ATLAS computing activities and developments in the Italian Grid cloud

L Rinaldi[1], A Annovi[2], M Antonelli[2], D Barberis[3], S Barberis[4], A Brunengo[3], S Campana[5], V Capone[6], G Carlino[6], L Carminati[4], C Ciocca[1], M Corosu[3], A De Salvo[7], A Di Girolamo[5], A Doria[6], R Esposito[6], M K Jha[1], L Luminari[7], A Martini[2], L Merola[6], L Perini[4], F Prelz[4], D Rebatto[4], G Russo[6], L Vaccarossa[4], E Vilucchi[2], on behalf of the ATLAS Collaboration

[1]INFN CNAF and INFN Bologna, V.Le Berti Pichat 6/2, IT-40127 Bologna, Italy
[2]INFN Laboratori Nazionali di Frascati, via Enrico Fermi 40, IT-00044 Frascati, Italy
[3]INFN Genova and Università di Genova, Dipartimento di Fisica, via Dodecaneso 33, IT-16146 Genova, Italy
[4] INFN Milano and Università di Milano, Dipartimento di Fisica, via Celoria 16, IT-20133, Milano, Italy
[5]CERN, CH-1211 Geneva 23, Switzerland
[6]INFN Napoli and Università di Napoli, Dipartimento di Scienze Fisiche, Complesso Universitario di Monte Sant'Angelo, via Cinthia, IT-80126 Napoli, Italy
[7] INFN Roma-1 and Università La Sapienza, Dipartimento di Fisica, Piazzale A. Moro 2, IT-00146, Roma, Italy

E-mail: `lorenzo.rinaldi@cnaf.infn.it`

Abstract. The large amount of data produced by the ATLAS experiment needs new computing paradigms for data processing and analysis, which involve many computing centres spread around the world. The computing workload is managed by regional federations, called "clouds". The Italian cloud consists of a main (Tier-1) center, located in Bologna, four secondary (Tier-2) centers, and a few smaller (Tier-3) sites. In this contribution we describe the Italian cloud facilities and the activities of data processing, analysis, simulation and software development performed within the cloud, and we discuss the tests of the new computing technologies contributing to evolution of the ATLAS Computing Model.

1. Introduction

The ATLAS experiment is one of the multi-purpose detectors at the Large Hadron Collider (LHC), located at CERN, Geneva (CH). Given the complexity of the experiment, the ATLAS detector produces huge amounts of data (several PB/year). In addition, an equivalent large sample of simulated data is necessary for each complete physics analysis. The organization of data management on such a large scale needed new computing paradigms, based on Distributed Grid Computing. The LHC data are managed by the World Wide LHC Computing Grid (WLCG) Collaboration [1]. The data are distributed around the world: each country involved in LHC experiments hosts in its computing centers a fraction of the data and grants access to physicists from all over the world. The sites are organized in different levels (Tier-0/1/2/3), each level having its own duties.

The Italian ATLAS Grid Computing cloud consists of many computing centers (Fig. 1): the Tier-1, located at the INFN CNAF laboratory (Bologna), four Tier-2s (Frascati, Milano, Napoli, Roma-1) and eight Tier-3s (Bologna, Cosenza, Genova, Lecce, Pavia, Roma-2, Roma-3, Trieste). The Italian sites are located in the Universities and National Institute of Nuclear Physics (INFN) departments and differ in capacities, setups and purposes. Recently the Universities of Witwatersrand and Johannesburg (ZA) and Thessaloniki (GR) joined the Italian cloud as new Tier-3s.

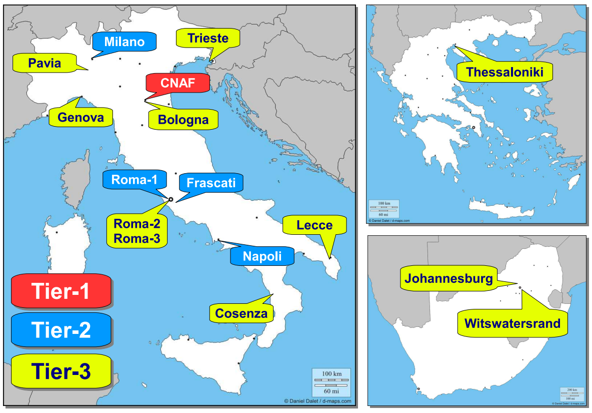

Figure 1. Geographical location of the Italian cloud sites.

