ATLAS Data Challenge 2: A massive Monte Carlo production on the Grid

Santiago González de la Hoz[1], Javier Sánchez[1], Julio Lozano[1], Jose Salt[1], Farida Fassi[1], Luis March[1], D. L. Adams[2], Gilbert Poulard[3], Luc Goossens[3], and DC2 Production TEAM (ATLAS Experiment)[3]

[1] IFIC- Instituto de Física Corpuscular, Edificio de Institutos de Investigación, Apartado de Correos 22085, E-46071 Valencia, Spain
{Santiago.Gonzalez, Javier.Sanchez, Julio.Lozano, Jose.Salt, Luis.March, Farida.Fassi}@ific.uv.es
http://ific.uv.es/

[2]Brookhaven National Laboratory
Upton NY 11973, USA
{dladams}@bnl.gov
http://www.bnl.gov/

[3] CERN, European Organization for Nuclear Research, 1211 Geneva 23, Switzerland
{Gilbert.Poulard, Luc.Goossens}@cern.ch
http://www.cern.ch/

Abstract. The study and validation of the ATLAS Computing Model started three years ago and will continue for few years in the context of the so-called Data Challenges (DC). DC1 was conducted during 2002-03; the main goals achieved were to set up the simulation data production infrastructure in a real worldwide collaborative effort and to gain experience in exercising an ATLAS wide production model. DC2 (from May until December 2004) is divided into three phases: (i) generate Monte Carlo data using GEANT4 on three different Grid projects: LCG, GRID3 and NorduGrid; (ii) simulate the first pass reconstruction of real data expected in 2007, and (iii) test the Distributed Analysis model. Experience with the use of the system in world-wide DC2 production of ten million events will be presented. We also present how the three Grid flavours are operated. Finally we discuss the first prototypes of Distributed Analysis systems.

1 Introduction

The ATLAS experiment [1] is a large detector for the study of high-energy proton-proton collisions at the Large Hadron Collider (LHC) [2], presently under construction at the European Organization for Nuclear Research (CERN) and scheduled to start operation in 2007. In the ATLAS Computing Model, after reduction of the data by the online trigger processor farms, the expected volume of data recorded for offline reconstruction and analysis will be of the order of 1 PB (10^{15} bytes) per year. Therefore, in 2002 a series of Data Challenges (DC's) were planned with the purpose of the

validation of the Computing Model, of the complete software suite, of the data model, and to ensure the correctness of the technical choices to be made. A major feature of the first Data Challenge (DC1) [3] was the development and the deployment of the software required for the production of large event samples required by the High Level Trigger and Physics communities, and the production of those large data samples involving institutions worldwide.

It should be noted that it was not possible to produce all the data at CERN, since the resources to perform this task on a reasonable timescale were not available.

The ATLAS collaboration decided to perform these DC's in the context of the LHC Computing Grid project, LCG [4], to which ATLAS is committed, but also to use both the middleware and the resources of two other Grid projects, GRID3 [5] and NorduGrid [6]. The job of the LCG is to prepare the computing infrastructure for the simulation, processing and analysis of the LHC data for all four LHC collaborations. The LCG scope spans both the common infrastructure of libraries, tools and frameworks required to support the physics application software, and the development and deployment of the computing services needed to store and process the data, providing batch and interactive facilities for the worldwide community of physicists involved in LHC. The main emphasis of the LCG project is the deployment of Grid technologies for the LHC computing. Both GRID3 and NorduGrid have similar approaches using the same foundations (GLOBUS) as LCG but with slightly different middleware.

Concerning the ATLAS data analysis model many important software components remain to be done. They will be based on the long term experience of previous experiments and on the emerging new technological breakthroughs. The development and integration of the detector specific reconstruction and physics analysis software, followed by their deployment to the Grid in large scale Data Challenges, will enable ATLAS to validate its Computing Model and to demonstrate its physics potential.

