Journal of Physics: Conference Series **219** (2010) 072011 doi:10.1088/1742-6596/219/7/072011

The ATLAS Tier-0: Overview and Operational Experience

Markus Elsing, Luc Goossens, Armin Nairz, Guido Negri

CERN, CH-1211 Geneva 23, Switzerland

Markus.Elsing@cern.ch, Luc.Goossens@cern.ch, Armin.Nairz@cern.ch, Guido.Negri@cern.ch

Abstract. Within the ATLAS hierarchical, multi-tier computing infrastructure, the Tier-0 centre at CERN is mainly responsible for prompt processing of the raw data coming from the online DAQ system, to archive the raw and derived data on tape, to register the data with the relevant catalogues and to distribute them to the associated Tier-1 centers. The Tier-0 is already fully functional. It has been successfully participating in all cosmic and commissioning data taking since May 2007, and was ramped up to its foreseen full size, performance and throughput for the cosmic (and short single-beam) run periods between July and October 2008. Data and work flows for collision data taking were exercised in several "Full Dress Rehearsals" (FDRs) in the course of 2008. The transition from an expert to a shifter-based system was successfully established in July 2008. This article will give an overview of the Tier-0 system, its data and work flows, and operations model. It will review the operational experience gained in cosmic, commissioning, and FDR exercises during the past year. And it will give an outlook on planned developments and the evolution of the system towards first collision data taking expected now in late Autumn 2009.

1. Introduction

The Tier-0 facility at CERN is responsible for the first-pass processing of the raw data received from the ATLAS [1] detector, for the archival of raw and derived data on the Tier-0 mass storage system, and for the distribution of the data from the Tier-0 to the Tier-1 centers around the world for further processing and analysis. The Tier-0 must provide high availability and short turn-around and response times.

The software for managing, orchestrating, executing and monitoring the Tier-0 internal data and work flows has been developed at CERN and been thoroughly tested since 2005. Since May 2007, the Tier-0 has been successfully participating in all cosmics, commissioning and single-beam data taking periods and, already in 2008, it proved to be prepared for the data and work flows to be expected during collision data taking. This article will give an overview of those Tier-0 activities in 2008.

2. Tier-0 in ATLAS

2.1. Tier-0 requirements

The Tier-0 receives “streamed” primary (raw) data in “byte-stream” format from the online data acquisition (DAQ) system. The stream types are: *calibration* (usually only partially built events for

© 2010 IOP Publishing Ltd

calibration and alignment purposes), *express* (subset of physics triggers, comprising about 5% of the full data rate, both for calibration processing and to provide a rapid alert on some high-profile physics triggers) and *main (bulk) physics*. The Tier-0 performs fast processing of the calibration and express streams, in order to arrive at (or provide the individual detector calibration and alignment groups with the necessary input to obtain) suitable calibration/alignment constants for first-pass processing of the main physics streams.

Bulk reconstruction/processing of the physics streams by the Tier-0 starts with a latency of 24 to 48 hours, and produces Event Summary Data (ESD), Analysis Object Data (AOD), primary Derived Physics Data (DPD), TAG files (containing event metadata only for fast selection), various n-tuple and histogram files, and log files as outputs. Small files have to be processed further (e.g., merging of the RAW, AODs, n-tuples and histograms), eventually the data (both the original and the derived ones) have to be archived on tape.

The Tier-0 registers all the produced data with the ATLAS Distributed Data Management [2] system (DDM) and the ATLAS metadata catalogue AMI [3]. TAG files are uploaded into various TAG Oracle databases hosted at CERN, BNL and several Tier-2 analysis centers. The Tier-0 is also responsible for supporting offline data quality monitoring (DQM), which runs in parallel with express stream and physics reconstruction, and creates web pages for the offline DQM shifters and experts.

Finally, selected streams have to be replicated to the CERN Analysis Facility (CAF), which serves as a platform for calibration/alignment processing and code development.

In order to accomplish those tasks, the Tier-0 infrastructure needs to cope with running about 10k jobs per day; with managing O(10k) permanent files per day (copying them to/from worker nodes, copying them to tape, replicating them to the CAF) and O(10k) temporary files (unmerged RAW, unmerged AODs, etc.) per day; with writing to mass storage disk at a rate of 880 MB/s, and to tape at 540 MB/s; and with reading from disk at about 1900 MB/s. The uploading of TAG files into each of the TAG databases is required to take place at data-taking rate of about 200 Hz.

