Massive Data Processing for the ATLAS Combined Test Beam

M. Dosil, A. Farilla, M. Gallas, V. Giangiobbe, and F. Orellana

***Abstract*—In 2004, a full slice of the ATLAS detector was tested for 6 months in the H8 experimental area of the CERN SPS, in the so-called Combined Test Beam, with beams of muons, pions, electrons and photons in the range 1 to 350 GeV. Approximately 90 million events were collected, corresponding to a data volume of 4.5 TB. The importance of this exercise was two-fold: for the first time the whole ATLAS software suite was used on fully combined real data. Besides, a novel production infrastructure was employed for the reconstruction of the real data as well as for a massive production of simulated events.**

This paper is a report on distributed Combined Test Beam Monte Carlo production with and without grid tools. In 2004, Monte Carlo production was started on the CERN LSF batch system and later it was, for the first time, performed on the LCG, with the simulation of about 4 Million events. In 2005, a more light-weight and user friendly system was tested on NorduGrid for the quick production of 210'000 photon events and 680'000 pion, muon and electron events. New Monte Carlo productions are already planed for the year 2006, using the new ATLAS production system.

***Index Terms*—Computer aided analysis, computer facilities, data management, data processing, distributed computing, nuclear physics, software verification and validation, user interfaces.**

I. Introduction

The experimental setup for the 2004 ATLAS Combined Test Beam (CTB) was composed of elements from all ATLAS subdetectors [1], [2] : Inner Detector (Pixel, Semiconductor Tracker and Transition Radiation Tracker), Calorimeters (Electromagnetic Liquid Argon Calorimeter and Hadronic Tile Calorimeter) and Muon System (Monitored Drift Tubes, Cathode Strip Chambers, Resistive Plate Chambers and Thin Gap Chambers).

The detector was tested for six months with beams of pions, muons, electrons and photons in the energy range of 1 to 350 GeV. About ninety million events with an average size of 50 KB each, were collected and stored on the CERN tape system CASTOR [4], taking up ~ 4.5 TB of space in total.

Manuscript received April 23, 2006; revised July 27, 2006.

M. Dosil is with Port d'Informacio Cientifica, Universitat Autonoma de Barcelona, Universitat Autonoma, edifici D, E-08193 Barcelona, Spain.

A. Farilla is with the Dipartimento di Fisica dell'Universita "Roma Tre" e Sezione INFN, I-00146 Roma, Italy.

M. Gallas is with CERN, CH-1211 Geneve 23, Switzerland.

V. Giangiobbe is with Laboratoire de Physique Corpusculaire, Universite de Clermont-Ferrand 24, F-63177 Aubiere, France.

F. Orellana is with the Departement de Physique Nucleaire et Corpusculaire, Universite de Geneve 24, Quai Ernest-Ansermet, CH-1211 Geneve 4, Switzerland.

Digital Object Identifier 10.1109/TNS.2006.882607

The 2004 CTB was an ambitious pre-commissioning exercise, since, for the first time, the complete software suite developed for the full ATLAS detector was used for real data. Moreover, Monte Carlo simulations were performed with the specifications of the already existing collected real data, so that the physics working groups could perform accurate comparison studies.

II. Data Production and Processing Tools

In order to manage simulation, digitization and reconstruction of such volumes of data, the production systems already used for the series of ATLAS Data Challenges [5] were employed. A number of adaptations and configurations were necessary in order for these systems to be usable for CTB data.

The metadata associated with each produced dataset, like beam energy, beam polarity, etc. was stored in the ATLAS metadata catalog (AMI) [6], which is used by physicists to find the run numbers corresponding to a given physics channel.

Ideally, when running production, it is desirable to work with groups of files making up a dataset, without having to deal with single files and computing jobs: it should be possible to start running a large-scale processing of data by specifying just the input dataset name, the output dataset name and a transformation script. At a later stage, after all jobs have finished, it should then be possible to publish information about the produced data to a central catalog (AMI). By the efforts described in the next section we explored how much of this functionality is available with the systems described above and how suited they are for activities like CTB production.

Apart from the metadata and runtime databases, running a production job typically requires access to four databases at CERN: Geometry information about the detector is stored in a MySQL database called NOVA. Information about the conditions of the detector (e.g., subdetector calibration) for a given time frame are stored as c++ objects streamed into ROOT [7] files. These files are registered in an Oracle database, the so-called conditions database. Information about the liquid Argon calorimeter is kept in two additional MySQL databases.

