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

Christos Anastopoulos, Jamie Boyd, James Catmore, *Johannes Elmsheuser*, Heather Gray, Attila Krasznahorkay, Josh McFayden, Chris Meyer, Anna Sfyrla, Jonas Strandberg, Kerim Suruliz, Timothée Theveneaux-Pelzer on behalf of the ATLAS collaboration

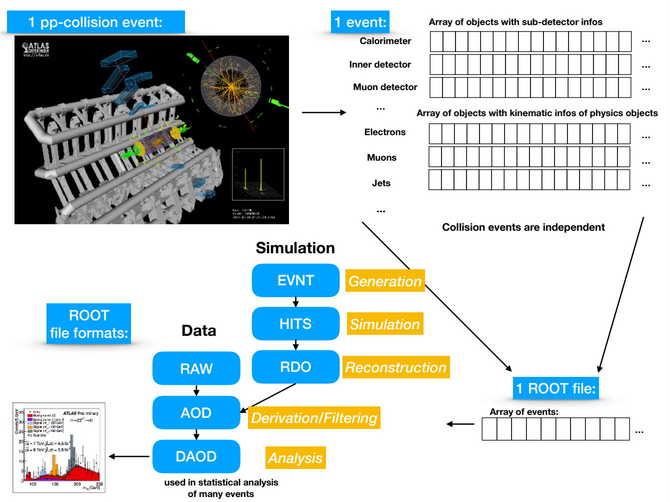
5 November 2019, CHEP 2019, Adelaide

Outline

ATLAS experiment analysis in LHC Run2 and resource usage

Recommendations of ATLAS experiment analysis model study group for Run3 (AMSG-R3)

Introduction: Simplified data analysis workflow for ATLAS



1 pp-collision event:

1 event:

Array of objects with sub-detector infos

Calorimeter ...

Inner detector ...

Muon detector ...

...

Array of objects with kinematic infos of physics objects

Electrons ...

Muons ...

Jets ...

...

Collision events are independent

Simulation

EVNT *Generation*

HITS *Simulation*

RDO *Reconstruction*

ROOT file formats:

Data

RAW

AOD *Derivation/Filtering*

DAOD *Analysis*

used in statistical analysis of many events

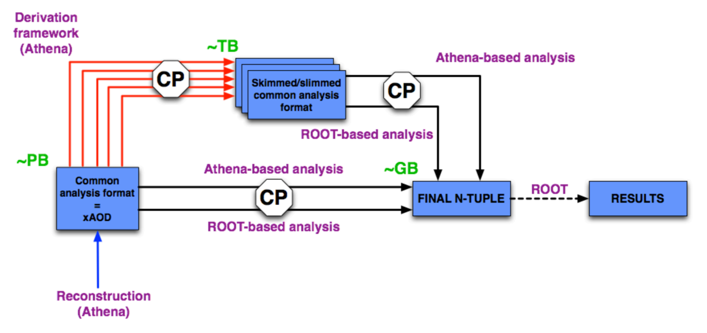
1 ROOT file:

Array of events: ...

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

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

ATLAS Run2 analysis workflows

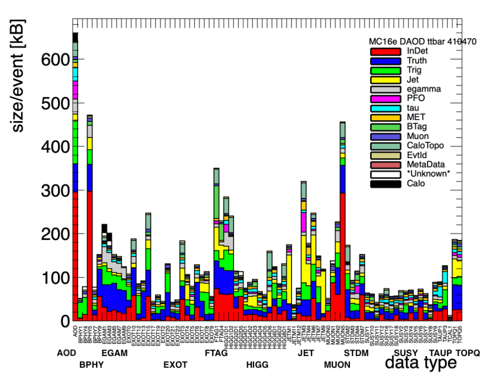


DAOD: highly successful in view of productivity of ATLAS, the Run 2 model has been expensive in terms of resources

- DAOD data formats used by almost all analysis in ATLAS - but additional group analysis post-DAOD
- Supposed to be ∼1% of size of data inputs
- 84 formats in current use, shared among similar physics final states,

