Evolution of the ATLAS analysis model for Run-3 and prospects for HL-LHC

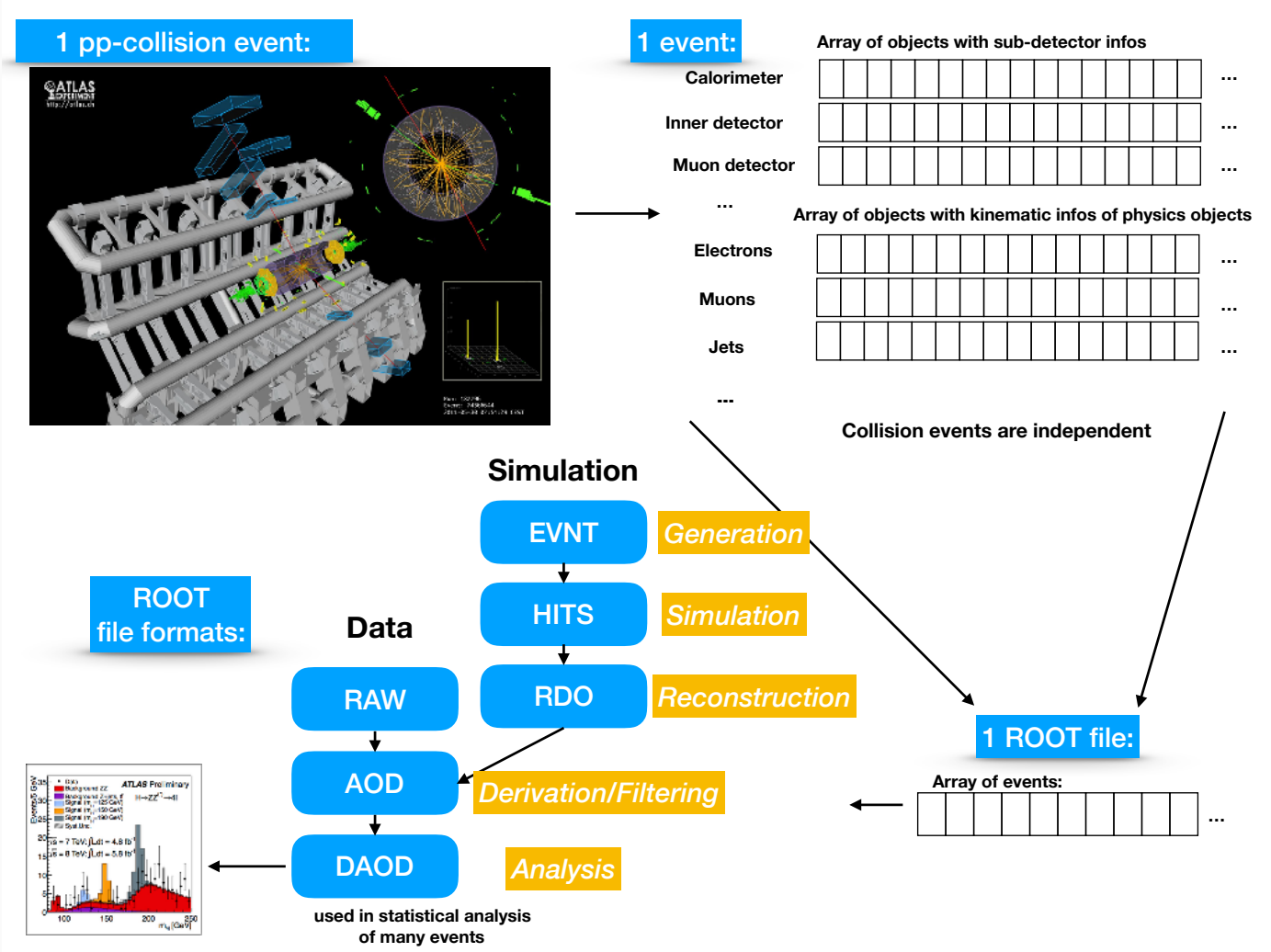
Johannes Elmsheuser, Alessandro Di Girolamo, Lukas Heinrich, TJ Khoo, Alison Lister, Zach Marshall on behalf of the ATLAS collaboration

