IOS PLANS (FY23)

PETER VAN GEMMEREN
For HEP-CCE/IOS group

Remote in IL
October 12th, 2022

Argonne National Laboratory is a
U.S. Department of Energy laboratory
managed by UChicago Argonne, LLC.

HEP-CCE/IOS

Overview: What we do

- Working areas for reports:
  - Investigate HDF5 as intermediate event storage for HPC processing
    - Saba gave a detailed presentation on this effort
  - Darshan for ROOT I/O in HEP workflows on HPC
    - Shane showed overview of the tool developments
    - Rui presented [new] example studies of ATLAS Simulation use-case
  - Investigate GPU friendly developments for data model
    - Amit has shown us the latest developments

INVESTIGATE HDF5 AS INTERMEDIATE EVENT STORAGE FOR HPC PROCESSING

Using ROOT Serialization to support existing Data Models

- *In FY23, IOS will complete and document the development of a ROOT serialization mechanism for HDF5 event data stores, either as a technical note or as a publication. Leading to this, In FY22 and early FY23, we will use the IOS test framework to test collective I/O effects with different HDF data store structures (by event, by batches, and using various methods to store navigational meta-data). We may also test multi-threading support in HDF5.*
- The intention of this effort was to enable more efficient/flexible execution of HEP workflows on HPC.
  - Of special interest: Multiple writer to single file
  - intermediate event storage ~ data that remains on HPC
  - E.g. Multistep workflows

FY23 PLAN: HDF5

PRESENTED YESTERDAY BY SABA

Next steps and plan for FY23

- Complete performance evaluation studies on Cori
  - Understand IO behavior, untimed calls that are actually constituting 50% of the IO time, use Darshan logs
- Move evaluation studies to Perlmutter soon
- HDF5 options to work with
  - Asynchronous IO
  - Multi dataset; we only have 3 datasets though
- Write a paper/report

Discussed yesterday: First look at ROOT TMPIFile as comparison

LONGER TERM PLAN: HDF5

PRESENTED YESTERDAY BY SABA

Future work

- Storing C++ objects (using ROOT’s reflection) directly to HDF5 without serialization
- Multi-threaded HDF5
- Any use of Subfiling for our work?
- HDF5 streaming services
- Explore direct storage access from GPUs
- Others not directly HDF5
    - RNtuple
    - C++ object design studies

From report: Consider Parquet, other I/O backends?

DARSHAN FOR ROOT I/O IN HEP WORKFLOWS ON HPC

Tool enhancements and workflow monitoring for ATLAS, CMS, DUNE

- *The report will include the results of IOS performance studies with Darshan, a performance analysis framework that HEP-CCE has enhanced and will continue to enhance to support HEP use cases. Specifically, there will be a study of the performance of the CMS event data store (for which preliminary results are already available) and the ATLAS data store (delayed by personnel changes, expected in FY22Q4) on HPC parallel file systems.*
- The intention of this effort was to develop the infrastructure and gain insight into I/O patterns of HEP workflows running on HPC.
  - Darshan Tool: Existing on HPCs, but suboptimal for HEP applications has been enhanced and provides capability to produce important observation for I/O behavior of HEP jobs.

FY23 PLAN: DARSHAN TOOL

PRESENTED YESTERDAY BY SHANE

Potential next steps with Darshan in IOS

Utilize new Darshan instrumentation modules to better understand I/O behavior of other IOS activities

- *HDF5 module*: insights into DUNE HDF5 usage, ROOT→HDF5 serialization efforts
- *DAOS module*: insights into ROOT's RNTuple DAOS backend

