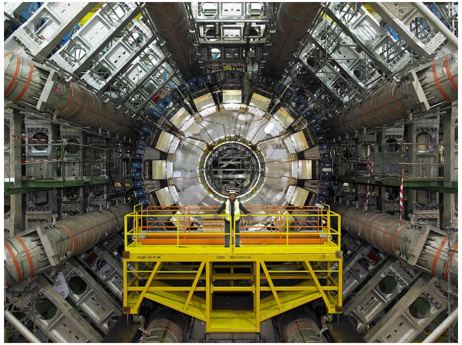
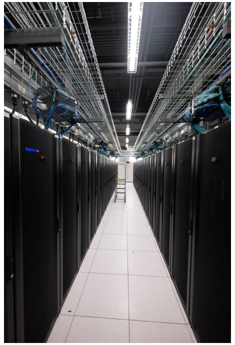
Analysis Infrastructure in ATLAS - LHCC Questions





The ATLAS Collaboration

Gordon Watts (U Washington)

Emma Torro (Valencia)

Status

Presentation to the LHCC isn't until 1st week of March

This is a work in progress

Brief History

ATLAS EXPERIMENT

1. LHCC asked the WLCG to come up with a series of review questions touching Analysis Facilities

2. WLCG solicited the questions from the experiments, and discussed the merged questions at the May Hamburg WLCG/HSF workshop.

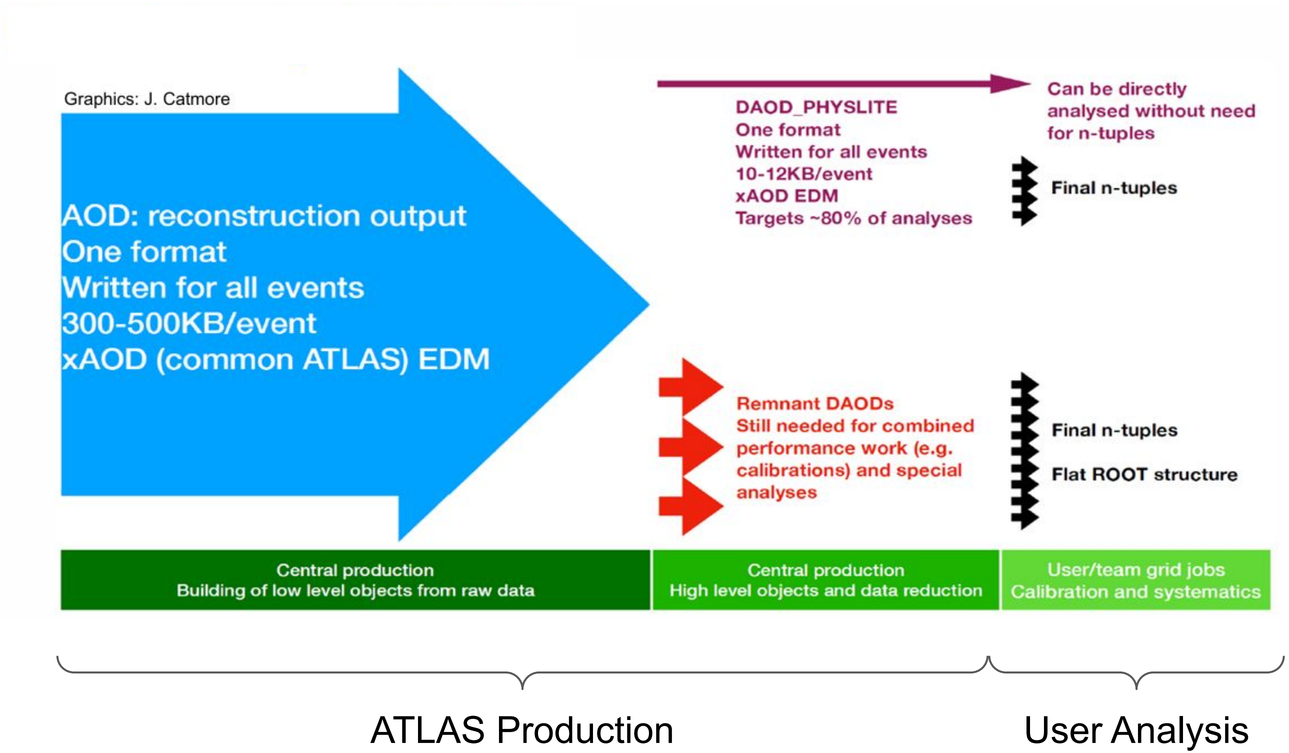
3. After editing and shortening them significantly, the LHCC asked each LHC experiment to present the answers at the March LHCC meeting

4. These are a draft of these answers from ATLAS. The goal of this session is to be coherent with CMS where possible, and understand differences where not.

1. Current Run 3 analysis model

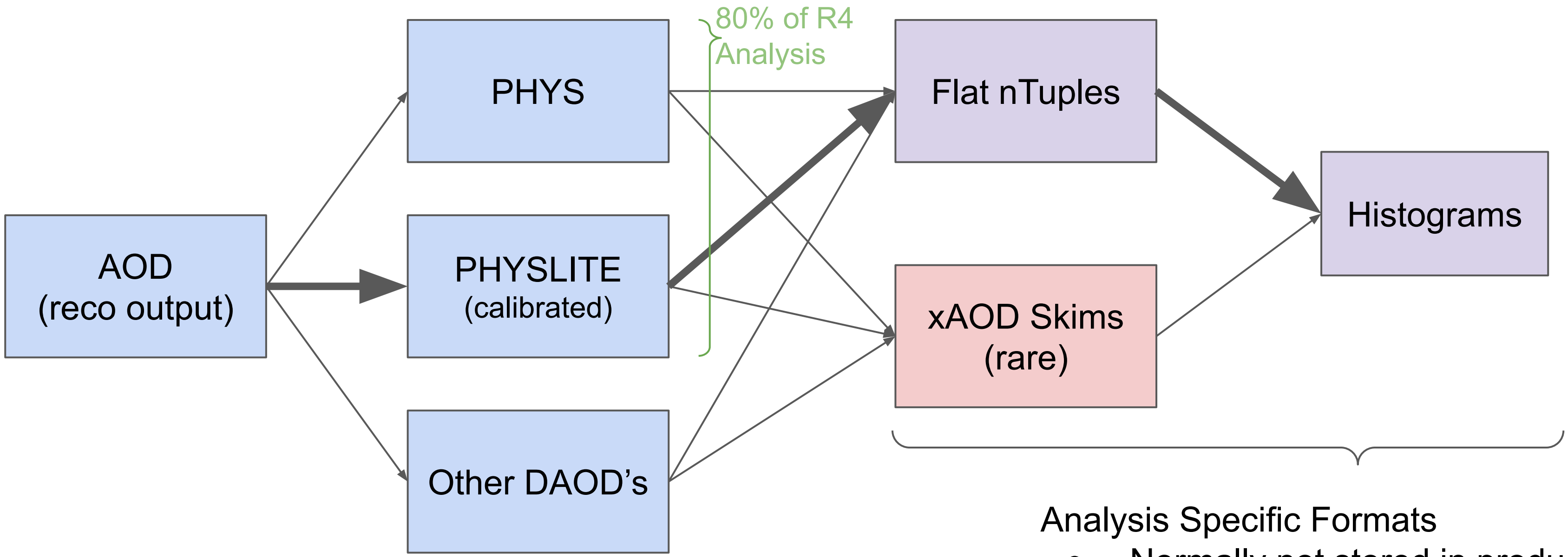
a) Main analysis workflows and data reduction steps, including how closely chained they need to be

