

# ATLAS S&C toward the HL-LHC challenges

HEP-CCE All Hands Meeting

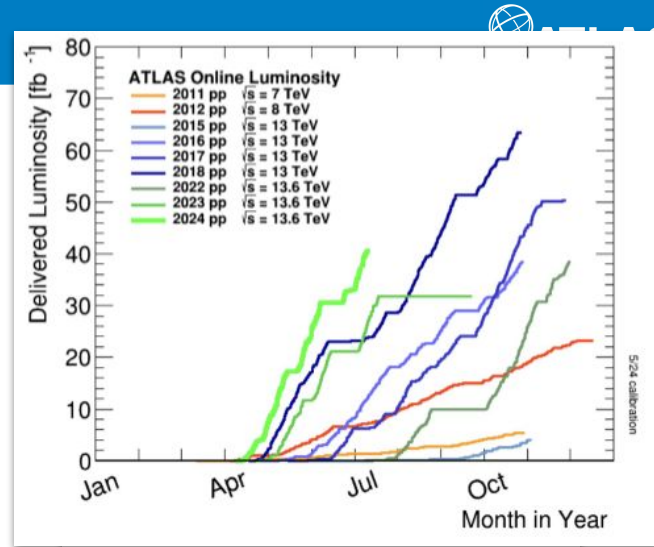
22th June 2024

Edward Moyse





# Introduction

- ATLAS is currently taking data for the LHC Run 3
  - We just published a [paper](#) on our Run3 software
- But in parallel, working hard to prepare for HL-LHC (2029, though discussion of a possible ~6 month delay)
  - Have previously mentioned our [roadmap](#) document
    - Work starting on TDR, planning to publish next year
  - Many technology demonstrators
    - See later...
  - Next gen trigger project is ramping up
    - <https://nextgentriggers.web.cern.ch/wp2/>



EUROPEAN ORGANISATION FOR NUCLEAR RESEARCH (CERN)

 Submitted to: EPJ C

 CERN-EP-2024-100  
10th April 2024

---

**Software and computing for Run 3 of the ATLAS experiment at the LHC**

The ATLAS Collaboration

The ATLAS experiment has developed extensive software and distributed computing systems for Run 3 of the LHC. These systems are described in detail, including software infrastructure and workflows, distributed data and workload management, database infrastructure, and validation. The use of these systems to prepare the data for physics analysis and assess its quality are described, along with the software tools used for data analysis itself. An outlook for the development of these projects towards Run 4 is also provided.

arXiv:2404.06335v1 [hep-ex] 9 Apr 2024

- ATLAS tracks progress with extensive list of milestones
- Have “Road to Run 4” meetings every ~6 months to check on progress

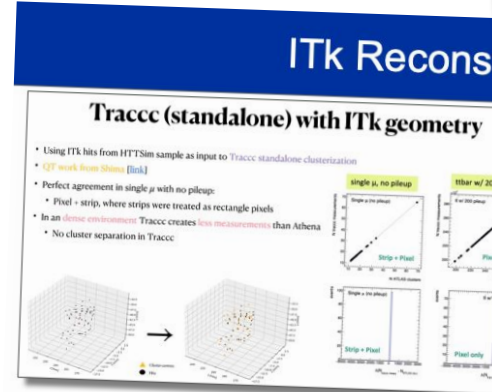
							September 2023	
	Category	Milestone	Deliverable	Description	Completion Date Baseline	Responsible	Current/Estimated Effort	Completion Date estimate
<b>Core Software, Heterogeneous Computing and Accelerators</b>								
B	CS	1		Pileup-digitization in AthenaMT production ready	12/2022	Beojan, John C., Tadej		12/2023
B	CS	1.1		Ensure reproducibility of MT production of presampled MB RDO files	6/2022	Beojan, John C., Tadej		6/2022
A	CS	2		Complete investigation of lossy compression techniques	12/2023	Serhan		12/2023
A	CS	2.1		<a href="#">Lossy compression of the ID track covariance matrix in the primary AODs</a>	12/2021	Thomas S.		12/2021
A	CS	2.2		<a href="#">Lossy compression of DAOD</a>	12/2021	James C.		12/2021
A	CS	2.3		<a href="#">Lossy compression of primary AODs</a>	12/2023	Peter vG		12/2023
B	CS	3		<a href="#">Implement I/O roadmap metadata recommendations</a>	12/2022	Peter vG, Vakho		12/2023
B	CS	3.1		Multi-threaded in-file metadata handling	6/2022	Maciej S.		6/2022
C	CS	3.2		Redesign of the metadata handling infrastructure (better support for fine-grained workflows)	12/2022	Maciej S.		12/2023
C	CS	3.3		Migrate current metadata framework to use single propagation tools per type	12/2024			
C	CS	3.4		Allow metadata tools to be executed by the SharedWriter via plug-in mechanism	12/2025			
C	CS	3.5		Evaluation of RNTuple for metadata storage and migration	12/2024			
A	CS	4		Evaluation of data formats well-suited for massively parallel I/O (HPCs)	3/2022	Peter vG		3/2022
A	CS	4.1		Storing intermediate EventService Simulation data in HDF5	3/2022	Peter vG		3/2022
B	CS	5		Migration to ROOT 7	12/2026	Marcin, Peter vG		6/2025