4 May 2021, HL-LHC Mini Workshop Analysis

Material

- ATLAS HL-LHC Computing Conceptual Design Report, CERN-LHCC-2020-015, esp. Section 8
- CPU and Disk resource projection plots, see link
- Evolution of the ATLAS analysis model for Run-3 and prospects for HL-LHC, https://doi.org/10.1051/epjconf/202024506014

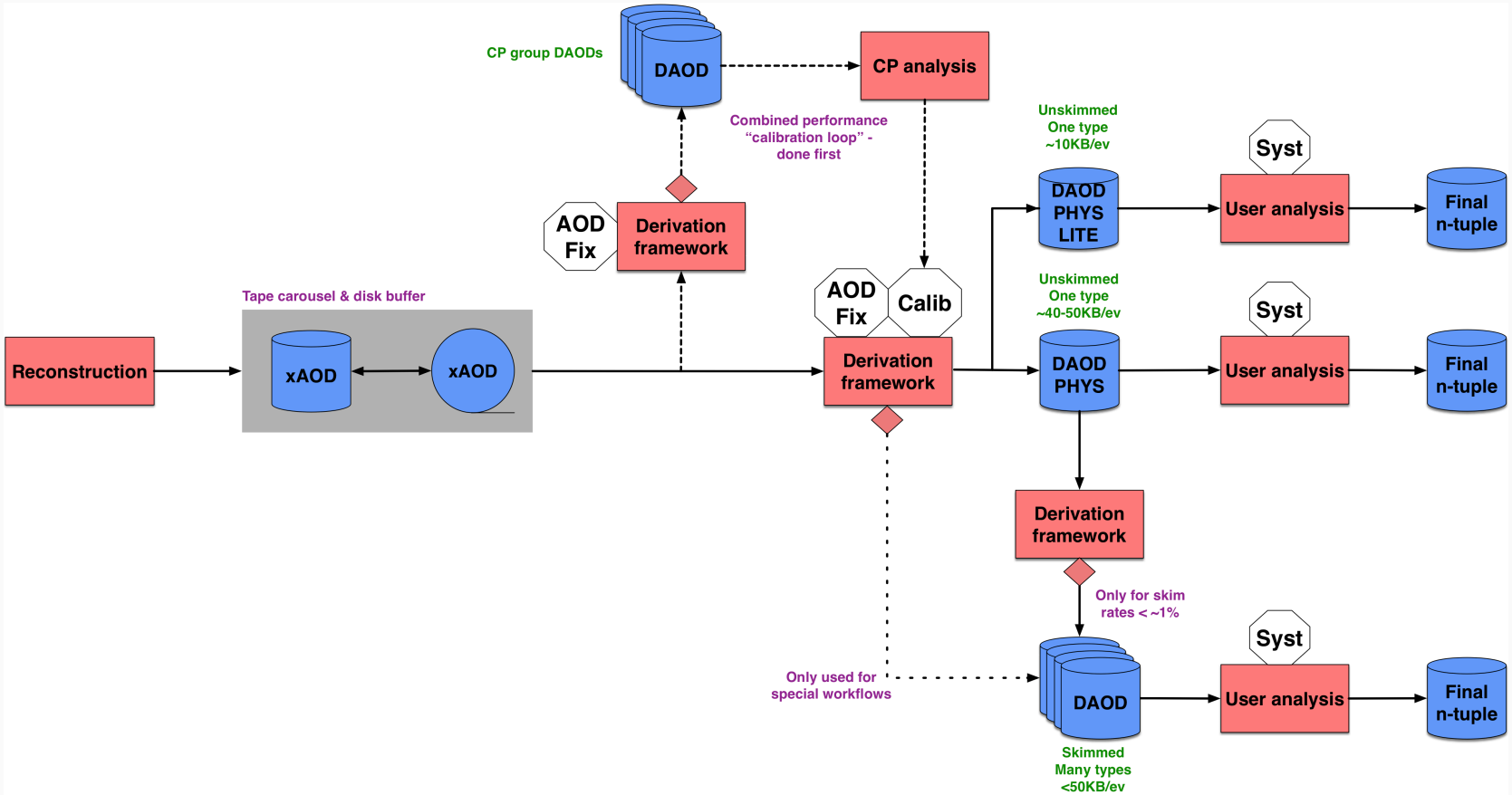
Introduction: Simplified data analysis workflow for ATLAS



