

Using TAGs to speed up the ATLAS analysis process

This article has been downloaded from IOPscience. Please scroll down to see the full text article.

2011 J. Phys.: Conf. Ser. 331 032007

(<http://iopscience.iop.org/1742-6596/331/3/032007>)

View [the table of contents for this issue](#), or go to the [journal homepage](#) for more

Download details:

IP Address: 95.172.225.198

The article was downloaded on 23/01/2012 at 08:07

Please note that [terms and conditions apply](#).

Using TAGs to speed up the ATLAS analysis process

W Ehrenfeld¹, R Buckingham², J Cranshaw³, T Cuhadar
Donszelmann⁴, T Doherty⁵, E Gallas², J Hrivnac⁶, D Malon³, M
Nowak⁷, M Slater⁸, F Viegas⁹, E Vinek⁹ and Q Zhang³ for the
ATLAS collaboration

¹Deutsches Elektronen-Synchrotron DESY, Notkestraße 85, 22607 Hamburg, Germany

²University of Oxford, Keble Road, Oxford OX1 3RH, UK

³Argonne National Laboratory, 9700 S. Cass Avenue, Argonne IL 60439, USA

⁴University of Sheffield, Hounsfield Road, Sheffield S3 7RH, UK

⁵University of Glasgow, Glasgow G12 8QQ, UK

⁶LAL, Univ. Paris-Sud, IN2P3/CNRS, Orsay, France

⁷Brookhaven National Laboratory, Upton, NY 11973, USA

⁸University of Birmingham, Edgbaston, Birmingham B15 2TT, UK

⁹CERN, CH - 1211 Geneva 23, Switzerland

E-mail: wolfgang.ehrenfeld@desy.de

Abstract. In the ATLAS experiment, Tag Data, or short TAG, are event-level metadata – thumbnail information about events to support efficient identification and selection of events of interest to a given analysis. TAG quantities range from detector status and trigger information to basic physics quantities, e. g. the number of loose electrons candidates and kinematic information for a limited number of these candidates sorted by their transverse momentum. The average TAG size per event is around 1kB, which is a factor 100 smaller than the Analysis Object Data (AOD) used for physics analysis. TAGs are primarily produced from AODs and stored in ROOT files. For easier access and usability TAGs are also stored in a database. Queries to the database can produce again TAG files. In a standard ATLAS analysis job, TAGs can be used to preselect events based on the TAG quantities before accessing the full AOD content. This allows for a significant speed up of the processing time. This paper will discuss the different analysis work flows using TAGs and compare them with other analysis work flows within ATLAS. Further, the performance for preselecting events using either directly AODs or TAG files is measured and compared. Peak performance is estimated on a single machine with local disk access, while more realistic performance is estimated using Grid like data access.

1. Introduction

The ATLAS experiment recorded roughly half a billion pp events within the first part of the run period 2010 (March to August), corresponding to an integrated luminosity of $\mathcal{L} = 3.4 \text{ pb}^{-1}$. In the second part of the run period 2010 (September to November) roughly 220 million pp events were recorded. This corresponds to an integrated luminosity of $\mathcal{L} = 46.7 \text{ pb}^{-1}$ for the full run period 2010. All detector systems including trigger and data acquisition and data processing and management performed as expected.

Three quarter of a billion events are a huge number, requiring similar large disc space to store them. For prompt analysis of the recorded data, the ATLAS collaboration has to ensure that

their users can analyse this huge amount of data efficiently and fast within the available CPU and storage resources.

It is instructive to consider two extreme analysis work flows before studying the real analysis work flows in more details: If enough storage is available and CPU is limited, one approach is to process data very seldom and write out as much as possible, hence minimising the CPU usage. The opposite extreme is to process data very often and only store minimal information such as the final results and some plots, if storage is the bottleneck. Realistically, both CPU and storage resources are limited and the user needs to find a compromise between both extremes. Further, a typical analysis work flow consists of at least two steps. The user starts from the experiments data formats and produces some intermediate data by selecting a reduced number of events and only storing the required event information for his work. In a second step, which usually is performed more frequently, the intermediate data is processed to perform the final analysis. The size of the intermediate data sample in terms of disc space strongly influence the turn around time of data analysis and in the long term the physics output of the experiment. The optimal working point is influence by many more aspects, including, but not only, the available CPU resources for the first and second step and the available storage resources for the intermediate data.