- ATLAS also has ~30 demonstrators
  - Idea is that these will help inform the TDR
- Most of these have been submitted to CHEP
- In the next slides, I will highlight a few...

Name / Brief Topic	Responsible	ATLAS Milestone(s)	Timeline	Google Doc Link
ACTS tracking: demonstrate that performance (physics, CPU) will be sufficient for HL-LHC needs	Noemi Calace, Andreas Salzburger	RE-4		<a href="http://cds.cern.ch/record/2845248">http://cds.cern.ch/record/2845248</a>
Performance advantage of running components of tracking algorithms on GPU and vs. of running end-to-end tracking on GPU	Antilia Krasznahorkay, Andreas Salzburger	RE-5	Next MS: Oct 22	<a href="https://docs.google.com/document/d/1TtP96bDc-8rZtOSsJNNVtj0cSG7pT1SwG2WYrG2z8edlPuspc/sharing">https://docs.google.com/document/d/1TtP96bDc-8rZtOSsJNNVtj0cSG7pT1SwG2WYrG2z8edlPuspc/sharing</a>
GNP Pipeline for Particle Tracking	Pablo Calafura, Jan Stark	RE-6		<a href="#">Doc</a>
Remote read of selected data vs copy of whole file	Ilija Vukotic, Paul Nilsson	DC-11 (relevant?)		
RNtuple fully capable of storing DAOD-PHYSLITE with advantages on storage footprint and I/O performance	Marcin Nowak	CS-5.1, AN-2		<a href="https://docs.google.com/document/d/1MeaB3mw82UD7ntmWYQM3vN0Ou_nkX-7DP6S8WtH0c/edit#">https://docs.google.com/document/d/1MeaB3mw82UD7ntmWYQM3vN0Ou_nkX-7DP6S8WtH0c/edit#</a>
Demonstrate that lossy compression preserves physics performance	Serhan Mete	CS-2		<a href="#">Lossy compression preserves physics performance Demonstrator - Google Docs</a>
Broaden DAOD-PHYS usability while limiting size increase for analysis by adding custom event augmentation.	Peter van Gemmeren	AN-5.4b		<a href="#">Broaden DAOD-PHYS usability by adding custom event augmentation performance Demonstrator - Google Docs</a>
Stress-test xcache by running many (or all) the US analysis jobs at a single site and have all inputs done through xcache.	Ilija Vukotic	DC-11 (kind of)		
HPX for scheduling tasks in distributed heterogeneous environment	Beojan Stanislaus, Julien Esseiva, Vahko Tsulaia	CS-13.1		<a href="#">Vertically Integrated Next Generation Task Scheduler for the Athena Framework - Google Docs</a>
Process data from multiple events on accelerators (event batching)	Beojan Stanislaus, Charles Leggett	CS-10.1		<a href="#">Event Batching Demonstrator</a>
FastChain demonstrator	FastChain group	SI-6		<a href="#">FastChain demonstrator doc, this is written as three demonstrators</a>
Adaptive data placement/job scheduling using Intel's Loihi2 neuromorphic computing platform	Ilija Vukotic	DC - no milestone		
Use ATLAS-Google site as "bursty" resources (demonstrate ATLAS bulk campaign [reprocessing or MC production] on this site)	Fernando Barreiro, Johannes Elmsheuser, Alexei Klimentov, Tadashi Maeno	DC-10.2		<a href="#">GoogleDoc</a>
ATLAS-Google site for analysis workflows and new technologies for physics analysis	Fernando Barreiro, Alexei Klimentov, Tadashi Maeno	DC-10		<b>ON HOLD</b>
ARM based PanDA queue setup and operation for Athena porting and Physics Validation	Fernando Barreiro, Johannes Elmsheuser	DC-6.1		<a href="#">GoogleDoc</a>
Recreate DAOD datasets on demand by PanDA using Data Carousel (Delete DAOD datasets from the lifetime model exception list and reproduce them on demand if needed)	Alexei Klimentov, Tadashi Maeno, Xn Zhao	New for DC-7 (was 11)		<a href="#">Recreate DAOD on demand demonstrator</a>
Using Xrootd to seamlessly integrate S3 storage with ATLAS DDM system, including Google and Amazon cloud S3 storage.	Andy Hanushevsky, Wei Yang	DC-10?	23Q4	
Large-scale AVMIL services across diverse resources, complex analysis workflows	Tadashi Maeno	DC - no milestone		<a href="#">AVMIL Services Demonstrator</a>

