Experience running a distributed Tier-2 in Spain for the ATLAS experiment

L March[1], S González de la Hoz[1], M Kaci[1], J del Peso[2], X Espinal[3], F Fassi[1], A Fernández[1], P Fernández[2], H Garitaonandia[3], A Lamas[1], M L Mir[3], L Muñoz[2], J Nadal[3], A Pacheco[3], J Pardo[2], J Salt[1], J Sánchez[1], S Shuskov[3]

[1] IFIC - Instituto de Física Corpuscular
(centro mixto CSIC – Univ. Valencia), E-46071 Valencia (Spain)
[2] UAM - Universidad Autónoma de Madrid
Dpto. de Física Teórica, 28049 Madrid (Spain)
[3] IFAE - Institut de Física d'Altes Energies
Facultat de Ciències UAB, E-08193 Bellaterra (Barcelona, Spain)

e-mail: Luis.March@ific.uv.es

Abstract: The main role of the Tier-2s is to provide computing resources for production of physics simulated events and distributed data analysis. The Spanish ATLAS Tier-2 is geographically distributed among three HEP institutes: IFAE (Barcelona), IFIC (Valencia) and UAM (Madrid). Currently it has a computing power of 430 kSI2K CPU, a disk storage capacity of 87 TB and a network bandwidth, connecting the three sites and the nearest Tier-1 (PIC), of 1 Gb/s. These resources will be increased according to the ATLAS Computing Model with time in parallel to those of all ATLAS Tier-2s. Since 2002, it has been participating into the different Data Challenge exercises. Currently, it is achieving around 1.5% of the whole ATLAS collaboration production in the framework of the Computing System Commissioning exercise. A distributed data management is also arising as an important issue in the daily activities of the Tier-2. The distribution in three sites has shown to be useful due to an increasing service redundancy, a faster solution of problems, the share of computing expertise and know-how. Experience gained running the distributed Tier-2 in order to be ready at the LHC start-up will be presented.

1. Introduction

The Large Hadron Collider (LHC) starts to operate in 2008 and will produce roughly 15 Petabytes (15 million Gigabytes) of data annually. When the LHC accelerator is running optimally, access to experimental data needs to be provided for 5000 scientists in some 500 research institutes and universities worldwide who are participating in the LHC experiments.

The Worldwide LHC Computing Grid collaboration (WLCG) [1] has been committed to develop, build and maintain a distributed computing infrastructure for the storage and analysis of data from the four LHC experiments. The WLCG is divided into three different Grid flavors):

- Grid3 / OSG [2]	USA
- NDGF / ARC [3]	Scandinavian countries + other countries
- LCG / EGEE [4]	most of European countries + Canada + far East

The WLCG Project implements a Grid to support the computing models of the experiments using a distributed four-tiered model, according to the MONARC [5] hierarchical model: Tier-0, Tier-1, Tier-2, Tier-3.

The role of the Tier-2 centres is to provide computational capacity and appropriate storage services for Monte Carlo event simulation and for end-user analysis. The Tier-2 centres will obtain data as required from Tier-1 centres, and the data generated at Tier-2 centres will be sent to Tier-1 centres for permanent storage. More than 100 Tier-2 centres have been identified and around 30 contribute for the ATLAS experiment.

The Spanish ATLAS Tier-2 is geographically distributed among three High Energy Physics (HEP) institutes: IFAE (Barcelona), IFIC (Valencia) and UAM (Madrid) and their resources are distributed among these three centres because these institutes are involved in the ATLAS experiment and want to contribute to the ATLAS Collaboration effort.

In section 2, a list of resources and services is given for the Spanish ATLAS Tier-2. In addition, network connectivity among sites is detailed also. A summary of the results obtained in the last exercises for distributed data management and Monte Carlo production is given in sections 3 and 4, respectively. On the other hand, the Spanish ATLAS Tier-2 is participating in Grid applications like Job Priorities [6] and ATLAS Event Filter (ATLAS trigger and Data Acquisition, TDAQ [7]). A summary of these activities is shown in section 5. Finally, the conclusions are given in section 6.