The above figures refer to the processing of the bulk physics streams only; it is estimated that calibration and express stream processing will add up to 25% additional load.

2.2. Tier-0 architecture

The Tier-0 system consists of two process entities – an instance of the Tier-0 Manager (TOM) [4] and an instance of the Production System (ProdSys) [5], and a database – an instance of the Production Database (ProdDB) [6]. A detailed presentation on the Tier-0 software suite was given at the previous CHEP conference [7].

TOM manages and "orchestrates" all the processes carried out at the Tier-0, like the definition of (raw and derived) datasets; the definition of (all types of reconstruction and merging) tasks and jobs; the registration of datasets with external catalogues; the replication of data inside CERN for further processing; and the clean-up of temporary files and the archiving of log files. TOM is very modular and easily configurable. Most of the processes derive from a single template. Processes can be loaded at initialization or during run time, and in a similar way their configurations can be dynamically changed during run time.

ProdSys consists of two processes: a facility-neutral supervisor (Eowyn), which interfaces to the ProdDB and is responsible for job processing, management and bookkeeping, and an executor plug-in to Eowyn (T0Executor), which interacts with the CERN-specific batch system technology (LSF [8]) and which is thus responsible for actually running the jobs on the Tier-0 worker nodes, for querying their status, retrieving their outputs, etc.

ProdDB is an Oracle database instance which persistifies the states of both TOM and ProdSys and stores associated logging and monitoring information, together with monitoring data gathered by regularly running cron (*acrontab*) jobs.

The Tier-0 interacts with four primary external entities:

- the SFOs ("Sub-Farm Output" manager instances), i.e., the output processes of the Event Filter ("level-3 trigger" in ATLAS), which transfer the data to the offline mass storage and

enter information about the data (runs, "luminosity blocks", files) into a dedicated database. The communication with the Tier-0 is established through a "handshake" mechanism between the SFO and Tier-0 databases.

- LSF, the CERN batch system, on which the Tier-0 runs all its (reconstruction, merging, archiving, etc.) jobs.
- CASTOR, the CERN mass-storage system [9].
- DQ2, the ATLAS Distributed Data Management system, with which the Tier-0 registers all raw and derived data for further export to external sites.

A schematic view of the Tier-0 architecture is shown in Fig. 1.

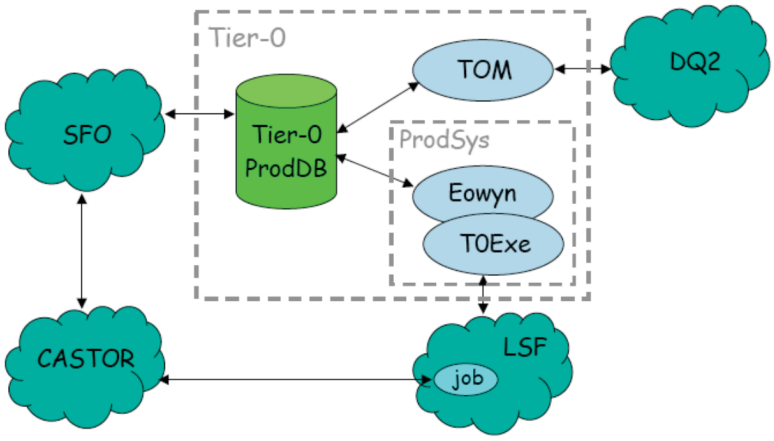

Figure 1: High-level view of the Tier-0 organisation and its components.

2.3. Tier-0 hardware

Currently the Tier-0 hardware consists of

- two CASTOR pools: *t0atlas* (38 disk servers, 200 TB, with tape back-end), for storing permanent (raw and derived) data and used in common with the SFOs and DQ2; and *t0merge* (15 disk servers, 80 TB, disk only) for storing transient data and log files;
- a dedicated LSF batch farm of about 200 nodes (about 1500 cores of each 2 kSI2k);
- an Oracle ProdDB instance;
- two Tier-0 server machines, one for production, one for development and as a spare.

This set of hardware has proven sufficient to tackle the "nominal" data and work flows and to meet the requirements as formulated in the ATLAS Computing Model [10].

