ATLAS Data Preparation in Run 2

PJ Laycock[1], MA Chelstowska[2], TC Donszelmann[3], J Guenther[4], A Nairz[2], R Nikolaidou[5], E Shabalina[6], J Strandberg[7], A Taffard[8], S Wang[9] on behalf of the ATLAS Collaboration.

[1]University of Liverpool, [2]CERN, [3]University of Sheffield, [4]University of Innsbruck, [5]CEA Saclay, [6]University of Göttingen, [7]KTH Royal Institute of Technology, Stockholm, [8]University of California, Irvine, [9]Academia Sinica, Taipei

E-mail: `paul.james.laycock@cern.ch`

Abstract. In this contribution, the data preparation workflows for Run 2 are presented. The challenges posed by the excellent performance and high live time fraction of the LHC are discussed, and the solutions implemented by ATLAS are described. The prompt calibration loop procedures are described and examples are given. Several levels of data quality assessment are used to quickly spot problems in the control room and prevent data loss, and to provide the final selection used for physics analysis. Finally the data quality efficiency for physics analysis is shown.

1. Introduction

Data preparation on the ATLAS experiment [1] is one of the five activity areas, together with detector operation, trigger, computing and software, and physics. The activity covers the first stage of the preparation of data for physics analysis and produces the primary physics analysis format, the Analysis Object Data (AOD), using calibrations derived in a Prompt Calibration Loop (PCL) running at the Tier-0 computing facility. The activity is also responsible for providing the luminosity measurement and data quality (DQ) assessment.

Data quality in the ATLAS control room uses a new hybrid software release that incorporates the latest offline DQ monitoring software for the online environment. This is used to provide fast feedback in the control room during a data acquisition (DAQ) run, via a histogram-based monitoring framework as well as the online Event Display (ATLAS Event Displays are discussed elsewhere in these proceedings). Data are recorded to several inclusive streams for offline processing at the Tier-0, including dedicated calibration streams and an "express" physics stream sampling approximately 2% of the primary physics stream. This express stream is processed quickly, allowing a first look at the offline DQ within hours of the end of a DAQ run.

The PCL starts shortly after an ATLAS DAQ run ends, nominally defining a 48 hour period in which calibrations and alignments can be derived using dedicated streams. The bulk processing of the main physics stream starts on expiry of the PCL, normally providing the primary physics analysis format after a further 24 hours. Physics data quality is assessed using the same monitoring packages, allowing exclusion down to a granularity of one "luminosity block" (approximately 1 minute). Meanwhile, the AOD is passed to the ATLAS Derivation Framework [2], providing data to users typically within 5 days of the end of a DAQ run, and on the same time scale as the DQ good run list.

1.1. LHC Performance

The performance of the Large Hadron Collider since data-taking started in 2009 has been consistent and good. Outages due to component failures have generally been limited in duration, which led to some confidence in predicting the LHC live time fraction (the fraction of time spent delivering collisions). Live time fractions of around 1/3 were consistently observed and used for predictions of computing storage requirements and this fraction allowed CPU requirements to be based on the average CPU processing time. The large ($\sim$2/3) fraction of downtime meant that prompt processing could keep up with data-taking even if peaks in demand resulted in temporary backlogs.

In 2016 this picture changed, as seen in figure 1. The LHC live time fraction approximately doubled, with prolonged periods of data-taking at 80% live time. Such a huge improvement in performance was very welcome from the point of view of the physics program, but resulted in backlogs for prompt processing which can be seen in the lower panel of figure 1. More and better[1] computing hardware was provided, more efficient software configurations were written, validated and commissioned and all this during the most successful data-taking period of the LHC to date. The details of the changes and improvements made are beyond the scope of this short document and instead an overview is given here of the principal workflows that the data preparation activity uses to quickly produce well calibrated and quality-assessed data that is passed on for further, more detailed physics analysis.

