Computing infrastructure for ATLAS data analysis in the Italian Grid cloud

A Andreazza1, A Annovi2, D Barberis3*, A Brunengo1, S Campana4, G Carlino5, C Ciocca6, M Cobal7a, M Corosu1, A De Salvo1, U De Sanctis7b, A Di Girolamo4, A Doria1, F Galeazzi9, MK Jha6, L Luminari8, A Martini2, L Merola1, E Musto5, F Pascolo7a, L Perini1, M Pistolese1, D Rebatto1, L Rinaldi6, L Vaccarossa1, E Vilucchi2 and R Vitillo10

1 INFN Milano and Università di Milano, Dipartimento di Fisica, via Celoria 16, IT-20133 Milano, 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 CERN, CH-1211 Geneva 23, Switzerland
5 INFN Napoli and Università di Napoli, Dipartimento di Scienze Fisiche, Complesso Universitario di Monte Sant’Angelo, via Cinthia, IT - 80126 Napoli, Italy
6 INFN-CNAF and INFN Sezione di Bologna, Viale Berti Pichat 6/2, IT-40127 Bologna, Italy
7a INFN Gruppo Collegato di Udine and Università di Udine, Dipartimento di Fisica, via delle Scienze 208, IT-33100 Udine
7b INFN Gruppo Collegato di Udine and Scuola Internazionale Superiore di Studi Avanzati, via Bonomea 265, IT-34136 Trieste, Italy
8 INFN Roma I and Università La Sapienza, Dipartimento di Fisica, Piazzale A. Moro 2, IT-00185 Roma, Italy
9 INFN Roma Tre and Università Roma Tre, Dipartimento di Fisica, via della Vasca Navale 84, IT-00146 Roma, Italy
10 INFN Pisa and Università di Pisa, Dipartimento di Fisica “E. Fermi”, Largo B. Pontecorvo 3, IT-56127 Pisa, Italy

* Corresponding author: Dario.Barberis@cern.ch

Abstract. ATLAS data are distributed centrally to Tier-1 and Tier-2 sites. The first stages of data selection and analysis take place mainly at Tier-2 centres, with the final, iterative and interactive, stages taking place mostly at Tier-3 clusters. The Italian ATLAS cloud consists of a Tier-1, four Tier-2s, and Tier-3 sites at each institute. Tier-3s that are grid-enabled are used to test code that will then be run on a larger scale at Tier-2s. All Tier-3s offer interactive data access to their users and the possibility to run PROOF. This paper describes the hardware and software infrastructure choices taken, the operational experience after 10 months of LHC data, and discusses site performances.

1. The Italian ATLAS Grid Cloud
The Italian ATLAS Grid cloud consists of several computing sites with very different capacities, set-ups and purposes, namely a Tier-1 site at CNAF (Bologna), four Tier-2 sites in Milan, Rome-1, Naples and LNF (Frascati), Grid-enabled Tier-3s in Genoa, Trieste and Rome-3, and finally non-Grid Tier-3s at all other ATLAS sites.
As defined in the ATLAS computing model [1], the Tier-1 site is used mostly for central production activities, such as simulation production and data reprocessing, but analysis jobs can also run if their input data are only available there. The Tier-2 resources are shared between simulation production and data analysis; Tier-3s are only dedicated to analysis but Grid-enabled Tier-3s can also contribute to simulation production in the background.

ATLAS data are sent from CERN and other Tier-1s to the CNAF Tier-1, from where part of the data (AODs, Analysis Object Data) are placed at Tier-2 sites, and other data formats are automatically dispatched to Tier-2s for analysis when jobs requesting them are submitted to the Italian cloud. The analysis model foresees that the outputs of group and user analysis tasks run on the Grid be brought "home" to a local Tier-3 facility for final (interactive) analysis; Grid-enabled Tier-3s have local disk storage where data of interest for a given community can be kept under their control. Non-Grid-enabled Tier-3s are essentially local computing farms that can be used for interactive data analysis with ROOT [2] or PROOF [3].

2. The CNAF Tier-1
The Italian Tier-1 is hosted at CNAF, the INFN laboratory for information technologies. The INFN Tier-1 serves all the LHC experiments and the other experimental activities carried on by INFN in Nuclear and Particle Physics.