In essence: several steps of data processing and then **data reduction**

First parts on Grid/Cloud/HPC - last step usually on local resources

Run3 analysis production workflows and formats



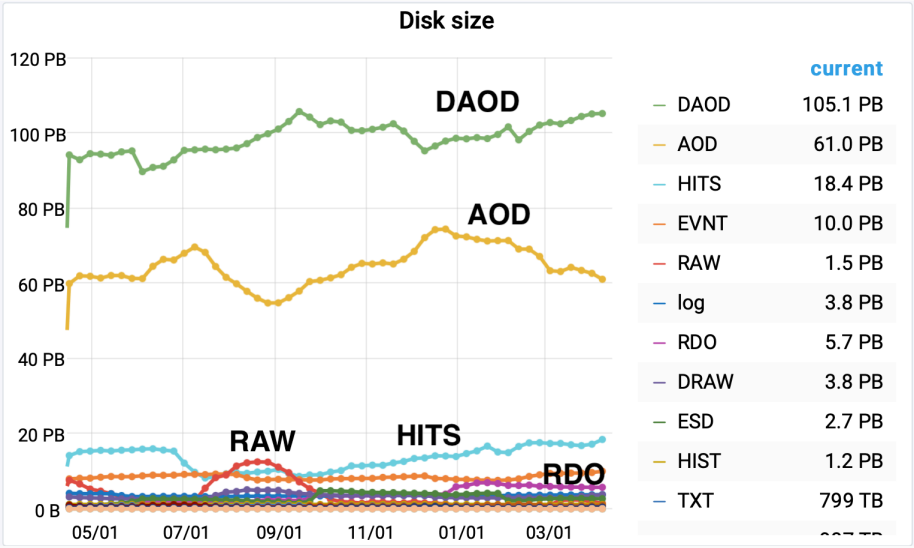
DAOD_PHYS:
50 kB/event, combined single DAOD format (for MC, but also DATA), AOD event data model (EDM)

DAOD_PHYSLITE:
10 kB/event, very condensed and calibrated objects, very important for HL-LHC, AOD or ntuple EDM, ideal for DOMA/XCache

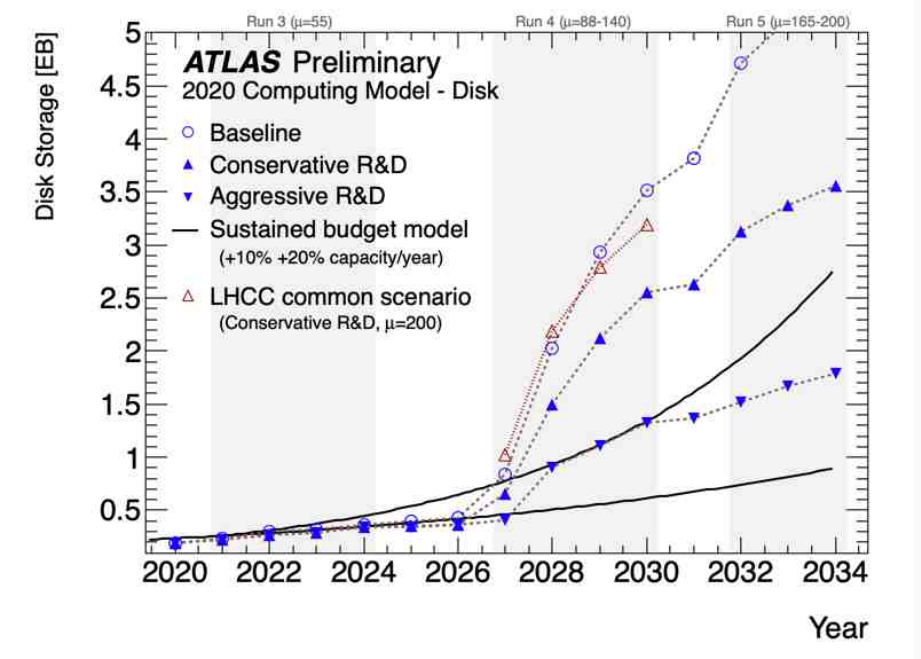
Remaining DAODs:
Significantly reduced number of additional DAOD types (10-20)

