Prompt reconstruction of LHC collision data with the ATLAS reconstruction software

Nick Barlow, for the ATLAS Collaboration
E-mail: nbarlow@mail.cern.ch

Abstract. Since the start of LHC operations in November 2009, the ATLAS detector has recorded many millions of collision events. With the luminosity increasing rapidly in 2010, it is a challenge for the reconstruction and data distribution efforts to provide physics-quality data to analysts around the world within a short timescale. In this note I will describe how a two-pass system, where calibration constants and corrections are calculated on a subset of the data before being used in bulk reconstruction, provides a robust and well-monitored mechanism for meeting these challenges.

1. Introduction
The LHC delivered its first $pp$ collisions, at 900GeV centre-of-mass (CM) energy, on 23rd November 2009, and within three weeks, overtook the Tevatron as the highest energy particle collider in the world, with 2.36TeV CM collisions. Throughout the proton physics run in 2010, the LHC has operated at an energy of 7TeV, and the maximum instantaneous luminosity has increased by several orders of magnitude. A challenge for ATLAS is to continue to record, process, and distribute the interesting events such that high quality, calibrated data are available to analysts throughout the world in a short timescale.

2. ATLAS computing model
Computing resources for ATLAS data reconstruction and analysis are arranged in a hierarchical model, with different Tiers. The “Tier-0” is a large farm of computers at CERN. There are several “Tier-1” centers around the world (typically one per country - a major national computing centre), and many more “Tier-2” centres, organized into “clouds” around Tier-1s. In the ATLAS computing model, the vast majority of end-user physics analysis is intended to be performed at Tier-2 centres, while reprocessing of previously recorded data is done at Tier-1s. However, the prompt processing of collision data, which is the subject of this note, is performed at the Tier-0, with the outputs then being transferred to Tier-1 and finally Tier-2 sites over the Grid[1].

3. Data format nomenclature
The output of the ATLAS trigger and data acquisition system (TDAQ) is a set of ByteStream or RAW files. These are the input to the prompt reconstruction. The ATLAS reconstruction software then produces types of output file, mostly based on the POOL/ROOT format [2][3]. These are:
• **ESD** (Event Summary Data): contains both detector-level hit information, and reconstructed quantities such as tracks, calorimeter clusters etc. Most early data analyses were performed on this format, as they required detailed understanding of the detector performance. Typical size is ≈ 1.5MB per event.

• **AOD** (Analysis Object Data): this is a subset of the ESD, containing the objects and quantities expected to be useful for most physics analysis. Typical size is ≈ 150kB per event.

• **dESD(m), dAOD(m)** (derived ESD/AOD (modified)): while the ESD and AOD have a one-to-one correspondance with the events in the original RAW files, many studies do not need all events, or all the information in an event. Several physics and performance groups have devised their own schemas for selected subsets of events (known as skimming, and only copying the useful subset of the event contents (known as thinning, stripping).

• **TAG**: contains event-level summary data, to enable selection of interesting events for downstream analysis. TAG files are produced by the production system in the format of simple ROOT TTrees, and these are later uploaded to a database, which can then be queried to find out how many events would pass different trigger and/or physics requirements.

• Various ROOT file formats are used for monitoring and analysis, sometimes known within the production system as HIST, NTUP, or D3PD.

4. ATLAS data streaming

Data, in all the formats described in the previous section, are organized into different streams, based on trigger chains. The streams are inclusive, meaning that the same event can appear in more than one stream. Broadly speaking, the streams can be categorized as calibration streams or physics streams. Some calibration streams are only used by specific subsystems, and can contain partial events, with only quantities of interest to that subsystem recorded. An important exception is the express stream, which is a subset of the physics data corresponding to a recording rate of about 10Hz, and which is used in the 36-hour calibration loop as described in section 7.

For early 2010 data, the physics streams were based on the hardware (Level-1) trigger: MinBias, MuonswBeam, L1Calo, L1CaloEM, CosmicCalo,CosmicCaloEM. As the luminosity increased, and the software High Level Trigger (HLT) was commissioned, ATLAS moved to the final streaming model for 2010: MinBias, EGamma, Muons, JetTauEtMiss, CosmicCalo.

5. Software

The software releases used for prompt reconstruction of ATLAS data are required to be exceedingly robust and stable. Therefore, new software features are generally placed in development releases, while the production release remains stable between full reprocessing campaigns. Indeed, since the first 7TeV collision data was recorded, this philosophy has been enforced in the frozen Tier-0 policy.

This policy dictates that no changes are permitted in the software if they would change the output of reconstruction. Both data and Monte Carlo simulation are processed with consistent releases, and the data being taken at any given moment can be combined with the output of the most recent reprocessing, with no need for analysts to worry about changes in the underlying reconstruction.

Changes that do not affect reconstruction results (e.g. those that improve monitoring or computing performance), or those that fix crashes in the reconstruction, are allowed under these policy, and these are included in Production caches - lightweight collections of a few software packages, built on top of a full ATLAS release, that can be built and deployed in a matter of hours if necessary.
There are nightly builds of candidate production and development releases, and extensive automatic tests of these builds.

- ATN (ATLAS Nightly Testing) tests run on a very small number of events. There are many of these tests, intended to find crashes when jobs are run with different configurations.
- RTT (RunTime Tester) tests, run on slightly larger samples of events, and more tools are available to analyse the output of the job.