This paper presents the use of ATLAS TAGs to speed-up data analysis at the first step. Further, this approach is compared to various other approaches within ATLAS to optimise the analysis work flow in terms of performance. The different ATLAS data formats are introduced in section 2 and give details about the content, size and purpose of the various data formats. In section 3 the TAG format and its use for fast event selection is described in more details. The performance of TAG based event selections is studied in section 4. Finally, the paper concludes with a summary and outlook in section 5.

2. ATLAS Data Formats

The ATLAS data flow and the various data formats are sketched in figure 1. From the RAW detector (or simulation) data different data formats are sequentially reconstructed. The Event Summary Data (ESD) format contains all reconstructed objects including low level information as hits/tracks and cells/clusters. The detailed information is aimed for detector and reconstruction performance studies. The average size is ~ 1500 kB per data events. The Analysis Object Data (AOD) format is derived from the ESD format. It is aimed at physics analysis and therefore contains mainly higher level reconstruction objects as jets, electrons, photons, muons, taus, missing transverse energy and more. The average size per data event is ~ 130 kB. This is much smaller than the ESD size and only possible by drastically reducing the content of the lower level information as hits and cells. The TAG format concludes this chain. It is derived from the AOD format and contains a thumbnail of the event information, including trigger information and performance and physics objects. The average size per data event is ~ 0.3 kB. For more details see section 3. Figure 2 illustrates the size per event for the various data formats. Recently, derived data formats for ESD (dESD) and AOD (dAOD) have been introduced. They are produced either from ESD or AOD, respectively, and are aimed at specific tasks. The idea is to reduce the overall dataset size and therefore improve the processing time by either skim events or slim, trim or thin the event content. One example is a dedicated γ -jet data sample derived from ESD for jet calibration studies. For more details on derived formats see reference [1].

The size of the full event sample recorded by the ATLAS experiment in the first part of 2010 ($\mathcal{L} = 3.4 \text{ pb}^{-1}$) is already quite big to analyse every single event. On the other side most data analyses do not need to look at every single event. Appropriate preselection cuts can be applied.

Within the ATLAS collaboration different options exists to reduce the overall dataset size for a given dataset: The different data formats as AOD and ESD have different levels of detail

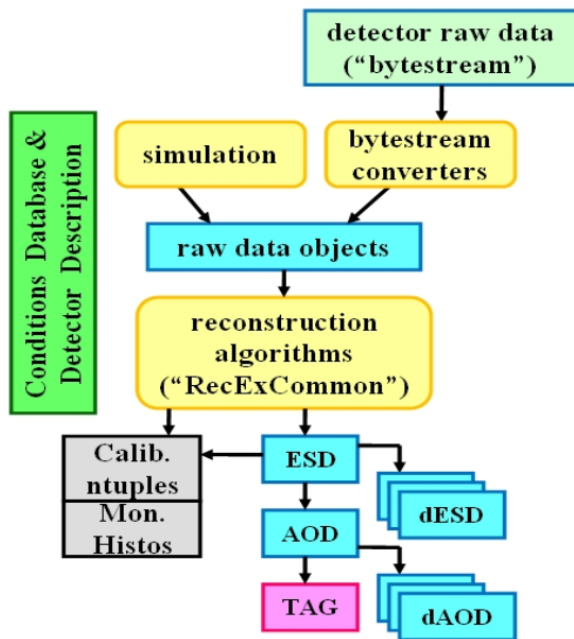


Figure 1. Schematic view of the ATLAS data flow for collision and Monte Carlo data. The different ATLAS data formats and their dependencies are sketched in the lower part.

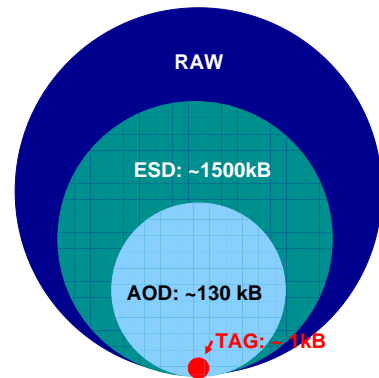


Figure 2. Illustration of the event size for the various ATLAS data formats.

and therefore a different size per event. Further reduction of the data content can be achieved with the derived data formats as dESD and dAOD. For example, the total size of the ATLAS data ($\mathcal{L} = 3.4 \text{ pb}^{-1}$) is $\sim 700 \text{ TB}$ for the ESD format and only $\sim 60 \text{ TB}$ for the AOD format.