AOD/DAOD CONTENTS

$t\bar{t}$ 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

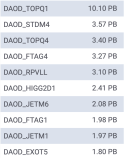
Lots of samples:

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

Example sample sizes:

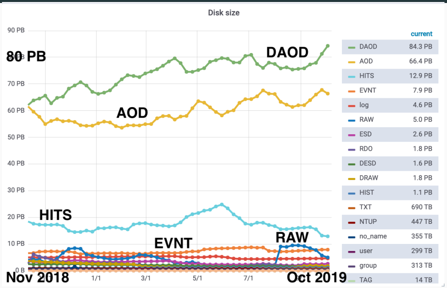
		MC16e	data18
AOD	logical [PB]	11.2	2.7
	disk [PB]	13.0	4.2
	evt [10^9]	17.178	12.108
DAOD	logical [PB]	9.9	6.1
	disk [PB]	13.4	12.7
	evt [10^9]	91.292	110.139

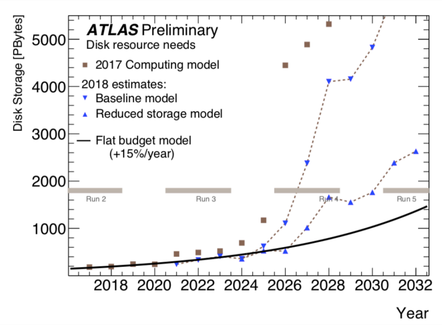
Top 10 DAOD:



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 & ATLAS disk space projections





- DISK: 223 PB, filled mainly with Analysis formats (AOD/DAOD)
- Only 1-2 replicas possible because of large sample sizes
- In addition TAPE ≈ 253 PB used and pledge of 315 PB

Run3: Initial assumption resources will be: 1.5 × (resources in 2018) Consistent with ”flat budget”

Outline

ATLAS experiment analysis in LHC Run2 and resource usage

Recommendations of ATLAS experiment analysis model study group for Run3 (AMSG-R3)

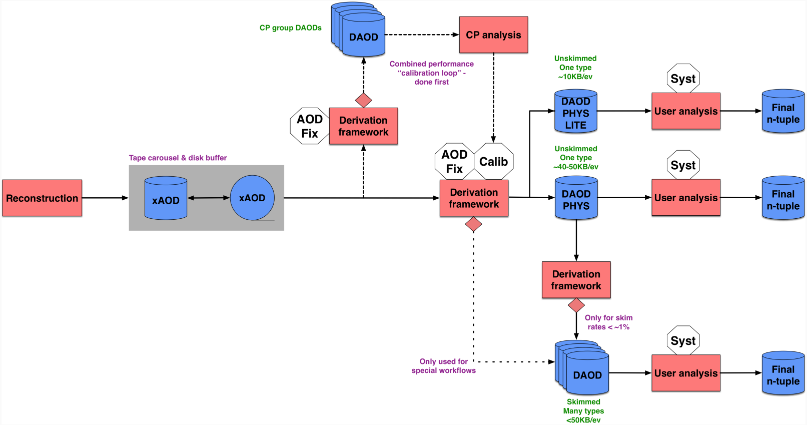
ATLAS Analysis model study group for Run3 (AMSG-R3) group mandate

- Analysis model study group for Run3 (AMSG-R3) formed in summer 2018, delivered set of recommendations for updated ATLAS Analysis/Computing model in June 2019
- Group mandate in essence:

  Collect options to save at least 30% disk space overall (for the same data/MC sample), harmonise analysis and give directions for further savings for the HL-LHC.