Both ATN and RTT test results are examined daily by dedicated software shifters, who then file bug reports to the appropriate trackers.

There are another two testing frameworks that run on larger samples of events, and have more detailed output. These are:

- TCT (Tier-0 Chain Test): this runs the full Tier-0 workflow, from RAW input, to produce ESD, AOD, and TAG files, and merge them. This test is also used to enforce the Frozen Tier-0 policy - by performing a byte-by-byte comparison of the outputs of two successive days TCT tests, it is possible to narrow down and reject any changes that alter the reconstruction output.
- BCT (Big Chain Test): run on several million events, in order to detect rare crashes.

6. Tier-0
The roles of the Tier-0 are as follows:

- First pass ESD, AOD, dESD, TAG file production.
- Production of histograms for monitoring and data quality assessment.
- First pass calibration and alignment processing.
- Upload TAG files into TAG databases
- Merging of RAW files.
- Registration of data files (RAW and derived) with ATLAS DDM (Distributed Data Management).

Performance of the reconstruction software running at the Tier-0 has been extremely stable with regard to crashes. Very occasionally, jobs may fail due to transient infrastructure problems (e.g. database connections, AFS problems), and these are automatically retried up to three times. Jobs that fail more than three times could be due to genuine software problems, but these have been exceedingly rare in the 2010 data taking. For example, out of the more than 160,000 reconstruction jobs run in October, 19 crashed more than three times, and seven of these were due to corruption in the RAW input files.

Over 4 million jobs have run at the Tier-0 since the start of physics data taking, including reconstruction, merging, and TAG uploading jobs.

Average reconstruction time per event varies according to software release, event content, and pile-up, but is usually of the order of 10 seconds (this includes the production of ESD and all derived file formats, plus execution of monitoring code and production of monitoring histograms). With this rate, and the computing resources available at the Tier-0, it would be possible to keep up with ATLAS data taking if the TDAQ were to output events at a rate of 200Hz, and the LHC was running 100% of the time. In reality, the LHC has been live for much less than 100% of the time (as is to be expected for a new accelerator), so ATLAS has taken advantage of this by increasing the recording rate, sometimes up to 300Hz, which means that triggers with lower thresholds can remain unprescaled.
7. 36-hour calibration loop and data availability.

For LHC collision data, one ATLAS “run” typically encompasses one LHC fill, though detector problems or other considerations may cause a run to be stopped and another one started mid-fill. Within one run, data are divided into luminosity blocks, each with a duration of two minutes.

The data are reconstructed in a two-pass process, illustrated graphically in Figure 1. The key component of this process is the Express Stream, as described in Section 4. For each run, the reconstruction is run on the express stream, and the outputs are sent to the CERN Analysis Facility (CAF). These data are used to assess data quality, and to derive calibration constants and corrections for that run. Some examples of specific calibration tasks that are performed, are beamspot determination, and dead or noisy channel finding in the calorimeters and silicon trackers.

The calibrated run is assessed at the daily Data Quality meeting, and calibration constants are uploaded to the conditions database, before the “bulk” processing of the other physics streams starts automatically, 36 hours after the end of the run. In exceptional circumstances, if calibration constants are not ready in time, it is possible for the start of the bulk processing to be delayed by manual intervention.

Figure 1. Diagram of the work flow of the 36 hour calibration loop.

Figure 2 illustrates the timescale at which data are available for offline analysis, starting from the moment the collision events occur in ATLAS. Note that with the exception of the bulk RAW → ESD processing jobs, all reconstruction tasks can begin as soon as the input files for a given luminosity block are available, rather than waiting for the whole run to finish the previous step. Typically, data are available in a form ready for final analysis, about 50-60 hours after the collisions took place.

For the first few months of collisions, the maximum beam intensities increased so rapidly as the LHC completed various commissioning steps, that the total integrated luminosity recorded by ATLAS could literally double overnight. Therefore, an accelerated strategy was put into place for the summer conferences, whereby the bulk processing was started manually as soon as
the calibration constants from the express stream processing were available. In this way, data from collisions that took place on a Monday, were reconstructed, analysed, and the results were approved, in time to be presented on the Friday of the same week.

**Figure 2.** Diagram showing the length of time after the end of an ATLAS run, before data in various formats become available.

8. Conclusions and outlook
In the original computing model, it was foreseen that only reprocessed data would be sufficiently well calibrated and understood to be used in published physics results. However, several factors: the excellent performance and reliability of the ATLAS reconstruction software and the Tier-0; the frozen Tier-0 policy; and the 36-hour calibration loop; have all combined to ensure that the output of the prompt processing is of excellent physics quality. Furthermore, the worldwide computing infrastructure has performed extremely well in ensuring the data are promptly distributed on the Grid and available for physics analysis within a few days of the data being taken.

Despite these impressive achievements, efforts are continuously ongoing to improve the system, in particular to increase the automation of various steps, such as the calibrations performed on the express stream data.

Overall, the extensive preparations in cosmic ray and single beam data, and the hard work of a large number of dedicated people, have resulted in a system that is ready for the even greater data volumes that will be delivered from LHC collisions in 2011.

[1] The LHC Computing Grid, http://lcg.web.cern.ch

[2] The POOL Project, http://pool.cern.ch

[3] R.Brun and F.Rademakers,
ROOT-An Object Oriented Data Analysis Framework, Nucl. Inst.&Meth. in Phys.Res.A389(1997)81-86.