1.1 Scientific originality and innovation

The high particle collision rate and the large event size in ATLAS make the offline computing much more difficult than in previous experiments, even comparing to CDF [7] and D0 [8] (two experiments which are currently running at the Fermilab laboratory in the United States). With respect to these two experiments, the event rate in ATLAS will be a factor of 50 and the event size will be eight times larger.

The offline computing will have to deal with an output event rate of 100 Hz, i.e. 10^9 events per year with an average event size of 1 Mbyte. This means that new algorithms for data reconstruction are needed in order to achieve the required reconstruction latencies and the necessary large reduction of the data volume.

The new Grid technologies will provide the tools to analyze all the data recorded by ATLAS and to generate the large "Monte Carlo" simulation samples required. They are expected to make feasible the creation of a giant computational environment out of a distributed collection of files, databases, computers, scientific instruments and devices.

2 ATLAS production system

In order to handle the task of ATLAS DC2 an automated production system [9] was designed. All jobs are defined and stored in a central database. A supervisor agent (Windmill) [10] picks them up, and sends their definition as XML message to various executors, via a Jabber server. Executors are specialised agents, able to convert the XML job description into a Grid specific language (e.g. JDL, job description language, for LCG). Four executors have been developed, for LCG (Lexor) [11], Nordugrid (Dulcinea) [12], GRID3 (Capone) [13] and legacy systems [14], allowing the Data Challenge to be run on different Grids.

When a LCG job is received by Lexor, it builds the corresponding JDL description, creates some scripts for data staging, and sends everything to a dedicated, standard Resource Broker (RB) through a Python module built over the workload management system (WMS) API. The requirements specified in the JDL let the RB choose a site where ATLAS software is present and the requested computing power is available. An extra requirement is a good outbound connectivity, necessary for data staging.

The actual executable is wrapped in a script that performs various tasks: Check the ATLAS software (s/w) installation on the worker nodes (WN); Download and install packages [15] for the required application; Set up the ATLAS s/w environment; Stage-in the input files, perform the transformation and Stage-out the results.

For data management, a central server, Don Quijote (DQ) [16] offers a uniform layer over the different replica catalogues of the 3 Grid flavors. Thus all the copy and registration operations are performed through calls to DQ. The s/w distribution is installed using LCG tools.

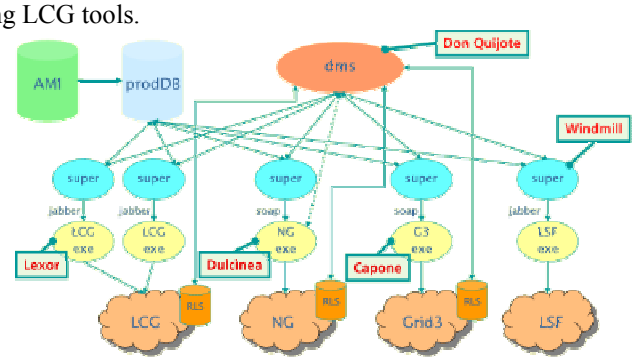

Fig. 1. The ATLAS production system consists of 4 components: The production database; The windmill supervisor; The Executors; Don Quijote, the Atlas Data Management System.

3 DC2 production phases

During the LHC preparation phase, ATLAS has large needs for simulated data which allow understanding the detector performance. These “Monte Carlo” simulations are

done using the full ATLAS chain which runs in the [17] Athena framework [18] and consists of:

- Event generation: (Pythia [19] and Herwig [20] generators), writing out the generated events to ROOT files [21], using POOL persistency [22]. The event size is 60 KB and the computer power required 156 KSI2K-s.
- GEANT4 simulation [23]: reading the generated events via POOL and running GEANT4 simulation, writing out the simulated hits from all ATLAS sub-detectors to ROOT files. The event size is1.9 MB and the computer power required 504 KSI2K-s.
- Digitization: reading in the simulated hits via POOL; writing out the RDO's (Raw Data Objects) to ROOT files. The event size is 1.9 MB and the computer power required 16 KSI2K-s

The relevant output information of the reconstruction will be stored in the form of ESD (Event Summary Data) and in a more compact form, more suitable for analysis, AOD (Analysis Object Data).

