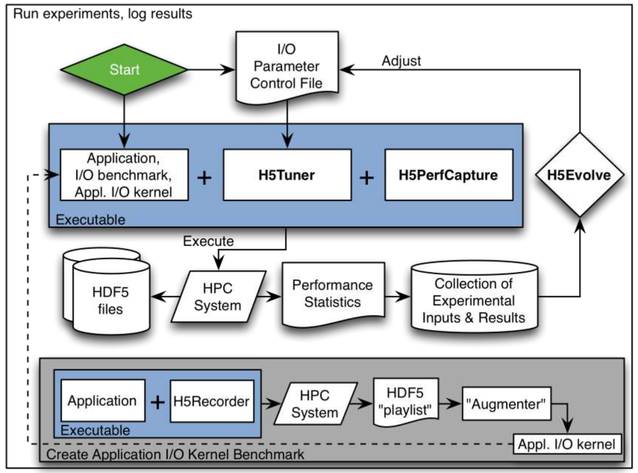
- Multiple layers of the I/O stack, each with its own parameters to be tuned, make the search space immense.
- Pruning search space requires expertise and heuristics.




## FRAMEWORK DEVELOPED

**H5Tuner:** Reads configuration file and adjusts I/O parameters without recompiling the application.

- ✧ **Stripe count** (Lustre): Number of OSTs over which a file is distributed.
  - ✓ 4, 8, 16, 24, 32, 48, 64, 96, 128, -1=all OSTs
- ✧ **Stripe size** (Lustre): Number of bytes written to an OST before cycling to the next.
  - ✓ 1, 2, 4, 8, 16, 32, 64, 128 *units = MB*
- ✧ **CB nodes** (MPI-IO): Maximum number of aggregators for collective buffering.
  - ✓ 1, 2, 4, 8, 16, 24, 32, 48, 64, 96, 128, 256
- ✧ **CB Buffer Size** (MPI-IO): Size of buffer for collective I/O. Currently set to stripe size.
- ✧ **Alignment (threshhold, boundary)** (HDF5): Objects >= threshhold size aligned to N*boundary address in HDF5 file.
  - ✓ (no alignment), (0,4), (0,16), (0,64), (0,256), (1,4), (1,16), (1,64), (1,256), (4,16), (4,64), (4,256), (16,64), (16,256) *units = KB*
- ✧ … other I/O parameters can be added

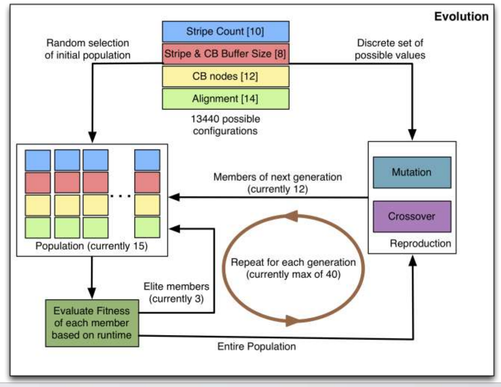



**H5PerfCapture:** An extension to Darshan* that records performance data for application, HDF5, and MPI-IO. *www.mcs.anl.gov/research/projects/darshan/

- Results used to further investigate I/O performance.

**H5Evolve:** An application built on Pyevolve* that uses Genetic Algorithms to narrow the huge space of possible configurations. *http://pyevolve.sourceforge.net

- Uses crossover and mutation functions to intelligently search for well-performing I/O parameter sets.
- Other approaches for pruning space could be plugged into framework.