2. Resources and services

The infrastructure provided by the ATLAS Spanish Tier-2 project aims to contribute around 5% to the whole ATLAS Computing effort. The computing resources available in September 2007 are shown in table 1:

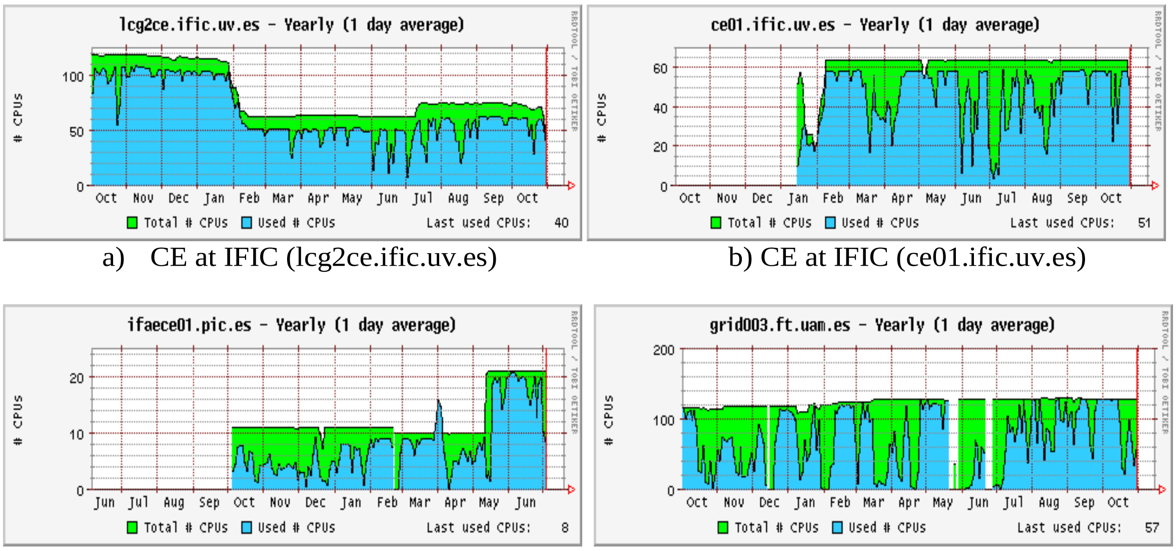
Equipment	Tier-2	IFAE	UAM	IFIC
CPU (kSI2k)	434	135	167	132
Storage (TB)	87	16	37	34 + 4.7 (tape front-end)

Table 1: Computing resources at the ATLAS Spanish Tier-2.

The total manpower for the Tier-2 is 14 FTE (by September 2007) to manage the different activities in which the Spanish ATLAS Tier-2 is involved.

Figure 1 shows the total number of CPUs at each CE per site and the number of CPUs that have been used running jobs during last year. Figures 1a (the oldest IFIC CE) and 1d (UAM CE) show statistics from October 2006 to October 2007. IFAE CE, figure 1c, shows statistics from October 2006 to June 2007 and, finally, figure 1b shows the newest IFIC CE statistics working from mid of January 2007 to October 2007.

Some periods of time these CEs have been stopped due to maintenance and upgrades. This has been communicated previously to the ATLAS LCG/EGEE collaboration and these CEs have been removed from the BDII in order to avoid getting jobs (figures 1c and 1d).



a) CE at IFIC (lcg2ce.ific.uv.es)

b) CE at IFIC (ce01.ific.uv.es)

c) CE at IFAE (ifaece01.pic.es)

d) CE at UAM (grid003.ft.uam.es)

Figure 1: Total number of CPUs per CE and site used during last year.

All these resources have been installed and configured using QUATTOR [8], a system administration toolkit for the automated installation, configuration and management of clusters and farms running UNIX flavors like Linux and Solaris.

UAM, in collaboration with other partner institutes is involved in the development and maintenance of the QUATTOR system.

