High Energy Physics Center for Computational Excellence

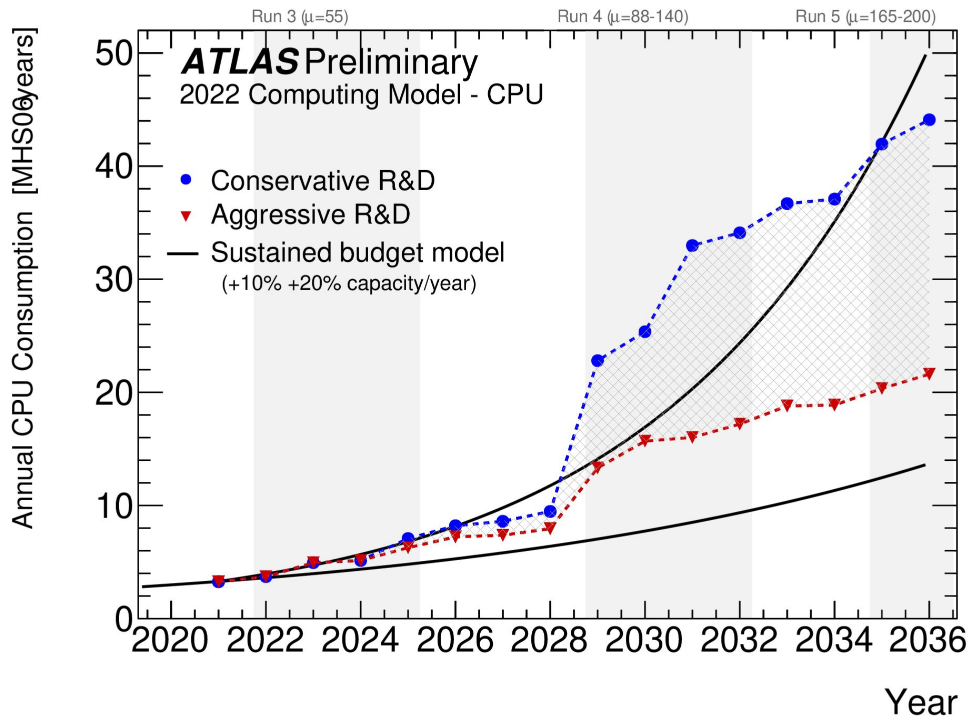
HEP-CCE/IOS -> HEP-CCE2/SOP:

Future Plans: RNTuple

Peter van Gemmeren (ANL) **for HEP-CCE**

ROOT RNTuple Workshop
November 6, 2023

Past: HEP-CCE, IOS



CCE past: Input/Output and Storage Activities

Measuring performance of ROOT I/O in HEP workflows on HPC systems

- Darshan a scalable HPC I/O characterization tool has been enhanced (including fork safety) and used to monitor HEP production workflows.

Investigate HDF5 as intermediate event storage for HPC processing

- Relying on ROOT to serialize complex Event Data Model used in Simulation/Reconstruction workflows
- Implementing Collective Writing to avoid potential merge step
- Mimicking framework for understanding scalability and performance of HEP output methods
  - Experiment agnostic tool allows scaling I/O beyond what is currently accessible by production and has uncovered/fixed bottlenecks in ROOT and frameworks.

HPC friendly Data Model

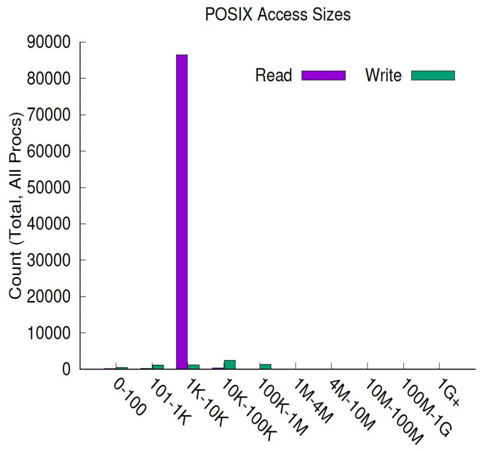
- Together with PPS team started investigating efforts to make data model more suitable for offloading to accelerator and storage on HPC.