Production Workflow



*J. Schaarschmidt

a) Main analysis workflows and data reduction steps, including how closely chained they need to be

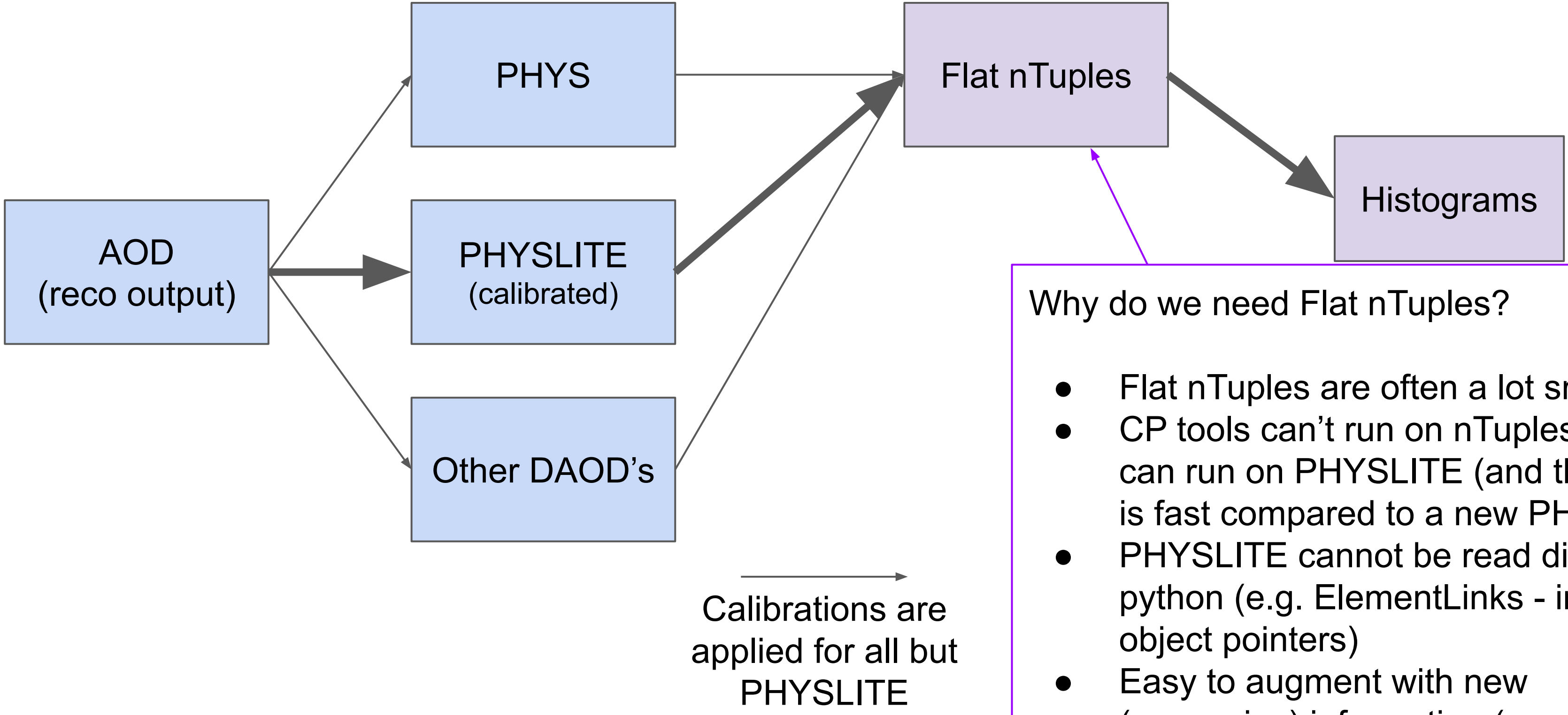


Analysis Specific Formats
- Normally not stored in production systems
- Some analysis group ntuples are large - but most small ("dark")

Calibrations are applied for all but PHYSLITE

All data is written in a **xAOD container**
- TTree format on Disk
- RNTuple capable now
- Can be read by uproot or RDataFrame, if carefully written

a) Main analysis workflows and data reduction steps, including how closely chained they need to be

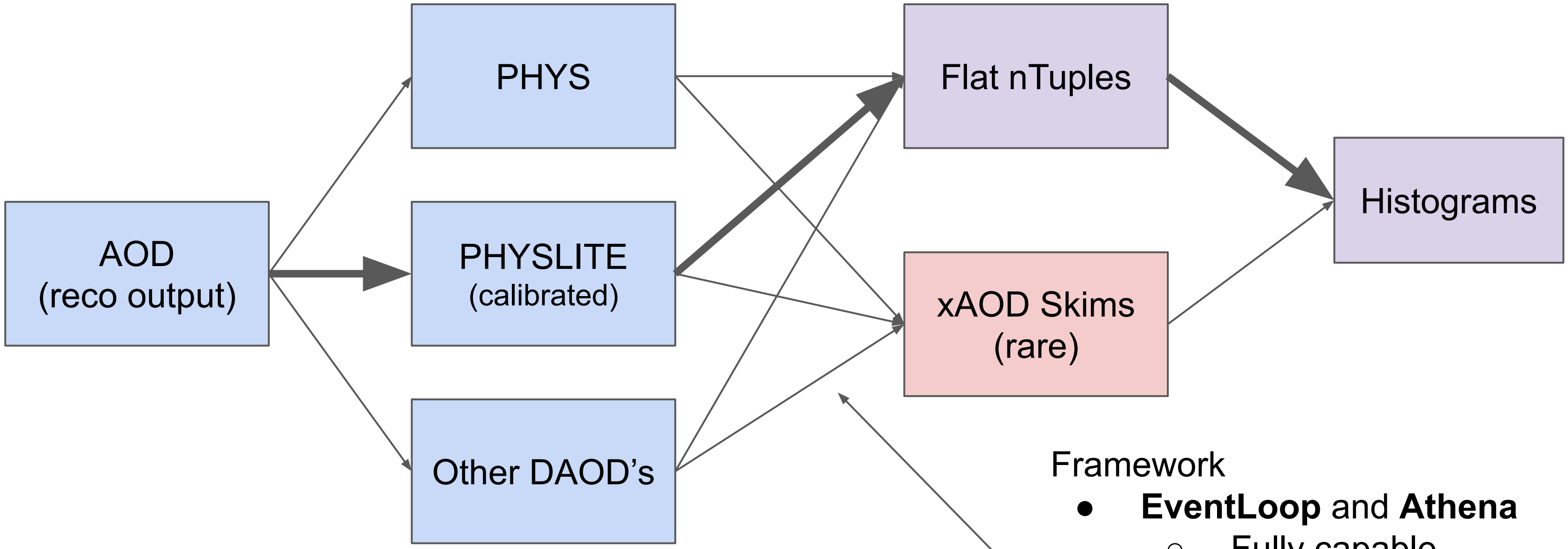


Why do we need Flat nTuples?

