Eur. Phys. J. C 54, 691–697 (2008)
DOI 10.1140/epjc/s10052-008-0575-9

THE EUROPEAN
PHYSICAL JOURNAL C

Special Article – Scientific Note

Analysis facility infrastructure (Tier-3) for ATLAS experiment

S. González de la Hoz[a], L. March, E. Ros, J. Sánchez, G. Amorós, F. Fassi, A. Fernández, M. Kaci, A. Lamas, J. Salt

Instituto de Física Corpuscular de Valencia, Centro mixto Universitat de València – CSIC,
Edificio Institutos de Investigación, Apartado de Correos 22085, 46071 Valencia, Spain

Received: 12 December 2007 / Revised version: 4 March 2008 /
Published online: 18 March 2008 – © Springer-Verlag / Società Italiana di Fisica 2008

Abstract. In the ATLAS computing model the tiered hierarchy ranged from the Tier-0 (CERN) down to desktops or workstations (Tier-3). The focus on defining the roles of each tiered component has evolved with the initial emphasis on the Tier-0 and Tier-1 definition and roles. The various LHC (Large Hadron Collider) projects, including ATLAS, then evolved the tiered hierarchy to include Tier-2's (Regional centers) as part of their projects. Tier-3 centres, on the other hand, have been defined as whatever an institution could construct to support their Physics goals using institutional and otherwise leveraged resources and therefore have not been considered to be part of the official ATLAS computing resources. However, Tier-3 centres are going to exist and will have implications on how the computing model should support ATLAS physicists. Tier-3 users will want to access LHC data and simulations and will want to enable their resources to support their analysis and simulation work. This document will define how IFIC (Instituto de Física Corpuscular de Valencia), after discussing with the ATLAS Tier-3 task force, should interact with the ATLAS computing model, detail the conditions under which Tier-3 centres can expect some level of support and set reasonable expectations for the scope and support of ATLAS Tier-3 sites.

1 Introduction

The ATLAS computing model [1] describes a hierarchical distributed virtual computing facility consisting of Tier-1 and Tier-2 computing centres, having certain specific memorandum of understanding (MOU) agreed roles and capacities, to be used for the benefit of ATLAS as a whole. ATLAS research program decides how these MOU pledged resources are used. In this model primary functions of the Tier-1 are to host and provide long term storage for, access to and re-reconstruction of a subset of the ATLAS RAW data, provide access to the event summary data (ESD) [2], analysis object data (AOD) [2] and TAG [3] data sets and support the analysis of these data sets. The primary functions of the Tier-2's are simulation (they provide the bulk of simulation for ATLAS), calibration, chaotic analysis for subset of analysis groups and hosting of AOD, TAG and some physics group samples.

Tier-3 sites are institution-level non-ATLAS funded or controlled centres/clusters which wish to participate in ATLAS computing, presumably most frequently in support of the particular interests of local physicists (physicists at the local Tier-3 decide how these resources are used). These are clusters of computers which can vary widely in size. An ATLAS Tier-3 task force [4] at CERN has been created to help to document requirements to facilitate setting up Tier-3 for ATLAS use.

The subject of the Tier-3 task force is to develop a model for Tier-3 and analysis facility (including CERN CAF) sites in ATLAS. Within the ATLAS model such sites will be used mostly for interactive or batch analysis of the so called DPD [2] (derived physics data) data sets, which have been produced from AOD data using distributed analysis tools. The definition of different possible DPD formats is been discussed in the analysis model group and will be physics working group or even analysis specific. It is up to the Tier-3 task force to propose possible Tier-3 configurations and software setups that match the requirements according to DPD analysis needs, as formulated by the analysis model group. The result of the task force should be:

- A set of physics analysis examples to motivate various sizes of ATLAS Tier-3s,
- A set of recommendations and documentation on how to setup a typical ATLAS Tier-3 centre at a university in order to provide a guideline for institutes joining ATLAS and/or starting now to set up their own ATLAS computing cluster and finally,
- A worked out proposal for a software infrastructure to operate such a compute and disk farm for interactive and batch analysis according to the needs of the proposed analysis model.

In this document we present the IFIC prototype setup which was discussed with the ATLAS experiment community.