2. Data Preparation

An overview of the Data Preparation workflows for the ATLAS experiment is shown in figure 2. The LHC provides collisions at a rate of 40 MHz, reduced to 100 kHz by the first level trigger system of ATLAS before further reduction using the high level trigger software down to 1 kHz for the final output rate for physics. In addition to this 1 kHz of rate for the main physics stream, there are several other lower rate physics streams and numerous calibration streams for the various subsystems. Most of these streams are then automatically processed using software predominantly based on the Athena framework [3] at the ATLAS Tier-0 computing facility, some of them several times, to produce the calibrations needed for the bulk reconstruction of the main physics stream used for further offline analysis. A key component of the prompt processing is the PCL, described in more detail in section 2.1, which derives updated calibrations to reconstruct the data and produce the AOD. Data quality is assessed at several levels during the procedures and is described in section 3, the result of which is a Good Run List (GRL) which is directly applied by analysers to exclude data that does not pass the DQ criteria. The GRL is used in conjunction with the output of the Derivation Framework [2], a key component of the Run 2 analysis model, that produces a vastly reduced version of the AOD suitable for analysers to process themselves and produce the final results used in publications.

2.1. Prompt Calibration Loop

The PCL workflows are normally completed within 48 hours[2] and are shown in figure 3. The first column of workflows, control room monitoring, is not part of the PCL as such, but plays a crucial role in fast DQ checks. In the ATLAS control room reconstruction jobs are run on a fraction of events from the express stream, producing DQ monitoring histograms within minutes of the data being recorded. These histograms are compared to a reference via a GUI, which in many cases also provides a basic statistical compatibility test. A dedicated DQ shifter in the control room checks the histograms regularly to spot problems, providing fast feedback in the

[1] The best performing Tier-0 hardware had SSDs and relatively large amounts (4GB) of physical memory per compute node, and it was this hardware configuration that was reinforced in numbers.

[2] There are occasions when the PCL is extended, for example after a technical stop or any other occasion where large mis-calibrations or mis-alignments can be expected.

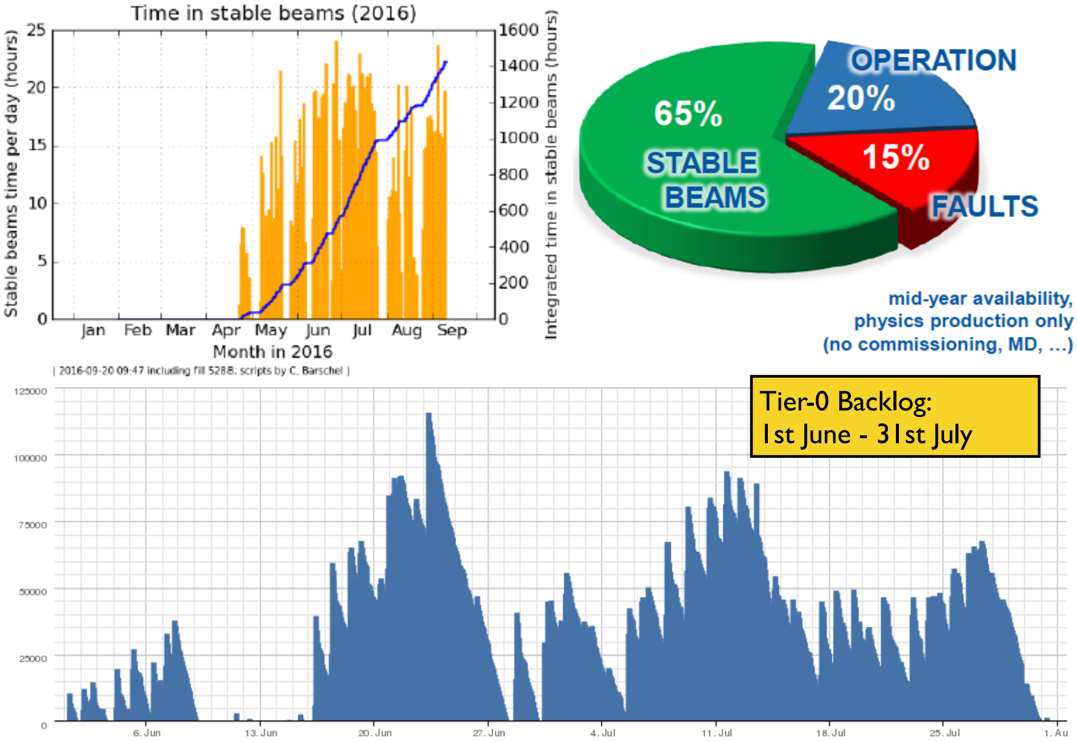


Figure 1. LHC Performance: Top panels show the time spent in delivering luminosity, as hours per day (left) and as an overall fraction (right). The bottom panel shows the backlog of reconstruction jobs at the Tier-0 as a function of time for the months of June and July in 2016.