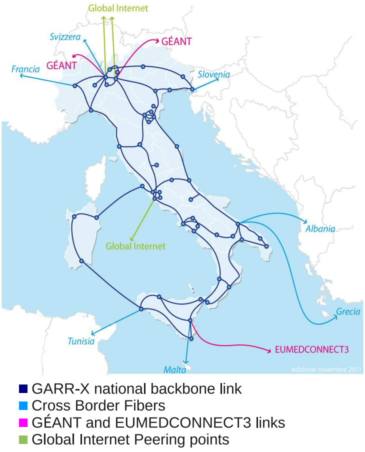

Figure 2. Topology of the GARR-X network.

According to the ATLAS Computing Model [2], the Tier-1 receives and stores permanently on tape a 10% fraction of ATLAS data in several formats: real RAW data and simulated HITS data, Event Summary Data (ESD), Analysis Object Data (AOD), Derived Physics Data (DPD). An additional 10% fraction of the AOD and DPD data dynamically placed by the ATLAS Production and Distributed Analysis (PanDA) System [3] is stored on the Tier-1 disk. The Tier-1 is mainly used for central production activities, such as simulation and data processing. The Tier-2 sites share the activities of production and analysis of the derived data samples dynamically placed in the sites. Tier-3s are dedicated to the data analysis, via grid or interactively, and code testing and development and in some cases they also contribute to the simulation production.

In the first period of the data taking activities, the data were pushed hierarchically from the CERN Tier-0 to the Tier-1s and eventually to the Tier-2s within the same cloud. In order to optimize the resource utilization, the data flow was recently modified and the data are now dynamically placed on sites, pulled by the Tier-1s and Tier-2s according to data popularity; only the more accessed datasets are effectively replicated. A further evolution of the computing model came from the high performance of the recent global network connections, which allow data transfers among Tier-1s and Tier-2s that are directly connected (T2D) to all Tier-1 and T2D sites, thus breaking the regional cloud boundaries.

2. The Tier-1 site

The Italian Tier-1 center is located at the CNAF laboratory, the INFN department for computing and telematic technologies, in Bologna. The INFN Tier-1 serves all LHC experiments and also other experiments in which INFN is involved.

The overall INFN Tier-1 storage capacity consists of 10 PB (of which 2.7 PB assigned to ATLAS) of fiber-channel SAN disk storage element and 17 PB (with 3.6 PB assigned to ATLAS) of on-line tape.

The ATLAS experiment has adopted at CNAF a storage system based on StoRM/GPFS/TSM. StoRM [4] is a storage resource manager for generic posix disk-based storage systems that decouples the data management layer from the underlying storage system characteristics, implementing the SRM interface version 2.2 [5]. StoRM works optimally with high-performance parallel file system like IBM-GPFS [6]. GPFS is installed on a storage Area Network (SAN) infrastructure interconnecting storage systems and disk servers; the data import/export throughput is managed by GridFTP servers. The farm worker nodes access the storage through a standard 10 Gb Ethernet network using the posix file protocol, and access the ATLAS experiment software via the CernVM File System (CVMFS) [7]. The IBM Tivoli Storage Manager (TSM) [8] system is used to manage data access on tape via the GEMMS interface [9], a complete Grid-enabled Hierarchical Storage Management (HSM) solution based on StoRM/GPFS and TSM, developed and deployed at CNAF.