The Phase 1 of the ATLAS DC2 started in July 2004 and it is divided into five parts: Event generation and detector simulation; Pile-up and digitization. Pile-up is the superposition of "background" events with the "signal" event and digitization data stores the response of the sensitive elements of the detector. The output, called byte stream data, looks like detector "Raw Data"; Data transfer to CERN (~35 TB in 4 weeks); Event mixing, events from different physics channels are "mixed" in "ad-hoc" proportions. For the Pile-up, the event size is 3.3 MB and the computer power required 144 KSI2K-s and for the Event mixing 3 MB and 5400 SI2K-s.

ATLAS is currently using 3 Grid flavors LCG [4], GRID3 [5] and NorduGrid [6] in different development states. The ATLAS DC2 collaboration finished the simulation part at the end of September 2004. 10 million events were generated and simulated using the three flavors. The contribution from each flavor is shown in Figure 2.

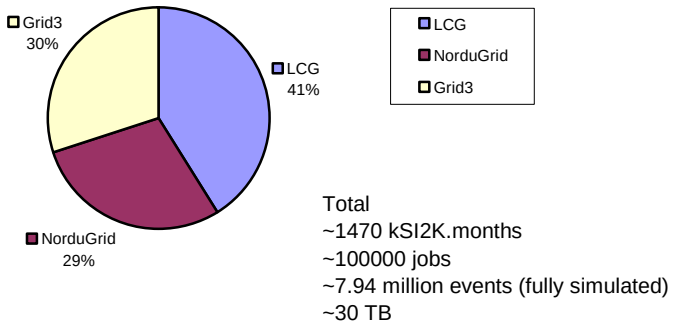

Fig. 2. The chart plots the contribution of each Grid flavor in the generation and simulation parts for the ATLAS Data Challenge.

3.1 DC2 production using GRID3

The GRID3 collaboration has deployed an international Data Grid. The facility is operated jointly by the U.S. Grid projects iVDGL, GriPhyN, PPDG, and the U.S. participants in the LHC experiments.

Fig. 3. Geographical distribution of GRID3.

The deployed infrastructure (see figure 3) has been in operation since November 2003 involving 27 sites, a peak of 2800 processors, work loads from 10 different applications exceeding 1300 simultaneous jobs, and data transfers among sites greater than 2 TB/day.

Figure 4 shows the jobs contribution to the ATLAS DC2 in the simulation part. A specific production system has submitted jobs to GRID3 sites at full scale using their shared facilities. Around 30000 jobs were finished successfully, 2.4 million of events and 8 TB were produced, and more than 0.5 million CPU-hours were consumed.

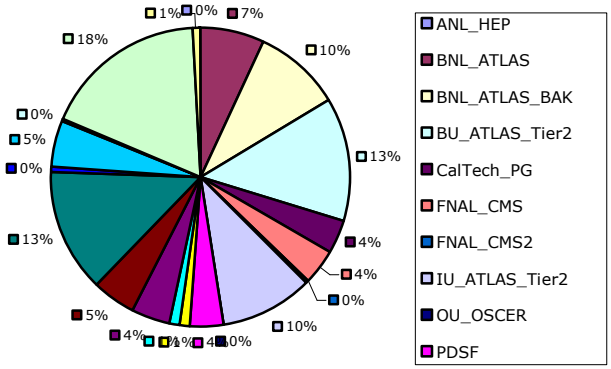

Fig. 4. Jobs contribution by site in GRID3 for the simulation part in DC2.

