Argonne NATIONAL LABORATORY

HEP-CCE: **S**torage **OP**timization

Peter van Gemmeren (ANL)
On behalf of the HEP-CCE/SOP group

High Energy Physics- Center for Computational Excellence

- Started as a 3 year (2020-2023) Pilot Project now **Base Program**
  - 6 Experiments (**Energy**, **Intensity** and Cosmic Frontiers)
  - 5 US National Labs (ANL, BNL, FNAL, LBNL & Oak Ridge joined)
- Pilot Project of HEP-CCE:
  - Address one major issue: Deploying Leadership Computing Facilities (LCF) to help future HEP computing challenges (Processing Cycles)
  - Activities:
    - **P**ortable **P**arallelization **S**trategies for High-Performance Computing Systems
    - Fine-Grained **I/O** and **S**torage on HPC Platforms, including Data Models and Structures
      - Demonstrated the capability of leveraging parallel I/O libraries to write HEP data into HPC native backends like **HDF5** (CHEP23-Link)
      - Enhance I/O Characterization tool **Darshan** and monitor HEP workflows (CHEP23-Link)

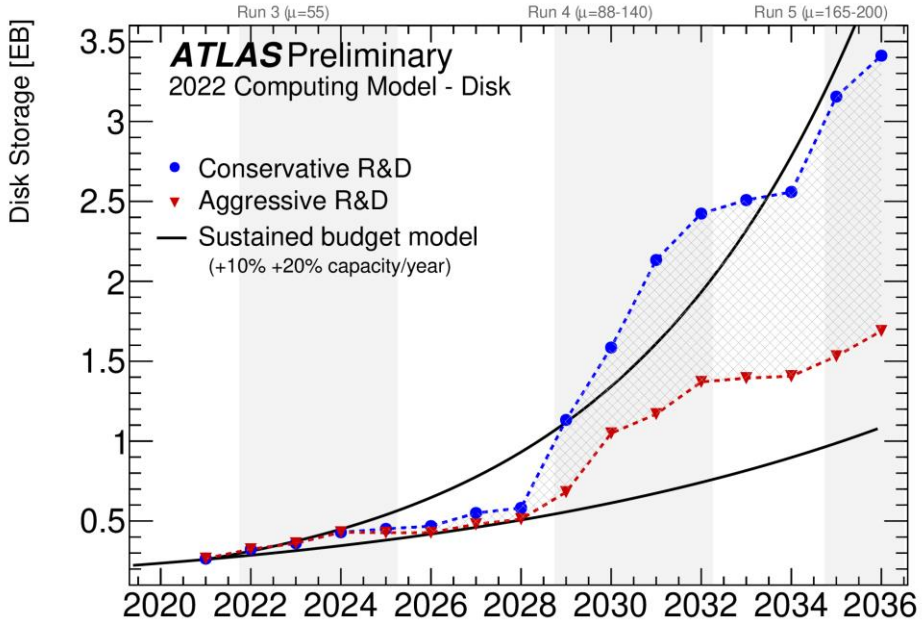
HEP-CCE2: IOS -> Storage Optimization SOP

After successful completion of Pilot Project and D.O.E. Review

HEP-CCE evolved as a Base Program and expanded its scope

Available **storage resources** can limit the physics reach of HL-LHC era experiments

- Optimizing Data Storage and Data Management
  - Investigate new storage backends and data volume reduction methods
    - Tracking and aiding the evolution of ROOT I/O, in particular **RNTuple**
    - Reduced Precision and Intelligent Domain-specific Compression Algorithms
    - **Object Stores** and Strategies for Data Placement and Replication
    - Optimized Data Delivery to HPC systems



ATLAS-Link

ROOT: From TTree to RNTuple

ROOT: HEP Community software used from data processing to physics analysis

- **TTree** as a storage backend that enables HEP experiments to use tools provided by ROOT ecosystem
  - Primary storage backend and I/O subroutine of HEP experiments for decades
  - Over Exabyte of data stored in TTree format
  - TTree evolved to address experimental needs and has been the backbone of HEP computational workflows
    - Now, supports persistence and I/O of complex experimental data
      - Decades of development made TTree outstanding in its support of C++ features
- However, TTree architecture predates recent overhaul in C++, modern programming paradigms and evolving computational landscape