The data transfer activity of the Tier-1 center demands a large bandwidth because of the very large computing farm and the central role that it plays in the regional computing cloud. The INFN Tier-1 infrastructure will be connected by Summer 2012 with an overall 40 Gbps link of GARR-X [10], the new optic fiber network of the university and research community provided by the Italian GARR Consortium and the European Géant Infrastructure [11], as shown in Fig. 2. An existing 20 Gbps link is part of LHC-OPN (Optical Private Network) [12], a private IP network connecting the Tier-0 and the Tier-1 sites and dedicated to WLCG traffic only. A dedicated 10 Gbps LHC-OPN link towards the German Tier-1 in Karlsruhe (from the GARR POP in Milan) using a cross-border fiber is provided by GARR in collaboration with Switch (the Swiss network provider), and is used for data transfer among CNAF and the Tier-1 sites at KIT (DE), SARA (NL) and IN2P3-CC (FR) and as a backup link to CERN in case of problems on the main T0-T1 connection. A 10 Gbps link that will be fully active by Summer 2012 is part of LHC-ONE (LHC Open Network Environment) [13], an open network topology built to provide a private network for Tier-1/2/3 traffic, designed to guarantee high performance and reliability in the data transfers, without requiring additional resources to those already provided by the various national research networks (NRENs), in particular new backbones. The INFN Tier-1 is also connected with a 10 Gbps link used for General IP traffic. This link is used to reach all other destinations (not expressly part of LHC-OPN and LHC-ONE communities).

Figure 3 shows the ATLAS data transfer throughput to the INFN Tier-1 from January 2011 to April 2012. The average throughput is 84 MB/s; daily peaks of 800 MB/s have been reached. The primary activities that contribute to the Tier-1 throughput are the data consolidation (transfer of reconstructed data between Tier-1 and Tier-2 sites), data export from Tier-0 and data subscription of private Users.

Data processing is carried out by 13000 CPU cores, managed by a LSF batch system. A computing power of 26 kHepSpec out of a total of 150 kHepSpec, equivalent to about 2200 job slots, is guaranteed to the ATLAS experiment. However the actual CPU consumption for ATLAS computing activities is on average 30% higher than the assignment, due to an opportunistic usage of the unused CPU resources.

The LSF batch system uses the ATLAS VOMS roles to define the job sharing [14]: 80% of the resource are assigned to the *production* VOMS role for simulation production, data reprocessing and group analysis and 20% is assigned to the *atlas*, *atlas/it* and *pilot* VOMS roles for Physics analysis. The ATLAS software installation and validation jobs run with highest priority using the *atlassgm* VOMS role.

The processing activities of the Tier-1 are shown in Fig. 4. In the year 2011 and in the first months of 2012, the primary activity at the Tier-1 was the Monte Carlo (MC) simulation and

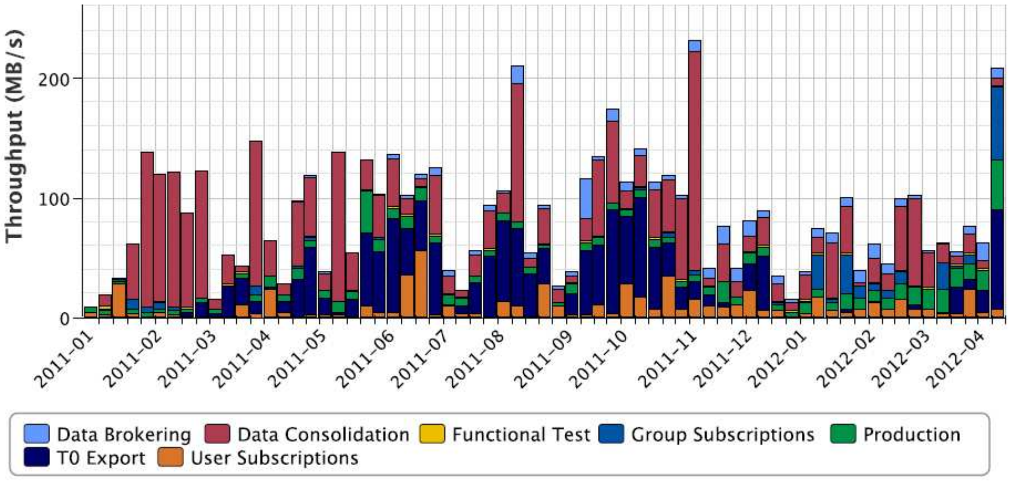

Figure 3. Tier-1 throughput rate by computing activity in 2011-2012.

reconstruction. During the 2010 Data Reprocessing campaign at the end of March 2011, the site resources were mainly dedicated to the Data Processing.