control room and minimising data losses. The second column of workflows in figure 3 is also for DQ assessment. The express stream and CosmicCalo stream (which samples events triggered by a cosmic ray signature in the calorimeter in empty bunches) are processed at the Tier-0 using the best available calibrations. This is not the final physics quality possible from the data as that requires important updates in the PCL, e.g. the position of the beamspot is not available at this time. Nevertheless the quantity and quality of data is sufficient to perform a first pass DQ assessment which is relevant to the PCL procedures, as the express and CosmicCalo streams provide a broad coverage for potential problems. Problems spotted from this DQ assessment may indicate that further investigation is needed and thus provides an early warning to the data preparation team that the PCL may need to be extended until the issue is resolved.

The third column of workflows represents the vast majority of the PCL procedures. There are many dedicated calibrations needed to maximise the detector and physics performance, ranging from correcting detector mis-alignments to determining the beamspot position. As the beamspot is particularly sensitive to the alignment of the inner detector (ID) this introduces an important dependency in the PCL workflows, i.e. the beamspot determination has to be performed after the ID alignment is finalised. Due to the increased constraints on Tier-0 computing resources in Run 2, a lot of effort was spent in optimising the various procedures and several new workflows were introduced. A good example is the improved ID alignment described in section 2.2.

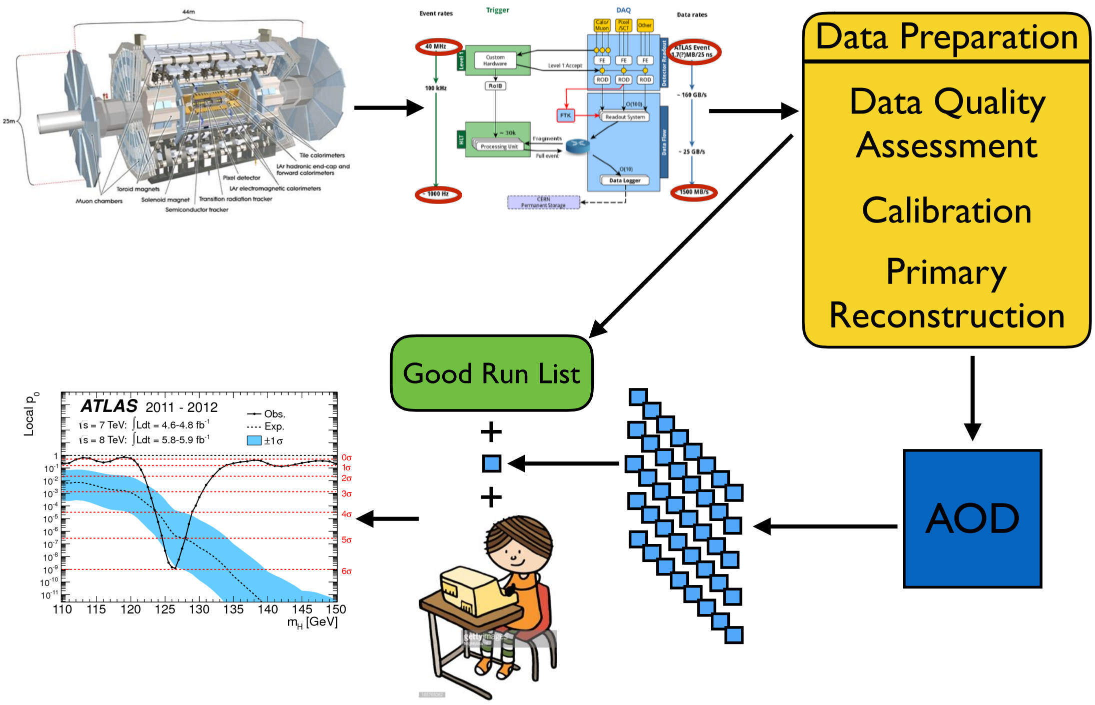


Figure 2. A summary of the main Data Preparation workflows involved in providing quality-assessed data for physics analysis, see text for details.

