Development and Setup of a Prototype System of Distributed Analysis for ATLAS Tier-2

Farida Fassi[a], Álvaro Fernández[a], Javier Sánchez [a], José Salt[a], Luís March[a], Mohammed Kaci[a], Gabriel Amoros[a], Xavier Espinal[b], Alejandro Lamas[a]

[a] IFIC – Instituto de Física Corpuscular, Valencia, Spain
[b] IFAE – Instituto de Física de Altas Energías , Barcelona, Spain

farida.fassi@ific.uv.es

Abstract

The ATLAS experiment currently under construction at CERN's Large Hadron Collider (LHC) presents data processing requirements of an unprecedented scale. ATLAS will accrue tens of petabytes of data per year, distributed around the world: the collaboration comprises more than 1800 physicists from 150 institutions in 34 countries. The Distributed Analysis (DA) has the goal of enabling ATLAS physicists to perform analysis on distributed data using distributed computing resources. Both data and resources are widely distributed throughout the world at CERN and at ATLAS Tier-1 and Tier-2 centers. Since DA is of strategic importance there is a large development activity going on in this area: the ATLAS production system has being evolving to support the analysis jobs which will have a seamless access to all ATLAS resources, as well as another activities that aim to support user analysis by submitting directly to the separate grid infrastructures (Panda at OSG, direct submission to LCG and Nordugrid). The test of DA functionality will be addressed in the final Service Challenge 4 (SC4), in which the system will exposed to the expected large number of final analysis users.

The Spanish ATLAS Tier-2 facility formed by IFIC, IFAE and UAM groups, is participating in several aspects of the Distributed Analysis System. In support of the ATLAS DA activities the IFIC Tier-2 centre has developed and deployed a local computational facility which comprises many service nodes, computational clusters and large scale disk and tape storage services. The resources contribute to a variety of activities such as the analysis centre facility for the next SC4 in which the technical aspects of DA will be tested and evaluated. In this paper we describe the ATLAS DA as well as we present our experience with the deployment, maintenance and operation of the mentioned DA prototype from the whole ATLAS collaboration and from the framework of the Spanish Tier-2 user's point of view.

I. INTRODUCTION

The ATLAS collaboration [1] is preparing for data taking and analysis at the CERN LHC [2], scheduled to start operating in 2007. Physics studies in ATLAS will require analysis of data volumes of the order of petabytes per year. The analysis will reply on the computing resources and the data will be distributed over the world-wide collaborating institutions. These will be collected together and shared in a coordinated way using grid technology that provides the infrastructure required to facilitate the distributed of data and the pooling of computing and storage resources between these institutions. Data for the large samples of simulated interaction, needed to well understand the detector behaviour, will also be distributed between multiple locations.

The grid-based ATLAS distributed analysis aims to deal with the challenge of supporting distributed users, data and processing enabling physicists to exploit the whole computing resource provided by the three ATLAS grid infrastructures: LCG [3], OSG [4] and Nordugrid [5]. DA must support all the analysis activities, including the simulated data production, hiding users from the complexities of the grid environment.

DA is fully described, with emphasis on the ATLAS strategy for distributed analysis in a heterogeneous grid environment. In the following sections we present the IFIC Tier-2 facility available for SC along with our experience gained from using Ganga.

II. DISTRIBUTED ANALYSIS

According to the ATLAS computing model [6], DA will enable users to submit jobs from any location helping them to effectively use the grid for performing their analysis activities. In addition, DA should satisfy the ATLAS analysis model requirement: data is distributed among several computing facilities and analysis jobs in turn routed base on the availability of relevant data. ATLAS activity within DA has several approaches to fully exploit its major grid deployments. Many of these activities are in prototype stage and are not yet deployed on a larger scale. Figure 1 shows the different distributed analysis sub-systems.

DA makes use of several different front-end and back-end systems that are described in the next sections, respectively.

Ganga[7]: Front-end for job definition and management. It is being developed to meet the needs of a user grid interface with the ATLAS and LHCb[8] experiments. Ganga provides user's access to generic distributed systems such as the all grid infrastructures supported by ATLAS; and local batch system, including LSF, PBS and Condor. It is planned to include also the ATLAS production system and further grid flavors like OSG and Nordugrid.