- Flat nTuples are often a lot smaller
- CP tools can’t run on nTuples, but can run on PHYSLITE (and that step is fast compared to a new PHYSITE)
- PHYSLITE cannot be read directly in python (e.g. ElementLinks - in-file object pointers)
- Easy to augment with new (expensive) information (e.g. ML inference)

1. Current Run 3 analysis model

a) Main analysis workflows and data reduction steps, including how closely chained they need to be



Calibrations are applied for all but PHYSLITE

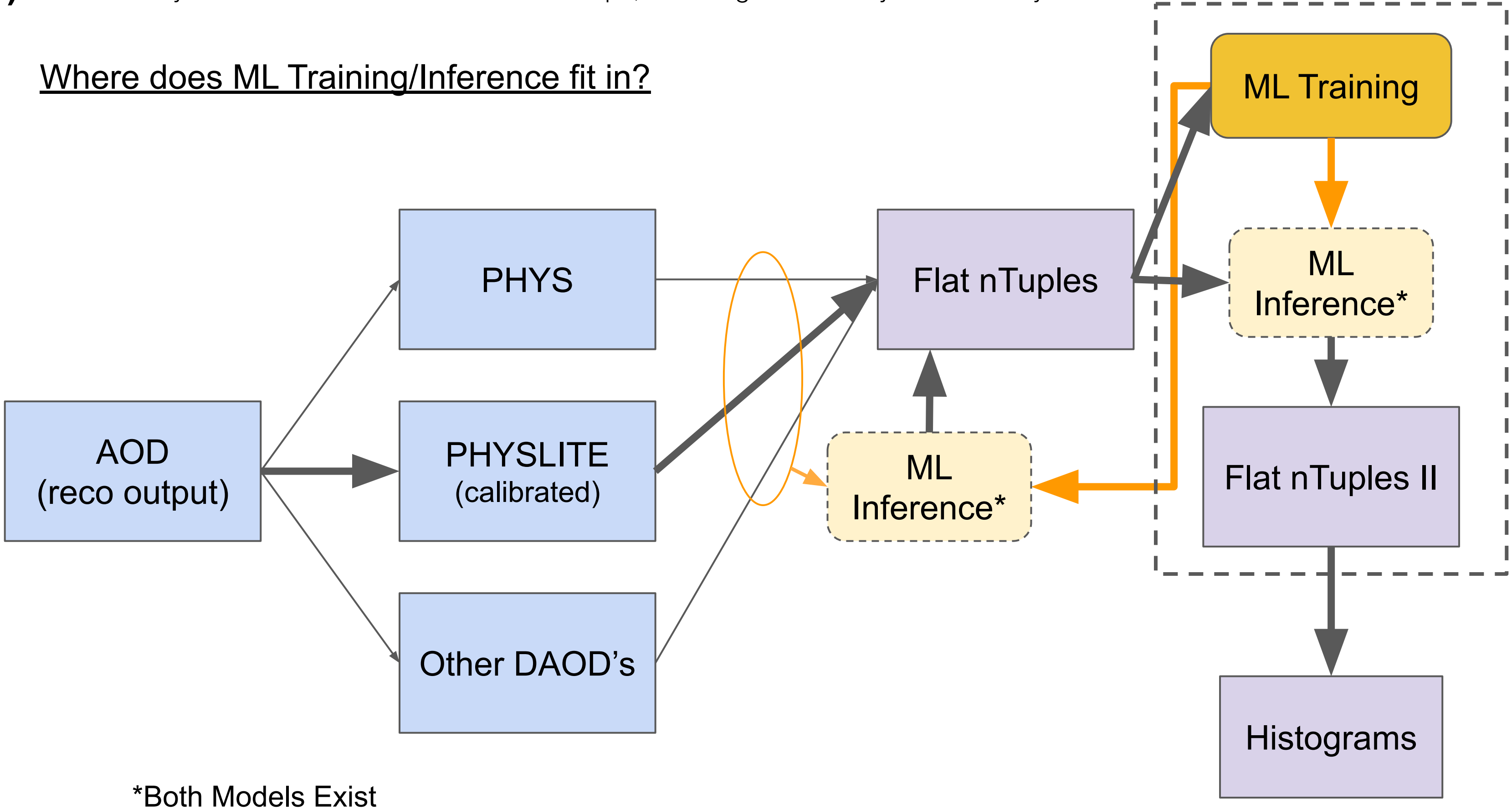
Framework

- **EventLoop** and **Athena**
  - Fully capable
- **Columnar** (python or RDF)
  - Still integrating calibrations and systematic errors, etc.

1. Current Run 3 analysis model

a) Main analysis workflows and data reduction steps, including how closely chained they need to be

<u>Where does ML Training/Inference fit in?</u>



*Both Models Exist

1. Current Run 3 analysis model

b) Data formats used for analysis, including their size and level of adoption (current and Run-3 final goal)

Run 3

Run 4

AOD → DAOD_PHYS → nTuples → Histograms

AOD → DAOD_PHYSLITE → nTuples / Histograms

AOD

- Size: ~400 (~500) KB/event for data (MC)
- Expected total size: ~120 PB in Run-3
- Usage:
  - Input for the derivation step
  - Input for calibrations calculation

DAOD_PHYS

- Size: ~25 (~35) KB/event for data (MC)
- Expected total size: ~30 PB in Run-3 (2 versions, 1 replica)
- Usage:
  - Main analysis format for Run-3
- Adoption: currently the most popular analysis format in Run-3

DAOD_PHYSLITE

- Size: ~9 (~13) KB/event for data (MC)
- Expected total size: ~12 PB in Run-3 (2 versions, 1 replica)
- Usage:
  - Lightweight format intended for Run-4
- Adoption: second most popular format, with increasing adoption as workflows adapt

nTuples

- Size: from few MB to TB per analysis, depending on the complexity and scale of the datasets used.
- Usage:
  - Final step before statistical interpretation.
  - Machine Learning (ML) workflows use nTuples for training, validation, and evaluation

Additional DAOD formats

- PHYS(LITE) are aggressively thinned
- some analyses and CP groups are forced to use a format with fewer events but more content per event

c) How much compute, storage and network **resources** are used for **Run-3 analysis**. Which fraction are pledged and which fraction are used in interactive mode (as opposed to batch)

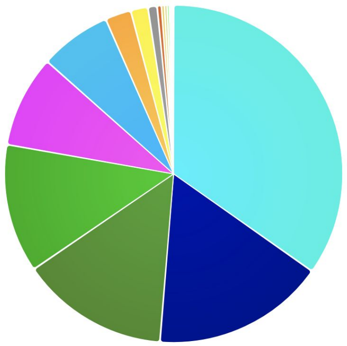
Compute resources:

a) ~10% of ATLAS CPU cycles are used for analysis workflows
   i) This is an educated guess: production facilities, T1 and T2's are monitored but T3's are not!
   ii) Analysis has heavy interaction with data management/caches/storage:
      (1) accesses 69% of the files on the grid
      (2) Produces 45% of the files
      (3) Workflows can be cyclical unlike production

b) Data-intensive workflows and large-scale processing (reconstruction, derivations) are done centrally: run on the GRID

c) High utilization of interactive systems for initial data exploration.