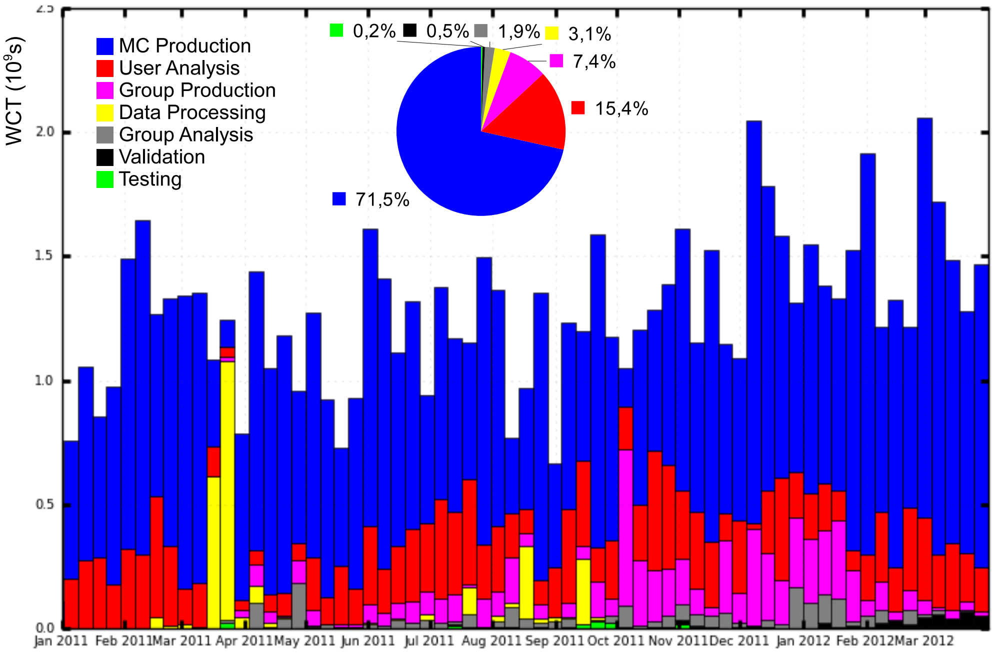

Figure 4. Tier-1 computing activities in 2011-2012. The pie-chart shows the relative usage of the computing resources.

3. The Tier-2 sites

The four Tier-2 facilities of the ATLAS Italian cloud are based at the Universities of Milano, Napoli and Roma-1 and at the INFN Laboratory in Frascati. The computing and storage resources dedicated to the ATLAS experiment that will be available at sites by the Summer of 2012 are summarized in Table 1.

The Tier-2 sites are connected to the INFN Tier-1 with 1 Gbps (Frascati) and 2 Gbps (Milano, Napoli, Roma-1) links. The Tier-2s link capacity will be increased to 10 Gbps by the end of 2012 when the GARR-X network will be operative.

The Napoli site is one of the 14 pilot sites that the ATLAS experiment has identified to start deployment and test of the new generation LHC-ONE network. In the near future, all Italian sites will join LHC-ONE.

Table 1. Capacity of the Italian Tier-2 sites in Q2-2012

Tier-2	Computing			Storage		
	Job Slots	HepSpec06	Batch System	Capacity (TB)	Storage Type	Bandwidth (Gbps)
Frascati	870	8300	PBS	420	DPM	10
Milano	1050	10900	PBS/Condor	1100	StoRM	10
Napoli	1200	12400	PBS	1100	DPM	10
Roma-1	1300	13100	LSF	1040	DPM	10

In the ATLAS Distributed Data Management (DDM) topology Milano, Napoli and Roma-1 sites are directly connected to all the ATLAS Tier-1 and T2D sites and the dynamic data placement occurs directly from all the Tier-1s and T2Ds, without passing through the INFN Tier-1. The amount of data destined to each site is determined by the site availability performance, evaluated using the ATLAS Site Status Board (SSB) [15].