3.2 DC2 production using NorduGrid

The NorduGrid project is established mainly across Nordic countries but includes sites from other countries. It was the first to reach production quality level and con-

tributed to a significant part of the DC1 production. It provides resources for DC2 and support for production on non-RedHat 7.3 platforms (e.g. Mandrake 8.0, Debian 3.0).

The NorduGrid resources range (see figure 5) from the original small test-clusters at the different physics-institutions to some of the biggest supercomputer clusters in Scandinavia. It is one of the largest operational Grids in the world with approximately 4000 CPU's, storage capacity of 14 TB, involving 40 sites and 11 countries.

Figure 6 shows the jobs contribution to the ATLAS DC2 in the generation and simulation part. Around 30000 jobs were finished and 2.4 million of events were produced.

Fig. 5. Geographical distribution of NorduGrid.

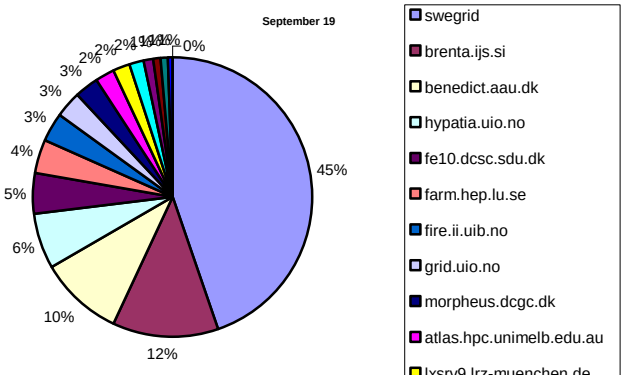

Fig. 6. Jobs contribution by site in NorduGrid for the simulation part in DC2.

3.3 DC2 production using LCG

The LCG project is built upon the most stable developments of the European Data Grid middleware (EDG) [24], the US Virtual Data Toolkit project (Globus) [25] and European DataTag [26] monitoring tools.

The requirements for LHC data handling are very large, in terms of computational power, data storage capacity and data access performance. It is not considered feasible to fund all of the resources at one site, and so it has been agreed that the LHC computing services will be implemented as a geographical distributed Computational Data GRID (see figure 7). This means that each service is using computing resources, both computational and storage, installed at a large number of Regional Computing Centres in many different countries, interconnected by fast networks. The deployed infrastructure has been operating since 2003 with 82 sites of 22 countries at peak of 7269 processors and a total storage capacity of 6558 TB.

Figure 8 shows the LCG jobs distribution in the generation and simulation part. LCG sites ran production systems at scale during this period using shared facilities. Around 40000 jobs were finished successfully and 3.1 million of events were produced.

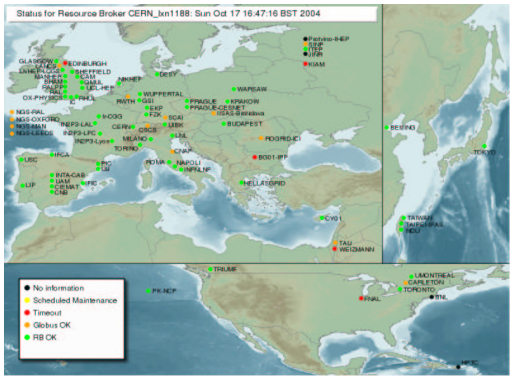

Fig. 7. Geographical distribution of LCG.