Among the current services in the Spanish ATLAS federated Tier-2 are the following ones:
IFIC: 2 SRM Interfaces, 2 CEs, 4 UIs, 1 BDII, 1 RB, 1 PROXY, 1 MON, 2 GridFTP, 2 QUATTOR
UAM: 1 CE, 1 SE (dCache), 1 MON, 1 LFCG (QUATTOR), 1 UI
IFAE : 1 CE, 1 MON, 1 SE, 4 UI

The network is provided by the Spanish NREN RedIRIS [9] and local providers for the 'last mille'. The links are 10 Gbps between POP's with alternate paths for backup. Connectivity from the sites (IFIC, IFAE, UAM) to network is at 1 Gbps.

3. Data transfer and data management

The Spanish ATLAS Tier-2 has participated into the framework of different ATLAS Distributed Data Management (DDM) exercises:

1) Service Challenge 4 (SC4 exercise)
2) Functional tests
3) And recently Cosmic Rays data taking (M4)

During SC4 exercise, from 25 October to 30 October 2006, four datasets of 18 GB each were transferred from the Tier-1 (PIC) to the Spanish Tier-2 using the DQ2 [10] tool.

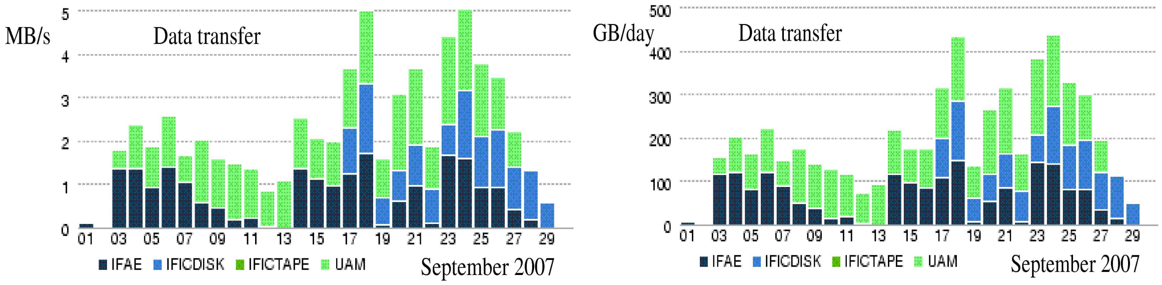
First transfer efficiencies to the Spanish Tier-2 centres were these ones: IFIC (100%), IFAE (98%) and UAM (25%). In case of the UAM, the Storage Element (SE) was failing due to a wrong configuration. Then, after fixing it, proceeded successfully.

The transfer back to the Tier-1 (PIC) was successful from the Spanish Tier-2, reaching a transfer rate of 5 MB/s in both directions during the exercise.

Last functional test, before the M4 exercise, was in the beginning of August 2007 (from 3 to 5 of August) where 0.5 TB of data was distributed into the Spanish Tier-2 according to this rate: IFIC 250 GB, IFAE 125 GB and UAM 125 GB. No problems were found in this exercise.

Finally, the most recent exercise, called M4, has taken from 25 August 2007 to 2 September 2007. The Tier-1 (PIC) received 13 TB of RAW data and 38 GB of ESD data. No AODs were produced at this exercise. The ESDs have been transferred to the Spanish Tier-2 centres according to this rate: IFIC 4GB (10%), IFAE 17 GB (45%) and UAM 17 GB (45%). IFIC needed to add new services and computer resources, then due to the computer room upgrade at IFIC during this exercise, it was not able to participate with the usual rate: IFIC (50%), IFAE (25%) and UAM (25%). Then, IFAE and UAM participated with a bigger contribution in this exercise.