CNAF provides overall 83k HepSpec06 [4] (to be increased to 100k in 2011) of computing power with 8000 CPU cores. The storage consists of a 7 PB fiber-channel SAN disk storage element and 8 PB of on-line tape. A computing power of 16k HepSpec06, a disk space of 1.8 PB and a tape space of 1.3 PB are assigned to the ATLAS experiment. The CNAF Tier-1 layout is shown in Figure 1 [5].

![Figure 1: layout of the CNAF Tier-1 computing services.](image)

The storage system solution for the ATLAS data is based on StoRM/GPFS/TSM. The StoRM [6] service is a storage resource manager for generic posix disk-based storage systems that decouples the data management layer from the underlying storage systems characteristics. It implements the SRM interface version 2.2 [7]. StoRM takes advantage from high-performance parallel file systems like GPFS (from IBM) [8]. The GPFS installation is based on a Storage Area Network (SAN)
infrastructure interconnecting storage systems and disk servers; the data import/export throughput is managed by GridFTP [9] servers, while the farm worker nodes access the storage through standard Ethernet network using the posix file protocol. The TSM (Tivoli Storage Manager) [10] system (also from IBM) is used to manage data access on tape. The GEMSS [11] interface, a complete Grid-enabled hierarchical storage management (HSM) solution based StoRM/GPFS and TSM has been developed and deployed at CNAF and it is now in production for ATLAS.

The Tier-1 farm has a LSF [12] batch system, which manages the computing resources for the ATLAS activities of data processing and reprocessing, simulation production and user data analysis. The job sharing is based on the ATLAS roles: 60% is assigned to the production role (data reprocessing and simulation) and 40% to the atlas, atlas/it and pilot roles (for user analysis).

3. The Tier-2s
The four Tier-2 sites are based at three Universities (Milan, Naples and Rome-1) and the INFN Laboratories in Frascati. They are all connected to the Tier-1 with at least 1 Gbps links that will be upgraded to 10 Gbps during 2011. On each site, the LAN topology consists of a star network, with a central 10 Gbps router, directly connected to the disk servers and the worker node switches. The site capacities are given in Table 1.

| Tier-2   | Job slots | HepSpec06 | Batch system | Capacity | Storage type | WAN   |
|----------|-----------|-----------|--------------|----------|--------------|-------|
| Frascati | 250       | 1700      | PBS          | 160 TB   | DPM          | 1 Gbps|
| Milan    | 680       | 4400      | PBS & Condor | 530 TB   | GPFS/StoRM   | 1 Gbps|
| Naples   | 700       | 4600      | PBS          | 550 TB   | DPM          | 1 Gbps|
| Rome-1   | 680       | 4400      | LSF          | 540 TB   | DPM          | 2 Gbps|

The Tier-2 sites provide computing capacity for simulation production and user analysis, and storage capacity for general analysis and for data that have been selected by physics and detector performance groups. These derived data are the input for further analysis on the Grid; the output of these jobs consists usually of histograms and/or ROOT n-tuples that are stored at Tier-3 sites for interactive analysis.

A fair-share system is essential for the proper usage of Tier-1 and Tier-2 resources. Italian sites implement on their batch systems a fair-share system based on VOMS [13] groups and roles, and on the concept of VOViews [14]. Figure 2 shows the number of jobs running at Tier-1 and Tier-2 sites during 2010; the analysis activities increased throughout the year together with the accumulated ATLAS data.

4. The Tier-3s
Potentially all ATLAS institutes could have a Grid enabled Tier-3, but the need of small Grid sites is limited to local or quasi-local testing of new simulation, reconstruction and analysis code and to the storage of derived datasets. The Tier-1 and the Tier-2s offer services to the local communities in addition to the pledged resources to the Collaboration. All institutes instead have a local computing facility that is used to run interactive analysis tasks, typically based on ROOT or on its parallelized version, PROOF.
4.1. The Grid-enabled Tier-3s

Three ATLAS institutes (Genoa, Rome-3 and Trieste) have set up Grid-enabled Tier-3s. These Tier-3s for ATLAS are part of larger Grid sites that serve several communities, so they are shared with other users (from the physics departments and elsewhere), thus reducing the system administration workload. They receive automatic software installations and are fully enabled to run Athena jobs, interactively, in local batch mode and as Grid jobs. They can be used to develop software and to test tasks on a small scale before submitting them to the Grid.