A. A GUI for the LSF Batch System

For production on LSF (LSF is a batch queuing system) at CERN, input and output data files were read and written directly from and to CASTOR. Moreover, the job scripts themselves and additional input files, like Athena [5] job options, compiled shared libraries and directories with code, were accessed directly from the file system, using AFS. A GUI was developed to manage submission and monitoring of jobs and to have validation and bookkeeping done automatically (see Fig. 1). The GUI is an extension of the Java program AtCom [8],

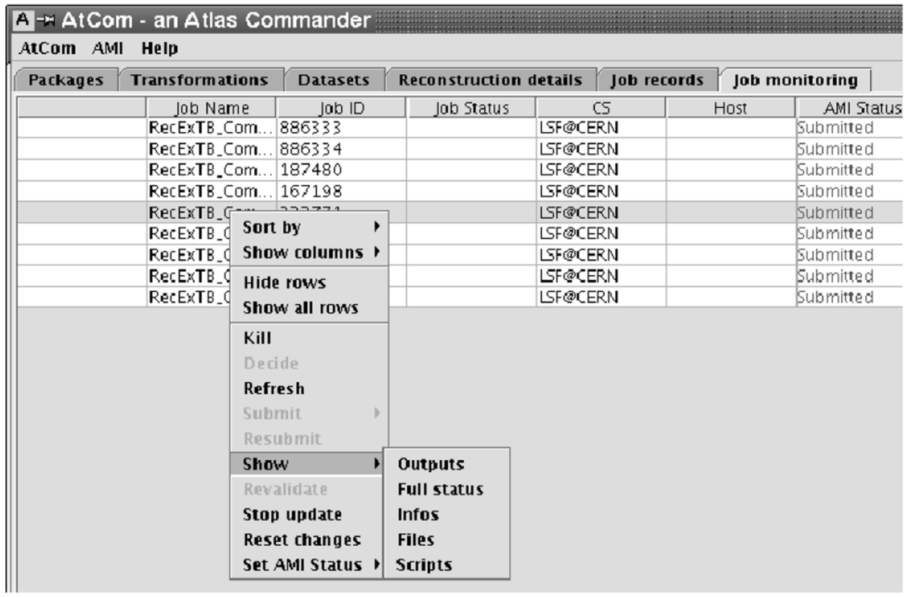

Fig. 1. Monitoring panel of the GUI for running CTB production job.

which was used to manage the production for the ATLAS Data Challenge 1 [9]. Effort was made to make the tool self contained, so users should not need a collection of other tools, like other MySQL interfaces, command-line tools for querying job status, etc. Also, effort was made to end up with one tool, applicable to both the reconstruction of real data as well as the simulation of Monte Carlo data, despite the different database schemas used. The result was a tool which was used by $\sim$ 10 users for processing millions of events (see below).

B. CTB Simulation Jobs on the ATLAS LCG Production System

In spring 2005, the first simulation CTB jobs were successfully run on the LHC Computing Grid (LCG) [10] using the ATLAS Data Challenge 2 production system [11]. The ATLAS Data Challenge 2 production system was made of four components: a central production database (prodDB), a data management system, a supervisor daemon and an executor daemon. The system distinguishes two levels of abstraction: a dataset is a group of files containing events with the same physics specifications. Individual input files are transformed into output files by means of a *job transformation*. The system also operates with the concept of a *task transformation*, describing the transformation of input datasets to output datasets.

CTB simulation jobs were registered in prodDB with a Python script. The supervisor took free jobs from the database and sent them out to the LCG executor it was connected with by exchanging XML messages. The executor then submitted the jobs to the LCG. After jobs finished, a cron job ran and filled the AMI metadata database from the CTB job entries in prodDB.

C. A GUI for NorduGrid

The relative ease with which new users were able to perform production tasks on the LSF batch system, using the mentioned GUI, and the amount of resources available on the grid systems in use by ATLAS, encouraged us to extend the GUI to a light-weight system for running production jobs on a Grid system. This was done using the plugin architecture of the original AtCom. We chose to start with NorduGrid, primarily because it too is light-weight and installable on several clusters of which we had control, avoiding reliance on too many externals. Several new features and work-arounds were also implemented:

- File copying moved from the job shell script to the Grid job description file.
- Reduction of the number of input parameters.
- Automatic job splitting according to number of events per output file.
- New panels for editing and deleting transformation, dataset and logical file records.