After the final trigger decision events are streamed into different datasets based on certain trigger decisions. For example, all events with an electron or photon trigger fired go into the Egamma stream. Further streams are the Muon, JetTauEtmiss and MinBias stream. The total ESD size for the different streams is 266, 160, 126 and 150 TB, respectively. The total AOD size for the different streams is 24, 15, 13 and 9 TB, respectively. By selecting the appropriate trigger stream the dataset size can be reduced by a factor 2 – 6.

As discussed above the derived data formats as dESD and dAOD do not only allow for a reduction of the data content but also allow an event selection, called skimming. This can be used to further reduce the data volume. The performance and physics groups are responsible to define dESD or dAOD selections for their tasks. For examples, the SUSY dAOD implement a loose event skimming. The total data volumes is only 7 TB.

TAG based event selection or skimming is another way to reduce the overall dataset size. The additional advantage is that the event skimming is done only from the very compact TAG format and hence allows for a quick skimming. The total size of the TAG format is only $\sim 140 \text{ GB}$.

3. TAGs

TAGs are event-level metadata. Or in other words, TAGs are thumbnails of the event content. On one side TAGs provide enough event information for fast event selection and on the other side the possibility to connect the selected event back to every ATLAS data format (AOD, ESD, RAW). Any standard framework job can be executed. Either loose event skimming for performance or physics analysis or dedicated event picking to study more details of a few events.

The TAG content consists of 295 attributes and 3 POOL tokens for back navigation to either

the AOD, ESD or RAW data format. As dESD and dAOD are produced orthogonal to the ESD to AOD to TAG chain, back navigation to the dESD and dAOD formats is technical not possible. The TAG content consists of basic event information as run, lumi block and event number. Further, detector status and trigger decisions are stored. For more performance related selections some limited hit/track and cell/cluster information is stored. Finally, for every physics object the total number of candidates is stored. For a certain number of candidates (jets - 6, electrons - 4, photons - 4, muons - 4, taus - 2), which are p_T sorted, more information is provided. This includes kinematic information as p_T , η and ϕ and more specific information, e.g. the tightness of an electron candidate or isolation of a muon candidate. Missing E_T and sum E_T are also contained in the TAG. There are two interesting additions: Every performance and physics group can encode up to 32 different event selections into a TAG word. Also, the official dESD and dAOD selections are encoded into a TAG word.

TAG based event selections are fully supported by the ATLAS distributed analysis tools, e. g. **GANGA** and **pathena**.

TAG files are flat ROOT files including back navigation information to the RAW, ESD and AOD data format. They can be read directly by the ATLAS software framework and a TAG selection can be performed before the full event is read-in and processed by the framework. This ensures that TAG based event selections can be quickly done. For more details see section 4. The size of the TAG content per event is 0.3 kB. The total size of the TAG format stored in the TAG files is ~ 140 GB ($\mathcal{L} = 3.4 \text{ pb}^{-1}$).

The TAG content is also stored in an Oracle database to gain from the advanced database query technology to have efficient and fast queries of the TAG content. The size of the TAG content per event is 1.7 kB, including extra space for additional indexing information. The total size of the TAG format stored in the database is ~ 900 GB ($\mathcal{L} = 3.4 \text{ pb}^{-1}$).

The user access to the database is done via a web interface, named iELSSI: the ATLAS interactive Event Level Selection Service Interface [2]. iELSSI is designed to easily and quickly develop interactively a TAG based event selection. It provides a query setup wizard, which guides the user through the different selection steps and displays the available options, therefore minimising user mistakes. The query results can be presented in various formats: starting from the number of selected events, a simple run and event number list with back navigation information or a full TAG file for all selected events. TAG content can also be displayed in histograms. The final result in form of TAG files can either be used directly by the user as described in section 3 or forwarded to the TAG skimming service, which can extract the selected events in the various ATLAS data formats.