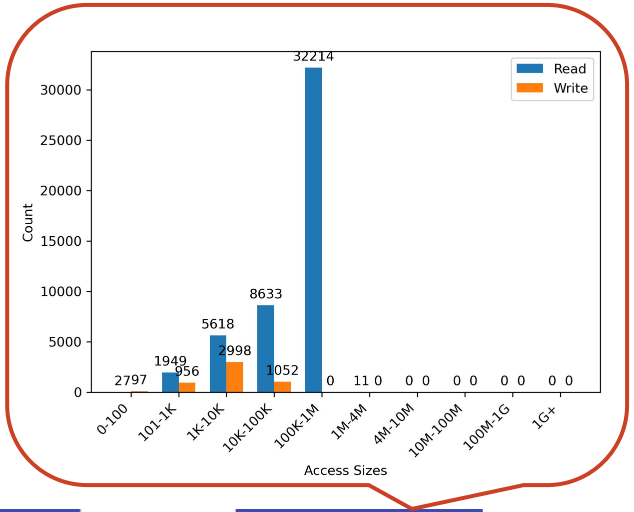
U.S. DEPARTMENT OF ENERGY | Office of Science
Argonne NATIONAL LABORATORY
Brookhaven National Laboratory
Fermilab
BERKELEY LAB Bringing Science Solutions to the World

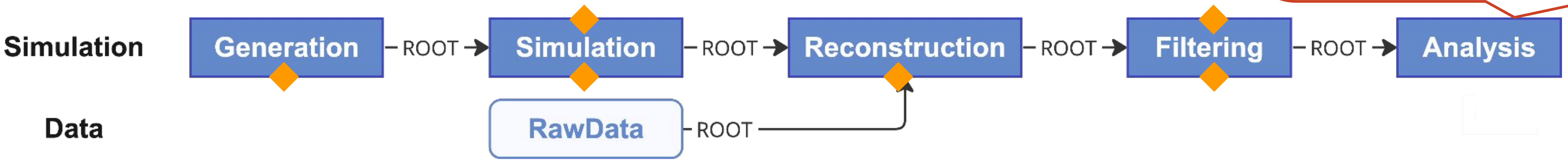
Case study: Access size

ATLAS EXPERIMENT

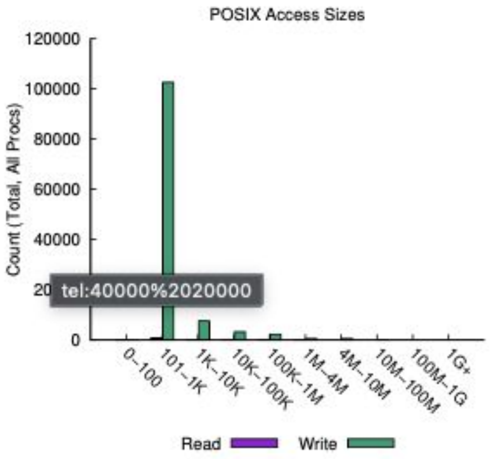
Broadwell on LCRC@ANL
GPFS
SDCC@BNL
GPFS







Haswell on Cori @Nersc
SSD + Lustre
100 events, 16 threads



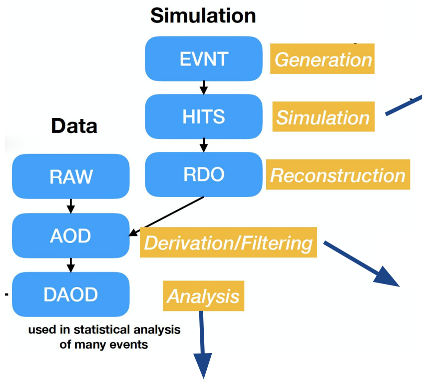
Small reads/writes at O(1KB)

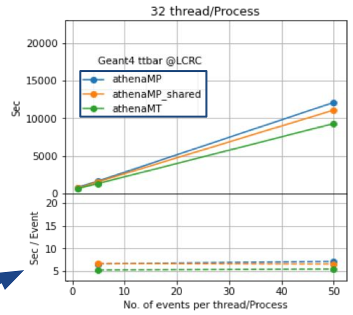
- All stages (marked) except ATLAS Analysis which is at O(100KB)
- Related to ROOT TTreeCache vector I/O support on certain FSes
- Potential bottleneck
- ROOT has a data sieving concept (overread) that might be taken advantage of