RNTuple, and upcoming HEP experiments

RNTuple: New Storage backend in ROOT version 7

- State of the art, HEP community supported storage and I/O subsystem
  - Address storage & I/O requirements of upcoming HEP experiments
  - Streamlined compared to TTree, provides limited data model support
    - ATLAS and CMS report **20-40% saving in their storage** (CHEP23-Link)
  - Use of modern C++ standards
    - Adoption of smart pointers, better error handling mechanisms, modern C++ libraries
- HEP experiments have to adopt RNTuple to stay current with ROOT
  - Adopt to new RNTuple API
  - May have to change the data model to be persisted in RNTuple

HEP-CCE: Tracking and aiding the evolution of … RNTuple

HEP-CCE will aid HEP experiments to adopt RNTuple

- Co-organized RNTuple Workshop:
  RNTuple Format and Feature Assessment (6-7 November 2023) · Indico (cern.ch)
- HEP-CCE is conducting RNTuple API review:
  Special CCE-SOP tele-conference: RNTuple API Review Kick Off (February 28, 2024) · INDICO-FNAL (Indico)
  - Aid the development of RNTuple as per the experimental requirements
  - Find common guidelines and recipes for experiments frameworks and data models to migrate to RNTuple
- ATLAS participation beyond HEP-CCE funded experts.
  - E.g.: Amit Bashyal (ANL), Doug Benjamin (BNL), Marcin Nowak (BNL), Serhan Mete (ANL), Scott Snyder (BNL), Rui Wang (ANL), myself (ANL)

Note: Since we are among friends

It's probably true, that ATLAS is most advanced on RNTuple at this point.

- We (in principal) can write/read **all our production data** to RNTuple
- Result of prior decisions in our framework (including **Transient/Persistent Separation, APR**)

That does not mean we won't profit from HEP-CCE

- At this time implementation is **not fully optimal**
- Not all production modes (e.g. **multi-process, multithreaded**) are supported, efficiently
- There are **missing features**, e.g. Indexing and Friends and **functionality**, e.g. Merging and Metadata

ATLAS may be in the best position to steer future work on RNTuple

Reduced Precision and Intelligent Domain-specific Compression Algorithms

Most experiment HEP data is stored compressed format using lossless compression, lossy compression are less common

- To reduce storage requirements further, experiments and ROOT are investigating means of reduced-precision storage as much of the data is derived from measurements with inherent uncertainties
- For derived data, **not RAW**
  - Under study for ATLAS PHYSLITE data, Potential **storage savings ~20-30%**
- Need trust-building/safeguarding validators, but may enable keep information down-stream.

IOS team has surveyed different tools developed by computer scientists:

- Hybrid Learning Techniques for Scientific Data Reduction with **MGARD**
- Compression of Scientific Data with **SZ**
- Statistical Similarity for Data Compression with **IDEALEM**

Object Stores and Strategies for Data Placement and Replication

- Numerous potential advantages for using in HEP:
  - **Reference** rather than copy **upstream data**, saving space
  - Allow **fine-grained versioning**, avoiding replication of unchanged objects
  - Facilitate **user-driven data augmentation**, to subset of events
  - These methods of referencing save storage space
- Object storage activities on HPC side as well, e.g. Distributed Asynchronous Object Storage (**DAOS**)
  - DAOS is an object storage service developed for use on **persistent memory** technologies as a **very high performance** online storage layer
    - Data model includes both key:value objects and array objects
    - Array objects can be used to streamline storage of large multidimensional arrays with record addressability
    - Access can be via POSIX or directly via **custom API**

Object Stores, DAOS, and RNTuple

ROOT's RNTuple supports DAOS

- Decoupling of namespace operations from data read/write is natural for ROOT data.
- Similar to key–value storage where the key is a UUID, but specifically tuned for low latency / high bandwidth workloads

HEP-CCE is studying RNTuple DAOS implementation using **Darshan**

- Darshan already provides initial support for characterizing DAOS storage access
- Building on: IOS has successfully used Darshan for current HEP workflows using ROOT
- Aligns with, and will benefit from, other activities to understand and tune DAOS use by team members

Outlook

Since becoming base program, HEP-CCE can contribute to a wider variety of challenges, including storage.

Need to ensure to be relevant to our Clients, **the experiments**, such as ATLAS, DUNE, and CMS

In my belief, that is best done by working in close collaboration sharing **expertis**e.

Acknowledgement

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)