- Latest ”ATLAS Computing Status and Plans: Report to the C-RSG” uses these recommendations
- Now it’s time for many ATLAS groups to work on the recommendations

New 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

today's DAODs:
Significantly reduce number of today's DAODs

AODs:
Larger fraction only available on TAPE

Summary of the AMSG-R3 recommendations

Formats	Introduce **DAOD_PHYS** with ∼50 kB/event Introduce **DAOD_PHYSLITE** with ∼10 kB/event and **calibrated objects** Significantly **reduce number DAODs** formats by **DAOD_PHYS(LITE)** in majority of analysis Allow exceptions for performance groups, B-physics (separate stream), long lived particle searches, soft QCD
Production	Use a **tape carousel model for AOD** inputs in parts of the DAOD production Increase usage of **docker/singularity containers** for analysis and group ntuple production and more like: changes in DAOD production policies, smarter replica placements, global Rucio file redirector
AOD/DAOD content	Significantly **reduced track, trigger, truth** information, use **calibrated objects** Apply **lossy compression** for most variables in AOD/DAODs where feasible and applicable

Simple disk space model with Run2 numbers

- Simple model of Run2 AOD+DAODs: 132 PB
  - 4 DAOD_PHYS+DAOD_PHYSLITE (MC+DATA) replicas
  - 0.5 AOD replica (aka TAPE buffer)
  - 50% of today's MC+DATA DAOD

	MC				Data			
	AOD	DAOD	DAOD PHYS	DAOD PHYS LITE	AOD	DAOD	DAOD PHYS	DAOD PHYS LITE
events	$3 \cdot 10^{10}$	$1 \cdot 10^{11}$	$3 \cdot 10^{10}$	$3 \cdot 10^{10}$	$2 \cdot 10^{10}$	$1 \cdot 10^{11}$	$2 \cdot 10^{10}$	$2 \cdot 10^{10}$
size/event [kB]	600	100	70	10	400	50	40	10
disk space [PB]	18.0	10.0	2.1	0.3	8.0	5.0	0.8	0.2
other versions	1.5	2	2	2	1.5	2	2	2
repl. fac.	0.5	1	4	4	0.5	2	4	4
Sum [PB]	13.5	20.0	16.8	2.4	6.0	20.0	6.4	1.6

- Sum: 85 PB
- Potential saving: 46 PB
  → allows room for more MC event production

Status of implementations: main AMSG-R3 recommendations

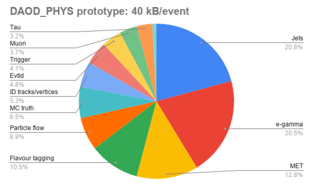
DAOD_PHYS:
target: 50 kB/event
prototype ready: 40 kB/event, significantly reduced trigger, MC truth and tracking info

DAOD_PHYSLITE:
target: 10 kB/event, prototype under preparation

Lossy compression:
Reduce precision of float elements by setting some digits of the mantissa to zero, allowing more efficient compression
Explore in parallel ROOT 6.18 Float16_t compression/truncation

Data carousel:
On demand reading from tape without pre-staging
Uses a rolling disk buffer with a to be tuned size
Rucio, FTS, dCache improvements work-in-progress

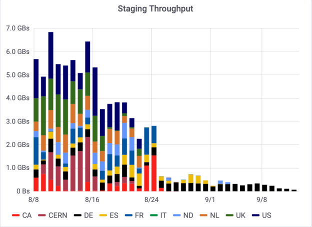
Containers:
PanDA uses OS containers for production and analysis and support user containers in place



$t\bar{t}$ MC, blind float to 7 bit mantissa compression:

Format	Compression ratio
AOD	0.72
DAOD_PHYS	0.75
DAOD_PHYSLITE	0.9

data18 reprocessing, Stage 7 PB within 2 weeks: 6 GB/s:



Very simple HL-LHC extrapolation for disk

	MC			Data			Sum
	AOD	DAOD	DAOD PHYSLITE	AOD	DAOD	DAOD PHYSLITE	
events (25-28)	$6.4 \cdot 10^{11}$			$1.5 \cdot 10^{11}$			
events / year	$2.13 \cdot 10^{11}$	$1.07 \cdot 10^{12}$	$2.13 \cdot 10^{11}$	$5.0 \cdot 10^{10}$	$2.5 \cdot 10^{11}$	$5.0 \cdot 10^{10}$	
size/event [kB]	1000	100	10	700	50	10	
disk [PB/year]	213.3	106.7	2.1	35.0	12.5	0.5	369.6