2.4. Monitoring and shift system

Extensive monitoring for the Tier-0 system is in place. Based on cron jobs collecting information from various sources (e.g., the Tier-0 database, AFS, Lemon [11] – the CERN IT monitoring tool, etc.), it records the information both in the Tier-0 database (for permanent reference) and in a Round Robin Database (RRD) [12] archive. RRD tools are then used to create graphs in a web-readable AFS directory.

All the information is accessible from a comprehensive set of monitoring web pages which are used both by system experts and shifters. Besides the monitoring web pages proper, shifters are also provided with all necessary web interfaces to interact with the system in a safe way, an electronic logbook, access to a dedicated mail account, documentation and manual TWiki pages. The shift system is well integrated with the Operation Task Planner (OTP), the official tool adopted by ATLAS for accounting shift activities within the collaboration.

In 2008, it was possible to gain experience with the shift system during a data-taking period of about four months, from July to October. During that period, the contents of the monitoring pages and interfaces got continuously improved, taking into account the feedback from the shifters.

So far, the team of experienced shifters consists of approximately 25 people; this number proved to be sufficient for running two 8-hour shifts per day. More shifters will be recruited once data taking requires 24/7 shift coverage; a rough estimation yields that about 35 people in total would be necessary for smooth operations.

3. Tier-0 activities in 2008

ATLAS Tier-0 test activities started already in 2005, with functionality and throughput test series throughout 2006 and 2007. Some of them were previously presented at a CHEP conference [13].
Since May 2007 ("milestone week" M3) the Tier-0 has been integral and indispensable part of all major data-taking exercises. The following sections will summarize the exercises with Tier-0 involvement during 2008.

3.1. Computing exercises

To prepare the offline computing and physics groups for collisions data taking (then foreseen to start in late summer or early autumn), two so-called "Full Dress Rehearsals" (FDRs) were carried out in the course of 2008. Their goal was to test, as realistically as possible, the full data processing and analysis chain, from data acquisition to the end-user analysis at remote sites.

FDR-1 was held in the first week of February. About 0.4 pb^{-1} of simulated data were used for the test. FDR-2 was held four months later, during the first week of June, using about 1.5 pb^{-1} of simulated and real cosmic data. Several follow-up exercises of FDR-2 took place during Summer 2008 (FDR-2a, FDR-2b, FDR-2c), aiming at re-processing some of the FDR-2 data with upgraded software and improved calibration.

In addition to the FDRs, the Tier-0 was also involved in a pure computing exercise called CCRC'08 (WLCG Common Computing Readiness Challenge), which was coordinated by CERN IT and took place in early summer 2008. Its purpose was to stress test the IT infrastructure through concurrent data taking and processing by several experiments (mainly ATLAS and CMS, but also short periods together with ALICE).

3.2. Cosmics and single-beam data taking

Data taking in 2008 was organized in roughly four periods:

- a commissioning period until August, consisting mainly of a sequence of "milestone" weeks (M6 in March, M7 in May-June and M8 in July-August), whose goal was to do combined data taking with as many detector systems participating as available/possible, and dedicated weeks for individual sub-detector commissioning (Inner Detector, calorimeters, Muon system);
- a continuous combined data taking period with cosmics from August to beginning of September;
- the short single-beam period around September 10th, when the LHC was operational;
- a long combined data taking period with cosmics from mid September to end of October, until the detector was shut down for repair and maintenance.

As already mentioned above, the Tier-0 was involved and running successfully in all those data-taking periods. Figs. 2 and 3 below show activity overview plots from Tier-0 and Lemon monitoring (raw data rate and running reconstruction jobs for the September-October combined cosmics period; CASTOR pool I/O from March 2008 to March 2009). As an example, in the period from June 24th to October 28th, online DAQ recorded and the Tier-0 processed 460 million (mainly cosmics) events, amounting to a total raw data volume of 1.1 PB. This is already about the data volume foreseen in the Computing Model for a comparable period of collisions data taking and demonstrates again that the Tier-0 is well prepared for the start-up of the LHC.