AODs:
Larger fraction only available on TAPE

ATLAS DISK SPACE STATUS AND PROJECTIONS

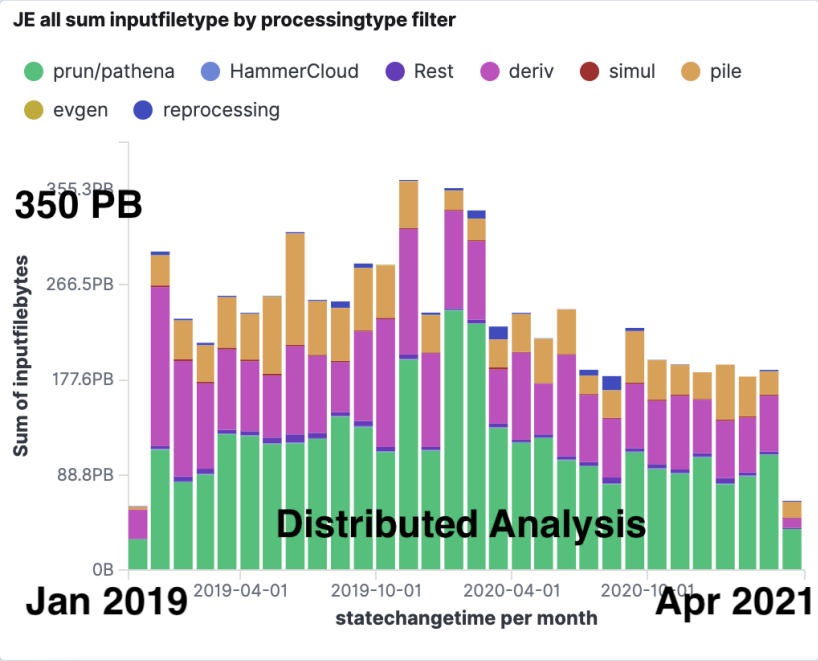


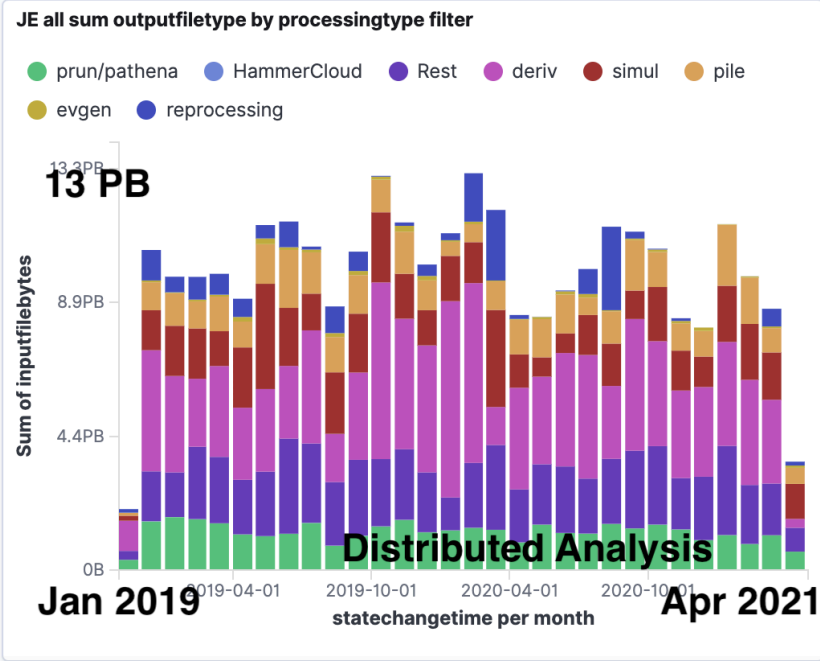
- DISK: 264 PB, filled mainly with Analysis formats (AOD/DAOD)
- Only 1-2 replicas possible because of large sample sizes
- In addition TAPE pledge of 330 PB



- **Run3:** within ”flat budget”
- **Run4:** challenging to stay within ”flat budget”

Processing input/output volumes PanDA in past ≈2 years





- Grid **input** processing volume ≈250 PB/month - 30-50% (≈100 PB/month) for analysis
- Grid **output** volume ≈10 PB/month - ≈2 PB/month for analysis
- Tier0 batch is not included here
- Distributed analysis users have relatively large freedom in workflow choices
- DAOD datasets largely distributed across Tier0/1/2 sites
- Extrapolations:
  - Run3: expect slightly higher numbers: more events, less formats
  - Run4: much more events should be balanced by smaller formats

