EPJ Web of Conferences **245**, 06014 (2020) https://doi.org/10.1051/epjconf/202024506014
CHEP 2019

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

Johannes Elmsheuser[1,*], *Christos* Anastopoulos[2], *Jamie* Boyd[3], *James* Catmore[4], *Heather* Gray[5], *Attila* Krasznahorkay[3], *Josh* McFayden[3], *Christopher John* Meyer[6], *Anna* Sfyrla[7], *Jonas* Strandberg[8], *Kerim* Suruliz[9], and *Timothee* Theveneaux-Pelzer[10]

[1]Brookhaven National Laboratory, Upton, NY, USA
[2]University of Sheffield, Sheffield, UK
[3]CERN, Geneva, Switzerland
[4]University of Oslo, Oslo, Norway
[5]Lawrence Berkeley National Laboratory, Berkeley, USA
[6]Indiana University, Bloomington, USA
[7]Universite de Geneve, Geneva, Switzerland
[8]KTH Royal Institute of Technology , Stockholm, Sweden
[9]University of Sussex, Brighton, UK
[10]Deutsches Elektronen-Synchrotron, Zeuthen, Germany

Abstract. With an increased dataset obtained during the Run-2 of the LHC at CERN, the even larger forthcoming Run-3 data and the expected increase of the dataset by more than one order of magnitude for the HL-LHC, the ATLAS experiment is reaching the limits of the current data production model in terms of disk storage resources. The anticipated availability of an improved fast simulation will enable ATLAS to produce significantly larger Monte Carlo samples with the available CPU, which will then be limited by insufficient disk resources. The ATLAS Analysis Model Study Group for Run-3 was setup at the end of Run-2. Its tasks have been to analyse the efficiency and suitability of the current analysis model and to propose significant improvements. The group has considered options allowing ATLAS to save, for the same sample of data and simulated events, at least 30% disk space overall, and has given recommendations on how significantly larger savings could be realised for the HL-LHC. Furthermore, suggestions were made to harmonise the current stage of analysis across the collaboration. The group has now completed its work: key recommendations will be the new small sized analysis formats DAOD_PHYS and DAOD_PHYSLITE and the increased usage of a tape carousel mode in the centralised production of these formats. This proceeding reviews the recommended ATLAS analysis model for Run-3 and the status of its implementation. It also provides an outlook to the HL-LHC analysis.

1 Introduction

The distributed computing system of the ATLAS experiment [1] at the LHC is built around the two main components: the workflow management system PanDA and the data manage-

*e-mail: johannes.elmsheuser@cern.ch

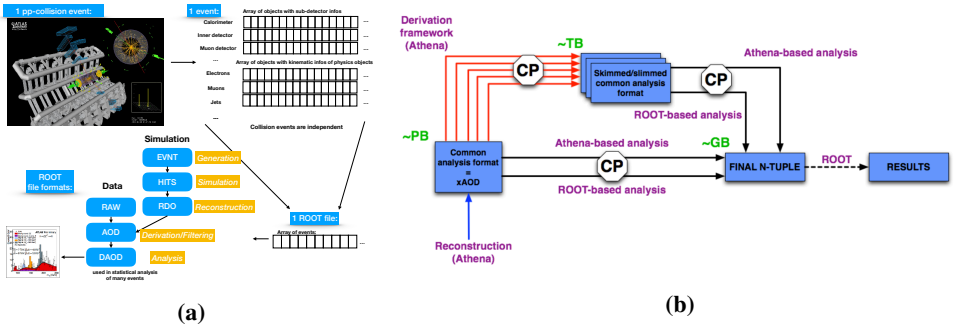


Figure 1: (a) The computing workflows in the ATLAS experiment. (b) The ATLAS Run-2 analysis workflow.

ment system Rucio [2]. It manages the computing resources to process the detector data at the Tier-0 at CERN, reprocesses it once per year at the Tier-1 and Tier-2 Worldwide LHC Computing Grid (WLCG) [3] sites and runs continuous Monte Carlo (MC) simulation and reconstruction. In addition continuous distributed analyses from several hundred ATLAS users are executed. The resources used are the Tier-0 at CERN and Tier-1/2/3 Grid sites worldwide, opportunistic resources at High Performance Computer (HPC) sites, cloud computing providers, and volunteer computing resources.