Pathena: Interface to the ATLAS offline software framework (Athena [9]) that is implemented in Python [10]. Pathena allows direct submission of the user-defined jobs to the Panda job management system [11]. Pathena approach is well aligned with Panda and Panda capabilities [12].

ATCOM [12]: Dedicated graphical user interface to the ATLAS production system. It allows user to create tasks and

define jobs in the database. Some activities are going on with the aim to improve it like front-end distributed analysis.

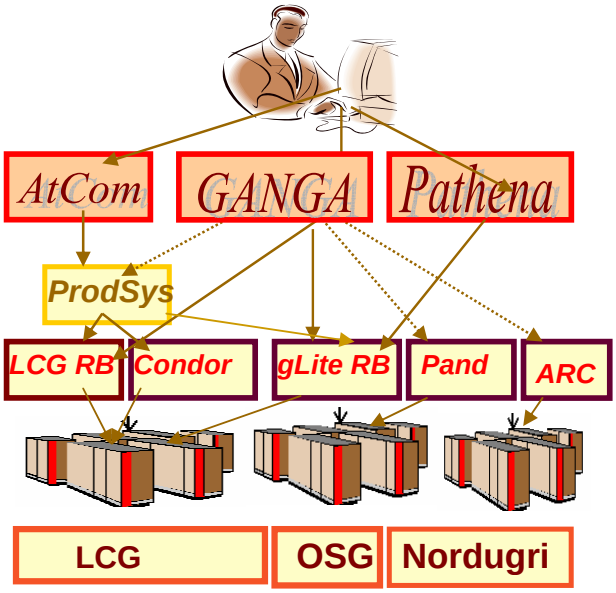

Figure 1: ATLAS Distributed Analysis strategy.

As mentioned ATLAS has adapted a strategy for DA in its heterogeneous grid environment by providing a series of backend job submission systems, which can be illustrated in the middle layer of figure 1.

LCG: Project to build and maintain data storage and analysis infrastructure for the physics community. The main DA target is to access to the LCG resources either via LCG Resource Broker (RB) or CondorG. Both systems showed a good functionality and performance in recent large-scale production exercise.

gLite: Job submission system based on gLite WMS (Workload management system), the EGEE middleware [13], that is the next grid middleware generation. The gLite WMS has a feature which is of particular impact from the distributed analysis point of view: bulk job submission along with improving support for output result retrieval.

ProdSys: ATLAS Production System [14] based in several grid flavours (LCG, OSG and Nordugrid). Jobs are supported by specific executors on the different infrastructure (LCG lexor and CondorG, OSG Panda and Nordugrid Dulcinea). It is being investigated [15] the possibility of submitting analysis jobs using Atcom as front-end.

Panda: Job processing system for production and analysis deployed in OSG. Its key design features are its tight integration with the distributed data management system [16], having a pilot jobs mechanism and a resource brokering.

ARC: Advance Resources Connector, ARC [17] based on the Globus Toolkit [18]. It is a light-weight solution, designed to support a dynamic, heterogeneous Grid facility. ARC client makes intelligent used of the distributed information and data available on the grid. ARC has a job delivery rate of approximately 40 jobs deliveries/min making it useful for distributed analysis. It offers user-friendly monitoring service [19]. Integration ARC into Ganga is currently under test.

III. IFIC ATLAS TIER2

The IFIC ATLAS computing resources are built on the Tier2 centre in Valencia that is a part of the Tier-2 Spanish federation, formed by 3 groups: IFIC, IFAE and UAM. It consists in a farm of 126 Worker Nodes (WN) disposed in racks specially prepared to the adaptation of communication equipment (Table 1). IFIC farm has Fast Ethernet of 100 Mbps. Five disk serves with a capacity of 5 TB and a tape robot with a potential capacity of 140 TB is also available. In terms of software, resources include ATLAS offline releases as well as the Grid infrastructure which is based on the most recent LCG-2 middleware. IFIC site contains all node types to make up a complete Grid. Core services are Resource Brokers (RB), Information Index (BDII), Proxy Sever (PXY) and Virtual Organization Management (VOMS). Resources are provided by the site information Services (GIIS), Computing Element (CE) with Worker Nodes and Castor-based Storage Element (SE) as a back-end to the mass storage facility. IFIC Tier-2 facility serves the dual purpose of providing resources to the ATLAS production and providing Spanish physicists with infrastructure to perform analysis jobs in grid.