2.2. ID Alignment

In late September 2015, a significant mis-alignment of the newly installed inner most layer of the ID, the insertable B layer (IBL), was observed. The layout of the ID is shown in figure 4 (top left). The IBL detector is closest to the beampipe and therefore plays a crucial role in determining the beamspot, which in turn has a large impact on physics performance. Such a significant mis-alignment is highly undesirable and must be corrected for and the experts quickly defined a correction procedure [4]. The correction worked, as can be seen in figure 4 (bottom panel) with blue showing the uncorrected residual and red showing the data after correction (the 1 micron offset from zero demonstrates the precision of the original alignment). However, the procedure took too much time, especially considering that the beamspot has to be determined after the ID alignment has been performed. Further studies were made to reduce the number of events processed and to limit the number of iterations used to achieve a stable and accurate result within (on average) 12 hours.

The procedure corrected average mis-alignment over a fill and were put into standard operation in 2015 shortly after the diagnosis of the problem. However, the IBL was also known to move within a fill, meaning that the resolution of the ID was still reduced (although unbiased), and this is clearly seen in figure 4 (top right). Further improvements [5] to the procedure were made in 2016, which finally allowed for the effect to be corrected to a granularity of one hundred luminosity blocks in the PCL.

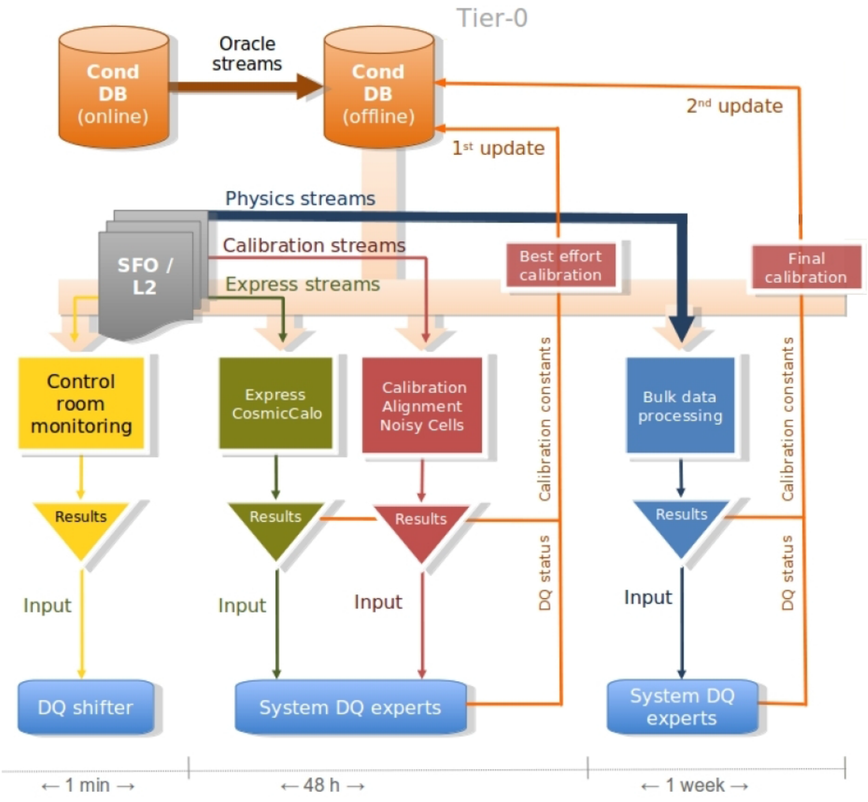


Figure 3. The workflows of the Prompt Calibration Loop, with timeframes increasing from left to right as indicated at the bottom of the figure. Further explanation can be found in the text.