The Napoli and Roma-1 groups had a leading role in developing and building the ATLAS muon sub-detectors. For this reason the two sites receive directly from the Tier-0 an additional flow of calibration stream data, which are used for the muon detector calibration and performance studies.

The Frascati site is used as main site for the Fast Tracker [16] simulations in Italy. It holds a replica of the Fast Tracker related datasets. The Fast Tracker jobs requires a few hundred GBs of RAM. In order to reduce the needed RAM to a few GBs per job, each event is distributed to 128 jobs that process only a small fraction of each event. When all jobs complete the processed events are merged to complete the simulation. In addition, the site has an active role in the development and promotion of a new tool for distributed analysis on the Grid: Proof on Demand [17].

The ATLAS data transfer throughput for each Italian Tier-2s is shown in Fig. 5.

The overall computing activities of the Italian Tier-2s are shown in Fig. 6. On average, 74% of the resources have been used for MC production, followed by a considerable fraction of user analysis jobs.

4. The Tier-3 sites

In the Italian cloud there are eleven Tier-3 grid-enabled facilities: eight sites are located in Italy (Universities and INFN Departments of Bologna, Cosenza, Genova, Lecce, Pavia, Roma-2, Roma-3, Trieste), two in South Africa (Johannesburg and Witwatersrand Universities) and one in Greece (Thessaloniki University). The computing and storage resources dedicated to

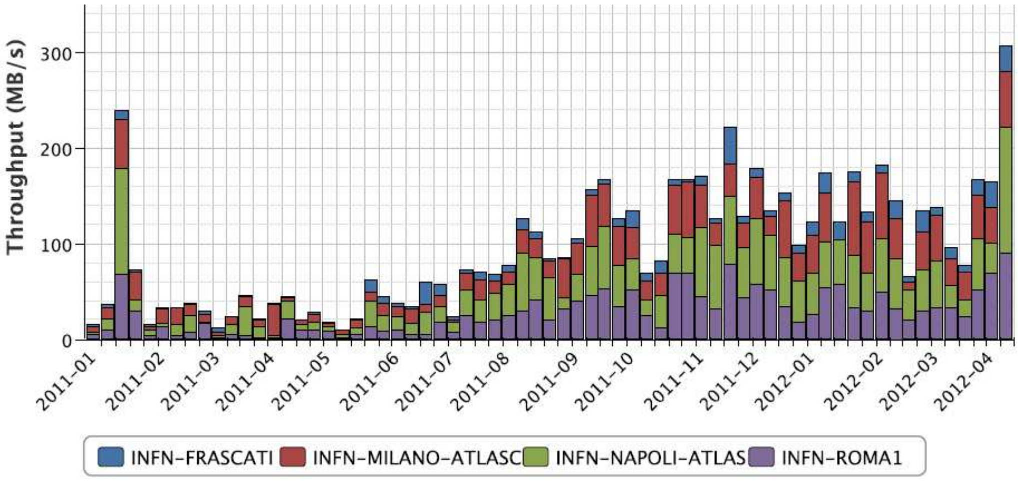

Figure 5. Tier-2s throughput rate by site in 2011-2012.