# Demonstrators(2)

- [Tracc](#) is the ACTS demonstrator for Tracking on accelerators
- Can run on the “OpenDataDetector” and gives reasonable results (even without optimisation)
- ... and also now on ATLAS’s ITk (HLK-LHC tracker)



A number of things remain to be improved / fixed, and will have to switch to the “fully-fledged” ITk geometry. But it’s already ~working.

### tracc

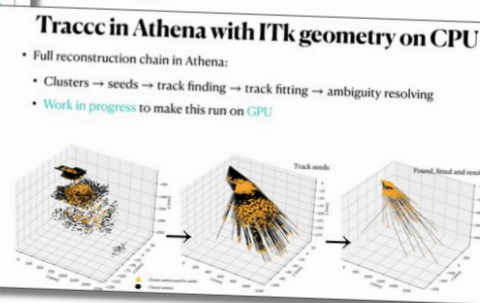
Demonstrator tracking chain for accelerators.

#### Features

Category	Algorithms	CPU	CUDA	SYCL	Alpaka	Kokkos	Futhark
Clusterization	CCL / FastSV / etc.	✓	✓	✓	●	●	✓
	Measurement creation	✓	✓	✓	●	●	✓
Seeding	Spacepoint formation	✓	✓	✓	●	●	●
	Spacepoint binning	✓	✓	✓	✓	✓	●
	Seed finding	✓	✓	✓	✓	✓	●
Track finding	Track param estimation	✓	✓	✓	✓	●	●
	Combinatorial KF	✓	✓	●	●	●	●
Track fitting	KF	✓	✓	✓	●	●	●
Ambiguity resolution	Greedy resolver	✓	●	●	●	●	●

✓: exists, ●: work started, ○: work not started yet

The relations between datatypes and algorithms is given in the (approximately commutative) diagram shown below. Black lines indicate CPU algorithms, green lines indicate CUDA algorithms, blue lines indicate SYCL algorithms, and brown lines indicate Futhark algorithms. Solid algorithms are ready for use, dashed algorithms are in development or future goals. Data types for different heterogeneous platforms are contracted for legibility, and identities are hidden.



## RNTuple demonstrator

- ATLAS has RNTuple support for all official formats, e.g. HITS, RDO, ESD, AOD, DAOD etc
  - [ACAT presentation](#)

## Event Augmentation

- Adds ability to add custom data for subset of events, for particular analysis
- On demand reading, in order not to impact other workflow
  - [ATL-SOFT-PROC-2023-003](#)

## GPU-based TopoCluster reconstruction

- Topo Cluster is one of the most resource demanding algorithms for HLTCalo in Run 2/3
  - [doi:10.1088/1742-6596/2438/1/012044](https://doi.org/10.1088/1742-6596/2438/1/012044)

### RNTuple fully capable of storing DAOD-PHYSLITE Demonstrator

#### Contributors

The ATLAS I/O team, lead Marcin Nowak (BNL), Serhan Metz (ANL), Peter van Gemmeren (ANL) plus - Doug Benjamin (BNL-SDCC), + likely others from BNL-SDCC, CERN IT/ ROOT RNTuple dev team: Jakob Blomer, Javier Lopez Gomez, Florine de Geus, Vincenzo Padulano