For high availability the TAG database is distributed over many Oracle instances hosted at different ATLAS institutes. iELSSI distributes transparently the TAG queries to the different database instances. For more details see references [3, 4, 5].

4. TAG Performance

In a first step the following ideal test case is studied to estimate additional overhead and processing speed up introduced by a TAG selection. It is based on a standard ATLAS job where a large fraction of the AOD content is stored in a plain ROOT ntuple. This is a typical user application. The default option is to produce a ntuple from all events in an AOD, without any TAG selection involved. This is compare to a job, where a variable amount of events is selected using the TAG mechanism. The selection rate S defines how many events are selected, e. g. $S = 1$ for every event, $S = 2$ for every second event, $S = 5$ for every fifth event and so on. For the purpose of simplicity and reproducibility the TAG selection is based on a random number.

In this test case roughly 15000 events are processed. Both TAG file (3.6 MB) and AOD file (2 GB) are stored on the local disk. This TAG file is only a subset of a standard ATLAS TAG

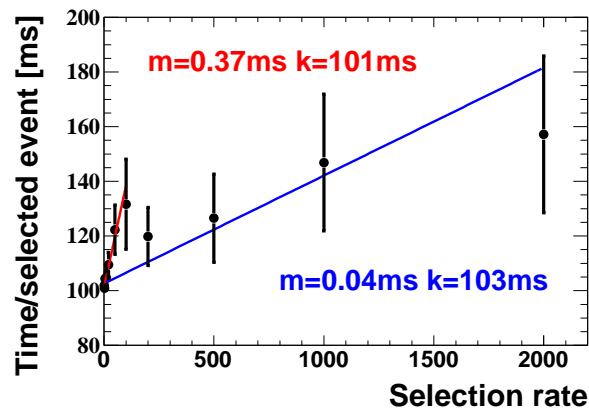


Figure 3. Processing time per selected event as a function of the selection rate for the ideal test case, where TAG and AOD file are read from local disk.

file containing only the events of the used AOD file. The standard TAG file is much bigger (around 40 MB) and contains events from many AOD files. The used machine has a modern 4 core Intel CPU. The absolute time measurement is not very important as the main goal of this study is to compare the case where no TAG selection is used to the TAG selection with different selection rates S . The output ntuple (0.9 GB for $S = 1$) is also stored on the local disk. The processing time is measured using the ATLAS framework monitoring tool (PerfMon) and the UNIX time command. Both approaches agree with each other and only the more detailed PerfMon measurements are shown in the following.

In figure 3 the processing time per selected event as a function of different selection rates S is shown. The error bars reflect the spread in processing time per event. Already for $S = 200$ the number of events is limited and larger statistical fluctuations can occur. For the processing time per event one would expect a linear behavior as the time per selected event includes the TAG query for S events and the event processing time for the selected event. Figure 3 shows two distinct groups of measurements. The measurements within each group can be described by linear line. In the case of low selection rates S , the fitted time per TAG selection is 0.37 ms and the time per event processing is 101.2 ms. The time for one TAG selection is negligible compare to the time for the full event processing. The selection rate S can directly interpreted as the speed-up factor. For $S = 2$ the speed-up factor is 1.99. For large S this is not fully true any more. For example, the case of $S = 200$ will give a total processing time of 470 ms, which is a factor 200 smaller than the total event processing time without TAG processing. A linear fit to the measurements for larger S give even better result: 0.04 ms for one TAG selection and 103 ms for the full event processing, which is consistent with the previous fit and a measurement from the case without TAG selection. The overall processing rate also needs to be compared to the overhead for job initialisation and finalisation, which is roughly 90 seconds. The point even is already reach for $S = 16$ for this test case. Nevertheless, there is no apparent reason to process more than one AOD per job.