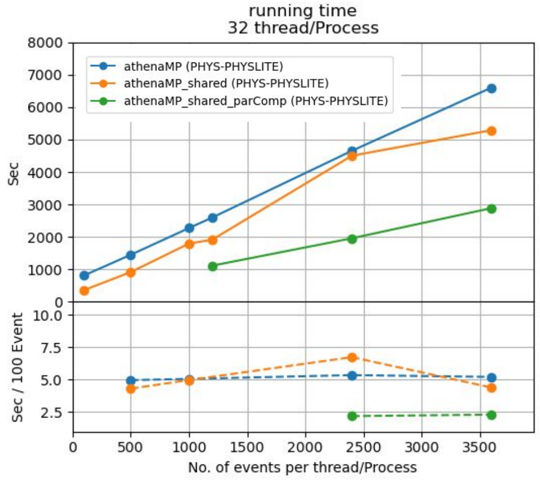
I/O and Storage: Recommendations

Work of the HEP-CCE/IOS team has resulted in

- Worthwhile insight to I/O behavior of HEP workflows
  - Including on HPC and for scales beyond current production.
- Fixes/enhancements to common software and experiments frameworks
  - Darshan included fork-safety and better filtering for I/O.
  - ROOT serialization bottleneck was fixed.
  - Patch to resolve the Athena library issue on DSO loading hooks which cause PyRoot crash when running with Darshan
- Prototype development of new functionality in collaboration with experiments:
  - ATLAS developed functionality to store their production data in HDF5





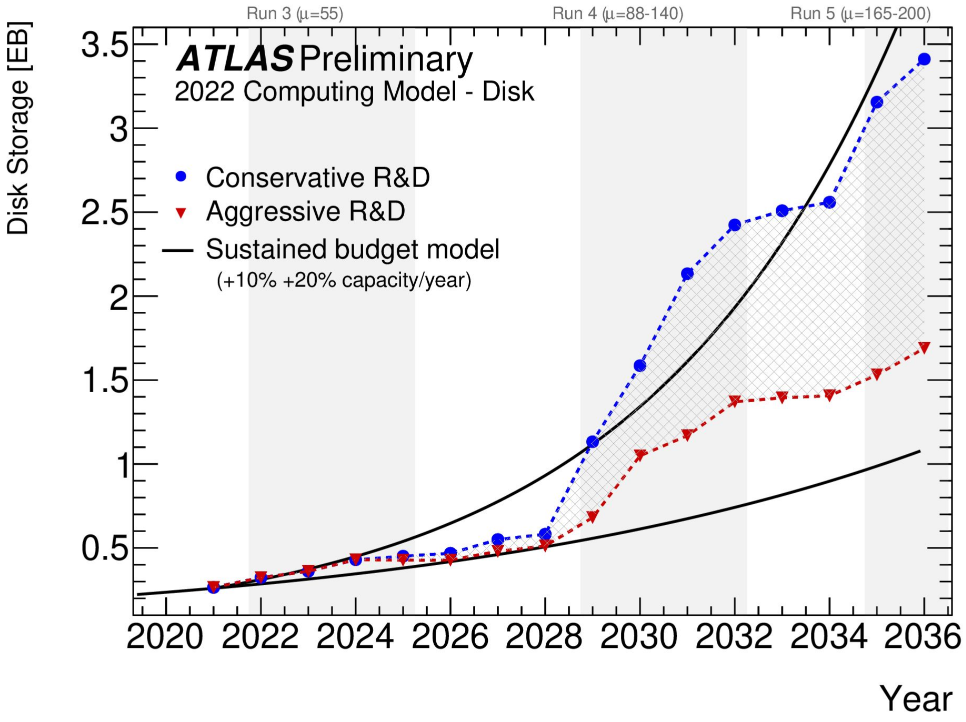



```
xAH_run.py --files InputFileList.txt
--inputList --nevents=0 --config
../sources/sh4b/config/minimal_commonCP.
py --daodPhys --submitDir
./sh4b-InputFileList-test --inputTag
*DAOD* --isMC --nevents=405000 direct
```