[a] e-mail: santiago.gonzalez@ific.uv.es

2 An example of an analysis in the framework of the Atlas analysis computing model

According to the ATLAS analysis computing model, analysis is divided into "group" and "on-demand" types. This analysis will be performed by physics groups on Tier-2 resources. This means that users from universities and institutes need some extra computing resources, to perform their own work and then contribute with their studies and algorithms to the group effort.

2.1 ATLAS computing model

The ATLAS computing model has a hierarchical model (event filter, Tier-0/Tier-1/Tier-2) with specific roles and responsibilities:

Event filter farm at CERN

- Assembles data at CERN intro a stream to the Tier-0 centre.

Tier-0 centre at CERN

- Data archiving: raw data to mass storage at CERN and to Tier-1 centres,
- Reconstruction, calibration and alignment processing: Fast production of ESD, AOD,
- Distribution: ESD, AOD to Tier-1 centres and storage at CERN.

Tier-1 centres distributed worldwide (10 centres)

- Data steward: Re-reconstruction of raw data producing new ESD and AOD,
- Coordinated access to full ESD and AOD (all AOD, 20%–100% of ESD depending on site).

Tier-2 centres distributed worldwide (around 30 centres)

- Monte Carlo simulation, producing ESD, AOD transferred to Tier-1 centres,
- On demand user physics analysis of shared datasets,
- Scheduled working group activities,
- AOD processing, TAG extraction.

Tier-3 centres refer to local compute resources, beyond Tier-1 and Tier-2 that are required to support physics analysis by researchers at universities and institutes.

These resources could range from workstations on each physicist's desk to computer farms, and could be operated as a shared facility with the institution own resources.

2.2 The Tier-3 structure

The main goal of the Tier-3 is physics analysis on site with seamless access to all ATLAS grid resources. For this purpose, analysis tools and access to data are very important, so these tools have to be installed in a Tier-3 infrastructure to facilitate the operations for users. Note that Tier-3 centres are outside the official infrastructure, but following the previous scheme, the specific role and responsibilities for a Tier-3 would be as follows:

- Interactive analysis
  - Prototype: reconstruction, calibration and algorithm selection,
  - Final selection: plots, studies, etc.
- Statistical analysis on AOD and derived physics data (DPD) [2].

In a Tier-3 infrastructure physicists from an institute or a university could perform physics analysis on site and have access to the different ATLAS simulation and analysis facilities: tools, data, etc. Actually, according to the ATLAS computing model, users should send analysis jobs to sites where data are available and extract relevant output as n-tuples or similar.

At IFIC the Tier-3 resources are being split into two parts:

- Some resources are being coupled to IFIC Tier-2 resources in a grid environment. These extra Tier-2 resources will be used preferently by Tier-3 users. While resources are idle, then they can be used by the ATLAS community.
- A computer farm to perform interactive analysis outside the grid framework.

As a starting point, in order to perform a Tier-3 prototype at IFIC, the user requirements (IFIC users) to perform analysis are taking into account. One of the requirements is to produce small Monte Carlo simulations and store the output data for further analysis. In next subsection, the requirements for such a production using the ATLAS full chain simulation are described, based on our experience.

2.3 Steps in the simulation of a physics sample

The Monte Carlo simulation production is going to be run at Tier-2 centres distributed worldwide according to the ATLAS computing model. In the following a real user example is given: We produced a private Monte Carlo simulation inside an ATLAS physics group to perform the study of b-tagging at very high p_{T}.

This private production consisted of three datasets. The physics processes were hadronic decays of Z_H in the Little Higgs model [5], where the Z_H decays into bbbar, uubar and ccbar. For each one of these hadronic decays a dataset was made. Each dataset consisted of 20 000 events, so the total number of events is 60 000. Each event was processed using the ATLAS full chain simulation (FullSim events):

- Generation,
- Simulation (Geant4) and digitization,
- Reconstruction,

Each step of this private production is detailed as follows:

Generation step

The generator (Pythia) and the generation options (one for each physics process) were validated by the Physics production managers; then those jobs were submitted to the ATLAS Tier-2 centres: 50 000 events per dataset were generated into 10 event generation output files (5000 events per file). Each generation file was approximately 193 MB in size.