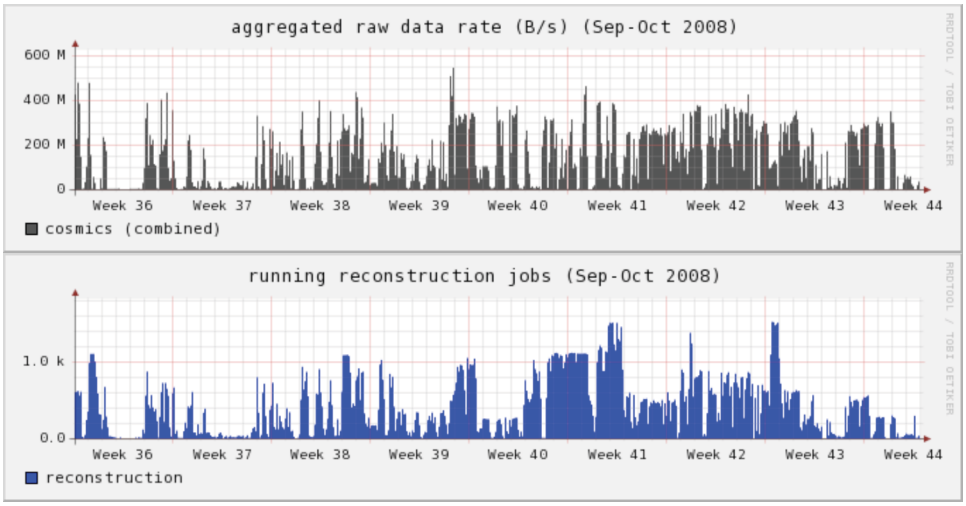

Figure 2: Cosmic raw data taken and reconstruction jobs run in September-October 2008

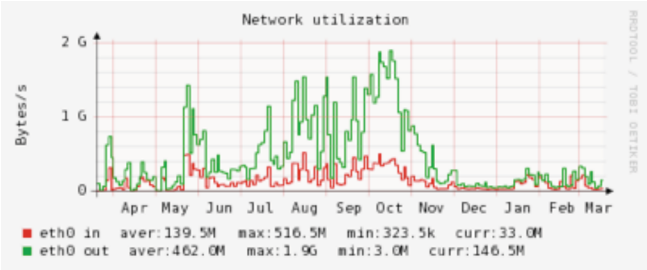

Figure 3: I/O rates on the main Tier-0/CASTOR production pool in the period March 2008 to March 2009

3.3. Tier-0 achievements

The main achievements of the Tier-0, apart from its availability, reliability and demonstrating its readiness for collisions data taking, in a more technical respect, were in particular

- the establishment of a "handshake" mechanism between the SFO and the Tier-0 databases, as only means of communication between the online and Tier-0 areas. This includes also a complicated replication procedure from the online SFO database (which is protected behind the firewall of the ATLAS online network) via Oracle streams to an off-line read-only replica; from there a Tier-0 process eventually pulls the relevant information (about runs, "luminosity blocks" and files) into the corresponding (writeable) tables of its own database.
- the exercising of the full processing chains/cycles for physics, express, and calibration and alignment streams during FDR-1 and FDR-2;
- the development and establishment of extensive monitoring and a fully functional shift system and infrastructure, together with the necessary web interfaces for changing the Tier-0 run configuration and to sign off datasets for processing;
- a successful partial re-design of TOM, to achieve more modularity and flexibility, which became necessary after experiences with the FDRs.

Both the computing exercises and the real data taking activities were also useful to spot out the limitations of the Tier-0 system or its design. It has to be pointed out, however, that often, in particular during the FDR exercises, new requirements were added that had not been there (or at least not clearly formulated) before. Those new requirements were usually quickly reacted to and taken into account, e.g. by partially re-designing TOM (as mentioned above).

4. Tier-0 readiness

In all 2008 activities, the Tier-0 has been working basically fine and reliably. The software (TOM, ProdSys) and the infrastructure (ProdDB, CASTOR, LSF) are in place and proved to be able to handle nominal data flows and carry out all necessary work flows.

The full processing chain has been thoroughly tested and is ready to cope with the expected data rate from the detector. This includes interactions with the offline SFO database for looking up new raw data; the definition of datasets, configuration and definition of tasks and jobs through TOM, and job execution, management and bookkeeping through ProdSys; the successful running of express-stream, bulk physics and calibration (proof-of-principle, during FDRs) processing, and the running of reconstruction, all sorts of merging and uploading jobs; the production of input (histograms) for data quality monitoring (DQM) groups and the creation of DQM web displays (which, in fact, turned out to be among the most "popular" Tier-0 services, also widely acknowledged by detector groups because of its reliability, low latency and fast turn-around); the registration of data with DDM and AMI; the replication of data necessary for calibration/alignment to the CAF; and the running of clean-up procedures (on CASTOR, AFS).