The ATLAS Run-2 analysis model has been highly successful in the view of the productivity of ATLAS, but it has been expensive in terms of resource usage. The ATLAS Analysis Model Study Group for Run-3 (AMSG-R3) setup at the end of Run-2 was tasked to analyse the efficiency and suitability of the current model and to propose significant improvements.

2 The ATLAS Run-2 analysis model

The analysis of ATLAS detector data and simulated events is a multi-step processing and reduction procedure. Figure 1 (a) shows a scheme of the workflows executed on the distributed computing infrastructure. The information of the LHC collision events detected by the different ATLAS sub-detectors are stored in ROOT files [4]. These files are centrally managed and processed in different MC simulation or data reconstruction steps and workflows before being eventually analysed by individuals. Figure 1 (b) shows the ATLAS Run-2 analysis workflow. The output of the data and MC reconstruction is stored in Analysis Object Data (AOD) files and grouped in datasets on the various Grid sites. These datasets are processed in the derivation framework which produces about 80 different derived AOD (DAOD) formats that contain a subset of events and reduced reconstruction information tailored for specific physics analysis and performance groups. These DAOD types are processed by many individual analysers in a random manner who produce very condensed individual ntuples for further processing or final physics results.

3 Run-2 format content and campaign sizes

Figure 2 shows the size/event distribution in units of kilobytes for an AOD and the different DAOD types derived from this. Events with $t\bar{t}$ final states generated in the 2018 MC simulation campaign are used. This process provides a representative physics final state in terms of the average event size. The size/event for the AOD is about 600 kB, whereas for the different

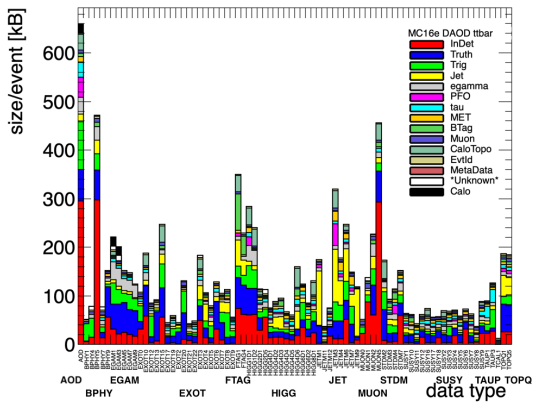


Figure 2. Distribution of size/event for AOD and DAOD formats for a $t\bar{t}$ MC sample.

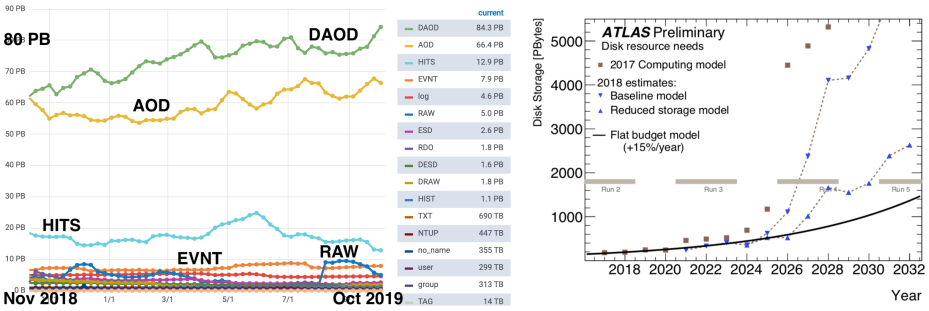
DAOD formats the values vary in the range of 40-450 kB depending on the type of the physics selection and the information retained. The majority of the event content is taken up by inner detector tracks, MC generator and trigger information across all formats. Table 1 shows the total sizes of the AOD and DAOD files for the 2018 proton-proton collision data and corresponding MC samples. The campaigns from previous years or heavy ion data taking are not shown here. The logical sizes in the table denote the dataset sizes without additional replicas on different Grid sites, whereas the disk size numbers include all additional replicas. Overall the Grid disk space is very heavily utilised as shown in Figure 3 (a) so that only 1-2 replicas of each dataset and campaign can be kept on disk. One notable effect of the derived DAOD model is that the number of events is significantly increased as illustrated by the difference between the number of events for the AOD and DAOD formats in the table.