Simulation and digitization step

In this case, two different steps from the ATLAS full chain simulation (Geant4 simulation and digitization) were run in the same job (grid job). First, Geant4 simulation was run and, once the simulation output files were produced, then the digitization was run. The generation output files were the input data for the Geant4 simulation. A usual simulation job is set to process only 50 events, since this is the most CPU-consuming step in the full chain. A simulation event usually took around 700 s in a computer with a performance of 1400 kSpecInt2000 s. The standard performance evaluation corporation (SPEC) is a non-profit corporation formed to establish, maintain and endorse a standardized set of relevant benchmarks that can be applied to the newest generation of high-performance computers SPEC designed CPU2000 to provide a comparative measure of compute intensive performance across the widest practical range of hardware. For each job a simulation output file is produced, called simul.HITS. Each simulation output file has a size around 88 MB.

As commented before, simulation and digitization are handled by a common job. Once the simulation is finished, the output is used as an input for digitization that produced as output a digit.RDO file. Each digitization data file contains 50 events and its size is about 138 MB. Digitization of an event only consumed 21 kSI2k s. The job submission scheme is displayed in Fig. 1, where the user is logged into a user interface (UI) and submits jobs using a resource broker (RB). Then the RB manages the jobs over all available grid resources according to the user job requirements.

Reconstruction step

Reconstruction is the last step in the ATLAS full simulation chain. The output data files from digitization (digit.RDO files) are the input for reconstruction. As all official ATLAS production, a reconstruction job contains 250 events, which means that a reconstruction job has as input 5 digit data files. Each reconstruction job produced 3 output files: ESD, AOD and NTUP, all of them with 250 events per file, but with different size: 760 MB for ESD, 134 MB for AOD and 37 MB for NTUP.

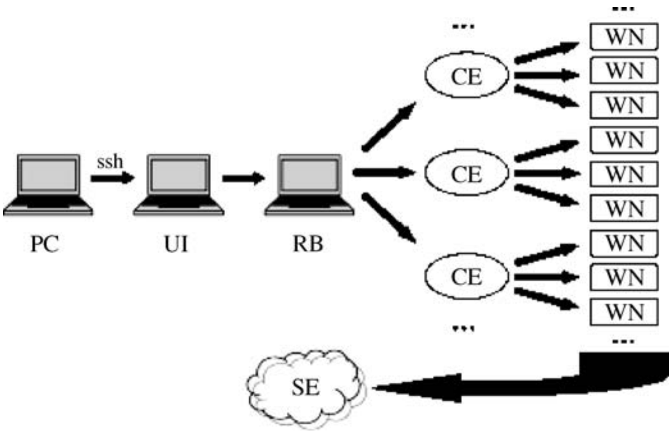

Fig. 1. Job submission scheme of the simulation private production

ESD are the first files produced using as input the digit data files and required around 160 kSI2k s. Then AOD were produced using ESD input files. They needed about 6 kSI2k s on the same computer. Finally NTUP (DPD) files were produced using AOD as input. Note that all these data files from this private production were stored at an IFIC storage element (SE) based on CASTOR.

This private Monte Carlo simulation production needs a storage capacity of 0.5 TB (see Table 1). A total of 60 000 full simulation events were produced. For future production, at ATLAS start up, larger production and additional local resources will be needed to perform similar private productions.

3 Towards a user analysis local facility: The Tier-3 prototype at IFIC-Valencia

ATLAS data taking is going to start on June 2008. For this reason, Tier-3 analysis facilities should be ready by that date.

Table 1. The number of events and size per output data file at each step of the Athena full chain simulation are shown. A total of 60 000 events were generated. The number of files per dataset, total number of datasets and total size for each Monte Carlo production step are provided. Finally, the total size of the production is shown as well

	Events per file	Size per file (MB)	# Files per dataset	# Dataset	Total size (GB)
Generation	5000	193	4	3	2.3
Simulation	50	88	400	3	105.6
Digitization	50	138	400	3	165.6
Reconstruction (ESD + AOD + NTUP)	250	931	80	3	233.4
Sum					469.9

IFIC (as many other centres, institutes and universities) has a Tier-3 prototype with particular goals and steps. As users at IFIC, our experience in Monte Carlo production and physics analysis is taken into account. The first steps towards a Tier-3 infrastructure at IFIC are described in this section.