Finally, extensive monitoring of all necessary system variables and a working shift system are in place. For shifters, the monitoring information is readily accessible, problems can be easily spotted, and all the necessary web tools for directly intervening with the Tier-0 system are available.

5. Work plan for 2009/10 and beyond

Based on the experience and shifters' feedback of last year's exercises and data taking, and with new manpower having become available, one of the goals is the development of a new shifters' interface and new monitoring pages. This has already started and a prototype is now (May 2009) ready. The fully functional final version of such a tool will be available for the collisions period in Autumn 2009. It will provide an easier overview on all running processes and their states, including alarms in case of problems, and provide easier means of intervention. Based on the ARDA [14] Dashboard [15] project, it actually develops a new web view mainly developed in Javascript with AJAX [16] technology.

New manpower has also become available for the development of an automated task/job management system for commissioning, calibration and alignment groups. The processing has been done until now by the individual groups themselves, but can benefit from the functionality and infrastructure which is already in place for the Tier-0 (ProdDB, task and job definition, job management, etc.). Some work flows, however, like the iterative ones of TRT calibration or Inner Detector alignment, are quite different from the (usually linear) Tier-0 ones, and require new developments. Requirements for such a task management system were defined in a brainstorming meeting with all the involved groups held at the end of January 2009. Work is in progress and a first prototype is expected to become available in late Spring 2009.

On the long term, the focus will be of course on maintenance and stable operation, with possible refinements of and extensions to the system, based on the operational experience from the cosmics and collisions run periods in 2009/10.

6. Conclusions

The Tier-0 has proven to work stably and reliably during all data taking and computing exercises in 2008. The system itself and all its components (hardware, software, monitoring, shift infrastructure, etc.) are mature and robust enough to handle the expected data rates and throughput for 2009/10 cosmic and collisions data taking.

Until now, experience with full-fledged processing work flows, including the demanding and involved ones of the "calibration loop", could only be gained in the context of the FDR exercises. Nevertheless we believe that the Tier-0 software suite is well prepared for all the work flows necessary

in collisions data taking, and that its design is modular and flexible enough to be adjusted quickly to new requirements.

Finally, the development of improvements and extensions of the existing system, e.g. of new monitoring pages, a new shifters' interface and a task management system for calibration and alignment processing, are well under way and expected to be operational well before the start of collision data taking in late Autumn 2009.

References

[1] https://www.cern.ch/ATLAS
[2] M Branco *et al*., 2008, *J. Phys.: Conf. Ser.* **119** 062017 (proceedings of CHEP 2007, Victoria, Canada), and https://twiki.cern.ch/twiki/bin/view/Atlas/DistributedDataManagement
[3] S Albrand *et al*., 2008, *J. Phys.: Conf. Ser.* **119** 072003 (proceedings of CHEP 2007, Victoria, Canada), and http://ami.in2p3.fr:8080/opencms/opencms/AMI/www/
[4] https://twiki.cern.ch/twiki/bin/view/Atlas/TierZeroManager
[5] https://twiki.cern.ch/twiki/bin/view/Atlas/ProdSys
[6] https://twiki.cern.ch/twiki/bin/view/Atlas/ProdDB
[7] Contribution to CHEP 2007, Victoria, Canada, http://indico.cern.ch/contributionDisplay.py?contribId=123&sessionId=24&confId=3580
[8] http://www.platform.com/Products/platform-lsf
[9] http://castor.web.cern.ch/castor/
[10] The ATLAS Collaboration, 2005, *ATLAS Computing Technical Design Report* (CERN-LHCC-2005-022, ISBN 92-9083-250-9)
[11] http://www.cern.ch/lemon
[12] http://oss.oetiker.ch/rrdtool/
[13] Contribution to CHEP 2006, Mumbai, India, http://indico.cern.ch/contributionDisplay.py?contribId=341&sessionId=8&confId=048
[14] http://lcg.web.cern.ch/LCG/peb/arda/
[15] J Andreeva *et al.*, 2008, *J. Phys.: Conf. Ser.* **119** 062008 (proceedings of CHEP 2007, Victoria, Canada), and http://dashboard.cern.ch/
[16] https://developer.mozilla.org/en/AJAX