Wall clock time



	Percent
MC Simulation Full	35%
MC Reconstruction	17%
MC Event Generation	14%
Group Production	12%
User Analysis	9%
MC Simulation Fast	7%
t0_processing	2%

1. Current Run 3 analysis model

c) How much compute, storage and network resources are used for Run-3 analysis. Which fraction are pledged and which fraction are used in interactive mode (as opposed to batch)

Storage resources:

a) AODs stored in T1, T2's – need access for some analysis work
b) Total data volume for analysis formats (DAOD) is ~30% of total storage:
   i) Total of PHYS + PHYSLITE + DAOD's is 79 PB on disks
   ii) Caching can help as well, though hasn't been successful in significantly reducing disk resource pressure.

Network resources:

We are evaluating this.
We do have expectations motivated by the WLCG Data Challenge.

Pledged vs. Interactive Usage:

c) highly variable, no good monitoring handles. Estimation: Interactive usage is ~10% of total compute
d) Majority of GRID/batch resources are pledged and managed via WLCG
e) Local batch systems and interactive environments often used for smaller or iterative workflows

d) Comment on **what is working well and what is not**, both from the point of view of users as well as providers (experiment S&C teams and sites).

- a) **What Works Well**
  - i) **For Users:**
    - (1) Streamlined workflows supported by robust DAOD formats, including centralized production
      - (a) Standardization efforts like DAOD_PHYSLITE and software frameworks
        - (i) reduce disk footprints and improve efficiency for storing systematic variations.
        - (ii) Common documentation, and bug fixes due to larger numbers of people looking at the same data (vs prior system with lots of formats).
    - (2) Centralized software tools (Athena, EventLoop, columnar, rucio) ensure consistency.
    - (3) Analysis frameworks such as TopCPToolKit in specific Working Groups streamline workflows
      - (a) The frameworks are supported by dedicated teams whose goal is to make it easy for the user
  - ii) **For Providers:**
    - (1) Mature distributed computing infrastructure (PanDA+Rucio and connections with WLCG).
    - (2) Effective use of data carousel and rucio for storage optimization.

1. Current Run 3 analysis model

d) Comment on what is working well and what is not, both from the point of view of users as well as providers (experiment S&C teams and sites).

- a) **Challenges**
  - i) **For Users:**
    - (1) New workflows imply a learning curve. Could be improved with the use of columnar-based flows that analysts already use like NumPy
    - (2) New workflows requires new tools that are not always supported by current interactive facilities
      - (a) dask, for example.
    - (3) Long and fragmented workflows. Too many different frameworks.
      - (a) ATLAS is encouraging the groups to move to only a few supported ones
      - (b) Current schemes for storing systematic variations do not scale well
    - (4) Limited resources for interactive analysis, especially at peak times (conferences).
    - (5) Some of the documentation and tutorials can be improved
    - (6) Handling GRID job failures at user level
      - (a) Improving error diagnostics would reduce delays and enhance reliability
    - (7) Easy access to GPU's for training, ML design, and hyper parameter tuning
  - ii) **For Providers:**
    - (1) Balancing resource allocation between analysis and other workflows (e.g., Monte Carlo simulation)

2. Future analysis model in Run 4 and Run 5

a) *Comment on which aspects of the current Run-3 analysis model* ***will not scale for Run-4***

- **a) Frameworks**
  - *i)* O(10)x increase in data will make some (currently used) end-user analysis patterns take too long
  - ii) Currently we have many analysis frameworks being used. Need to move to support only a few
  - iii) Need standardised workflows for ML training and inference in Athena. ONNX is likely the best way forward
  - iv) Access to the Cross-section DB in ATLAS is mostly limited to only C++. Would be good to add python interface
  - v) ElementLinks and similar functionality are not useful/flexible enough to support the EventLoop paradigm and the columnar paradigm
- **b) Systematics**
  - *i)* currently their calculation can require an excessively large amount of disk space
- **c) Size**
  - *i) Full PHYSLITE production is getting large.*
    - (1) This will reduce the number of production campaigns per year
    - *(2)* Will probably need a common way to skim PHYSLITE to run it locally
    - (3) More effort to be invested in reducing the size of PHYSLITE.
  - *ii)* Data placement on TAPE has to be improved
  - iii) The number of residual DAOD formats (other than PHYS/LITE) should be kepts small
- *d)* **Documentation** *can always be improved*

2. Future analysis model in Run 4 and Run 5

b) *Describe the relevant **changes in the model and their impact** in resources: policies for number of versions and replicas, fraction of data which is managed vs. unmanaged (e.g. caches), remote vs. local data access, batch vs. interactive cpu/gpu access, need of access to external DBs, or any other*

Computing Model for Run 3, 4, 5, …

- All of our expectation plots are made with this
- Usage assumptions and full simulation are maintained in gitlab
- Does not include any analysis assumptions - just production.
- But it does model numbers of:
    - Replicas: 2.4 for current year, 1 for others *
    - Versions: 2 *

Caching Policy

- Currently ATLAS devotes ~10% of its disk resource to caching
- Willing to change this, but need experience with use cases to gauge change.

Remote vs Local Access

- GRID analysis jobs will use the current deployment policy
- Remote direct access is possible, but expected to remain small (caching if it increases)
- Must monitor user patterns…

* Details in backup

b) *Describe the relevant **changes in the model and their impact** in resources: policies for number of versions and replicas, fraction of data which is managed vs. unmanaged (e.g. caches), remote vs. local data access, batch vs. interactive cpu/gpu access, need of access to external DBs, or any other*

Log information from US ATLAS Shared T3's:

- UChicago
- BNL
- SLAC

Batch, ssh, jupyter

Currently ssh dominates

- Will this mix change as ATLAS enables other forms of access?
- As Run 3 analysis really gets underway?

2. Future analysis model in Run 4 and Run 5

ATLAS EXPERIMENT

b) *Describe the relevant **changes in the model and their impact** in resources: policies for number of versions and replicas, fraction of data which is managed vs. unmanaged (e.g. caches), remote vs. local data access, batch vs. interactive cpu/gpu access, need of access to external DBs, or any other*

GPU Access

- Shared T3's often have GPU's
- GRID has GPU's

We do not have accurate monitoring information for their use, however (very small).

- Many people run on "private" clusters

Common wisdom: GPU requirements will increase for Run 4 - especially as we employ foundation models

Selecting GPU memory

Here is a table of our GPUs and their availability.