Users at IFIC could get access to ATLAS collaboration Tier-2 resources and Tier-3 resources (extra Tier-2 resources) from their own desktops or laptops (see Fig. 2). These extra Tier-2 resources would be used preferently by IFIC users.

Otherwise, if these resources are available, they could be used, in addition to the IFIC Tier-2 resources, by ATLAS collaboration. On the other hand, other special requirement setup could be deployed for IFIC users as well, like a PC farm to perform interactive analysis.

An individual physicist can get access to the ATLAS software and analysis tools using its desktop or laptop. These software and analysis tools are already installed at IFIC for a first test. These tools are installed in a distributed networked file system, called AFS [6], where IFIC users have access, in the same way as at CERN, in order to avoid installing them on each desktop or laptop. AFS has several benefits over a traditional networked file system, in particular in the areas of security and scalability. The software already installed is the following:

- ATLAS software: Athena, Atlantis [7], etc.
- Distributed Analysis tools: GANGA [8].
- Other useful tools like ROOT or a grid user interface environment.

Users can use their desktop as a private user interface and developer station. In this way, the initial steps shown in Fig. 2 have already started, like the access to ATLAS grid resources and tools, as commented before, but for instance several ATLAS software releases and other useful tools have also to be installed.

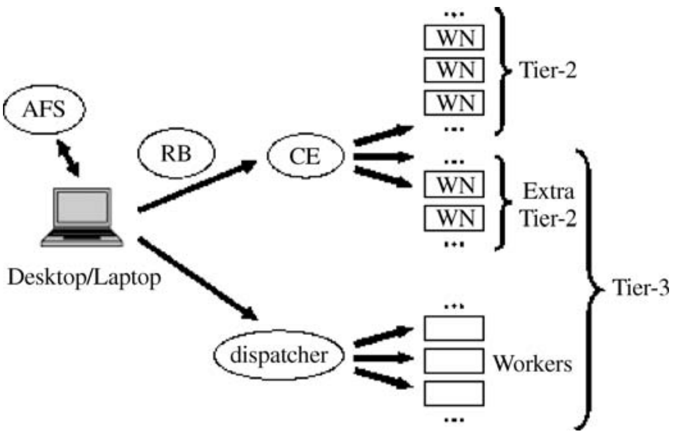

Fig. 2. User access from desktops or laptops to the ATLAS resources to perform physics analysis at IFIC. The IFIC Tier-3 resources will be coupled to the IFIC Tier-2 resources in a grid environment, and considered as extra Tier-2 resources. In addition, a PC farm will be deployed to perform interactive analysis tests in a non-grid environment

IFIC Tier-3 Grid resources are going to be coupled to IFIC Tier-2 resources, but they will be separated clearly using queues with different priorities for the ATLAS Collaboration and IFIC Tier-3 users. This means that IFIC Tier-2 resources are used by the whole ATLAS Collaboration, in order to match the ATLAS resource requirements (disk storage and CPU). Extra IFIC Tier-2 resources (Tier-3) are going to be used by IFIC users to run local simulation and analysis productions and to store data interesting for analysis:

- Production of Monte Carlo samples of special interest for the local institution,
- AOD private productions for further analysis and
- AOD analysis.

Having access to the ATLAS software and Grid resources, users can perform local checks, run small grid or non-grid test jobs and develop their analysis code before submitting larger simulation or analysis productions to the Tier-2 or Tier-1 centres. The development of analysis code could motivate also a local copy of a small number (perhaps a few thousands) of ESD, AOD or RAW data events.

3.1 Interactive analysis

In addition to the IFIC Tier-3 grid resources, a PROOF (parallel ROOT facility) [9] farm is going to be deployed for interactive analysis of n-tuples. Direct access to ESD or AOD is not required, but just direct access where these n-tuples are generated. It is clear that a Tier-3 infrastructure must be partly inside the Grid, in order to get the data (ESD, AOD and DPD), and partly outside the grid for interactive analysis of n-tuples (DPD). At this point, we should note that this infrastructure, Tier-3 grid and non-grid resources, is going to use the same storage element, having access to the data in both cases.