Figure 2a (left picture) shows the data transfer in MB/s from the Tier-1 (PIC) to the Spanish Tier-2 centres during September 2007, reaching a maximum peak of 5 MB/s. Figure 2b (right picture) shows the same picture as 2a, but in GB/day. This amount of data transferred in GB/day is also shown in figure 2b in the same period of time as figure 2a, i.e. in September 2007. These data transfers come from different ATLAS exercises, like the M4 exercise and datasets from the ATLAS Monte Carlo production (Computing System Commissioning [11] exercise).



a) b)

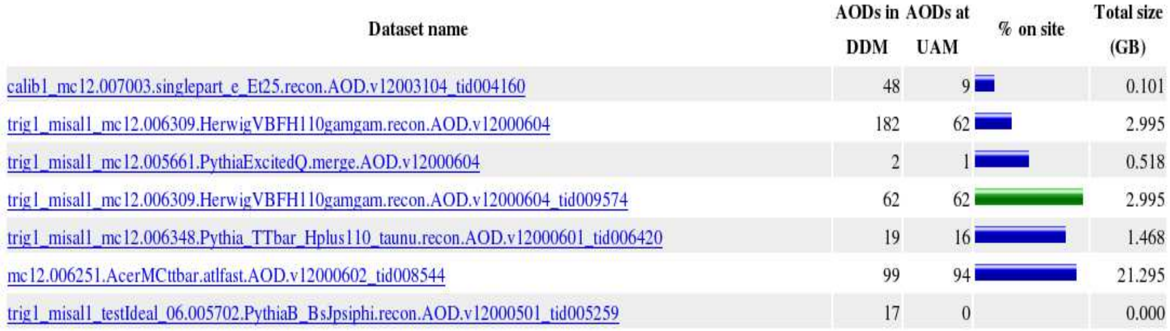
Figure 2: Data transfers in MB/s (a) and GB/day (b) from the Tier-1 (PIC) to the Spanish Tier-2 centres (IFAE, IFIC and UAM) in September 2007.

On the other hand, a control of this data that is stored at the Spanish Tier-2 centres is necessary for users of the ATLAS physicist community. For this purpose, a web site [12] is maintained by IFIC and provides information about the datasets which are registered at the PIC LFC catalogue and, at the same time, at the DQ2 catalogue.

In this web site, the information displayed is:

- The type of datasets: HITS, RDO, ESD, AOD, TAG, etc.
- Number and size of files stored at the local site and the total number of files registered in DDM for each dataset.
- Location of the datasets and files at the ATLAS Spanish Tier-2 sites.

An example of the web display is shown in figure 3, where a list of AOD data sets are shown: name of dataset, number of files per dataset registered in DDM, number of files per data set stored at UAM and the total size of the data set stored at UAM.



Dataset name	AODs in DDM	AODs at UAM	% on site	Total size (GB)
calib1_mc12.007003.singlepart_e_Et25.recon.AOD.v12003104_tid004160	48	9	▬	0.101
trig1_misal1_mc12.006309.HerwigVBFH110gamgam.recon.AOD.v12000604	182	62	▬	2.995
trig1_misal1_mc12.005661.PythiaExcitedQ.merge.AOD.v12000604	2	1	▬	0.518
trig1_misal1_mc12.006309.HerwigVBFH110gamgam.recon.AOD.v12000604_tid009574	62	62	▬	2.995
trig1_misal1_mc12.006348.Pythia_TTbar_Hplus110_taunu.recon.AOD.v12000601_tid006420	19	16	▬	1.468
mc12.006251.AcerMCttbar.atlfast.AOD.v12000602_tid008544	99	94	▬	21.295
trig1_misal1_testIdeal_06.005702.PythiaB_BsJpsiphi.recon.AOD.v12000501_tid005259	17	0		0.000

Figure 3: Screenshot of the web page with data sets and files stored at the ATLAS Spanish Tier-2.

4. Computing System Commissioning and Monte Carlo production

The production of simulated data for the ATLAS experiment is one of the main activities of the Spanish ATLAS Tier-2. In the framework of the Computing System Commissioning (CSC) (test of the ATLAS Computing Model), the following results are shown:

1) Statistics of the Monte Carlo production jobs run at the Spanish ATLAS Tier-2 since January 2006 and its contribution to the whole ATLAS LCG/EGEE collaboration.
2) Statistics of the Monte Carlo production jobs managed by an Eowyn/Lexor [13] ProdSys instance (a Supervisor/Executor instance from the ATLAS Production System) installed and run at IFIC from 25 January 2006 to 7 August 2006.