GPU	Count	Avail.	Memory (MB)	Max CPU Req	Max Mem Req (GB)
NVIDIA-A100-SXM4-40GB-MIG-1g.5gb	28	13	4864	67	286
NVIDIA-GeForce-RTX-2080-Ti	23	16	11264	44	160
NVIDIA-GeForce-GTX-1080-Ti	8	0	11264	0	0
Tesla-V100-PCIE-16GB	4	1	16384	11	3
NVIDIA-A100-SXM4-40GB	4	0	40960	0	0

UChicago GPU availability…

Often zero!

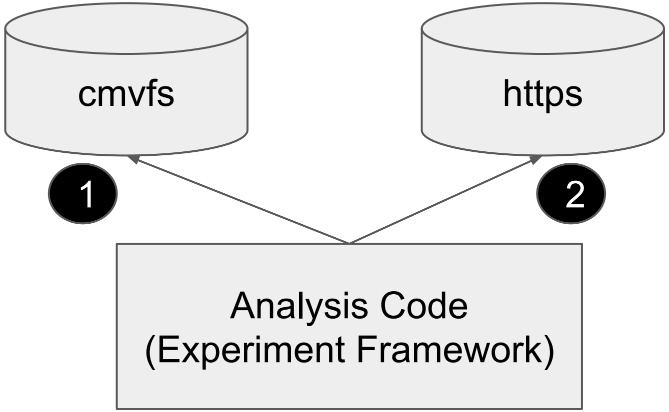
b) *Describe the relevant **changes in the model and their impact** in resources: policies for number of versions and replicas, fraction of data which is managed vs. unmanaged (e.g. caches), remote vs. local data access, batch vs. interactive cpu/gpu access, need of access to external DBs, or any other*

Significant work has been done to architect ATLAS analysis software to not require DB access!

- Constants are gotten from cvmfs or http (identical data is served).

Constantly tested by running in many different environments

- GRID
- Users' laptops (e.g. VM on mac, wsl2 on windows, etc.)
- Containers



- Fail over to 2 only if 1 fails
- Pulled files from 2 are cached locally

We are not expecting big changes in access patterns from this method of analysis

2. Future analysis model in Run 4 and Run 5

ATLAS EXPERIMENT

b) *Describe the relevant* ***changes in the model and their impact*** *in resources: policies for number of versions and replicas, fraction of data which is managed vs. unmanaged (e.g. caches), remote vs. local data access, batch vs. interactive cpu/gpu access, need of access to external DBs, or any other*

Other Tools

We have a series of other smaller tools

1) WebAPI tools
    a) E.g. AMI, our meta data database for all production samples. What to know the generator parameters for your file?
    b) Sometimes data is extracted and stored on cvmfs to reliability
    c) Otherwise, webAPI

Bigger Tools

1) Production Tools
    a) e.g. event summary and Event Picking Tools
    b) Service with web interface
    c) Not tools in middle of analysis chain - tend to be at start
    d) Use web to interact, results downloadable (or in rucio)

2. Future analysis model in Run 4 and Run 5

b) *Describe the relevant changes in the model and their impact in resources: policies for number of versions and replicas, fraction of data which is managed vs. unmanaged (e.g. caches), remote vs. local data access, batch vs. interactive cpu/gpu access, need of access to external DBs, or any other*

Interactive (Columnar) Access

- Not clear what fraction of interactive analysis will exist in Run 4 yet
- Currently, in R2 and R3, this has mostly been invisible to most facilities
- Larger datasets and desire to run at interactive speed (e.g. scaling) will likely drive more users to bigger facilities like a shared T3

Driving Metric for Interactive Access:
Time-to-plot

Resource Impact

- High IOPS disks
- Local storage for cached data
- Caches for external data
- Interactive tools (e.g. JupyterHub, ServiceX, etc.)
- Horizontal Scaling (e.g. DASK)
- Easily sharing resources between tasks in facility (e.g. T2 <-> T3)

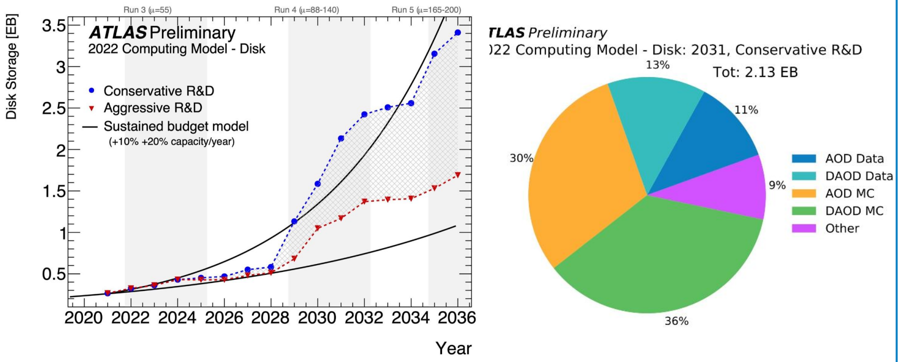
Actual Impact: depends on number of simultaneous users
Avoid pre-designing by using a flexible substrate like k8s?

c) ***Annual volume*** *expected for the different data formats, both data and MC.*

a) 7(10) EB expected including disk and tape by the end of Run 5 with aggressive (conservative) R&D

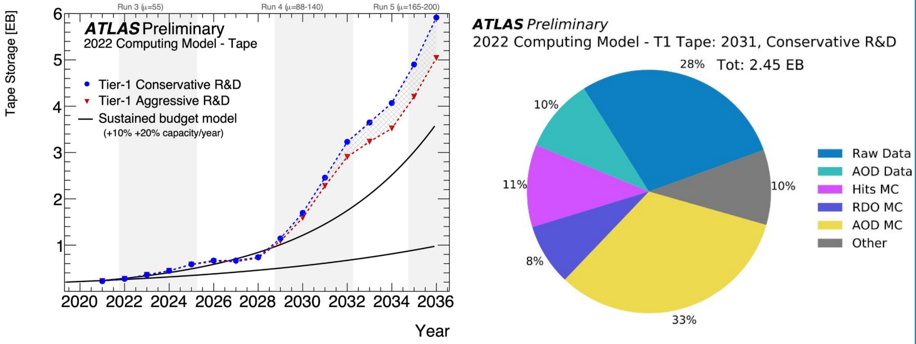
Disk

- Mainly DAOD and AOD
- Expecting to need x2 or x3 more MC than data to obtain physics goals



Tape

- 30% Raw data
- 45% AOD, stored in tape while not being used to make derivations



3. Managing the evolution of the Analysis Infrastructure

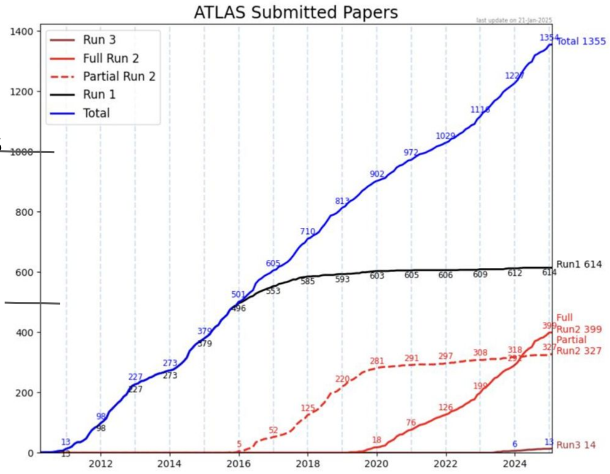
a) *Describe the **user requirements for analysis in HL-LHC** and the processes that will be used to track their evolution in the next few years*

The **current ATLAS computing** infrastructure is **serving our physics goals well**.