Datasets are defined by directly filling the fields of the database record. For each dataset, jobs can be defined. This is done by filling a form with typically 15 values (depending on the schema in use), allowing the use of the special variable $i which is iterated over, and $1, $2, . . . , to access the fields of the dataset record. The interaction with the Grid system is done via an intermediate remote shell started on a server with the NorduGrid client software installed and file system access to the production scripts (e.g., lxplus.cern.ch). After initial tests, we observed the following:

- The failure rate is very low (a few percent).
- Despite the shell interaction with the Grid system, the GUI appears responsive.
- The job submission time is of the order of 1.2 seconds per job when on a fast connection. The submission time increases on a slow connection. All in all, working with more than a few thousand jobs is not manageable.
- Typing in dataset definitions is still too involved. Given the run number, most remaining parameters, like beam energy, etc. should be taken automatically from other databases.

- Filling in the job record form should not be necessary. In most cases, reasonable default values suffice.

We addressed the last two points by providing functionality for creating new datasets on the basis of existing ones, either by cloning or by using the existing ones as input datasets for the new ones. Also, the specification of the location of input and output files was moved from the logical file record to the dataset record, thus simplifying the job definition.

III. Data Production and Processing

In this section we report on the actual data production and processing carried out with the various tools described in the previous section.

A. Real Data Reconstruction on LSF

The reconstruction of CTB data was performed by a package (RecExTB) that integrated all the subdetectors' reconstruction algorithms through different steps. The program gave as output Event Summary Data (ESD)[1] and a ROOT ntuple needed for physics analysis.

For the massive reconstruction of real data, the CERN LSF batch system was used through the aforementioned GUI. Jobs were intentionally not submitted to resources outside of CERN since, at that time, no replication of the conditions database was foreseen and the central database server had access restrictions prohibiting worldwide access.

A subset of 400 good runs, corresponding to approximately 25 Million events, was selected by the CTB community to be used for physics studies, and reprocessing of this data was performed with all major ATLAS production releases (8.8.0, 9.1.2, 10.0.2, 10.3.0, 10.4.0). Large-scale real data reconstruction was carried out twice in 2005 with two different software releases. Half of the runs were processed with the fully combined set-up and half of them with combined calorimetry only. The average CPU time per event was of the order of 1.5 KSI2K seconds (1.5 seconds on a 1000 Specint2000 reference machine [12]). All the runs were split into logical files of maximum 10'000 events each and for each job two types of output files were produced and stored on CASTOR: ROOT files with the combined ntuples (20 KB/event) and POOL [13] files with ESD (40 KB/event). For later user analysis, all this information was kept in the AMI metadata database. These runs were rather successful, with an overall failure rate of about 2%. New reprocessing of the data is already foreseen in 2006 with the next major ATLAS software releases.

B. Monte Carlo Data Production on the LCG

In preparation for the June 2005 ATLAS Physics Workshop held in Rome, a big Monte Carlo data production was carried out using the ATLAS production system and the resources of the LCG.

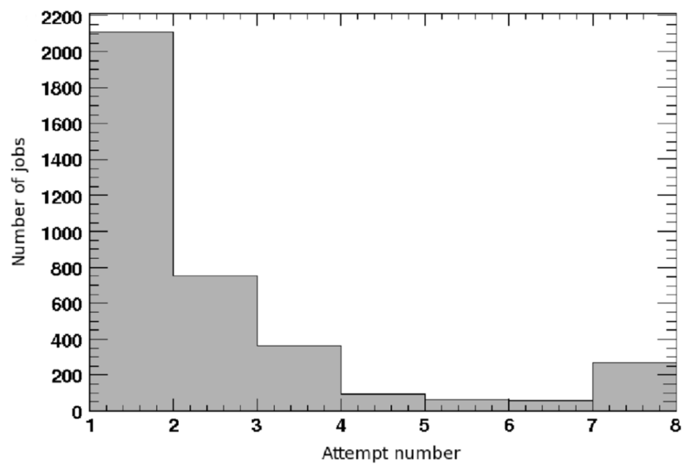

Fig. 2. Distribution of the attempt number that a CTB simulation job reached on the LCG.

The simulation of the CTB setup was performed within a simulation framework based on the Geant4 toolkit [14], [15]. The package "CTB_G4Sim" performed the simulation and digitization of the events. The output of the digitization was reconstructed with the same RecExTB package that was used for real data reconstruction.