Interactive analysis is a very important issue for physics analysis. Usually, interactive analysis with DPD n-tuples are analysed using ROOT. For this reason a PROOF farm with few powerful PCs (outside grid environment) is going to be installed. One important issue is that this farm must be well connected to the storage element in order to get a fast access to the data.

The motivation to use PROOF is to provide an alternative and dynamic approach to end-user high energy physic analysis on distributed systems. Typical analyses are a continuous refinement cycle of implementation of algorithms, running over some dataset (collections of independent events, e.g. 350 TB/year) and making improvements. Exploiting intrinsic parallelism is the only way to analyze the data in a reasonable time. PROOF is a system for the interactive analysis of very large sets of ROOT data files on a cluster of computers. It speeds up the query processing by employing inherent parallelism in event data.

In a performance test of a PROOF farm in Wisconsin [10] with 8 x Intel 2.66 GHz cores machine, 16 GB memory and 8×750 GB on RAID5, 200 files were copied on disk. Using one session 12 752 events were processed per second and using two sessions the rate almost doubled to 25 695 events per second. Finally, using eight sessions, 95 135 events per second are reached. This result shows the scalability of PROOF.

Similar results have been obtained in Munich [11] with 10 nodes dedicated to PROOF analysis using two dual core processors with 2.7 GHz and 8 GB Ram. The data were stored locally on each node. For these reasons, the IFIC Tier-3 is going to be equipped with this infrastructure in a short term future.

3.2 Data access to the storage element

Data access performance tests have been made using the IFIC local storage element, and a b-tagging physics sample, consisting of $Z_H \rightarrow$ bbbar. AODs and DPDs for these events have been analyzed using the ATLAS framework (Athena) and ROOT, respectively. These data have been accessed in different ways: local disk, RFIO (CASTOR) and Lustre [12].

The storage hardware used for Lustre was the following:

- Two disk servers (2xSUN X4500) with a net capacity of 34 TB.
- A switch gigabit CISCO Catalyst 4500was used for connectivity.
- Grid access was provided with a SRM and a GridFTP server.
- One metadata server (MDS) Lustre server with redundancy RAID1.

Lustre is a storage-architecture for clusters. The central component is the Lustre file system, a shared file system for clusters. The Lustre file system is available for Linux and provides a POSIX-compliant UNIX file system interface.

Figure 3 shows the results of the data access performance test using different forms of local storage, the Athena framework to analyze the data and 38 AOD's as input files. Each AOD has 250 events (9500 in total) and the average size per AOD was around 160 MB. As Fig. 4 shows, the slowest way to access these data is using local disk, analyzing around 6 events/s. RFIO and Lustre performance were around 7 events/s. In general, the bandwidth used was around 4 MB/s.

Figure 4 shows the results of the data access performance test using different ways of local storage, the ROOT analysis tool to analyze the data and just one DPD with 9500 events as input data file. The size of this DPD is around 178 MB and contains all data from the previous AOD's (9500 events), extracted from these AOD's. As Fig. 4 shows, the slowest way to access this data is RFIO, in this case, analyzing around 98 events per second. Lustre needed around 101 events per second and for the fastest way, using the local disk, the performance was around 102 events per second The bandwidth used was around 2 MB/s, lower than in Fig. 3. This is due to the data structure and data size, since the event size in a DPD is much smaller tan in a AOD.

As shown in Figs. 3 and 4, the results obtained for data access performance test using AOD's and DPD'a are different and show that DPD's are analyzed faster than AOD's. A fast data access for analysis is very important. On the other hand, in both cases the results obtained with Lustre are better than those with RFIO, and similar to those obtained with local disk access. Moreover, Lustre allows a very flexible reconfiguration of computing nodes (i.e batch and PROOF) without intervention on the storage and a real */path/to/my/files* access is used together with a very high performance. For that reason, Lustre could be a possible access data mode to support a PROOF farm for interactive analysis.