- Production computing, analysis on the GRID, batch, and local AF's in institutions around the world.

1000 Papers

500 Papers



The [HSF AF White Paper](#) has a list of questions to consider:

- Ability to perform **fast research iterations** on large datasets interactively
- Ability to **convert interactive to batch**-schedulable workloads
- Ability to interact with the WLCG and **scale outside of the facility** on occasion
- Ability to efficiently **train machine learning** models for HEP
- Ability to **reproducibly** instantiate desired software stack
- Ability to **collaborate in a multi-organisational team** on a single resource
- Ability to **move analyses to new facilities**
- Ability to efficiently **access collaboration data** as well as make intermediate data products available to the team
- Ability to express interdependent **distributed computations** at small and large scales

3. Managing the evolution of the Analysis Infrastructure

ATLAS EXPERIMENT

a) *Describe the user requirements for analysis in HL-LHC and the processes that will be used to track their evolution in the next few years*

ATLAS has not tracked analysis requirements formally up to now

- They have been either small enough fraction of production
- Institutions have built (shared) T3's or other ad-hoc services from non-pledge resources
- Monitoring depends on the site, and is not aggregated centrally.

What Does ATLAS Need to do?

- Build a few benchmark analyses
  - Working potentially with other experiments and IRIS-HEP
  - Explore aspects of data size, ML, workflow, interactive, batch, GRID and how they fit together (see list) →
  - We can't have one true way: need to build a library of techniques that work in concert.
- Tracking will be done by Facility and Analysis groups in ATLAS
  - E.g. Analysis Model Group, Core Software Group, Distributed Analysis, etc.

The HSF AF White Paper has a list of questions to consider:

- Ability to perform fast research iterations on large datasets interactively
- Ability to convert interactive to batch-schedulable workloads
- Ability to interact with the WLCG and scale outside of the facility on occasion
- Ability to efficiently train machine learning models for HEP
- Ability to reproducibly instantiate desired software stack
- Ability to collaborate in a multi-organisational team on a single resource
- Ability to move analyses to new facilities
- Ability to efficiently access collaboration data as well as make intermediate data products available to the team
- Ability to express interdependent distributed computations at small and large scales

3. Managing the evolution of the Analysis Infrastructure

ATLAS EXPERIMENT

b) *Comment on which **new technologies** or emerging paradigms you expect to be needed or have a relevant impact on the future Analysis Infrastructure and which mechanisms can be set up to manage this evolution as new technology will appear (e.g. ML, GPUs/FPGAs, etc)*

Demonstrators

- ATLAS projects tracked by S&C developing prototypes for new techniques in all of S&C for HL-LHC
- Most have been submitted as talks to CHEP (a requirement)
- Now bringing them to a close as we write our TDR
- This list is formed by looking at the work from the demonstrators and spin offs

Hardware Related Technologies

- GPUs including enough to train in parallel (ganged training for large models and hyperparameter-scans)
- Storage for interactive use (IOPS)
- Caching infrastructure
- Cloud-like facilities

Software Related Technologies

- RNTuple
- RDataFrame
- Python ecosystem
- Federated Identity for access to all ATLAS-wide resources
- How to efficiently schedule GPU’s for both training and inference
- Athena with GPU support
- Cloud technologies (e.g. k8s) to make facilities more flexible

3. Managing the evolution of the Analysis Infrastructure

ATLAS EXPERIMENT

b) *Comment on which new technologies or **emerging paradigms** you expect to be needed or have a relevant impact on the future Analysis Infrastructure and which mechanisms can be set up to manage this evolution as new technology will appear (e.g. ML, GPUs/FPGAs, etc)*

Demonstrators

- ATLAS projects tracked by S&C developing prototypes for new techniques in all of S&C for HL-LHC
- Most have been submitted as talks to CHEP (a requirement)
- Now bringing them to a close as we write our TDR
- This list is formed by looking at the work from the demonstrators and spin offs

Paradigm Shifts & Community

- Ability to train ML models both very large ones (e.g. reconstruction) and many small ones (unfolding, SBI, etc.)
- Interactive Analysis
- Caching (policy, etc.)
- ATLAS is good at managing production facilities - integrate with analysis infrastructure?
- Documentation, support, tutorials, for ATLAS AI resources
- New techniques like Simulation Based Inference, etc.

3. Managing the evolution of the Analysis Infrastructure

ATLAS EXPERIMENT

b) *Comment on which new technologies or emerging paradigms you expect to be needed or have a relevant impact on the future Analysis Infrastructure and which **mechanisms can be set up to manage this evolution** as new technology will appear (e.g. ML, GPUs/FPGAs, etc)*

Mechanisms To Manage This Evolution

During R&D:

- Mostly prior to TDR
- Sources of ideas and projects (and people): inside ATLAS, external groups (e.g. IRIS-HEP, HEP-CCE, SWIFTHEP, etc.)
- Cross cutting: can involve software and hardware!
- Currently using the **Demonstrator** program
- Expected to continue at some smaller level after the TDR as we move forward to Run 4+.

During HL-LHC:

- Mostly post-TDR
- ATLAS maintains *milestones, a risk registry,* and metrics
- There is a well functioning internal group structure
  - Analysis Model Group, Core Software, Distributed Analysis, etc.
  - This group structure is reviewed for efficiency periodically

3. Managing the evolution of the Analysis Infrastructure

ATLAS EXPERIMENT

c) *Describe the plans to develop specific **use cases that can be used to benchmark** different building blocks of the Analysis Infrastructure so that a **comparison** can be made between different implementations.*

There are several already in existence:

- **Data Challenge**: a series of progressive data movement exercises designed to emulate HL-LHC required data flows (WLCG)
- **Analysis Grand Challenge**: An integrated analysis benchmark (ttbar) designed to test the easy of use and efficiency of tooling (IRIS-HEP, and many others)
- **Analysis Language Benchmarks**: simple benchmarks for performing HEP simple data analysis (IRIS-HEP, and many others)

ATLAS (and the community?) needs to **develop a few benchmark analyses**

- Different data sizes
- Different workflows (ML, not ML, fitting, etc.)
- And any other appropriate axis

These analyses can be used to:

- Develop integrated benchmarks
- Synthetic benchmarks to test components of the system individually.

d) *Comment if you think that support for analysis workflows in Run-4 will need specialized infrastructure different from the Grid. If so, please describe what features that Analysis Infrastructure will need to provide to expand the one in the Grid*

Plain old ntuples in various formats

Often run on:

- Local university/lab cluster
- Batch (e.g. lxplus, or T3, etc.)
- Sometimes on GRID, but rarely
- The further down the line, the more likely you are to run on non-GRID resources

Other resources are already in heavy use!