Assumptions:

- DAOD: 5*AOD events, use DAOD_PHYS(LITE) as in AMSG-R3
- no extra versions & no replication - this will increase the volume by a factor 2-4
- Average size/event and no pile-up dependence assumed here

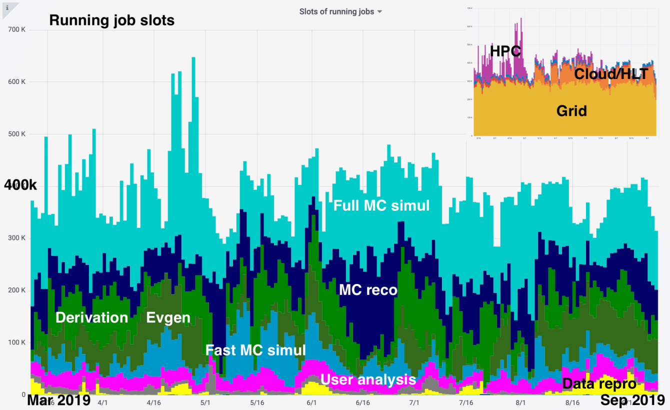
→ More DAOD_PHYSLITE and less DAOD usage, AOD with tape carousel will reduce disk capacity needs

Summary and Conclusions

- ATLAS Run2 analysis model very successful but expensive w.r.t. disk space usage
- For Run3: significant disk usage reduction planned with new formats DAOD_PHYS, DAOD_PHYSLITE and tape carousel
- Without something similar to DAOD_PHYSLITE, analysis at HL-LHC very difficult
- Development work in many ATLAS software, computing and physics areas on-going

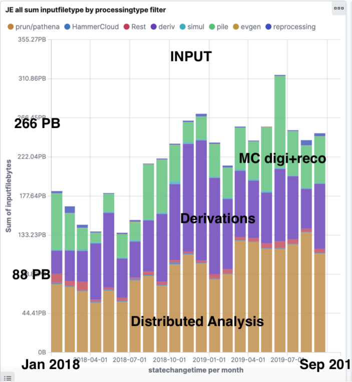
BACKUP

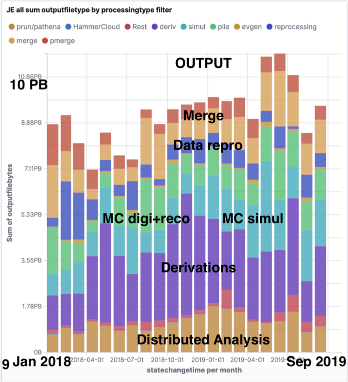
CPU USAGE



- 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

Processing input and output volumes PanDA in past 17 months



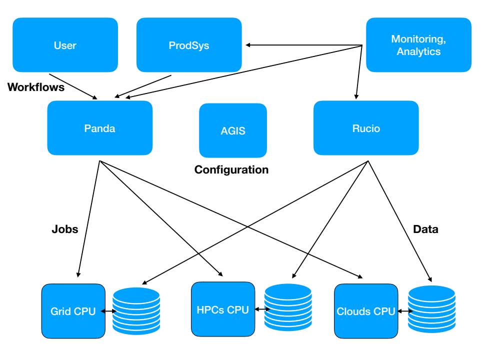


- Grid **input** processing volume ≈200-250 PB/month - 30-50% derivation production, 30-50% analysis
- Copied to worker node - files might be accessed multiple times on the worker node (digi-reco)
- Grid **output** volume: ≈ 8-9 PB/month of which 2-5 PB/month derivation production
- Tier0 batch is not included here and adds to the input/output volumes

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)


User
ProdSys
Monitoring, Analytics
Workflows
Panda
AGIS
Rucio
Configuration
Jobs
Data
Grid CPU
HPCs CPU
Clouds CPU