HL-LHC analysis R&D scenarios

- **Baseline**: new data formats foreseen for Run 3, AthenaMT, but otherwise continues in largely the same way as in Run 2.
- **Conservative R&D**: R&D for Run3 successful: data carousel, fast track reconstruction, lossy compression, most of detector simulation with fast simulation
- **Aggressive R&D**: New developments with very significantly improve the speed or storage volumes. For analysis almost universal adoption by the physics groups of DAOD_PHYSLITE. Faster full and fast simulation, porting of code to GPUs

Status for Run3

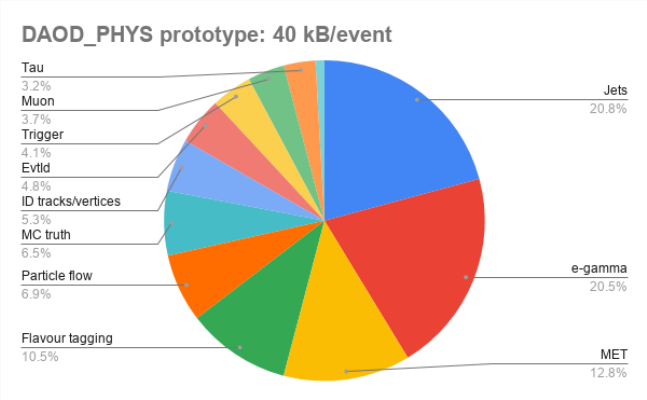
DAOD_PHYS:
Available and under commissioning for Run3

DAOD_PHYSLITE:
Advanced prototype available and more work w.r.t. systematic handling needed towards Run3

Lossy compression:
Under commissioning for Run3 in DAOD_PHYS

Containers:
Analysis and production releases and user container in production, can be used e.g. on PanDA or in local cluster

Data carousel:
In production - popular AODs and HITS kept in a disk buffer, and others staged from TAPE on-demand



Q&A: very simple HL-LHC extrapolation for events and disk

	MC			Data		
	AOD	DAOD PHYS	DAOD PHYSLITE	AOD	DAOD PHYS	DAOD PHYSLITE
events / year	$2 \cdot 10^{11}$	$2 \cdot 10^{11}$	$2 \cdot 10^{11}$	$7 \cdot 10^{10}$	$7 \cdot 10^{10}$	$7 \cdot 10^{10}$
size/event [kB]	1000	50	10	700	50	10
disk [PB/year]	200	10	2	49.0	3.5	0.7

Assumptions:

- no extra versions & no replication - this will increase the volume by a factor 2-4
- More disk space is needed for additional DAOD flavours for combined performance groups and special physics analysis
- Average size/event and no pile-up dependence assumed here

→ More DAOD_PHYSLITE and less DAOD usage, AOD with tape carousel will reduce disk capacity needs

Q&A II - Computing model

Production frequency	produce new DAOD_PHYS and DAOD_PHYSLITE versions several times per year
Processing rate	DAOD_PHYSLITE and ntuples as often as necessary
Anticipated CPU rate	DAOD_PHYS: 10-100 Hz (nominal w/o systematics) DAOD_PHYSLITE/ntuple: 100-1000 Hz (nominal w/o systematics) Columnar format with potentially higher rates
Events/month processed	DAOD_PHYS/LITE see above Fast histogramming as often as necessary
Analysis pipelines	For production: derivation production Individual analysis: requires proper environment like e.g. Analysis facility Analysis Preservation/RECAST with REANA is already a requirement for e.g. BSM analysis
Analysis diversity	Expect 70-80% can use DAOD_PHYS/LITE but special needs for extra formats/samples in e.g. b-physics or Combined Performance groups
Programming models	Declarative vs. procedural: support both Automating systematics, calibrations: central code for both are used in all individual analysis and is planned to be used in DAOD_PHYSLITE production how to best handle systematics in central production is under discussion and requires R&D

Q&A III - Impact of R&Ds

GPU	mainly projected to be used in production (Simul and Reco) and ML for analysis - but are there new trends ?
Analysis facilities	Active R&D within WLCG/DOMA on-going Presentation with ADC plans will be presented at vCHEP Some prototypes at small scale available e.g. in the US, Cloud R&D on-going
Jupyter Notebooks	Ideal for code prototyping but requires seamless integration with a data facility in case of bulk processing needs Unclear how many users can be served simultaneously
Trends	ROOT remains a corner stone of analysis Python analysis ecosystem (see below) Columnar data format for analysis: R&D on-going Real-time analysis R&D on-going, but offline analysis still needed
Python ecosystem	ROOT is one pilar of analysis But also support for data science tools inside/outside of HEP Important for AI/ML tools and data formats Training of new students
Languages	C++, Python

