Fine-grained I/O and Storage (IOS)

Peter Van Gemmeren
High Energy Physics Division
Argonne National Laboratory
gemmeren@anl.gov

Rob Ross
Mathematics and Computer Science Division
Argonne National Laboratory
rross@mcs.anl.gov

IOS Participants, initial list, will grow

High Energy Physics

- Philippe Canal (FNAL)
- Oliver Gutsche (FNAL)
- Christopher Jones (FNAL)
- Michael Kirby (FNAL)
- Matti Kortelainen (FNAL)
- Peter Van Gemmeren (ANL)
- Kevin Pedro (FNAL)
- Brett Viren (BNL)
- Torre Wenaus (BNL)

Computer Science

- Suren Byna (LBL)
- Matthieu Dorier (ANL)
- Rob Latham (ANL)
- Rob Ross (ANL)
- Saba Sehrish (FNAL)
- John Wu (LBL)

Fine-grained I/O and Storage

Traditionally:

- Events have been grouped into (many) files
- Often processing is broken up by splitting up collection of files
- Multi-stage workflows pass data through files also

In HPC:

- File access overheads are high(er) relative to cost of computation and communication.
- Larger files tend to help amortize costs, but force reorganization of data
- Lots of ongoing work in alternatives to files for passing data within workflows

What are we trying to accomplish?

- Explore options for improving absolute performance and parallelism of I/O during workflows
  - Alternatives to use of files
  - Connection with parallelization strategies work – data organization should have good “impedance match” with what is needed for computation
- Understand implications of options for event data models and representations
  - Allow for events to be segmented into smaller regions to speed up processing that can scale poorly with high multiplicity
  - Mapping back to traditional file-based representations at end of workflow
- Demonstrate promising options in real-world HEP workflows

Plan of Work

Phase I: Preparation

– Document existing implementations for participating experiments

– Define a set of representative synthetic benchmarks

– Discuss viability of alternatives for HPC workflows

Phase II: Prototyping

– Develop proof-of-concept prototype(s)

– Work with PPS team to ensure efficient mapping to memory constructs

Phase III: Benchmarking and reporting

– Run experiments using synthetic benchmarks on relevant platforms, refine prototypes

– Develop recommendations for experiments and engage in dialog on outcomes

Near Term

- Get to know one another!
  - Give short presentations on background topics with Q&A
  - Learn each others’ language

Topic Ideas

- Run and benchmark I/O for HEP production workflows on HPC
  - Adapt to and utilize technologies such as Parallel File Systems (PFS)
  - Develop a set of benchmarks to ground discussion and experimentation
- Use of non-POSIX storage for LHC Analysis Data
  - i.e., not proper parallel file systems
  - DataWarp – very close to a PFS, easy to use
  - Distributed Asynchronous Object Storage (DAOS) – can mimic a PFS, but has richer multi-dimensional capabilities we might use (e.g., look like a table store)
  - Other key-value options?
- Investigate optimization of (persistent) Event Data Model (EDM) in cooperation with Portable Parallelization group

Communication

- Mailing list:

– https://lists.anl.gov/mailman/listinfo/cce-ios

– cce-ios@lists.anl.gov

- Recurring calls:

– TBD