Table 1: Sums of dataset samples sizes for different production campaigns in 2018 for data and MC samples, in AOD and DAOD formats.

		MC16e	data18
AOD	logical [PB]	11.2	2.7
	disk [PB]	13.0	4.2
	events [10^9]	17.2	12.1
DAOD	logical [PB]	9.9	6.1
	disk [PB]	13.4	12.7
	events [10^9]	91.3	110.1

4 CPU Usage and ATLAS disk space projections

Figure 3 (a) shows the distribution of the different ATLAS data formats over the time range of November 2018 to October 2019. The majority of the Grid disk space is used by the AOD and DAOD formats in the orders of 60 and 80 PB, respectively. Figure 3 (b) shows the projection of the ATLAS disk space needs over the coming years for the LHC Run-3 and the HL-LHC. For Run-3 the disk space needs match the projections of a so-called "flat budget" increase of the resources by 15% each year due to the technology advancements. For the HL-LHC the projections of the ATLAS needs are significantly over the yearly flat budget increase. ATLAS is therefore investing significantly in methods to reduce the disk space needs in several areas as described in the next section.



(a) **(b)**

Figure 3: (a) Distribution the ATLAS data formats on disk from November 2018 until October 2019. (b) Projections of ATLAS disk space needs for the next years [5].

5 ATLAS analysis model study group for Run-3 (AMSG-R3)

The AMSG-R3 was formed within ATLAS in the summer 2018 and delivered a set of recommendations for an updated ATLAS analysis and computing model in June 2019. In essence the group mandate was as follows: collect options to save at least 30% disk space overall (for the same data and MC samples), harmonise the analyses and give directions towards further savings for the HL-LHC. With the planned enhanced usage of fast MC simulation and even fast full event generation, simulation and reconstruction chain, more events can be simulated but demand at the same time more disk resources, which will not be available given flat budget projections.

As described in the previous sections, in the current Run-2 model a rather diverse and large number of DAOD formats are centrally produced. These formats are further skimmed and slimmed by many analysers. Several of these DAOD formats are rather large in size. Overall the AOD and DAOD event information contains lots of low level quantities for all physics objects to allow calibrations and systematic studies very late in the analysis chain. This allows for very flexible physics object definitions but increases the format sizes significantly. This is especially the case in the trigger, MC generator and the inner detector tracking areas.

The AMSG-R3 recommends therefore to significantly reduce the number of DAOD formats and introduce instead a new single DAOD_PHYS targeted for all physics analysis. In addition a new smaller DAOD_PHYSLITE format will be introduced that contains already calibrated physics objects and will be centrally produced with frequent updates, typically every few months. A larger fraction of the AODs will be removed from disk and staged-in back from tape storage on demand in a so called data carousel mode of operation [6]. The AOD and DAOD format contents will be reduced in size in different domains and a lossy float variable compression will be applied where the detector resolution and physics precision allows for this. Figure 4 shows the new production workflows and formats recommended by the AMSG-R3. Table 2 summarises the AMSG-R3 recommendations in different areas.

Table 3 provides an overview of a simple disk space model with the Run-2 numbers using the AMSG-R3 recommendations. As a rough Run-2 input parameter an initial sum of 132 PB of disk space used for AOD and DAOD formats before the AMSG-R3 recommendation is assumed. The new simple model introduces four replicas of DAOD_PHYS and DAOD_PHYSLITE for data and simulated events and only half of the AODs are kept on disk. Some DAOD formats are kept for physics performance group developments and their