The Monte Carlo production for the 2004 CTB was divided in two parts: a preproduction phase and a production one. During preproduction, more than 1.7 Million Monte Carlo events were simulated, digitized and reconstructed using the developed GUI for running on the LSF batch system and AMI for bookkeeping. The goal of this phase was two-fold: to test the different software pieces in order to get a stable version and on the other hand, to compare the first reconstructed Monte Carlo events with the ones from real data reconstruction. The mean processing time per event ranged between $\sim$2 KSI2K seconds for muon simulation and $\sim$50 KSI2K seconds for photon simulation.

In May 2005 the production phase started: a total sample of 392 runs corresponding to almost 4 million events were simulated using the ATLAS production system. This simulation exercise was important because it was the first time ATLAS events were simulated with the same conditions as those of the good "real" runs selected by the CTB physics community. Outputs of simulation were stored on the CERN storage element. 57% of the jobs finished successfully on the first attempt, as can be seen in Fig. 2, while the rest had to be rerun 1.7 times on average. The job failures were due to temporary failures of central services of both the ATLAS production system and the LCG.

Digitization and reconstruction jobs were performed using the GUI for the LSF batch system, since the CTB conditions database was not replicated worldwide as stated before.

C. Monte Carlo Data Production on NorduGrid

Monte Carlo production on NorduGrid was performed with the simulation framework described above. For job submission,

[1]From [5]: ESD refers to event data written as the output of the reconstruction process of the raw data. ESD is intermediate in size between the raw data and Analysis Object Data (AOD). Its content is intended to make access to raw data unnecessary for most physics applications. AOD is a reduced event representation, derived from ESD, suitable for analysis. It contains physics objects and other elements of analysis interest. Both ESD and AOD have an object-oriented representation and are stored in POOL [13] ROOT files.

the GUI for NorduGrid was used (see above). The clusters involved were those forming the Swiss subset of NorduGrid [16].

Initially, 21 datasets with 210'000 photon events were simulated and digitized with release 10.4.0 of the ATLAS software.

- Output data files were copied directly from the clusters to CASTOR at CERN.
- Log files were copied to a directory in AFS at CERN.
- Bookkeeping information on the produced data was kept on the AMI MySQL server in Grenoble.
- Runtime information on the running jobs was also kept on the AMI MySQL server in Grenoble.
- Conditions and geometry data was read from central databases at CERN.
- The average simulation time per event was 25 KSI2K seconds.
- The average event size was 57 kB for the simulated data and 46 kB for the digitized data.

To summarize, this small production employed technology proven to work in previous productions, but on a set of clusters outside of CERN. A number of issues were encountered. Some of them were of system-administration nature, like:

- Network configuration. The downloading of input files via http failed on one cluster due to a proxy misconfiguration.
- Batch system configuration. On one cluster, a non-standard batch system (SUN Grid Engine) was used, requiring some configuration and bug fixing of the grid middleware.
- Software installation. The necessary ATLAS software had to be installed by hand on all clusters.

Other issues were encountered with the ATLAS software, like getting Athena jobOptions right, compiling and including the right additional code and patching bugs in the Athena software. These problems could certainly be alleviated by improving the usability of the grid and ATLAS software, but a discussion of this is beyond the scope of this paper. Of more immediate relevance are the following operational issues:

- Accessing the conditions and geometry databases at CERN was hampered due to their limit of 500 simultaneous connections. Although the current production rarely exceeded 150 simultaneous jobs, the limit of 500 connections was sometimes exceeded when other productions were also running. Increasing the number of allowed connections to a few thousands would fix the problem for now, but in the long run, a more appropriate solutions would probably be to replicate and decentralize the databases.
- Copying output simulated data to CASTOR at CERN and reading it back via gridftp introduced unnecessary complications. The reading was done with a different grid certificate (belonging to a different virtual organization) than that used for writing the data. This caused a different CASTOR stager to be used and the data to be reported as not found.
- Having bookkeeping and runtime information on a server in Grenoble was an unnecessary complication.

To address these, it was decided to:

- Switch to using static replicas of the 4 central databases at CERN: these replicas were all hosted on a MySQL server running on a machine at the university of Geneva.
- Copy and read data files to and from a disk-based gridftp server instead of the CASTOR gridftp server. The replication of output files to CASTOR was then done manually in a separate step.
- Have bookkeeping and runtime information on a local MySQL server and replicate the bookkeeping information to the AMI server in Grenoble after the end of production.