#### Description

With the next major ROOT release...

changes

success

IT/Tr

at the

RNTuple

manne

ROOT

factor of

20%"

only pro

The prin

team, esp

especially

but not li

benchma

the ROOT

long-term

RNTuple

technolog

stores (Cap

the team

### Broaden DAOD-PHYS/PHYSLITE usability by adding custom event augmentation performance Demonstrator

#### Contributors:

The I/O team, lead Peter van Gemmeren, plus experts from AMO (Nils Krummack, Lukas Heinrich, Jackson Burzynski...) and Derivation (James Cammer...) plus thd for data management support

#### Description

We demonstrate the capability to augment standard production DAOD-PHYS (or PHYSLITE) with custom data for a subset of events. Augmentations are only added to the subset of events that are of interest to the analysis use-case requesting them, thereby minimizing storage impact. On demand reading of these augmentations will ensure there is virtually no other performance impact on workflows using the regular content of the production DAOD. For HL-LHC, we propose developing functionality to store these augmentations in separate files, increasing flexibility on where and how long this data can be stored.

#### Program of work:

ATLAS Trigger Software / ATR-26697

Trigger Calorimeter Cluster Building for HLTCalo EF (Phase II) in a

Edit Add comment Assign More In Progress

#### Details

Type:

Epic

Priority:

Medium

Affects Version/s:

None

Component/s:

HLT Calo

Labels:

phase-II

Epic Name:

Trigger Calorimeter Cluster Building for HLTCalo EF (Phase II) in a GPU

#### Description

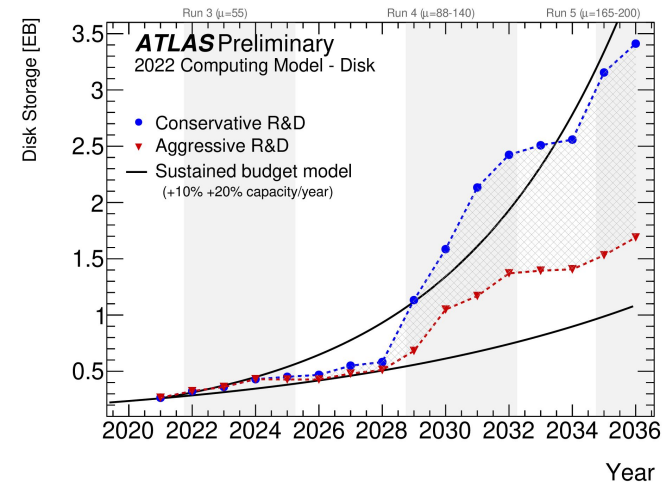
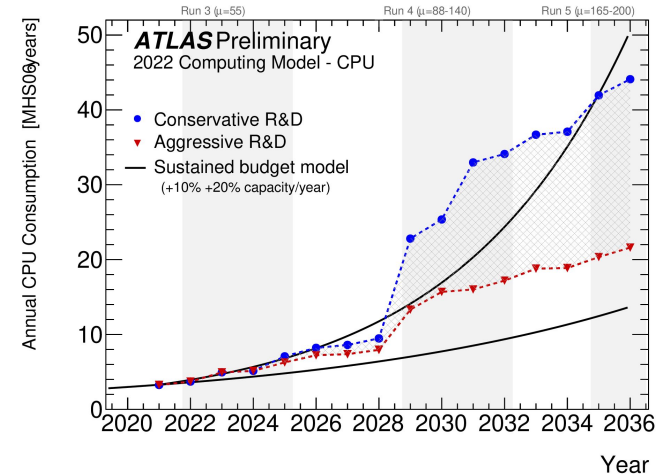
As described in the Twiki (<https://twiki.cern.ch/twiki/bin/view/Atlas/HLTcaloGPUFGA>), ATLAS is pursuing a proper development of the Topological Clustering Algorithm in technologies other than pure CPU. This EPIC unites the different parts of the task. JIRA development tickets here will take care of specific parts:

- Development of the TrigCaloCluster Maker (1st/3rd of the builder).
- Tests to perform comparisons at this first step and a quick how to run them in the lxplus GPU cluster.
- Comparisons between CPU and GPU clusters.
- Documentation about the implementation of this first step.
- Development of the TrigCaloCluster Splitter (2nd/3rd of the builder).
- Tests to perform comparisons at this second step and a quick how to run them in the lxplus GPU cluster. Comparisons between CPU and GPU clusters.
- Documentation about the implementation of this second step.
- Development of the TrigCaloCluster Moment Maker (3rd/3rd of the builder).
- Tests to perform comparisons at this third and a quick how to run them in the lxplus GPU cluster. Comparisons between CPU and GPU clusters.
- Documentation about the implementation of this third step.

More details about the Topo Cluster algorithm can be found here: <https://twiki.cern.ch/twiki/pub/Atlas/HLTcaloGPUFGA/HLTcaloEF20231110.pdf>

More details about the Noise Thresholding can be found here: <https://twiki.cern.ch/twiki/pub/Atlas/HLTcaloGPUFGA/HLTcaloEF20231110.pdf>


- Hopefully these computing model plots are familiar now!
- Will be updating these soon™, adding GPUs for the first time
  - Requires estimates from each area for how much work they think can be done on a GPU
  - Requires model for incorporating GPUs into the resource estimates



## General






- ATLAS is grateful for HEP-CCE's awareness of our timeline and needs
  - e.g. in Paolo's [talk](#) at a recent ATLAS S&C week, he linked our R2R4 milestones to HEP-CCE goals
- Focus on usability, and real-world applications is great!

**HEP-CCE and ATLAS Road 2 Run 4**



Many areas of active or potential collaboration (refer to [R2R4 milestones](#))

	R2R4 milestone	CCE activity
RE	3 New Reconstruction EDM	EDM for heterogeneous computing and columnar data formats
RE	5 Accelerator and machine learning (R&D)	Inference + ACTS/traccs as a Service, Scaling up training and inference for large models (GNNs, transformers)
EG	1 Assessment of CPU costs using Run3 MC setup	by Run 4, CCE MadGraph on GPU, + Pepper NLO can help move some of
EG	2 Development of nominal MC setups	EvGen workloads to GPU systems

 Office of Science
 






## PAW

- Good example of activity that is very nicely focused on real world applications
- Tasks and plan for next year(s) looks both interesting and feasible/realistic
- “Our” Tracc is getting closer to production quality.
  - We think that this might be good to add this as a mini-app?
- We also support the idea of using an ATLAS HPC workflow
  - We regularly run on HPC systems, so we know this works well!
  - **Update:** I gather from Charles’ [slides](#) earlier this is happening? Great!

## HEP-CCE Phase 2 Plan

HEP-CCE

- In HEP-CCE Phase 2, our goal is to provide the experiments with both a validated, ready-to-use portability solution and a suite of portability tools that can be integrated into their production systems.
  - To reconcile different services and tools provided by HEP and ASCR.
  - To reduce the operation and maintenance overhead of deploying HEP workflows on HPC systems
- Building on the experience of **PPS** and **CW** groups in HEP-CCE Phase I, we will have two main tasks in Phase 2:
  - **Task 1:** apply lessons learned in PPS to help HEP experiments develop portability solutions in their applications
  - **Task 2:** develop portable, experiment-agnostic, workflow overlays to interface existing HEP workflows with HPC centers

10



## Year 1 Plan

HEP-CCE

### Task 1 - Application Portability:

- Develop a **cookbook** for portability layers based on Phase 1 findings
- Outreach to experiments for portable solution implementation (**workshops/hackathons**, followed by regular office hours)
  - Understand the experiments’ timescales for portable accelerator uses
- Create **mini-apps** based on two of the Phase I PPS testbeds that can be executed at NERSC, OLCF and ALCF, preferably with the same software environment (FCS, p2)
- Use mini-apps to extract **figures of merit** for ASCR facilities and LCFs to use as baselines

### Task 2 - Workflow Portability:

- Complete **survey** of existing HEP experiment workflow technologies on HPC; also look into workflow technologies used by **other experiment facilities such as light sources**.
  - Find commonalities between experiment workflow systems
- Explore the needs of HEP in terms of **ML workflows/pipelines and microservices** (synergistic with the distributed ML activity)
- Investigate common layers and interfaces (batch scheduler, policies, pilots, ...) to facilitate portability and interoperability across ASCR facilities in collaboration with **IRI testbeds**
- Create **2 representative HEP experiment workflows** to run two different HPC systems. Candidates include: LSST/DESC, LZ, **DUNE**, LHC Experiments (ATLAS/CMS).