Table1: Worker Nodes characterisation.

Athlon K7	63 WNs	63 WNs
CPU	1.2 GHz	1.4 GHz
RAM	1 Gbytes	1 Gbytes
Hard Disk	40 Gbytes	40 Gbytes

IV. SERVICE CHALLENGE 4 PROTOTYPE

The fundamental aim of the Service Challenge (SC) program is to achieve the goal of a production quality world-wide grid that meets the requirements of the ATLAS experiment in term of functionality and scale. DA will be tested for the first time within SC4, where the different DA components will be run in an environment that is as realistic as possible:

- Random job submission accessing data at Tier-1s (some) and Tier-2s (mostly)
- Tests of performance of job submission, distribution and output retrieval

IFIC Tier-2 is contributing in the preparation of DA SC4. This requires a set of resources, where some of them are being put in place on the preproduction service. Next, the steps which are being performed are described.

A. Dada transfer

The Distributed Data Management (DDM) [16] system provides a set of services to move data between grid-enabled computing facilities. In addition to maintain a series of databases to track these data movement. Access to the distributed data is managed by DDM efficient tools. This is crucially important for distributed analysis; hence analysis

jobs are routed based on the data availability in the sites analysis facility.

IFIC Tier-2 is contributing in data transfer task for SC4. It was asked to put in place resources such as a SRM-based Storage Element. A FTS (File Transfer Server) channel to and from the associated Tier-1 (PIC) should be set up on the Tier-1 FTS server with a FTS client software running in the IFIC User Interface (UI). As part of the Tier-2 SC4 milestones, several tests were performed to exercise the FTS channel, between PIC Tier-1 and IFIC Tier-2. Connectivity at high bandwidth from PIC Tier-1 to IFIC Tier-2, with a target rate of ~43Mb/s was demonstrated as well as the FTS channel stability linking the two sites.

As mentioned above the distributed analysis model requires that analysis jobs lend where data exists. To satisfy this with the goal to preparing for the distributed analysis program, a series of activities was started aiming to distributed data using the subscription mechanism from Tier-0 to Tier-1 and then from Tier-1s to its associated Tier-2s, willing to take part in DA SC4. For that purpose, a Disk-only area of 4 TB capacities was put in place and dedicated storage endpoints were created at IFIC Tier-2 for the data flow between PIC Tier-1 and IFIC Tier-2. Figure 2 shows data transfer between PIC-Tier1 and Spanish Federation Tier-2s

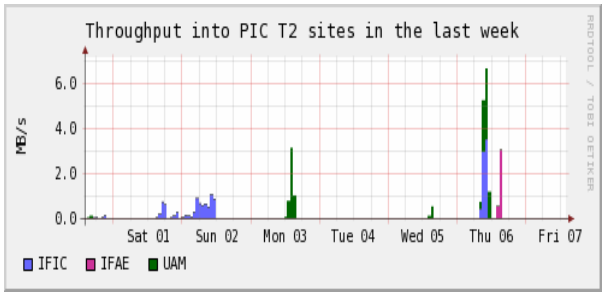

Figure 2: Data transfer from T1 to the its associated T2s in the first week of July 2006

B. Job Priority

As it is mentioned previously, gLiteRB and CondorG/LCG RB are the main job submission systems that ATLAS plans to use for submitting jobs to sites for SC4. On the current infrastructure a site has typically a configuration with one CE per Virtual Organization (VO). In addition, CE is the common entry point for ATLAS production and analysis jobs. This configuration is suitable for large scale very long lasting production jobs, whilst causes a long waiting time for analysis short jobs that require much more less CPU consuming and few time execution.

On the other hand, several batch system configurations scenarios were considered such as specific slots for analysis jobs and CPUs that will to be kept with the production jobs until analysis jobs arrive. These scenarios were excluded because of 1) resources will be wasted when no analysis jobs are arriving on the site and 2) some required resources do not support the model, respectively[20]. ATLAS is putting in place a system for the management of priorities on the LCG/EGEE infrastructure to support that analysis jobs could be performed in parallel to the production ones. The implementation was based on the use of the VOMS (Virtual Organization Management System) mechanism to assign role to a job, and then the role will direct the job to a specific share of the batch system [21]. Driven by the ATLAS requirements, the system, in an initial phase, support limited subsets of roles to control job priority (table 2).