3. Data Quality Assessment

Data quality assessment is made at several levels during the data preparation workflows, as previously discussed. The final DQ assessment required for physics analysis relies on a dedicated DQ Monitoring (DQM) infrastructure which is documented in detail elsewhere [6]. The infrastructure automates many checks based on detector slow control status and DAQ conditions. A team of DQ shifters and experts look at the histograms as presented by the DQM Server and they summarise the results in a database that records DQ problems down to a granularity of one luminosity block. The global DQ assessment then combines these various DQ problems using logic determined by the Data Quality group to produce the final GRL used in physics analysis. The luminosity calculation for a dataset, which is also provided by the data preparation group, is corrected for this loss using centrally provided tools.

3.1. Data Quality Efficiency

The final DQ efficiency (defined with respect to the data recorded by ATLAS when the DAQ and detector final state machine report that they are ready for physics) is shown in the table in figure 5. The individual efficiency for all of the subsystems is very high and generally better than 99%, with the exception of the toroid magnet system that experienced multiple failures. For those analyses that do not rely on the toroid magnet, the DQ efficiency for the ICHEP 2016 dataset was 98%, reducing to 91% if the toroid is required.

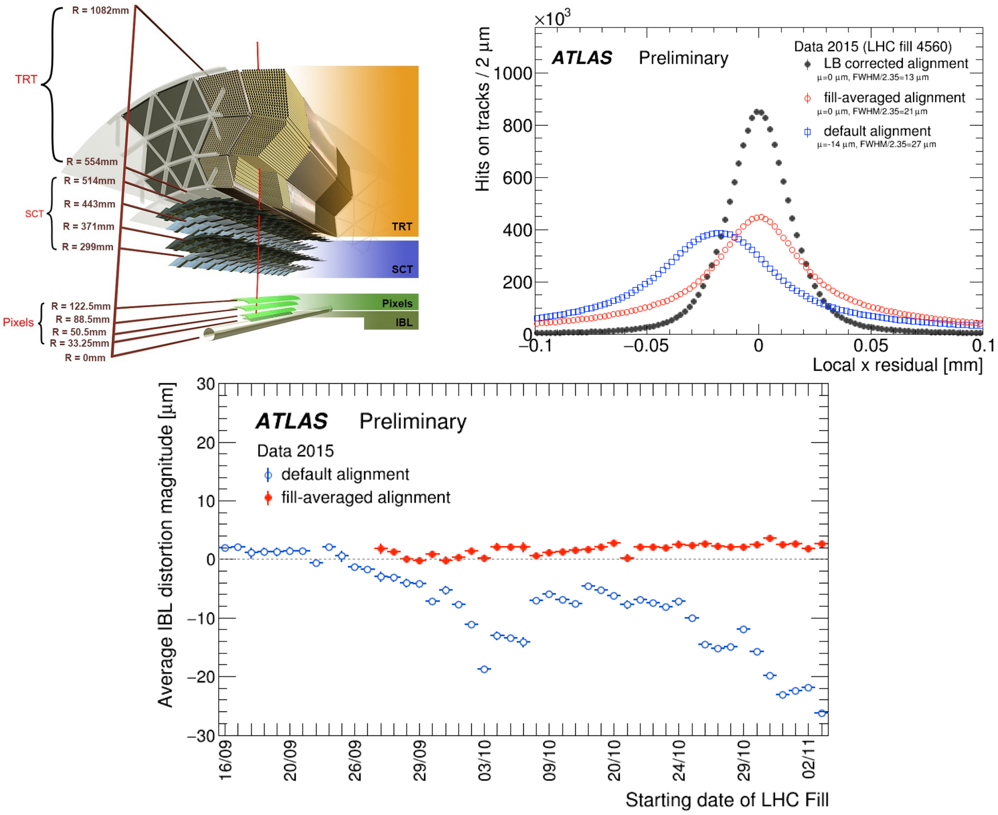
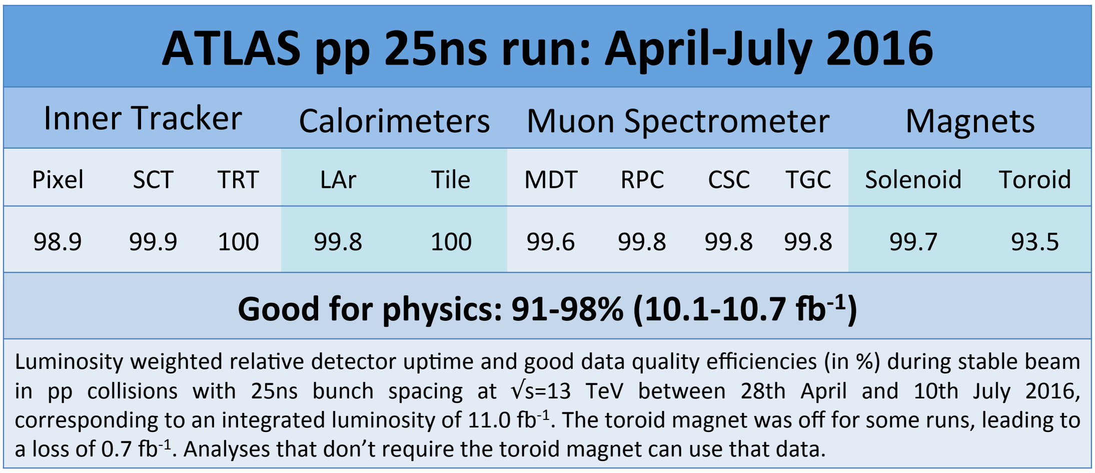