The storage is divided about 50-50 between purely local areas and SRM managed Grid storage, which includes an area (“hotdisk”) for database releases and conditions data files, an area (“scratchdisk”) for the temporary outputs of local Grid jobs and for data import/export from the Grid, and an area (“localgroupdisk”) for the long-term storage of data that are useful to the local community (this is effectively the “home directory” of local Grid users). All sites use GPFS to manage the storage and StoRM as SRM interface; in this way all data in the Grid storage areas are directly visible to all interactive and batch worker nodes, as GPFS is mounted on all nodes. Files can be imported to a given SRM storage area and analysed directly with an interactive job; there is no need of separate storage servers for grid and non-Grid users, which leads to easier system management and easier life for users. In addition, direct access to the Grid file system is much more intuitive for non-expert users.

ATLAS data analysis demands a stable and high-performance service for physicists to carry out their scientific research. As single ROOT-based analysis jobs can read in excess of 20 MB/s of input data, data analysis is an input-output intensive activity that places large demands on the network, central storage, and local disks at computing facilities. Data throughput is of the highest importance at Tier-3s, as they run only analysis jobs. In order to achieve the maximum throughput of 1 Gbps per CPU box with the current Gigabit Ethernet technology, it is mandatory to have a well-designed LAN with a 10 GE switch and redundant network connections.

The HammerCloud [15] framework was developed to allow automated submission of large numbers of user analysis jobs to ATLAS computing sites using the Ganga [16] distributed analysis tool. This allows computing sites to test their performance against a load from a sample ATLAS user analysis tasks in a repeatable manner, which also provides metrics of success for comparison. Along with several other Italian sites, Genoa has used HammerCloud analysis stress tests consisting of approximately 320 jobs to test the site configurations and performance of their storage infrastructure. Figure 3 shows the results from HammerCloud running on input files of two different sizes. The input files were AODs from simulated and real data; simulated AODs were smaller in size in comparison to the AODs from real data. The algorithms that were run on these datasets were also different (more...
complex on real data). The CPU/WCT (CPU usage efficiency ratio) were 96% and 73% for AODs of smaller and larger sizes respectively. The lower efficiency values for data AODs was due to larger transfer time needed to store the job outputs. In both cases the performance was at the level of the best Tier-2 sites.

4.2. The PROOF clusters

Within the local activities that can be done in a Tier-3, the analysis of ROOT [2] n-tuples (data and simulation) is for sure the most important one. Therefore Tier-3s are the natural places to use PROOF [3], the “Parallel ROOT Facility” available in the recent versions of ROOT (from 5.22 onwards).

The configuration we tested as prototype was set up locally using a GPFS as file system, a cluster with 3 multi-core machines (8 cores each), a single 1 Gbps network switch and ROOT n-tuples as data. Different types of analyses of increasing complexity have been written to test the scalability, the I/O speed and the event rate. The main results we obtained are that:

- For a typical analysis macro with hundreds of branches to be read, algebraic operations on their contents and hundreds of histograms to fill, the possible bottleneck is the network switch (here 118 MB/s) rather than the storage I/O speed;
- The reading speed increases linearly with the number of cores to within 10-20% and depends on the complexity of the analysis program. The linearity improves when the computational part dominates over the I/O operations. For a typical analysis program with 24 cores available we reached the maximum rate of 116 MB/s, which was very close to the bandwidth limit.
- In case of the cut-based analysis macro, the majority of time is spent by ROOT in reading the event rather than in computation, suggesting to read only the branches used in the analysis.

The main conclusion of this work is that Tier-3 sites as described above are well equipped to support PROOF analysis of large and complex n-tuples, provided the LAN capacity is well matched to the computing and storage power.