Q&A IV - continued

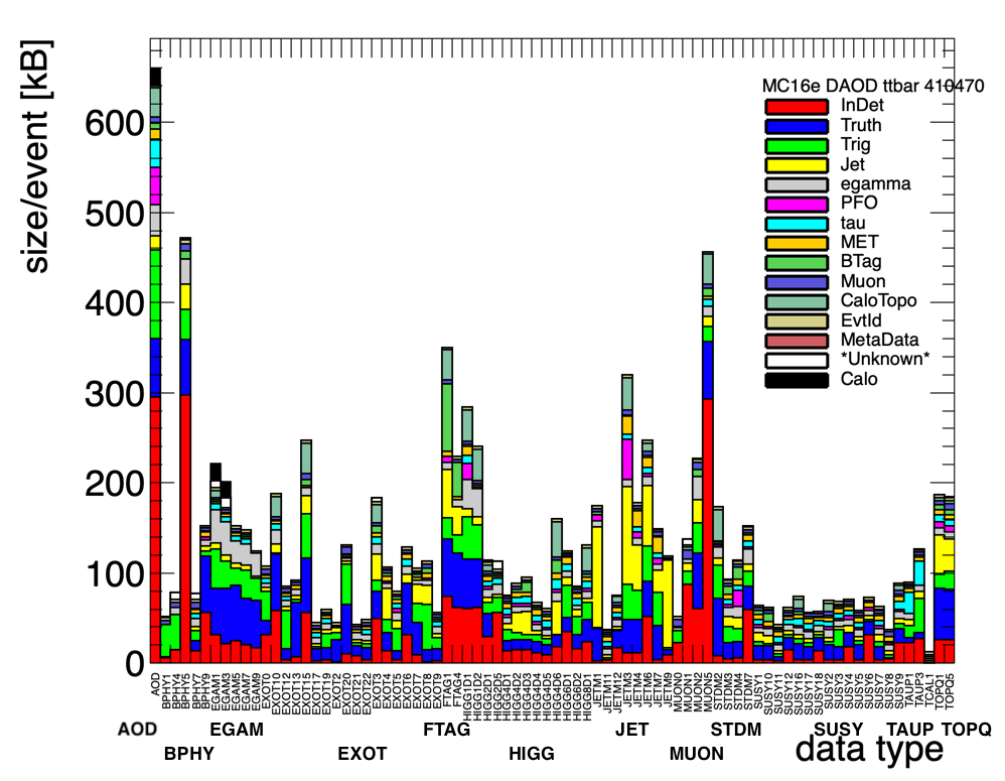
Special workflows	Large statistics analysis with DAOD_PHYS/LITE not best choice sometime $\gamma\gamma \rightarrow WW$ production or LongLivedParticle searches in a two-step workflow already working today: a) Initial reconstruction b) RAW event picking of candidate events c) Re-reconstruction with special settings d) Analysis
Lifetime model	All formats (except RAW, EVNT) have a lifetime on disk typically 6-24 months keep only 2 versions of DAODs

- Key components for HL-LHC:
  - Multi-step data reduction
  - Compact data format(s)
  - Calibrations and smart systematic handling
  - Smart integration of ML and emerging technologies like dedicated facilities or new data handling

BACKUP

AOD/DAOD contents !!! OLD !!!

$t\bar{t}$ MC, 1 AOD, 79 DAODs



General AOD/DAOD content:

- Lots of low level quantities for all physics objects in DAOD to allow calibrations and systematics very late in analysis chain
- Allows very flexible object definitions but increases format sizes significantly

Lots of AOD/DAODs infos:

- Tracks/InDet, MC truth, Trigger dominate size

Lots of samples:

- Only 1-2 replicas possible because of large sample sizes
- Many event duplication from AOD to DAOD