IFIC Tier-2 is taking part of this activity by providing pre-production service (PPS-IFIC). The implementation and configuration of the system was done on PPS-IFIC CE and it works as follows:

- Resources are selected with VOMS credentials
- Ranks are computed using VOView[21] specific information
- Mixing atlas jobs with different roles are executed with the corresponding priorities

Table 2: VOMS Roles defined by ATLAS

VOMS roles	Description
/atlas	All atlas VO users
/atlas/Role=production	Used for Production activity
/atlas/Role=software	Used to software installation
/atlas/Role=lcgadmin	Vo management operation

C. Ganga experience

Although Ganga is still in the development phase it already has functionality that makes it useful for physics studies. An analysis job example for a user task based on the ttbar algorithm was exercised using Ganga. The ttbar reconstructed algorithm is provided by the Physics Analysis Tools (PAT) [22]. Ganga provides a set of ATLAS-specific features such as application configuration based on the Athena framework and input data location based on DDM. It can be run either on the command line, with Python scripts or through a graphical interface. The IFIC Tier2 infrastructure was used to process jobs using our CE with dedicated queues for analysis jobs. The processing is started within a few minutes on all Worker Nodes behind this CE. After job termination program logging and error messages are automatically stored in a job archive together with the job configuration. Also jobs were sent to several LCG sites. In this case the waiting time to get the job executing was very long because of the CE queues were occupied by the production job. Hence, the deployment of the job priority mechanism is relevant important to take full advance from the whole grid infrastructure for distributed analysis. Concerning to Ganga, in terms of configuring, submitting, monitoring and output retrieving has demonstrated a good performance. However, error handling and recovery of failed jobs in the user analysis code needs to be improved by an automatic error parsing.

V. CONCLUSIONS.

The IFIC Tier-2 facility for ATLAS experiment offers functional resources to Spanish physicists willing to perform distributed analysis using grid. In order to face the challenges of real data analysis, the facility needs more functionality and

performance as well as growing in size. To provide good analysis facility, we have therefore to continue installing and testing available software tools to acquire expertise in using and debugging the different pieces of distributed analysis system.

REFERENCES

[1] ATLAS: http://atlas.web.cern.ch/Atlas/index.html
[2] CERN-LHC http://www.cern.ch/lhc
[3] http:/lcg.web.cern.ch/LCG
[4] http://www.opensciencegrid.org/
[5] http://www.nordugrid.org/
[6] ATLAS Computing Technical Design Report, CERN-LHCC-2005-022
[7] GANGA: http://ganga.web.cern.ch/ganga/
[8]LHCb Collaboration ; http://lhcb.web.cern.ch/lhcb/
[9] http://cern.ch/atlas-proj-computing-tdr/Html/Computing-TDR-21_htm#pgfld-1019542
[10] http://www.python.org/
[11] PANDA : http://twiki.cern.ch./twiki/bin/view/Atlas/Panda
[12] https://twiki.cern.ch/twiki/bin/view/Atlas/PandaReviewMar06
[13] http://egee-intranet.web.cern.ch/egee-intranet/gateway.html
[14] ATLAS Production System; http://uimon.cern.ch/twiki/bin/view/Atlas/ProdSys
[15] Gonzalez de la Hoz et al., “The ATLAS Strategy for Distributed Analysis on several Grid Infrastructure”
[16] http://atlas.cern.ch/Atlas/GROUPS/DATABASE/project/ddm/
[17] http://www.nordugrid.org/middleware/
[18] http://www.globus.org/
[19] http://www.nordugrid.org/monitor/
[20] https://uimon.cern.ch/twiki/bin/view/Atlas/ShortQueuesForDistributedAnalysis
[21] http://egee-intranet.web.cern.ch/egee-intranet/NA1/TCG/wgs/priority.htm
[22] https://twiki.cern.ch/twiki/bin/view/Atlas/PhysicsAnalysisTools