Figure 3: Results of HammerCloud tests run on the Genoa Tier-3 in Summer 2010. Top: small input AOD files with simulated events. Bottom: larger input AOD files with real events. Left: job efficiency. Centre: CPU efficiency. Right: event throughput.
5. Networking
The CNAF Tier-1 is the most demanding site in terms of bandwidth, both because of its very large computing farm (compared to Tier-2s) and its role in the computing model. It has to receive data (RAW, AOD, ESD, etc.) from Tier-0, distribute data to all the Tier-2s for processing, and collect their output. For the connection with the Tier-0 at CERN, it uses a 10 Gbps dedicated link provided by the Italian GARR Consortium [17] and the European Géant infrastructure [18]. An additional 10 Gpbs link towards the German Tier-1 in Karlsruhe is provided by GARR in collaboration with Switch for the Tier-1–Tier-1 connectivity. Both these connections are part of the LHC-OPN (Optical Private Network), a private IP network connecting the Tier-0 and Tier-1 sites of LCG and dedicated to WLCG traffic only. Moreover, GARR provides CNAF with a 10 Gbps link to the national backbone for the connection to the Tier-2s.

The Tier-2s use internally a LAN with a 10 Gbps backbone, in order to provide fast data access to the computing nodes. The estimated bandwidth consumption is in the order of 1 Gpbs every eight job slots. The connectivity towards other sites is currently at a lower speed. Rome-1 has a 3 Gbps link with GARR, Naples a 2 Gbps one, Milan and Frascati 1 Gbps. All of them will eventually upgrade to 10 Gpbs in the next few months, when the GARR-X project (the new optical network infrastructure by GARR) will be operative. This will allow them to scale with the planned upgrades of their computing farms (more jobs means more data to transfer).

The Tier-3s usually have no dedicated network infrastructure. They are hosted by the computing center of several INFN bases, and share their network connectivity.

6. Conclusions
The Italian ATLAS computing group has provided the ATLAS Collaboration with an integrated system for data production and analysis, based on the INFN Grid middleware and infrastructure, with the ATLAS Distributed Computing tools on the top layer. This setup has been progressively tested over the years and has been in full operation for the 2010 LHC run. The performance is satisfactory and no bottleneck has been observed in data transfer or production and user job execution. The forthcoming improvements to available network bandwidths will allow the integrated usage of all Tier-3 facilities as if they were a single large computing centre.

References
[1] Jones R and Barberis D 2008 The ATLAS computing model J. Phys.: Conf. Ser. 119 072020
[2] Antcheva I et al 2009 ROOT — A C++ framework for petabyte data storage, statistical analysis and visualization Comput. Phys. Commun. 180, 12 2499-2512
[3] Ballintijn M, Roland G, Brun R and Rademakers F 2003 The PROOF distributed parallel analysis framework based on ROOT Proc. Conf. for Computing in High-Energy and Nuclear Physics (CHEP 2003) (La Jolla, California)
[4] HepSpec06 definition: http://w3.hepix.org/benchmarks/doku.php
[5] Ronchieri E et al 2011 INFN-CNAF Tier-1 Storage and Data Management Systems for the LHC Experiments Data Taking Proc. Conf. for Computing in High-Energy and Nuclear Physics (CHEP 2010) (Taipei, Taiwan)
[6] StoRM: http://stormforge.cnaf.infn.it
[7] SRM 2.2 specification: https://sdm.lbl.gov/srm-wg/doc/SRM.v2.2.html
[8] GPFS: http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/index.jsp?topic=com.ibm.cluster.gpfs.doc/gpfsbooks.html
[9] GridFTP: http://www.globus.org/alliance/publications/papers/dataMgmt.pdf
[10] TSM (Tivoli Storage Manager): http://www-01.ibm.com/software/tivoli/products/storage-mgr/
[11] Cavalli A et al 2010 StoRM-GPFS-TSM: A new approach to hierarchical storage management for the LHC experiments J. Phys.: Conf. Ser. 219 072030
[12] LSF: http://www.platform.com/workload-management/high-performance-computing
[13] VOMS: http://www.globus.org/grid_software/security/voms.php
[14] Vilucchi E et al 2009 Activities and performance optimization of the Italian computing centers supporting the ATLAS experiment Nuclear Science Symposium Record, IEEE: 516-522
[15] Van der Ster D et al 2010 Functional and large-scale testing of the ATLAS distributed analysis facilities with Ganga J. Phys. Conf. Ser. 219 072021
[16] Mościcki JT et al 2009 GANGA: a tool for computational-task management and easy access to Grid resources Comput. Phys. Commun. 180, 11 2303-16
[17] GARR: http://www.garr.it/eng
[18] European Géant network: http://archive.geant.net