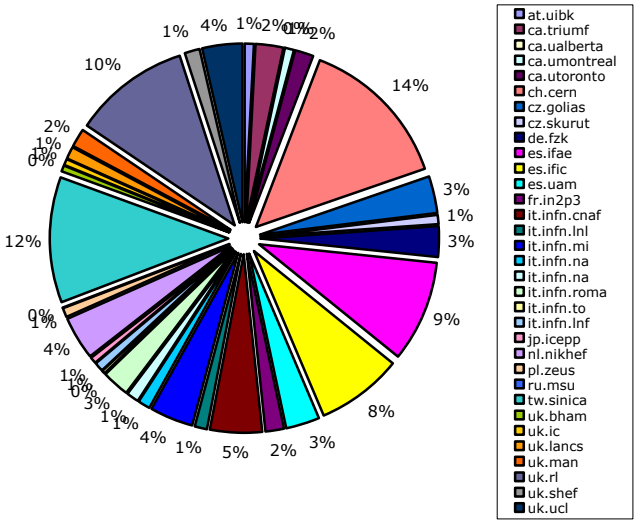

Fig. 8. Jobs contribution by site in LCG for the simulation part in DC2.

4 Experiences

By design, the production system was highly dependent on the services of the Grids it interfaces to. The beta status of the implementation of these services has caused troubles while operating the system. For example, the Globus Replica Location Services (RLS) [27], the Resource Broker and the information system were unstable at the initial phase. But it was not only the Grid software that needed many bug fixes, another common failure was the mis-configuration of sites.

On the other hand, since the beginning of DC2 to the end of September around 6 TB of data have been moved using Don Quijote servers. The automatic production system has submitted about 235000 jobs belonging to 158000 job definitions in the Database, producing around 250000 logical files and reaching approximately 2500-3500 jobs per day distributed over the three Grid flavors.

5 Distributed Analysis system

The Grid infrastructure will be used also by the physicists to perform the analysis of the reconstructed data. The ADA [28] (ATLAS Distributed Analysis) project aims at identifying those software components which will allow the end-user to take benefit from the Grid. The goals are rather challenging since they should cover a wide variety of applications involving production and analysis of large distributed data samples. It is possible to identify a generic way in which the user will interact with the Grid resources: she will define an input dataset which will be processed by an application whose behavior will depend on a set of parameters and/or user's code and it will be transformed into an output dataset. Those elements will characterize the job which will be sent to a high level service in charge of the job submission and output management.

The aforementioned software components should provide a way to implement the different processing elements: input dataset, application, task (user's source code and parameters), and output dataset and submission service. All these components make up the Analysis Job Description Language (AJDL) [29]. The submission component involves different tasks including the localization of resources, the splitting of the jobs, the staging of input data and the recollection of the output data.

The ADA architecture is sketched in figure 9. Command line interfaces allow the user to interact via AJDL with the analysis services (AS) which give access to distributed computing power.

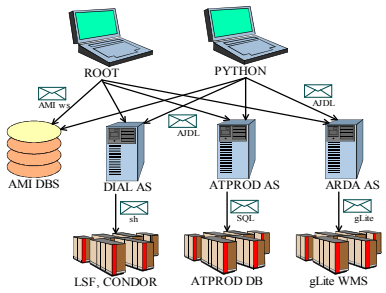

Fig. 9. Main current elements of the ADA schema, including services and interfaces.

DIAL [30] (Distributed Interactive Analysis of Large datasets) is a project within ADA which strives to demonstrate the feasibility of distributed analysis. It implements the AJDL components in C++ classes which can be interfaced through the ROOT [21] analysis framework or the Python [31] command line interface.

The ATPROD (ATLAS PRODuction) system has been used for the DC2 ATLAS data production. Work has started to build an interface allowing users to make their own small scale productions.

ARDA [32] (A Realization of Distributed Analysis for LHC) is another project to deliver a prototype for distributed analysis. The underlying middleware is based on the EGEE [33] gLite [34] package. User interfaces are at this moment under development.

Currently, a fully operational system based on the DIAL analysis service is available; it accesses a Condor [35] based PC cluster for long running jobs and a LSF [36] based cluster providing interactive response, both located at BNL (Brookhaven National Laboratory). A prototype analysis service based on gLite has been recently installed for testing purposes.