Concerning the first point, a statistics study of the Monte Carlo production jobs and their associated wall time run at the Spanish ATLAS Tier-2 per month in the ATLAS CSC exercise is shown in figure 4. On top of figure 4, the number of jobs finished successfully at the three sites from January 2006 to June 2007 is shown. At bottom of figure 4, the associated wall time for the ATLAS production jobs (on top of figure 4) in the same period of time is shown.

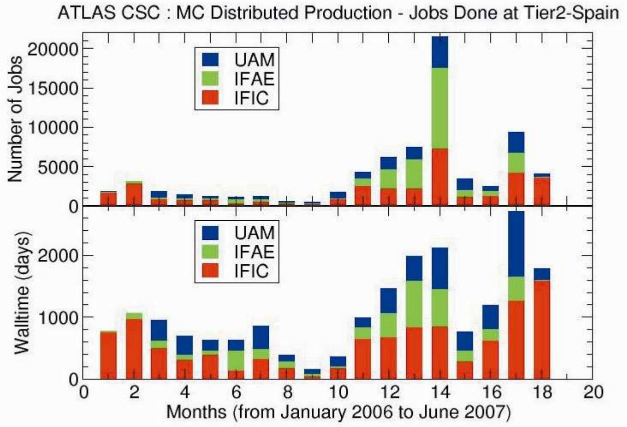


Figure 4: Monte Carlo production jobs and associated wall time per month in the ATLAS CSC exercise run at the Spanish ATLAS Tier-2.

In addition to figure 4, figure 5 shows the contribution of the Spanish ATLAS Tier-2 into the whole ATLAS LCG/EGEE collaboration by running ATLAS Monte Carlo production jobs (processed successfully) and their associated wall time used from January 2006 to June 2007. The top ratio shows the percentage (out of 100) of the average contribution of the ATLAS Spanish Tier-2 resources to the CSC exercise, also shown in table 2.

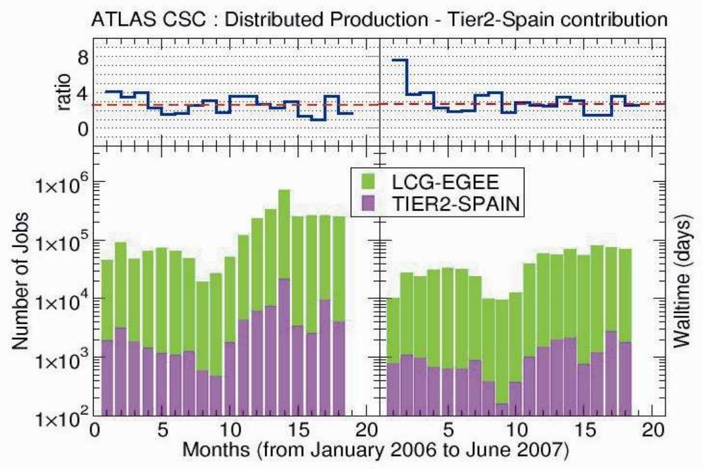


Figure 5: Contribution of the ATLAS Spanish Tier-2 resources into the whole ATLAS LCG/EGEE collaboration for the ATLAS CSC exercise.

The main results extracted from figures 4 and 5 are shown at table 2:

	Tier-2	LCG/EGEE	Contribution
Total jobs done	84833	332087	2.55 %
Total wall time used (days	22670.21	818282.15	2.77 %

Table 2: Contribution of the ATLAS Spanish Tier-2 resources to the ATLAS CSC exercise.

On the other hand, concerning the second point, a daily statistics study of the ATLAS Monte Carlo production jobs processed successfully and managed by an Eowyn/Lexor ProdSys instance is shown in figure 6. This ProdSys instance has been installed and run at IFIC from January 25th 2006 to August 7th 2006. The total number of jobs processed by the whole ATLAS LCG/EGEE collaboration in the same period of time is also shown in figure 6.