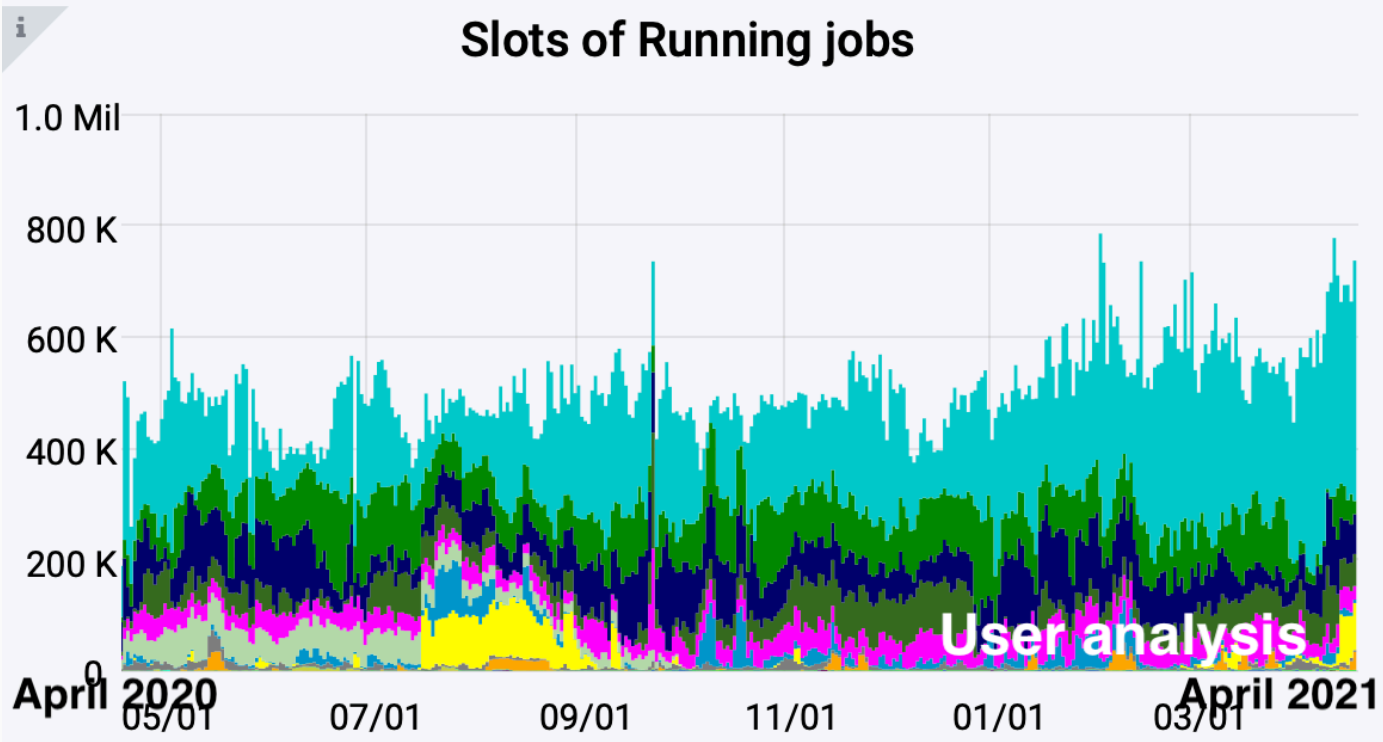
Example sample sizes:

		MC16e	data18
AOD	logical [PB]	11.2	2.7
	disk [PB]	13.0	4.2
	evt [10^9]	17.178	12.108
DAOD	logical [PB]	9.9	6.1
	disk [PB]	13.4	12.7
	evt [10^9]	91.292	110.139

Top 10 DAOD:

DAOD_TOPQ1	10.10 PB
DAOD_STDM4	3.57 PB
DAOD_TOPQ4	3.40 PB
DAOD_FTAG4	3.27 PB
DAOD_RPVLL	3.10 PB
DAOD_HIGG2D1	2.41 PB
DAOD_JETM6	2.08 PB
DAOD_FTAG1	1.98 PB
DAOD_JETM1	1.97 PB
DAOD_EXOT5	1.80 PB

CPU USAGE



- CPU pledge of 3125 kHS06
- 10-20% of analysis share on the Grid/Cloud - not HPC - mainly single core serial processing payloads
- Very diverse inputs and processing payloads in analysis
- In addition lots of final analysis happens on local batch farm or computers on individual ntuples

ATLAS DISTRIBUTED COMPUTING OVERVIEW

The ATLAS distributed computing system is centered around:

- **Workflow management system**: PanDA
- **Data management system**: Rucio
- Many **additional components**: AGIS, ProdSys, Analytics, ...
- **Resources**: WLCG grid sites, Tier0, HPCs, Boinc, Cloud
- **Shifters**: Grid, Expert and Analysis (ADCoS, CRC, DAST)