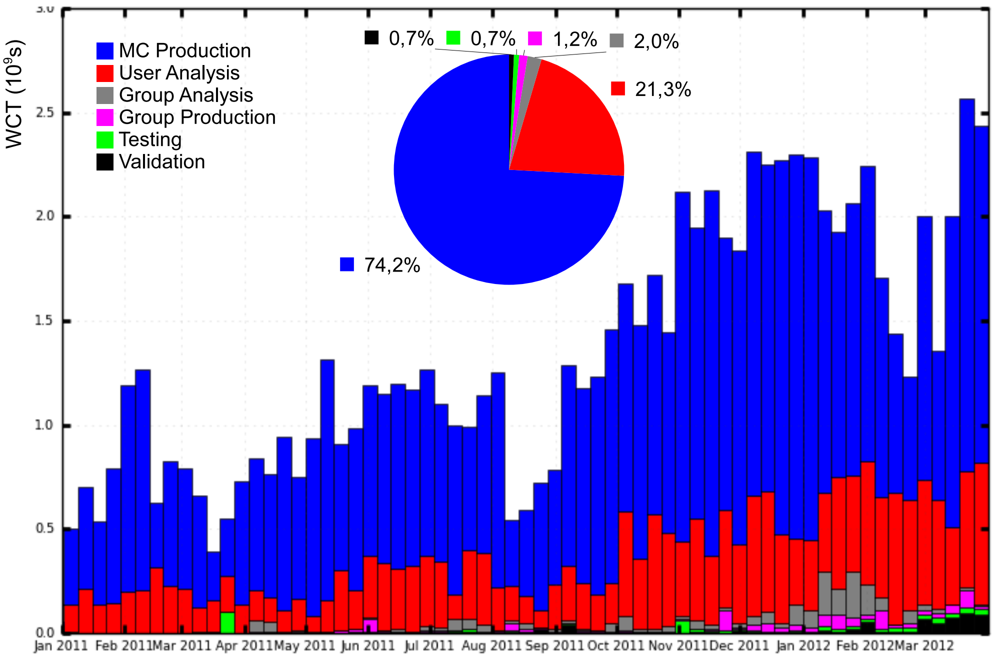

Figure 6. Tier-2s computing activities in 2011-2012. The pie-chart shows the relative usage of the computing resources.

the ATLAS experiment are summarized in Table 2. The Tier-3s are connected through the local institute network that for the Italian sites is the General IP Research network provided by GARR. The sum of all the Tier-3s resources is almost comparable to an ordinary Tier-2 site, although the duties and the activities are completely different.

Unlike the Tier-1 and Tier-2 sites, which have to manage the usage of their resources according to the ATLAS experiment requirements, the Tier-3s resources are not pledged and the site settings can be tuned on the necessity of the local community. The Tier-3 sites have been designed for local analysis, which is carried out interactively on the farm or via grid jobs, and testing purposes. In particular, several sites have the Storage Element (SE) based on a posix file system, allowing direct access to the data that can be analyzed interactively. As secondary and background activity, many sites run ATLAS production jobs too.

As shown in Fig. 7, the ATLAS data transfers to the Tier-3s consist of MC production (for those sites enabling that activity) and subscriptions by local users that copy their data produced in other sites.

Table 2. Capacity of the Tier-3 sites in 2012.

Tier-3	Computing			Storage	
	Job Slots	Batch System	Capacity (TB) (grid+local)	Storage Type	WAN (Gbps)
Bologna	100	LSF	40	StoRM/GPFS	10
Cosenza	30	PBS	30	StoRM/XFS(GPFS)	0.1
Genova	90	LSF	21+23	StoRM/GPFS	1
Johannesburg	50	PBS	20	DPM	0.1
Lecce	110	LSF	20+16	StoRM/NFS	0.1
Pavia	40	PBS	7	DPM	1
Roma-2	20	PBS	20	StoRM/NFS	0.15
Roma-3	140	PBS	37+30	StoRM/GPFS	1
Thessaloniki	60	PBS	4	DPM	1
Trieste	8	LSF	13+3.5	StoRM/GPFS	1
Witwatersrand	50	PBS	26	DPM	0.1