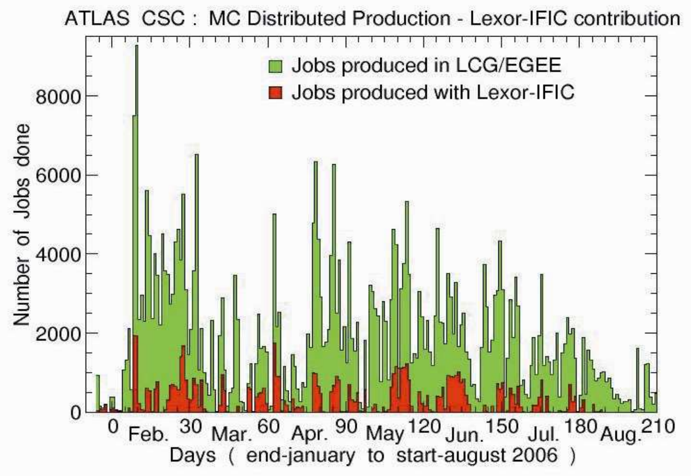


Figure 6: Daily statistics study of CSC production jobs processed successfully.

In addition, figure 7 shows the cumulative number of ATLAS jobs processed successfully and managed by the Eowyn/Lexor ProdSys instance run at IFIC in the same period of time as figure 6.

The cumulative number of jobs processed successfully by the whole ATLAS LCG/EGEE collaboration is also shown in figure 7. The top ratio shows the percentage (out of 1) of the average contribution of this ProdSys instance into the whole ATLAS LCG/EGEE collaboration by managing ATLAS MC production jobs for the CSC production, also shown in table 3.

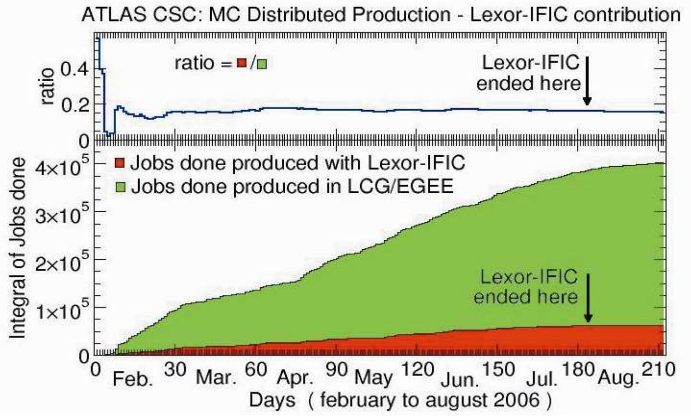


Figure 7: Contribution of successful ATLAS jobs managed by the Eowyn/Lexor ProdSys instance at IFIC into the whole ATLAS LCG/EGEE collaboration for the CSC production.

The main results extracted from figures 6 and 7 are shown at table 3:

	Eowyn/Lexor instance	ATLAS LCG/EGEE	Contribution
Total jobs done	62910	393714	16 %

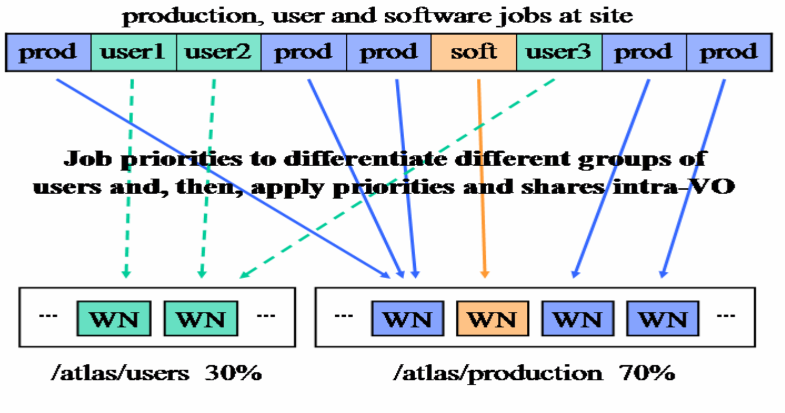
Table 3: Contribution of the Eowyn/Lexor ProdSys instance into the whole ATLAS LCG/EGEE collaboration by managing ATLAS production jobs for the CSC production.