Job ID	19765
# Processes	1
Run time (s)	1318.6418

Darshan Monitoring of different ATLAS workflow steps

Now: Transitioning to HEP-CCE2, IOS -> Storage Optimization (SOP)



HEP-CCE/IOS: STORAGE AND COMPRESSION

Future Priorities

- The past cycle of HEP-CCE has been mainly focused on making HEP applications make [efficient] use of High Performance Computer
- This addresses the crucial need for **CPU cycles** expected for HEP experiment at the HL-LHC, DUNE and beyond.
- HEP, however, faces similar challenges for **disk and tape storage**, which also need to be addressed
  - Additional compute cycles may help, but won't solve this issue

HEP-CCE/IOS: STORAGE AND COMPRESSION

RNTuple, very brief, relevant experts are in the room.

- **RNTuple** part of **ROOT7** has been implemented in ATLAS/CMS for most derived production analysis products
  - Promises significant **storage savings** and I/O speed up
  - Limited Data Model support vs. TTree
    - Streamlined design of RNTuple will require leaner approach than TTree
- **HEP-CCE Role**:
- Adjustments to complex Simulation/Reconstruction Data Models
  - Development of techniques to hide complexity from persistence
    - Synergetic to HPC friendly data model work
  - Performance Testing and Optimization, e.g. using Darshan monitoring and I/O mimicking
- Consolidate requests for additional functionality to ROOT

Experiment status of RNTuple

ATLAS and CMS can store most derived analysis product in RNTuple

- Both experiments see **very significant storage reductions**
- Depending on:
  - Data Model/Data Product
  - Compression Algorithms
  - Storage Parameter
- Preliminary, numbers:
  - CMS nano-AOD: **30-40% reduction in file size**
  - ATLAS DAOD-PHYS: **20-30% reduction in file size**

On HL-LHC scale, these savings would correspond to ExaBytes of Disk and Tape

- Main contributors to space savings:
  1. More compact representation of collections and bools
  2. Data encoding optimized for better compression ratio (byte-splitting, delta encoding, etc.)

Data Model support of RNTuple

Streamlined RNTuple, will not support full C++ data models (as TTree does)

- To achieve better performance than TTree, RNTuple design choices made it more streamlined and reduced support for very complex data model features.

Complex production data models will need redesign

- HEP-CCE will:
  - provide generalized templates and guidelines for developing data models that can be stored in RNTuple
    - This effort is synergistic with the design of HPC-friendly data model
  - identify possible limitations and coordinate areas of improvement for RNTuple while it is still in the experimental stage

Type	Examples	EDM Coverage			RNTuple Status
PoD	`bool, int, float`	Flat n-tuple	Reduced AOD	Full AOD / RECO	Available
Vector<PoD>	`RVec<float>`				Available
String	`std::string`				Available
Nested vector	`RVec<RVec<float>>`				Available
User-defined classes	`"TEvent"`				Available
User-defined collections	`"TCudaVector"`				Available
`stdlib` collections	`std::map, std::tuple`				Avail. / Testing
Variadic types	`std::variant, std::unique_ptr`				Avail. / Testing
Intra-event references	`"&Electrons[7]"`				In design
Low-precision floating points	`Float16_t, Double32_t`	*Optimization benefitting all EDMs*			Testing
	Custom precision and range				In design
	Precision cascades ACAT'22				In design

Data model support by RNTuple.

Jakob Blomer, Philippe Canal, Axel Naumann, Javier Lopez-Gomez, Giovanna Lazzari Miotto, "ROOT's RNTuple I/O Subsystem: The Path to Production," CHEP, May 2023. https://indico.jlab.org/event/459/contributions/11594/attachments/9389/13620/rntuple-chep23.pdf

Thank you

This work was supported by the U.S. Department of Energy, Office of Science, Office of High Energy Physics, High Energy Physics Center for Computational Excellence (HEP-CCE). This research used resources at Argonne Leadership Computing Facility, NERSC and BNL Scientific Data and Computing Center.

U.S. DEPARTMENT OF ENERGY | Office of Science

Argonne NATIONAL LABORATORY

Brookhaven National Laboratory

Fermilab

BERKELEY LAB Bringing Science Solutions to the World