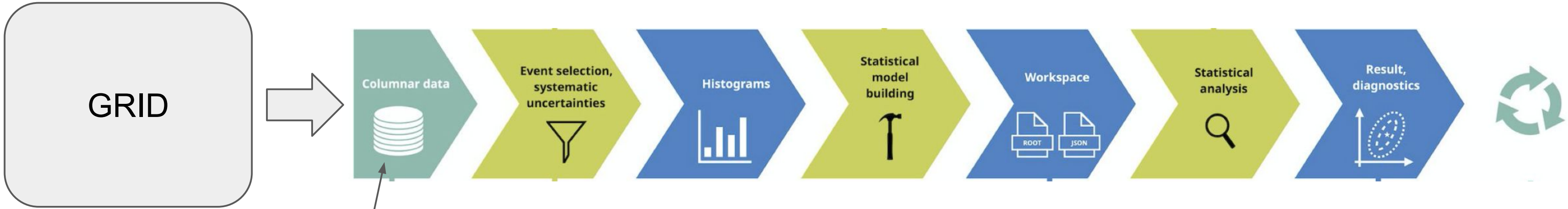
*Picture from IRIS-HEP AS

3. Managing the evolution of the Analysis Infrastructure

ATLAS EXPERIMENT

d) *Comment if you think that support for analysis workflows in Run-4 will need specialized infrastructure different from the Grid. If so, please describe what features that Analysis Infrastructure will need to provide to expand the one in the Grid*

For specifics see lists in 3b)



Plain old ntuples in various formats

We already have issues of access

- For larger analyses large compute facilities are already required

We already know having centralized access will improve this

- Benchmarks will tell us how much more we need (see blueprint by iris-hep)

*Picture from IRIS-HEP AS

3. Managing the evolution of the Analysis Infrastructure

d) *Comment if you think that support for analysis workflows in Run-4 will need specialized infrastructure different from the Grid. If so, please describe what features that Analysis Infrastructure will need to provide to expand the one in the Grid*

For specifics see lists in 3b)

Plain old ntuples in various formats

Combining:

- The list of new software and hardware technologies in 3b
- The changing access patterns of current users - that does not play to the GRID's strengths (interactive, ML, etc.)

ATLAS will need infrastructure beyond the GRID.

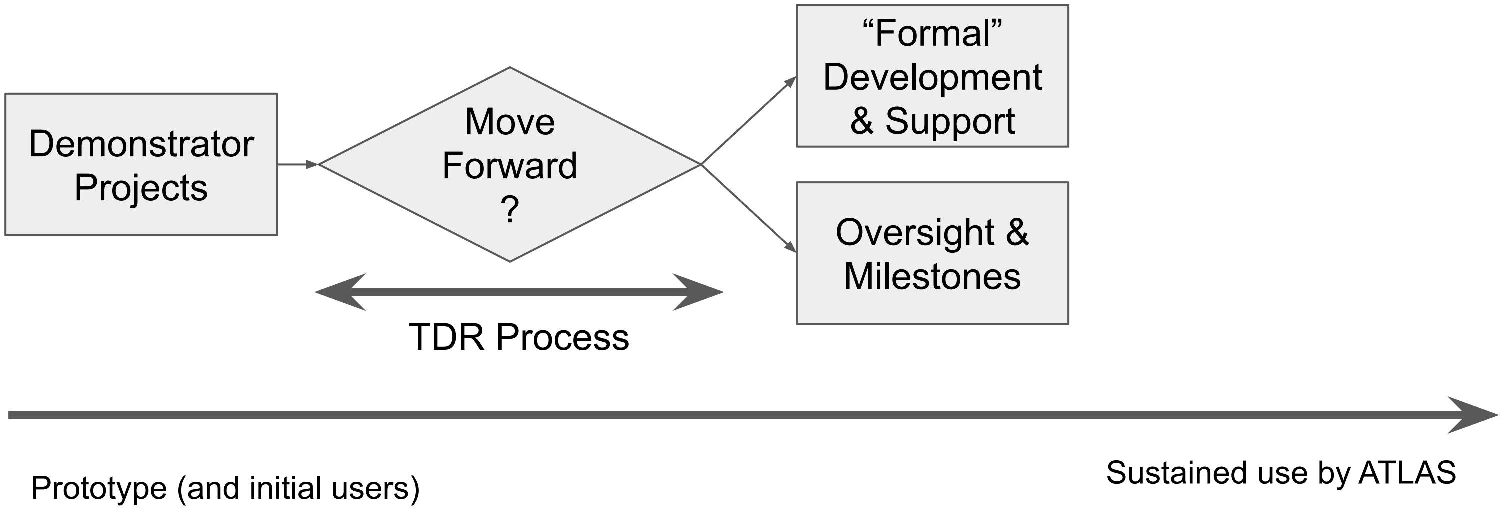
*Picture from IRIS-HEP AS

3. Managing the evolution of the Analysis Infrastructure

e) *Describe the current status and the **R&D** work that is underway **towards** implementing relevant **Analysis Infrastructure** functionality*

ATLAS has:

- Hardware and Facilities Related Research and Development
- Analysis Software Research and Development



3. Managing the evolution of the Analysis Infrastructure

e) *Describe the current status and the R&D work that is underway towards implementing relevant Analysis Infrastructure functionality*

Facilities Research

- Modern Shared T3 (Analysis Facility)
  - UChicago (IRIS-HEP)
  - Uses k8s as substrate
  - Actively serving all ATLAS members (federated access)
  - Has batch, ssh, JupyterHub, ServiceX, GPU's, etc.
  - BNL also has a Shared T3 with many of these services
- DESY Shared T3
  - More traditional with VMs'
  - Not for all ATLAS members, valuable experience(s).
  - Has batch, interactive, etc.
- SWAN, etc.