5. Applications: Job Priorities and Event Filter

The ATLAS Spanish Tier-2 is involved in two ATLAS Grid applications: Job Priorities and ATLAS multi-level Trigger and Data Acquisition system (TDAQ) using Grid resources. Both applications are explained in the following.

The first application is the mechanism of Job Priorities that is investigated to differentiate groups of Grid users based on VOMS [14] groups and roles (see figure 8), e.g. /atlas/Role=production, for example. This mechanism applies priorities and shares intra-VO at the LRMS [15] (i.e. 50% production, 50% rest of users), publish the related information using VoViews and at the WMS [16] level, match jobs (VOMS proxy FQAN) with VoViews, to better schedule jobs according to published policies.

IFIC is participating in the configuration and deployment of Job Priorities.


production, user and software jobs at site
prod
user1
user2
prod
prod
soft
user3
prod
prod
Job priorities to differentiate different groups of users and, then, apply priorities and shares intra-VO

WN
WN

WN
WN
WN
WN

/atlas/users 30%
/atlas/production 70%

Figure 8: Job Priorities mechanism.

The current deployment at production site, ce01.ific.uv.es, is like this:

- ATLAS (70%): users (50%), production (50%), software: rest with high priority
- IFIC (30%): no ATLAS users

After installing and testing it, some design problems have been found. Actually, group/role matching several VoViews can be matched, but local mapping at the site (LCMAPS [17]) may map incorrectly.

On the other hand, the second application is the ATLAS multi-level Trigger and Data Acquisition system (TDAQ) using Grid resources.

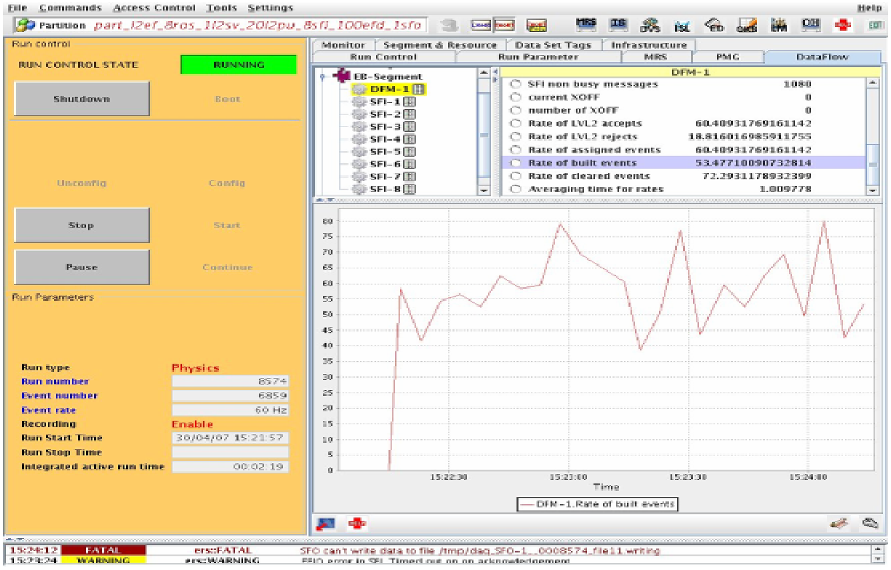
ATLAS will reach a high frequency of collisions: 40 MHz, where most of it is background. The ATLAS TDAQ reduces the rate to 200 events per second (the interesting ones) and transfers data to mass storage for later analysis.

The ATLAS TDAQ has been designed to use more than 200 CPUs. It is an interactive parallel application and during the current development stage it is crucial to test the system on a number of CPUs of similar scale. A dedicated farm of this size is difficult to find, and can only be made available for short periods.

Many large farms have become available recently as part of computing grids, leading to the idea of using them to test the TDAQ. However, the task of adapting the TDAQ to run on the Grid is not trivial, as the TDAQ system requires full access to the computing resources it runs on and real-time interaction. Moreover the Grid virtualises the resources to present a common interface to the user.