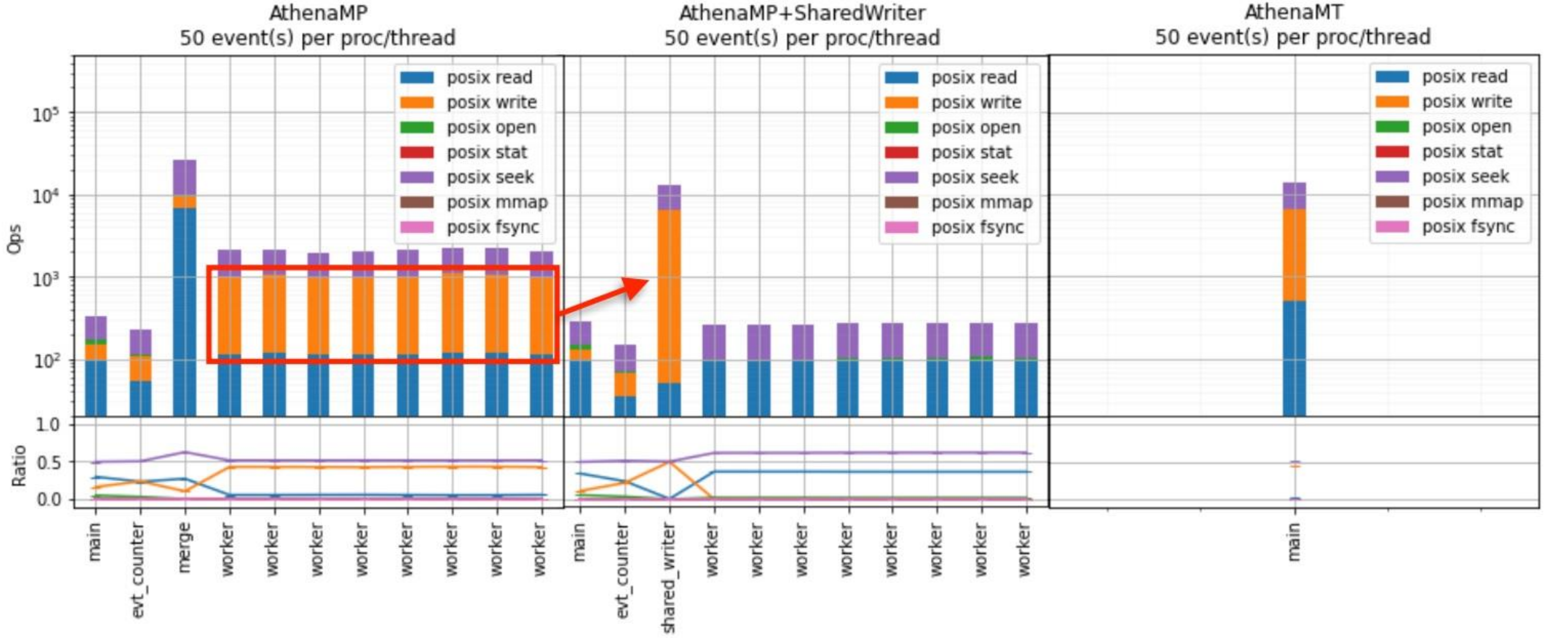
Utilize PyDarshan log analysis tools and extend them to help analyze I/O characteristics of HEP workflows

FY23 PLAN: EXPERIMENT MONITORING WITH DARSHAN

PRESENTED YESTERDAY BY RUI, FOR ATLAS (E.G.)

POSIX Operations

- In AthenaMP each worker does the writes while SharedWirter took over these



DARSHAN FOR ROOT I/O IN HEP WORKFLOWS ON HPC

Continued, consider this ground work for longer term effort

- *The new Darshan features will enable IOS to characterize the performance of a test HDF5 event data store, including collective I/O effects, by FY23Q2. Darshan is expected to support the Distributed Asynchronous Object Storage (DAOS, to be used by ALCF Aurora) in FY22Q4. Once this becomes available, IOS will begin studying the performance of ROOT RNTuple in comparison with traditional analysis formats with the goal of including preliminary results in the final report.*
- Converging with HDF5 work.
- Some new effort from BNL (Doug B., et al)

INVESTIGATE GPU FRIENDLY DEVELOPMENTS FOR DATA MODEL

In coordination with PPS

- *Also in FY 22/23, IOS will complete a survey of existing accelerator-friendly data model approaches (from the experiments, but also in HSF, IRIS-HEP, etc.) and prepare a simplified HDF5 toy DUNE-RAW data model that does not rely on ROOT serialization and can be offloaded to co-processors.*
- This effort started somewhat late, but has made good progress.

FY23 PLAN: DATA MODEL

PRESENTED YESTERDAY BY AMIT

Future Works

- Further Work on 2D arrays needed
- Collective I/O Implementation
  - For HDF5 related I/O only
- Effect of precision on performance
- More customized data models that are closer to HEP data models currently used.
  - Build on the top of 1D and 2D arrays that the framework currently supports.
- Ideally would like to minimize (or remove at all) any CUDA API calls when needed.
- AOB

Follow up with discussion with PPS

Discussion