6 Summary

The generation and Geant4 simulation of events foreseen for ATLAS Data Challenges 2 have been completed using 3 flavors of Grid technology. They have been proven to be usable in a coherent way for a real production and this is a major achievement.

On the other hand, this exercise has taught us that all the involved elements (Grid middleware; production system; deployment and monitoring tools over the sites) need improvements.

Between the start of DC2 in July 2004 and the end of September 2004, the automatic production system has submitted about 235000 jobs. These jobs were approximately evenly distributed over the 3 Grid flavors. Overall, they consumed ~1.5 million SI2K months of cpu (~5000 cpu months on average present day cpu) and produced more than 30TB of physics data.

ATLAS is also pursuing a model for distributed analysis which would improve the productivity of end users by profiting from Grid available resources. Generic software components have been identified and several implementations of useful tools are be-

ing developed. An important aim is to provide the users with simple interfaces (ROOT, Python) that facilitate their interaction with the Grid infrastructure.

Acknowledgements

The authors would like to thank the many people involved in ATLAS DC2.

References

1. http://www.cern.ch/atlas
2. http://www.cern.ch/lhc
3. R. Sturrock et al. "ATLAS Data Challenge 1", CERN-PH-EP-2004-028, CERN, Apr 2004
4. http://lcg.web.cern.ch/LCG/
5. "The Grid 2003 Project." http://www.ivdgl.org/grid2003/index.php
6. "NorduGrid." http://www.nordugrid.org
7. http://www-cdf.fnal.gov
8. http://www-do.fnal.gov
9. L. Goossens, "Production System in ATLAS DC2", CHEP04, Interlaken, contr. no. 501
10. http://heppc12.uta.edu/windmill/
11. D. Rebatto, "The LCG Executor for the ATLAS DC2", CHEP04, Interlaken, contr. no. 364
12. R. Gardner, "ATLAS DC Production on Grid3", CHEP04, Interlaken, contr. no. 503
13. X. Zhao, "Experience with Deployment and Operation of the ATLAS production System and the Grid3", CHEP04, Interlaken, contr. no. 185
14. J. Kennedy, "Legacy services within ATLAS DC2", CHEP 2004, Interlaken, contr. no. 234
15. http://physics.bu.edu/pacman/
16. M. Branco, "Don Quijote", CHEP04, Interlaken, contr. no. 142
17. http://atlas.web.cern.ch/atlas/groups/software/DOCUMENTATION/ATLSIM/atlsim.html
18. http://atlas.web.cern.ch/atlas/GROUPS/SOFTWARE/OO/architecture/General/index.htm
19. http://www.thep.lu.se/~torbjorn/Pythia.html
20. http://hepwww.rl.ac.uk/theory/seymour/herwig/
21. R. Brun & F. Rademakers, "ROOT, An Object Oriented Data Analysis Framework",Proceedings AIHENP'96, Lausanne, Sep. 1996, Nucl. Inst. & Meth.A389 (1997) 81-86.
22. D. Duellmann, "The LCG POOL Project – General Overview and Project Structure" 2003
23. http://geant4.web.cern.ch/geant4/
24. "The European DataGrid project." http://eu-datagrid.web.cern.ch/
25. "Globus Toolkit." http://www-unix.globus.org/toolkit/
26. "Data TransAtlantic Grid." http://datatag.web.cern.ch/datatag
27. A. Chervenak, "Giggle: A Framework for Constructing Replica Location Services", SC02
28. http://www.usatlas.bnl.gov/ADA/
29. http://www.usatlas.bnl.gov/ADA/dels/ajdl/
30. D. L. Adams, "DIAL: Distributed Interactive Analysis of Large Datasets", CHEP03, UCSD
31. http://www.python.org/
32. http://lcg.web.cern.ch/LCG/peb/arda/
33. http://public.eu-egee.org/
34. http://glite.web.cern.ch/glite/
35. http://www.cs.wisc.edu/condor/
36. http://www.platform.com/products/LSFfamily/