3. Managing the evolution of the Analysis Infrastructure

e) *Describe the current status and the R&D work that is underway towards implementing relevant Analysis Infrastructure functionality*

The ATLAS Demonstrator Program (Analysis Related)

- **RNtuple** fully capable of storing DAOD-PHYSLITE with advantages on storage footprint and I/O performance
  - Completed successfully
- Broaden DAOD-PHYS usability while limiting size increase for analysis by adding custom **event augmentation**.
  - Technical work done, use cases under consideration.
- The IRIS-HEP Analysis Grand Challenge, but on ATLAS internal data
  - Initial demonstrator working, now incorporating systematic errors

Other Software Efforts

- Columnar tests, in both python and RDF
- Calibration and Systematic Error tool development

Some of this work requires integrated development of facilities and software

Much of this work is already in milestones for ATLAS

Backup

Details on DAOD sizes

Size estimation in Run 3

Run 3	Per version, per year (PB)	2 versions, per year (PB)	2 versions, 5 years (PB)
PHYS	3	6	30
PHYSLITE	1.2	2.4	12

Number of versions and replicas used in the computing model:

Current year	data	MC
Versions	2	2
Replicas	2.4	2.3
Earlier years	**data**	**MC**
Versions	2	1.5
Replicas	1	1

1. Current Run 3 analysis model

ATLAS EXPERIMENT

a) Main analysis workflows and data reduction steps, including how closely chained they need to be

- a) Workflows overview
  - i) From the reconstruction stage, Analysis Object Data (AODs) are formed
    - (1) ROOT-based data structure compatible with RNTuple. In certain output configurations this can be read directly with RDataFrame and with UpROOT. This container is called xAOD
    - (2) They contain:
      - (a) physics objects: electrons, muons, etc
      - (b) sets of calibrations: provided by ATLAS Combined Performance Groups. Deriving them involves a physics analysis itself. Need much of the data available in the AODs.
  - ii) Derived datasets (DAODs), generated from AODs.
    - (1) The derivation step includes event and object filtering (reducing data volume) and application of calibrations
    - (2) In Run-3 there are two main derivation formats (PHYS and PHYSLITE) aimed at covering 80% of the analyses. The number of additional formats used for analyses with special needs (Long-lived particles, B-Physics, Trigger-level analyses, etc.) has been reduced from Run 2 to Run 3.
  - iii) Most analyses create nTuples from the DAODs:
    - (1) Finer filtering of events and physics objects
    - (2) Calculation of specific variables including Machine Learning (ML) integration, increasingly used for object identification, event classification, and systematic uncertainty estimation. ML workflows also involve new training datasets and introducing additional data flow requirements for storage, distribution, and preprocessing.
    - (3) Histogram production and statistical analysis: Final steps executed on small nTuples or histograms derived from DAODs.
      - (a) Note: the above is somewhat EventLoop-based, outputting ntuples. A columnar workflow typically will not produce a slimmed/skimmed ntuple as it's designed to be fast/performant and re-run over the inputs instead. Could run either directly on DAOD_PHYSLITE or the ntuple outputs from above – and produce histograms.
    - (4) Full integration of systematics into a columnar analysis is currently work in progress.

1. Current Run 3 analysis model

ATLAS EXPERIMENT

a) *analysis workflows and data reduction steps, including how closely chained they need to be*

- Chaining and coupling
  - Workflows are typically linearly chained
    - different algorithms are scheduled with defined inputs/outputs.
    - A “scheduler” determines how to best distribute or execute the job, effectively parsing a DAG, which presumably can be multi-threaded or multi-node (multi-processor).
    - some particular dataflows (e.g. involving ML or analysis optimisation) could require repeated processing
  - Consistency ensured in data formats and calibrations.
    - Reconstruction and derivation are part of the ATLAS-managed production system,
    - Everything else is done by the analysis group.
    - Coordinating and automating workflows across this boundary introduces a point of friction.
      - Effort to move to only a few analysis frameworks within the Collaboration

1. Current Run 3 analysis model

ATLAS EXPERIMENT

c) How much compute, storage and network resources are used for Run-3 analysis. Which fraction are pledged and which fraction are used in interactive mode (as opposed to batch)

Example of resource usage for the US ATLAS Analysis Facilities last 12 months

Counting every user
- UC - AF
- SLAC - AF
- BNL - AF

US ATLAS users ← lxbatch

- Batch
- Jupyter
- SSH

https://atlas-kibana.mwt2.org:5601/s/analysis-facility/app/dashboards#/view/8bb58440-6145-11ed-afcf-d91dad577662?_g=(filters:!(),refreshInterval:(pause:!t,value:0),time:(from:now-12M,to:now))

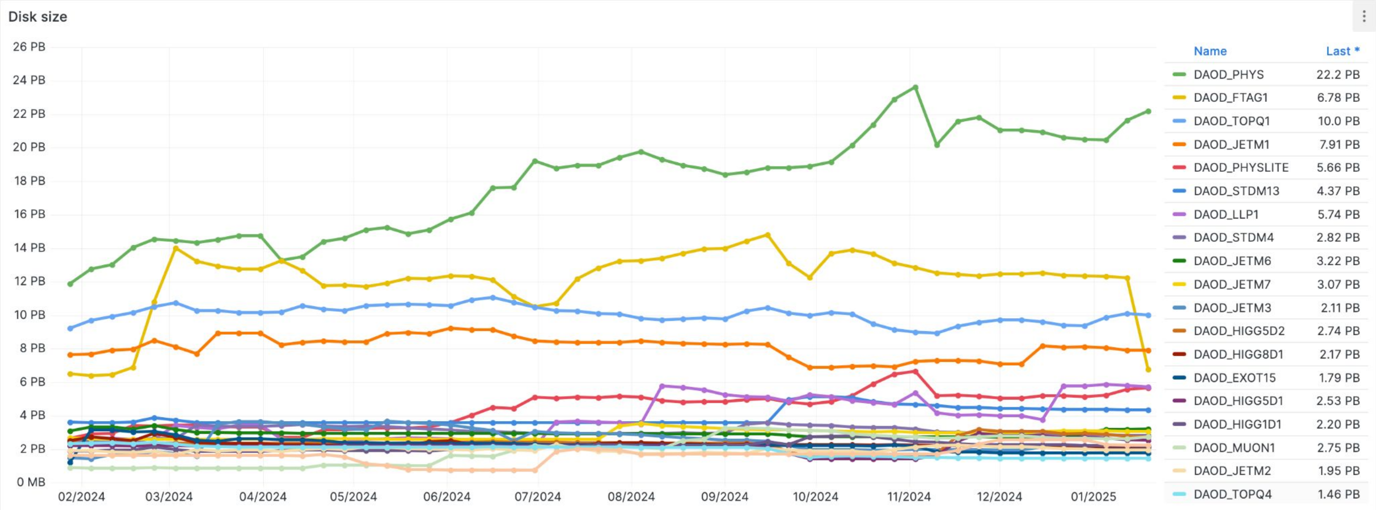
1. Current Run 3 analysis model

b) Data formats used for analysis, including their size and level of adoption (current and Run-3 final goal)

- Current total disk Derivation sizes:
  1. PHYS: 22PB
  2. PHYSLITE: 5.7PB
  3. Other derivation formats: ~50PB (several Run 2 analyses ongoing, still keeping their derivations)

Derivations size/event for the latest production

https://monit-grafana.cern.ch/d/eb091fee-47cd-430c-b2c1-3ba4006b2bf6/ddm-dataset-and-event-sizes?orgId=17	PHYS (kB/event)	PHYSLITE (kB/event)
Average data17 - data24	23	8
Average mc20, mc23	33	13.5
Average data, mc	29.6	11.6



- Work ongoing to migrate to RNTuple and REvent
- Preliminary studies show this will have a large impact in size reduction

https://monit-grafana.cern.ch/d/2jNUrRQmz/ddm-global-accounting-historical?orgId=17&var-datatype=All&var-datatype_grouped=DAOD&var-campaign=All&var-prod_step=All&var-project=All&var-scope=All&var-groupby=datatype