The ATLAS Tier-3 task force proposes the following resources for a Tier-3:

- Local access protocol (that works with ATLAS Software, ROOT, etc),
- Load balancing,

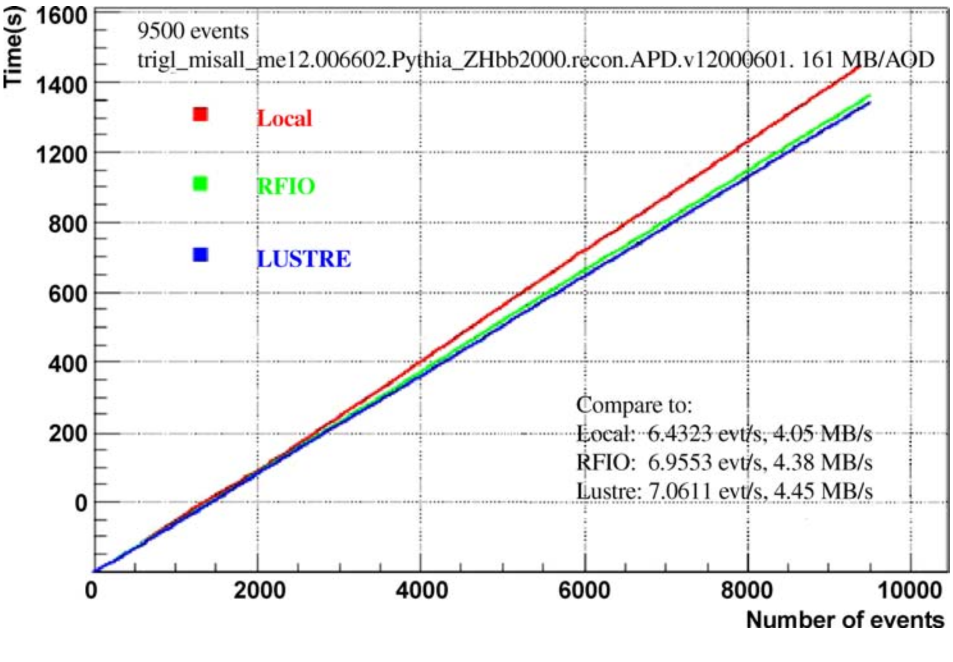

Fig. 3. Data access performance test using the IFIC local SE, the Athena framework and AOD's as input data files

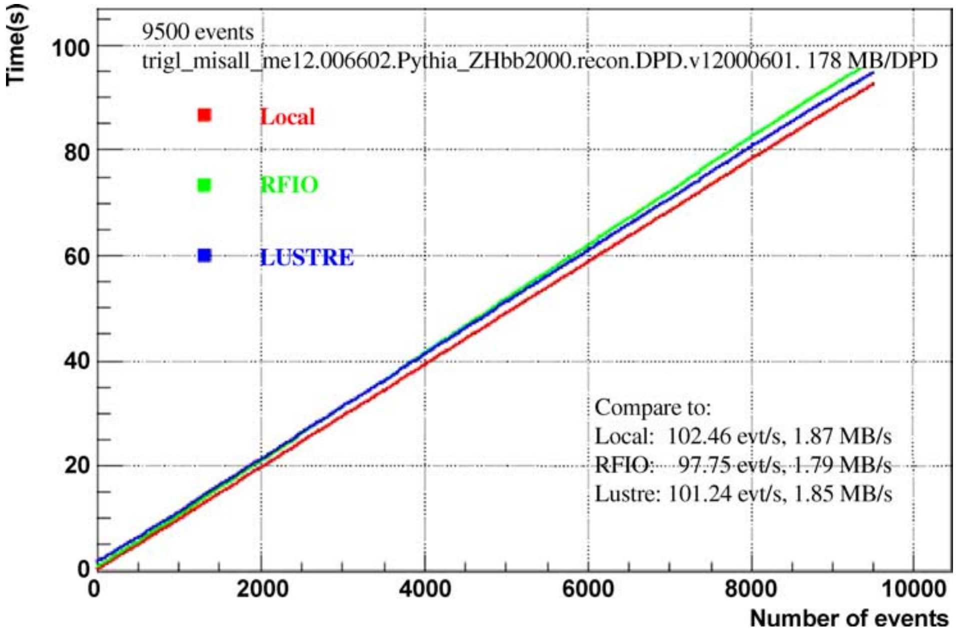

Fig. 4. Data access performance test using the IFIC local SE, ROOT analysis tool and one DPD as input data file

- Secure when external facing, and
- POSIX access. Some places turn this option off (e.g. dCache [13] in Lyon) as it has security and performance issues. Felt it is convenient for users interactively but less useful from worker node jobs like in a Tier-2.