The Tier-2 cluster in Manchester was successfully used to run a full TDAQ system on 400 nodes using this implementation, see figure 9. This scheme also has great potential for other applications, like running Grid remote farms to perform detector calibration and monitoring in real-time, and automatic nightly testing on the TDAQ.

This is an ongoing project where IFAE has participated actively.



Partition part_l2ef_8ros_112sv_2012pu_8sfi_100efd_1sfo

RUN CONTROL STATE RUNNING

Run type	Physics
Run number	8574
Event number	6859
Event rate	60 Hz
Recording	Enable
Run Start Time	30/04/07 15:21:57
Run Stop Time	
Integrated active run time	00:02:19

DFM-1

○ SFI non busy messages	1080
○ current XOFF	0
○ number of XOFF	0
○ Rate of LVL2 accepts	60.40931769161142
○ Rate of LVL2 rejects	18.816016985911755
○ Rate of assigned events	60.40931769161142
○ Rate of built events	53.47710090732814
○ Rate of cleared events	72.2931178932399
○ Averaging time for rates	1.009778

```
15:24:12  FATAL    ers::FATAL    SFO can't write data to file /tmp/daq_SFO-1__0008574_file11.writing
15:23:24  WARNING  ers::WARNING  EFIO error in SFI. Timed out on on acknowledgement
```

Figure 9: Screenshot of the TDAQ system on Grid.

6. Conclusions

The experience gained by the three teams (IFAE, IFIC and UAM) running the ATLAS Spanish distributed Tier-2 is quite relevant and allows to improve the efficiency of the various services provided for the Spanish physicist community as well as for the whole ATLAS collaboration in the near future.

Acknowledgements

Thanks to IFAE and UAM groups involved in the ATLAS Spanish Tier-2 project.

References

[1] Worldwide LHC Computing Grid (WLCG): http://lcg.web.cern.ch/LCG/

[2] The Grid 2003 project: http://www.ivdgl.org/grid2003/
Open Science Grid (OSG) project: http://www.opensciencegrid.org/

[3] Nordic Data Grid Facility (NDGF): http://www.ndgf.org/

[4] Enabling Grids for E-science in Europe (EGEE): http://www.cern.ch/egee

[5] The MONARC project: http://monarc.web.cern.ch/MONARC/

[6] ATLAS Job Priorities: https://twiki.cern.ch/twiki/bin/view/Atlas/JobPriorities

[7] ATLAS Trigger and Data Acquisition (TDAQ):
https://twiki.cern.ch/twiki/bin/view/Atlas/TriggerDAQ

[8] QUATTOR system administration toolkit: http://quattor.web.cern.ch/quattor

[9] Red Española de I+D (RedIris): http://www.rediris.es/

[10] ATLAS Distributed Data Management (DDM/DQ2):
https://twiki.cern.ch/twiki/bin/view/Atlas/DistributedDataManagement

[11] Computing System Commissioning (CSC):
https://twiki.cern.ch/twiki/bin/view/Atlas/ComputingSystemCommissioning

[12] Datasets at the ATLAS Spanish Tier-2: http://ific.uv.es/atlas-t2-es/

[13] Eowyn supervisor: https://twiki.cern.ch/twiki/bin/view/Atlas/EowynSupervisor
Lexor executor: http://lxmi.mi.infn.it/~rebatto/lexor/lexor.html

[14] Authorization System for Virtual Organizations:
http://grid.cesga.es/eabstracts/voms.pdf

[15] Local Resource Management System (LRMS), like Condor, PBS and LSF:
Condor High Throughput Computing: http://www.cs.wisc.edu/condor/
Portable Batch System (PBS): http://en.wikipedia.org/wiki/Portable_Batch_System
Load Sharing Facility (LSF): http://wwwpdp.web.cern.ch/wwwpdp/bis/services/lsf/

[16] Workload Management System (WMS): http://glite.web.cern.ch/glite/wms/

[17] Local Credential MAPing Service (LCMAPS): http://www.sysadmin.hep.ac.uk/wiki/LCMAPS