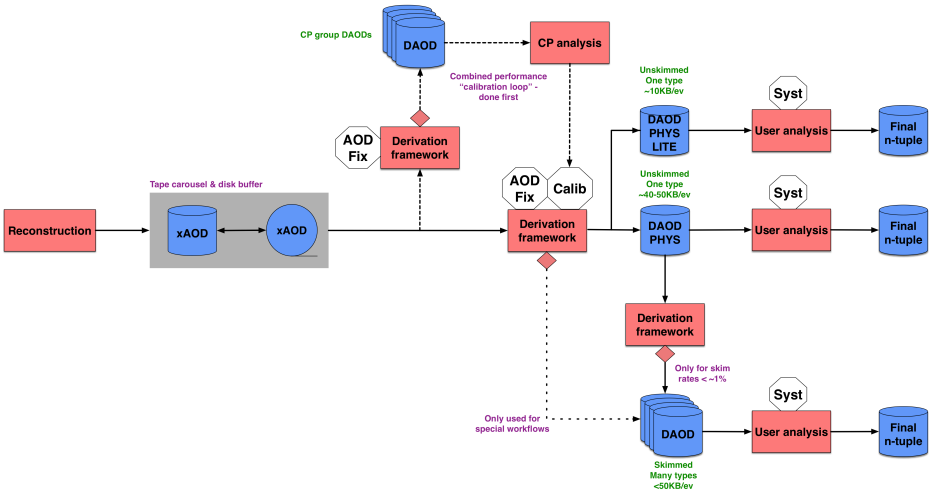


Figure 4: The new production workflows and formats recommended by the AMSG-R3.

Table 2: Overview of the AMSG-R3 recommendations in different areas.

Formats	Introduce DAOD_PHYS with ~50 kB/event Introduce DAOD_PHYSLITE with ~10 kB/event and calibrated objects Reduce the number of DAOD formats by DAOD_PHYS(LITE) in the majority of analyses 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 Changes in DAOD production policies, smarter replica placements, global Rucio file redirector
AOD/DAOD content	Significantly reduced track, trigger, MC generator information Use calibrated objects Apply lossy compression for most variables in AOD/DAODs where feasible and applicable

total size is assumed to be 50% of the current size. The new total sum of all formats is 85 PB, which amounts to a total saving of 46 PB. This extra space allows room for more MC event production and storage of these extra events in Run-3.

Table 4 shows an overview of a very simple disk space model extrapolation for the HL-LHC using the AMSG-R3 recommendations. It uses the assumption that there are five times more DAOD events than AOD events. The DAOD_PHYS and DAOD_PHYSLITE formats are used for the vast majority of the analysis. No pile-up dependence of the event sizes is taken into account. No extra format versions and no extra replicas are considered here in the overall disk space usage. If included these two will increase the disk space usage by a factor in the range of two to four. The disk capacity needs will be reduced if the DAOD_PHYSLITE format is used much more often so that the storage of other DAOD formats can be reduced.

Table 3: Overview of a simple disk space model with Run-2 numbers and the AMSG-R3 recommendations applied as described in the text.

	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

Table 4: Overview of a very simple disk space model extrapolation for the HL-LHC applying the AMSG-R3 recommendations as described in the text.

	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	$2.1 \cdot 10^{11}$	$1.1 \cdot 10^{12}$	$2.1 \cdot 10^{11}$	$5.0 \cdot 10^{10}$	$2.5 \cdot 10^{11}$	$5.0 \cdot 10^{10}$	
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

6 Summary

The ATLAS analysis model has been highly successful in the view of the productivity of the ATLAS experiment at the LHC, but the Run-2 model has been expensive in terms of resource usage. The Analysis Model Study Group for Run-3 has been setup in the past year and collected options to save disk space and harmonise the analyses across ATLAS. It recommends to significantly reduce the content and the number of AODs and DAODs which are the two formats taking more than 70% of the disk space today. The reduction in DAOD formats and content will be compensated by creating two common analysis formats, the DAOD_PHYS and DAOD_PHYSLITE, where the latter contains already calibrated physics objects. The wider usage of this second format should also allow the extra needed disk space savings for the HL-LHC runs. Prototypes and pre-production versions of most of the recommendations already exist and are at present under testing and physics validation.

References

[1] ATLAS Collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider,* JINST **3** S08003 (2008).

[2] J. Elmsheuser et al. [ATLAS Collaboration], *Overview of the ATLAS distributed computing system,* EPJ Web Conf. **214** 03010 (2019).

[3] Worldwide LHC Computing Grid project, URL http://cern.ch/lcg [accessed 2020-01-15]

[4] ROOT, Version 6.18/04 available from https://root.cern.ch/downloading-root [accessed 2020-01-15]

[5] ATLAS Computing and Software - Public Results, URL https://twiki.cern.ch/twiki/bin/view/AtlasPublic/ComputingandSoftwarePublicResults [accessed 2020-01-15]

[6] X. Zhao et al., *ATLAS Data Carousel*, These Proceedings (2020).