This proved to make the infrastructure more stable and for the next production, failure rates became very low (a few percent). The failures were always due to problems connecting to overloaded gridftp servers. The production consisted of 68 reference datasets, selected by the CTB community, containing 680'000 pion, muon and electron events. In the first run, these events were simulated and digitized with release 11.0.2 of the ATLAS software. The infrastructure put in place should make it an easy task to reprocess these events with coming releases of the ATLAS software.

IV. Conclusions and Outlook

Various tools and systems have been explored for producing and processing CTB data. Compared to the ATLAS Data Challenges and the ongoing centralized ATLAS production, this is a smaller, parallel effort, serving the needs of specific physics communities.

From the computing perspective, the result has been both to provide feedback to the ATLAS software and production teams and to put in place a light-weight system for small to medium scale production.

New CTB Monte Carlo productions are already scheduled for 2006, both using the new ATLAS distributed production system on the three grid flavors and, for producing smaller number of events, the graphical tool described above. Such productions will be of capital importance for finishing the real data - Monte Carlo comparison studies before the start of ATLAS data taking in 2007.

Acknowledgment

The authors would like to thank the physicists of the ATLAS CTB group for a fruitful collaboration. Thanks are also due to the ATLAS production group, which provided several of the tools used, and to the ATLAS database group and the LCG/ARDA group for a stimulating collaboration.

References

[1] ATLAS Collaboration, ATLAS detector and physics performance: Tech. Design Rep., 1, 1999, pp. 9–14, CERN/LHCC.

[2] ATLAS Collaboration, ATLAS detector and physics performance: Tech. Design Rep., 2, 1999, pp. 99–15, CERN/LHCC.

[3] M. V. Gallas, "ATLAS detector simulation: Status and outlook," in *Proc. IEEE Nuclear Science Symp. Conf. Rec.*, Puerto Rico, 2005, vol. 2, pp. 990–994.

[4] J.-P. Baud, B. Couturier, C. Curran, J.-D. Durand, E. Knezo, S. Occhetti, and O. Barring, in *CASTOR Status and Evolution*, 2003, arxiv.org: cs.OH/0305047.

[5] G. Duckeck, ATLAS Collaboration, ATLAS computing: Tech. Design Rep., 2005, CERN-LHCC-2005–022.

[6] S. Albrand and J. Fulachier, "ATLAS metadata interfaces (AMI) and ATLAS metadata catalogs," in *Proc. Computing in High Energy and Nuclear Physics (CHEP '04)*, Interlaken, Switzerland, 2004, pp. 494–497.

[7] "ROOT—An object oriented data analysis framework," *Nuclear Instrum. Methods Phys. Res. A*, vol. A389, no. 1–2, pp. 81–86, 1997.

[8] V. Berten, L. Goossens, and C. Tan, in *ATLAS Commander: An ATLAS Production Tool*, 2003, arxiv.org: hep-ex/0305089.

[9] G. Poulard, "ATLAS data challenge 1," in *Proc. Computing in High Energy and Nuclear Physics (CHEP '03)*, La Jolla, CA, 2003, arxiv.org cs.DC/0306052.

[10] J. Knobloch, [LCG Collaboration] LHC computing Grid: Tech. Design Rep., CERN/LHCC 2005-024.

[11] L. Goossens and K. De, "ATLAS production system in ATLAS data challenge 2," in *Proc. Computing in High Energy and Nuclear Physics (CHEP '04)*, Interlaken, Switzerland, 2004, vol. 2, pp. 943–946.

[12] Standard Performance Evaluation Corporation [Online]. Available: http://www.spec.org

[13] D. Duellmann, "The LCG pool project, general overview and project structure," presented at the 2003 Computing in High Energy and Nuclear Physics (CHEP '03), La Jolla, CA, arxiv.org: physics/0306129.

[14] S. Agostinelli, "GEANT 4: A simulation toolkit," *Nuclear Instrum. Methods Phys. Res. A*, vol. A506, pp. 250–250, 2003.

[15] J. Allison, "Geant 4 developments and applications," *IEEE Trans. Nucl. Sci.*, vol. 53, no. 1, pp. 270–278, Feb. 2006.

[16] S. Gadomski, C. Haeberli, and F. Orellana, "Prototype of the Swiss ATLAS computing infrastructure," in *Proc. Computing in High Energy and Nuclear Physics (CHEP '06)*, Mumbai, India, 2006.