The Physics Analysis Tools project for the ATLAS experiment

Bruno Lenzi for the ATLAS collaboration

CEA / Irfu / SPP, Centre de Saclay, F-91191 Gif-sur-Yvette, FRANCE

Abstract. The Large Hadron Collider is expected to start colliding proton beams in 2009. The enormous amount of data produced by the ATLAS experiment (≈1 PB per year) will be used in searches for the Higgs boson and Physics beyond the standard model. In order to meet this challenge, a suite of common Physics Analysis Tools has been developed as part of the Physics Analysis software project. These tools run within the ATLAS software framework, ATHENA, covering a wide range of applications. There are tools responsible for event selection based on analysed data and detector quality information, tools responsible for specific physics analysis operations including data quality monitoring and physics validation, and complete analysis toolkits (frameworks) with the goal to aid the physicist to perform his analysis hiding the details of the ATHENA framework.

Keywords: Analysis tools, ATLAS, LHC
PACS: 29.85.-c

INTRODUCTION

The Large Hadron Collider (LHC) will resume its operations this year, and collisions between proton beams are expected by the end of 2009. The ATLAS experiment, in the search for Higgs boson and signs of physics beyond the Standard Model, will record approximately 1 PB of data per year.

The analysis of this enormous amount of data will be a great challenge for the ATLAS collaboration, composed of more than 2000 physicists. With the starting point and constrains given by the ATLAS computing and event data model, the Physics Analysis Tools (PAT) project was established. A group composed of 2 conveners and about 20 permanent contributors was formed with the aim of developing, maintaining and documenting tools and even complete frameworks used by the community in their analysis. An overview of the tools supported in the context of this project is given in this document.

THE ATLAS SOFTWARE FRAMEWORK AND EVENT DATA MODEL

ATHENA[1] is the standard framework for simulation, reconstruction and physics analyses in ATLAS. It is an implementation of the component-based architecture Gaudi responsible for handling the configuration and execution of several C++ packages through python scripts called *jobOptions*. It takes care of the execution order, data flow and persistification issues. The description of some components typically used on analyses is given below and a simplified scheme of their relations in the ATHENA framework is presented in Fig. 1.

- *Algorithm:* application building block, visible and controlled by the framework, performing a well-defined configurable operation. Runs once per event, calling *tools* and *services*, reading and usually producing data.
- *Service:* globally available software, for common tasks such as data access and message printing.
- *Tool:* lightweight piece of code to execute a specific task one or multiple times per event. Shared and owned by *algorithms* or *services*.
- *Data object:* object-oriented representation of particles (muon, electron) or detector information (cells).

The data formats handled by ATHENA and foreseen in the ATLAS Event Data Model are the following:

- **RAW** data: contains the output of the ATLAS detector, produced by real or simulated events after the High-Level Trigger. It comes in the "bytestream" format as they are delivered from the detector, rather than object-oriented format. The size of each event is approximately 1.6 MB.

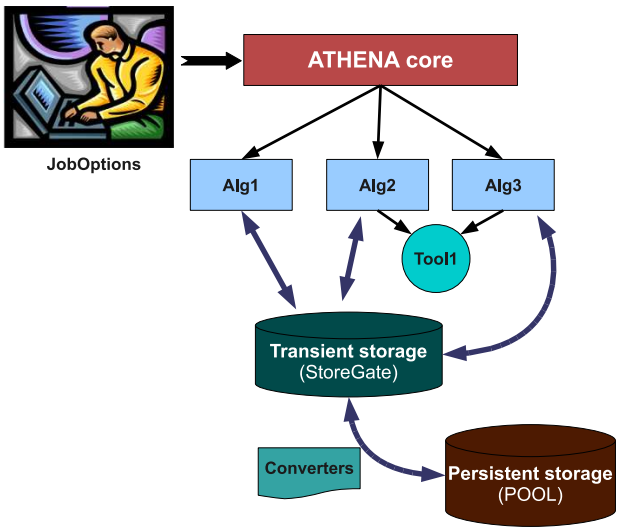

FIGURE 1. Simplified scheme of the ATHENA framework and relations between components.

- **Event Summary Data (ESD)**: holds the output of the reconstruction process. Both detector information and combined reconstruction objects like muons, electrons and jets are stored at this stage. An object-oriented format - POOL / ROOT [2, 3] - is adopted, and the typical event size is 1 MB.
- **Analysis Object Data (AOD)**: a subset of the ESD, with the physical objects used in analysis and few detector objects to allow track-refitting, isolation studies and others. Also stored in POOL / ROOT format, the nominal event size is of the order of 100 KB.
- **Derived Physics Data (DPD):** contains a small subset derived from the AOD / ESD, specific for an analysis or performance group. More than one derivation is possible, in which the data is reduced by removing unnecessary containers, selecting objects and dropping information from those objects. User-data can be added in the process, and in the final stage of derivation a flat ROOT tuple can be produced.
- **TAG**: event-level meta-data containing a thumbnail to efficiently identify and select events for a specific analysis. Can be either ROOT files or databases which are replicated and can be accessed online. Advanced queries can be made and ROOT files, histograms and tables can be produced.