Some potential options could be Lustre, GPFS (general parallel file system [14]), dCache, etc. But it is not clear that GPFS is free to be used, so it is possibly not a reasonable solution for a Tier-3. Lustre on the other side could be used for free and shows a nearly equivalent behaviour as the local storage. In general dCache performs not as well as Lustre, Local disk or GPFS because it doesn't have a POXIS access, but it is easy to setup and maintain.

The aim of the present work being done at the Tier-3 of IFIC is to provide a flexible computing resource; where the physicist could make an analysis work in an easy and convenient way. The Tier-3 will provide interactive and non-interactive CPU, enough storage capacity and bandwidth to the data sources from a Tier-2 (Spanish federated Tier-2) and other resources abroad.

4 Summary and conclusions

In this paper we have tried to develop a model for Tier-3 and analysis facility sites in ATLAS. Within the ATLAS model such sites will be used mostly for interactive or batch analysis of the so called DPD data sets, which have been produced from AOD data using Grid distributed analysis tools. The definition of different possible DPD formats is being discussed in the ATLAS Analysis Model group and will be physics working analysis specific. At IFIC in Valencia we are proposing a possible Tier-3 configuration and software setup that matches the requirements according to the DPD analysis needs as formulated by the ATLAS analysis model group. The results of this first evaluation are:

- Some local resources, beyond Tier-1 and Tier-2, are required to do physics analysis in ATLAS.
- These resources could consist of workstations on each physicist's desk or computer farms, and could be operated as a shared facility provided by the institution own resources.

Support from the Tier-1 and Tier-2's to such Tier-3 centres in terms of expertise (installation, configuration, tuning, troubleshooting of ATLAS releases and the Grid stack) and services (data storage, data serving, etc.) is recommended.

We envisage the following examples as typical uses of a Tier-3:

- Interactive analysis of n-tuples. It does require access to the data when these n-tuples are generated.
- Development of analysis code. This would motivate a local copy of a small number of data.
- Running small local test jobs before submitting larger jobs to the Tier-1 or Tier-2 via the grid. This would motivate similar copies of the data as above.
- Running skimming jobs of the Tier-1 and Tier-2's via the grid, and copying the skimmed AOD back to the Tier-3 for further analysis. The output of this skim must be a very small subset of the AOD – of order a few percent.
- Analyzing the above skimmed data using the ATLAS framework (Athena).
- Production of Monte Carlo samples of special interest for the local institution.

References

1. D. Adams et al., The ATLAS Computing Model, ATL-SOFT-2004-007, CERN, 15 Dec 2004
2. ATLAS Event Data Model, https://twiki.cern.ch/twiki/bin/view/Atlas/EventDataModel

3. Tag for Event Selection, https://twiki.cern.ch/twiki/bin/Atlas/TagForEventSelection
4. ATLAS TF https://twiki.cern.ch/twiki/bin/view/Atlas/AtlasComputing?topic=Tier3TaskForce
5. G. Azuelos et al., Exploring Little Higgs models with ATLAS at the LHC, Eur. Phys. J. C **39**, s2, s13–s24 (2005) [hep-ph/0402037]
6. Andrew File System (AFS), http://en.wikipedia.org/wiki/Andrew_file_system
7. Atlantis event display for ATLAS, http://www.hep.ucl.ac.uk/atlas/atlantis
8. GANGA, http://ganga.web.cern.ch/ganga/
9. PROOF: the Parallel ROOT Facility, http://root.cern.ch/twiki/bin/view/ROOT/PROOF
10. PROOF tests at Wisconsin university, http://www-wisconsin.cern.ch/∼nengxu/proof
11. Test results using PROOF: https://twiki.cern.ch/twiki/bin/view/Atlas/Tier3TaskForceProof
12. Lustre Cluster File System: http://wiki.lustre.org
13. dCache system for storing and retrieving huge amounts of data: http://www.dcache.org/
14. General Parallel File System: http://www.almaden.ibm.com/storagesystems/projects/gpfs/