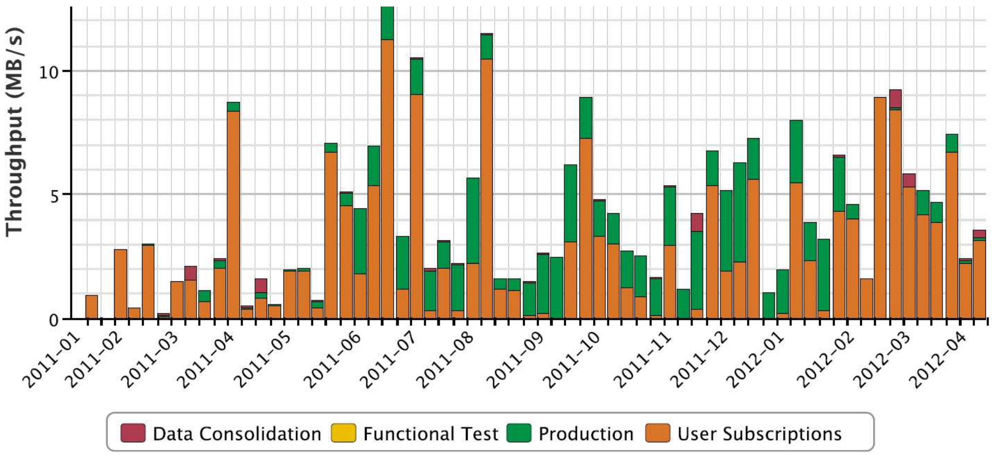

Figure 7. Tier-3s throughput rate by activity in 2011-2012.

5. Conclusions

The Italian ATLAS Grid Computing cloud provides to the ATLAS experiment a substantial fraction of computing infrastructures. An efficient use of the resources has been fulfilled during the LHC data taking period 2011/2012. The CNAF INFN Tier-1 and the four Tier-2 sites have been used for ATLAS data storage, processing and simulation and a part of the resources have been dedicated to user analysis. The Tier-3 facilities provided instead resources for grid and interactive analysis. The completion of the new high-bandwidth network connection by the Summer of the 2012 and the assignment of new computing resources will boost the computing activities of the Italian cloud, allowing more efficient and flexible data distribution and management.

References

[1] I. Bird et al., LHC Computing Grid Technical Design Report, CERN-LHCC-2005-024, http://lcg.web.cern.ch/LCG/tdr/ (2005)

[2] Jones R and Barberis D, *J. Phys.: Conf. Ser.* **119** 072020

[3] T. Maeno, "PanDA: Distributed Production and Distributed Analysis System for ATLAS". *J. Phys.: Conf. Ser.* (2008) **119**, 062036

[4] StoRM: http://storm.forge.cnaf.infn.it

[5] SRM 2.2 specification: https://sdm.lbl.gov/srm-wg/doc/SRM.v2.2.html

[6] GPFS: http://www-03.ibm.com/systems/software/gpfs/

[7] A De Salvo et al, "Software installation and condition data distribution via CernVM File-System in ATLAS", Proc. Conf. for Computing in High-Energy and Nuclear Physics (CHEP 2012) (New York, USA)

[8] TSM (Tivoli Storage Manager): http://www-01.ibm.com/software/tivoli/products/storage-mgr/

[9] Cavalli A et al, "StoRM-GPFS-TSM: A new approach to hierarchical storage management for the LHC experiments" *J. Phys.: Conf. Ser.* (2010) **219** 072030

[10] GARR: http://www.garr.it/eng

[11] GEANT: http://www.geant.net/pages/home.aspx

[12] LHC-OPN: https://twiki.cern.ch/twiki/bin/view/LHCOPN/WebHome

[13] LHC-ONE: http://lhcone.net/

[14] A Doria et al, "Deployment of job priority mechanisms in the Italian cloud of the ATLAS experiment", *J. Phys. Conf. Ser.* (2010) **219**, 072001

[15] E Magradze et al, "Automating ATLAS Computing Operations using the Site Status Board", Proc. Conf. for Computing in High-Energy and Nuclear Physics (CHEP 2012) (New York, USA)

[16] A Andreaniet al, *IEEE Trans. on Nucl. Sci.* (2010) **59** 348

[17] E Vilucchi et al, "Enabling data analysis á la PROOF on the Italian ATLAS-Tier2's using PoD", Proc. Conf. for Computing in High-Energy and Nuclear Physics (CHEP 2012) (New York, USA)