The support of the two lightweight formats described above - DPDs and TAGs - are one of the tasks of the PAT group, which provides tools for their production and usage. Filtering the data is an important step of the analysis process, reducing the storage and processing demands. The usage of these formats has been exercised successfully during the detector commissioning phase.

READING AND ANALYSING THE ATLAS DATA

The standard framework for physics analysis in ATLAS is the ATHENA framework. The great advantage of this approach is the possibility to use all the functionalities provided by the framework in the default input formats. This includes the tools used in the event reconstruction for fitting, calibration, track extrapolation, geometry and magnetic field mapping, just to cite some examples.

On the other hand, the user has to insert his code in a rather rigid structure, described in the previous section, and the performance hardly achieves the standards of plain ROOT data format for instance. Several alternatives are supported by the Physics Analysis Tools project, and some are described in this section.

AthenaROOTAccess

AthenaROOTAccess is a widely used alternative to the ATHENA framework as it allows accessing POOL data (ESD / AOD / DPD) directly from ROOT. Upon initialization - that can be done in C++, python or with the ROOT interpreter CINT - a transient tree object (TTree) is set up and the analysis proceeds as with standard ROOT code using the object-oriented classes from ATHENA.

When a specific object is requested, the persistent data is brought to the transient format using the same POOL converters as in ATHENA, which are invoked automatically on demand. The main limitation of the framework is the lack of components such as services and some tools, which provide access to databases and detector description. As a consequence, data that relies on this infrastructure is not readable.

PyAthena

PyAthena allows the retrieval and creation of ATHENA components in python and the access of C++ components through dictionaries. Being a high-level dynamic typed and interpreted language, python is very powerful for interactive use and debugging. Additionally, it offers the possibility to change and reload a specific module at runtime, avoiding not only the compilation step but also the initialization, which can speed-up considerably the development process.

ROOT tuples and other frameworks

On top of the ATHENA framework, other software was developed in order to standardize and simplify common tasks, such as the configuration of the various analysis packages, data conversion into a set of object-oriented classes which is more adapted to a specific analysis, adding of analysis-specific information (user-data) and production of ntuples for the final analysis. The details of each framework goes beyond the scope of this document, but examples can be found in references [4, 5].

Tools for producing ntuples are also provided by the PAT group, corresponding to the last stage of Derived Physics Data. These tools run in ATHENA and thus can profit from the latest calibrations and developments, making the production simple and reproducible. ROOT is the most popular framework for analysis in High-Energy Physics. Being very portable, it can be installed in the main operational systems and allows the analysis of small amounts of data on virtually any computer. All the previous alternatives require a local installation of the ATHENA software.

COMMON ANALYSIS TOOLS

The PAT group is particularly active in the collection, documentation and integration of tools which are common to many analysis groups. A non-exhaustive list includes packages for data handling, for tuning Monte Carlo simulations with respect to detector performance, tools for measuring and bookkeeping the detector performance and utilities to associate trigger, reconstruction and simulation objects.

A concrete example is the unification of the tools responsible for isolation calculations. Isolation is a measurement of the detector activity around a given particle and it is used for background rejection. It is widely applied in Standard Model analyses and Higgs searches, where information from the inner detector and the calorimeters allows the separation between leptons produced on the decay of vector bosons (W and Z), and leptons coming from semi-leptonic decays of heavy-flavour mesons.

The calculation of isolation variable for different objects such as muons, tracks, electrons and photons present specific challenges which can hardly be met by a unique and simple software. On the other hand, the final information requested by the user is basically common. This allowed the development of common interfaces, leaving the particularities of the calculations for each object to a separate tool.

CONCLUSIONS

With the approach of the collision data taking period at the ATLAS experiment, the Physics Analysis Tools project has the challenging mission of managing part of the analysis software which is used by more than 2000 physicists in the collaboration. On top of the computing and event data model, the PAT group provides ways of filtering, reading and analysing the ATLAS data.

Lightweight data formats are used for efficient event selection and tools for their production are provided. Several alternatives to the ATHENA framework are also supported, including the use of ROOT tuples, reading POOL data with ROOT, using the python language for analysis and interactive debugging, and also complete frameworks that run on top of ATHENA. Common tools within the standard framework are maintained, aiming at code optimization and re-usability. Physics Analysis Tools is an ongoing project and serves as a forum for different communities within the experiment to address their requests and experiences in the software aspects of the analyses.

ACKNOWLEDGMENTS

The author would like to thank the conveners and the members of the ATLAS Physics Analysis Tools group for their help. This work is partially supported by the European Commission, through the ARTEMIS Research Training Network, contract number MRTN-CT-2006-035657.

REFERENCES

1. Atlas computing: Technical design report, Tech. rep., CERN, Geneva (2005).
2. ROOT - An Object Oriented Data Analysis Framework (2009), URL `http://root.cern.ch`.
3. POOL - Pool Of persistent Objects for LHC (2009), URL `http://pool.cern.ch`.
4. K. Cranmer, A. Farbin, and A. Shibata, Eventview - the design behind an analysis framework, Tech. Rep. ATL-SOFT-PUB-2007-008, CERN, Geneva (2007).
5. A. Shibata, Topview - atlas top physics analysis package, Tech. Rep. ATL-SOFT-PUB-2007-002, CERN, Geneva (2007).