In a second step the test case is updated to a more realistic data storage technology. The AOD file is read from a local dCache system, similar to what a Grid job would do. The time per selected event is shown in figure 4. A linear fit to the measurements for $S = 2 - 50$ gives 8.8ms for one TAG selection and 178 ms for the full event processing. In general, the overall picture is quite different from the ideal test case. The processing time per event is much higher than before. This can be due either to the slower dCache access or caching effects for local disc access. Further, there is large increase in processing time between $S = 1$ and $S = 2$, which is

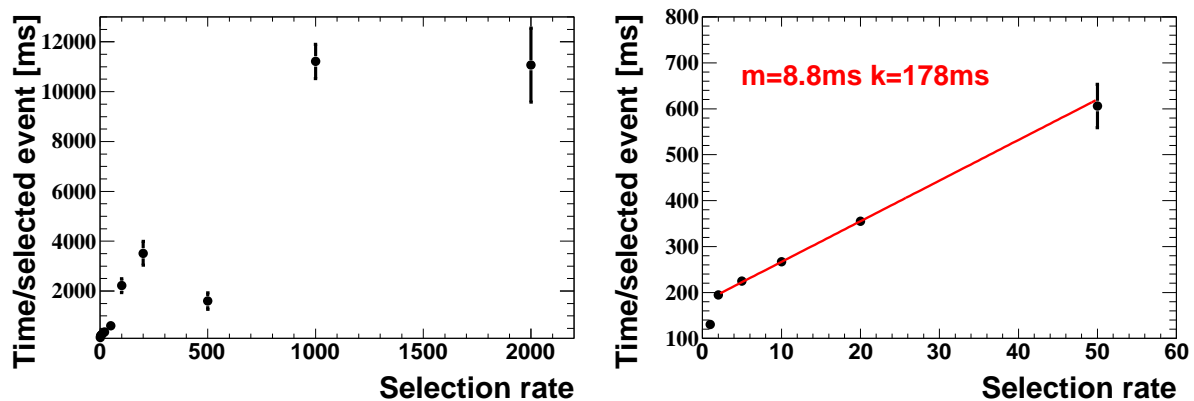


Figure 4. Processing time per selected event as a function of the selection rate for a more Grid like test case, where the small TAG file is read from local disc and the large AOD file is read from a dCache system.

not expected. This can be due to the same reasons, but further investigations are needed to better understand this feature. Still, the time for one TAG selection is smaller than the total processing time per event, although the ratio is smaller than before. The selection rate S does not directly reflect the speed-up factor anymore. For $S = 1000$ the total processing time is nine seconds. This is a factor 20 faster compared to the processing time without TAG selection. This is not as impressive as in the ideal test case but still quite good.

A next step would be to use large scale tests as HammerCloud [6] to estimate the speed-up factor for TAG selections. It is also important to use standard TAG files, which contain the TAG information from more events than from one AOD file. Another important test case is a dESD selection based on the corresponding TAG information. As discussed in section 2 the ESD size per event is much larger than the AOD size per event which could result in a different behavior. It is natural to extend this test case to a physics selection based on AOD files.

5. Summary and Outlook

A fast turn around for data analysis is essential to achieve prompt physics results. This can be obtained by a combination of fast event processing and small data sets. This paper has discussed the different approaches within ATLAS to achieve this goal. In more details the TAG based event selection has been presented as one way to speed up the event selection process. The presented performance studies showed the large speed-up factors for ideal and more realistic test cases. TAG based event selection is one method to speed up the ATLAS analysis process.

References

- [1] Köneke K for the ATLAS Collaboration 2010 Distributing and Storing Required Data Efficiently by Means of Specifically Tailored Data Formats in the ATLAS Collaboration These proceedings (CHEP2010)
- [2] Zhang Q for the ATLAS Collaboration 2010 Engineering the ATLAS TAG Browser These proceedings (CHEP2010)
- [3] Vinek E for the ATLAS Collaboration 2010 Composing Distributed Services for Selection and Retrieval of Event Data in the ATLAS Experiment These proceedings (CHEP2010)
- [4] Hrivnac J for the ATLAS Collaboration 2010 ATLAS Tags Web Service Calls Athena via Athenaem Framework These proceedings (CHEP2010)
- [5] Buckingham R for the ATLAS Collaboration 2010 Metadata aided Run Selection at ATLAS These proceedings (CHEP2010)
- [6] van der Ster D, Elmsheuser J, Ubeda Garcia U, Paladin M 2010 HammerCloud: A Stress Testing System for Distributed Analysis These proceedings (CHEP2010)