13



[Charles's talk at BNL workshop](#)

## SOP

- Again, effort seems nicely focussed on real-world practicality
  - i.e. actionable results
- ATLAS is expecting RNTuple to be vitally important in Run-4, so we strongly support the SOP efforts here e.g.
  - RNTuple [workshop](#)
  - RNTuple API [review](#)
- Interested to see the outcome of other investigations, e.g. metadata, compression and object stores

### Argonne **Persistifying the Complex Event Data Model of the ATLAS Experiment in RNTuple**



Alaettin Serhan Mete (Argonne), Marcin Nowak (Brookhaven), Peter Van Gemmeren (Argonne)

- ATLAS has been using ROOT's TTree storage backend for about two decades
- In LHC Run 4 (2029), ROOT's main I/O subsystem will be RNTuple
  - In a nutshell, a more modern and (compute and storage-wise) efficient technology
- ATLAS has made significant progress for adopting RNTuple for its event data
  - **All applicable ATLAS data formats can be written into RNTuple seamlessly**
  - Both reading and writing are supported on the official software framework (Athena) side
  - Everything is handled by the I/O infrastructure with no change needed for the client code
- Preliminary estimates suggest **20+% storage savings** in some analysis formats
- Getting production-ready still needs a number of key milestones reached:
  - Finalizing/adopting a number of in-progress RNTuple work, e.g., fast merging etc.
  - Updating standalone tools used by the production system for metadata access, file validation etc.
  - Running large-scale stress tests and performing detailed validation studies
- ATLAS will use the rest of Run 3 and the Long Shutdown 3 to deliver these!

HEP-CCE

### Compression Framework

Working on a **test framework** that generates (or takes as input) raw data, applies intelligent compression using the tools mentioned above and writes compressed data (including in RNTuple).

Perform tests to **measure fidelity and usability** of the compressed data downstream (raw data remains loss-less).

- DUNE FD Trigger raw data will be large waveform (~ few GBs) where some of the above tools could perform well in terms of compression and data fidelity
  - Faster and easier way to inspect data than accessing original raw data
  - Could be useful in some further **ML based analyses** that could be compute intensive but does **not require full precision of the data**
  - Needs collaboration with DUNE stakeholders, compression developers and the ROOT team

## SIM

- Not so obvious what has been happening in this area
  - Perhaps the work could be better publicised?
  - (I heard from Julien Esseiva about work on optical physics/Orange)
- ATLAS has been working on Celeritas integration
  - Planning [hackathon](#) in October

## SML

- Recent [meeting](#) discussing Simulation based inference in ATLAS
  - (“ATLAS Data Analysis using a Parallel Workflow on Distributed Cloud-based Services with GPUs” [paper](#))
- Given expected increasing importance of ML in Run4, of definite interest to ATLAS

## Introduction

Celeritas provides a GPU accelerated simulation of EM showers

- Recent report [here](#)

Several presentations were given on Celeritas, which has the ability to run simulation of complex geometries and was validated against G4

- CMS Run-3 geometry (2018)
- TileCal test beam
- Full ATLAS recent hold up due to EMEC custom solid (now resolved??)

ATLAS use of Celeritas either

- Standalone application (geometry passed via gdmf)
- Integration with FullSimLight (demonstrated about 1 year ago)

To benefit from Celeritas we need to transfer hits to the Sensitive Detectors

- Store them and pass them on to the rest of the chain
- Requires athena integration
- Note: Similar requirements also for AdePT, a common interface under development (Ben, link to github [CelerAdePT](#))

ATLAS Simulation meeting

Celeritas integration

Davide Costanzo <sup>2</sup>

[Slides from Davide Costanzo](#) (in ATLAS simulation meeting)

ATLAS strongly supports the HEP-CCE efforts

Some areas apparently already very active, some .. there is apparently work ongoing, but not-so well publicised (regular summary talks are very useful).

We look forward to seeing