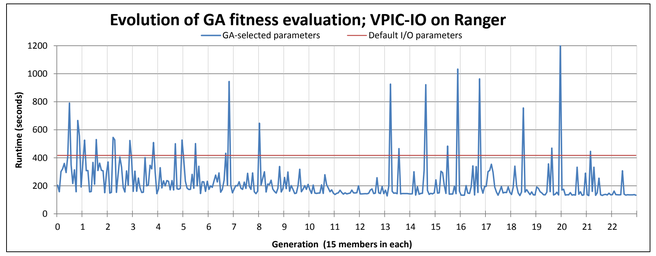

## EXPERIMENTAL SETUP

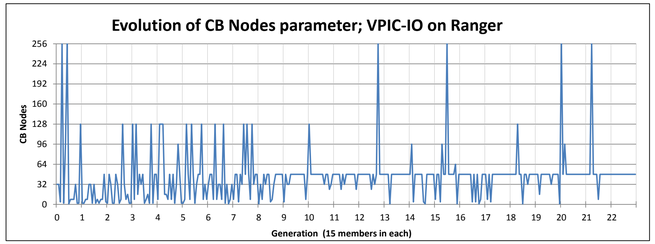
- **Hopper:** NERSC's Cray XE6 system peak performance of 1.28 PFLOPS/sec, 153,216 compute cores, 156 OSTs
  - ✓ Peak I/O Bandwidth: 35 GB/sec
- **Ranger**: TACC's Sun Constellation Cluster peak performance of 579 TFLOPS, 62,976 compute cores, 160 OSTs
  - ✓ Peak I/O Bandwidth: 35 GB/sec

- **VPIC-IO:** I/O replay of Vector Particle-In-Cell (VPIC), a plasma physics code with 1D I/O.
- **VORPAL-IO:** I/O replay of VORPAL, a particle accelerator code with structured I/O with different dimensions on each rank.
- **GCRM-IO:** I/O benchmark for Global Cloud Resolving Model (GCRM), a global atmospheric model with structured I/O with the same dimensions on all ranks.

## RESULTS

**Configurations used for results shown:**
Ranger: 512 cores; 32 nodes; Hopper: 512 cores; 32 nodes
**Amount of data written:** (VPIC, GCRM, VORPAL)
Ranger: 128 GB, 163 GB, 61 GB Hopper: 128 GB, 130 GB, 128 GB




**Tuned Parameters, Runtimes, and Speedup of Tuned over Default**

| System | Ranger | | | Hopper | | |
|---|---|---|---|---|---|---|
| **Application** | **VPIC-IO** | **GCRM-IO** | **VORPAL-IO** | **VPIC-IO** | **GCRM-IO** | **VORPAL-IO** |
| **Parameter** | **Tuned Sets of Parameters Identified by H5Evolve** | | | | | |
| Stripe Count | 96 | 96 | 32 | 64 | 96 | 96 |
| Stripe Size CB Buffer Size | 128 MB | 1 MB | 8 MB | 2 MB | 4 MB | 32 MB |
| CB Nodes | 48 | 64 | 64 | 24 | 128 | 48 |
| Alignment (thrsh, bndry) | 4 KB, 16 KB | 0 KB, 64 KB | 0 KB, 256 KB | 0 KB, 64 KB | 0 KB, 4 KB | 16 KB, 256 KB |
| **Description** | **Measured Runtime (seconds)** | | | | | |
| Default Parameters | 417.50 | 498.21 | 391.72 | 320.32 | 472.18 | 425.76 |
| Minimum Observed | 127.92 | 84.29 | 103.99 | 18.84 | 54.99 | 68.37 |
| Maximum Observed | 1205.89 | 1485.36 | 959.51 | 1183.63 | 1968.91 | 1277.85 |
| Tuned Set | 127.92 | 84.29 | 103.99 | 19.23 | 55.61 | 68.37 |
| **Speedup** | 3.26x | 5.91x | 3.77x | 16.66x | 8.49x | 6.23x |


**Contact Information:** koziol@hdfgroup.org


## CONCLUSIONS AND FUTURE WORK

**Conclusions:**

- Speedups of 3.3x – 16.7x were measured using auto-tuned I/O parameters compared to default values for three application-based I/O benchmarks on two HPC systems
- Converged parameter sets differ across applications and systems.
- Auto-tuning can improve parallel I/O performance without hands-on optimization

**Future Work:**

- Conduct additional experiments with more codes
- Further refinement of GA methods used
- Automatically create I/O kernel benchmarks from full applications by developing H5Recorder


This work was 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. Some architectural concepts were drawn from I/O benchmarking work with LLNL under Contract B592421. H. Luu is partially supported by NSF grant 0938064. This research used resources of the National Energy Research Scientific Computing Center and of the Texas Advanced Computing Center.