Figure 4. The inner detector alignment: Top left shows a schematic of the ID showing the IBL; top right shows uncorrected (blue), fill-average-corrected (red) and final-corrected (black) IBL hit residuals; bottom shows the uncorrected (blue) and fill-average-corrected (red) IBL residual as a function of time in 2015.



ATLAS pp 25ns run: April-July 2016										
Inner Tracker			Calorimeters		Muon Spectrometer				Magnets	
Pixel	SCT	TRT	LAr	Tile	MDT	RPC	CSC	TGC	Solenoid	Toroid
98.9	99.9	100	99.8	100	99.6	99.8	99.8	99.8	99.7	93.5
Good for physics: 91-98% (10.1-10.7 fb^{-1})										

Luminosity weighted relative detector uptime and good data quality efficiencies (in %) during stable beam in pp collisions with 25ns bunch spacing at √s=13 TeV between 28th April and 10th July 2016, corresponding to an integrated luminosity of 11.0 fb^{-1}. The toroid magnet was off for some runs, leading to a loss of 0.7 fb^{-1}. Analyses that don't require the toroid magnet can use that data.

Figure 5. The Data Quality efficiency for physics analysis for the ICHEP 2016 dataset for collisions recorded in 2016, further details are given in the inset.

4. Conclusion

The data preparation workflows for Run 2 have been presented in the context of the impressive performance of the LHC in Run 2. The increase in live time fraction of the LHC in 2016, by a factor of two on average, placed huge demands on the prompt processing workflows employed by the data preparation group to quickly produce well calibrated data together with timely DQ assessment. Many workflows were improved and optimised and new procedures were implemented to mitigate new problems, in particular correcting for the movement of the innermost layer of the inner detector. The success of the work is well summarised by the high DQ efficiency for physics analysis, and the wealth of physics results presented at conferences and published in physics journals.

References

[1] Aad G *et al.* (ATLAS) 2008 *JINST* **3** S08003

[2] Catmore J, Cranshaw J, Gillam T, Gramstad E, Laycock P, Ozturk N and Stewart G A 2015 *J. Phys. Conf. Ser.* **664** 072007

[3] Calafiura P, Lavrijsen W, Leggett C, Marino M and Quarrie D 2005 The athena control framework in production, new developments and lessons learned *Computing in high energy physics and nuclear physics. Proceedings, Conference, CHEP'04, Interlaken, Switzerland, September 27-October 1, 2004* pp 456–458 URL `http://doc.cern.ch/yellowrep/2005/2005-002/p456.pdf`

[4] Butti P (ATLAS) 2016 *Nucl. Part. Phys. Proc.* **273-275** 2533–2535

[5] Peña J J 2016 *JINST* **11** C11036

[6] Onyisi P 2015 *J. Phys. Conf. Ser.* **664** 062045